Annotation automatique des lieux dans l'oral spontané transcrit

Hélène Flamein[1]
(1) LLL, UMR 7270, UFR LLSH, 10 Rue de Tours, 45065 ORLEANS cedex 2
helene.flamein@univ-orleans.fr

RESUME

Cet article a pour but de présenter une démarche généraliste pour l'annotation automatique des lieux dans l'oral transcrit. Cette annotation est effectuée sur le corpus ESLO (Enquête SocioLinguistique à Orléans) et suppose une réflexion sur les caractéristiques propres à la désignation d'un lieu à l'oral. Avant d'expliciter la méthode employée pour traiter automatiquement notre corpus, nous présenterons le travail préparatoire de la constitution d'une convention d'annotation et d'un corpus de référence indispensable pour l'évaluation du système.

ABSTRACT

Automatic annotation of places in the transcribed oral
The aim of this article is to present a general methodology for the automatic annotation of places in the transcribed oral. This annotation is carried out on the corpus ESLO (Enquête SocioLinguistique à Orléans – Sociolinguistic Investigation in Orléans) and presupposes a reflection on the characteristics proper to the designation of an oral place. Before explaining the method used to automatically process our corpus, we will present the preparatory work for the constitution of an annotation convention and a corpus of reference indispensable for the evaluation of the system.

MOTS-CLES : Détection automatique de lieux, Perception des lieux, Traitement automatique du Langage, Entités nommées, ESLO, Corpus oral

KEYWORDS: Automatic detection of places, Place-perception, Natural Language Processing, Named entities, ESLO, Oral corpus

1 Introduction

Cette étude s'inscrit dans un travail de thèse qui s'intéresse à la perception des habitants d'une ville. L'objectif à terme est de présenter l'analyse automatique de la perception des lieux dans le discours transcrit de l'oral par différents locuteurs grâce aux techniques du TAL. Ces techniques permettent notamment de détecter automatiquement les lieux ainsi que les expressions qui les accompagnent dans les transcriptions. Nous appuyons notre travail sur le corpus ESLO2 (Enquête SocioLinguistique à Orléans[1]), composé de 460 heures d'enregistrements répartis en une vingtaine de modules qui présentent chacun des situations de communication différentes.

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

Après avoir fait le rappel du cadre théorique encadrant notre démarche, nous présenterons la méthode suivie pour le repérage des lieux. Ses différentes étapes seront développées : l'exploration manuelle du corpus, l'élaboration de conventions d'annotation, la constitution d'un corpus de référence ainsi que notre expérience d'annotation automatique. A la fin du processus d'annotation, les lieux détectés seront projetés sur la carte d'Orléans grâce à leurs coordonnées géographiques présentes dans les métadonnées des ressources spécifiques constituées.

2 Etat de l'art

L'utilisation de corpus annotés est une pratique fondamentale en TAL et ce depuis les années 90. Le corpus arboré de l'anglais de l'Université de Pennsylvanie créé entre 1989 et 1994 ou Penn Treebank fait référence avec ses annotations morpho-syntaxiques et syntaxiques en inspirant la création d'autres ressources similaires telles que le corpus arboré du français présenté par Abeillé et al. (2003).

Depuis les années 1990 et la dernière série des conférences américaines MUC[2] (Message Understanding Conferences), la question de la reconnaissance des entités nommées est incontournable dans le domaine du TAL. Selon (Ehrmann, 2008), les entités nommées correspondent à « toute expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus ». Elles 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 et Sekine, 2009). De nombreux outils d'annotation automatique sont dédiés à la reconnaissance des entités nommées (REN) comme l'un des modules du Stanford NER[3] (Finkel et. al, 2005) ou le POLYGLOT NER[4] présenté dans Al-Rafou et. al (2015). La REN est devenue une tâche indépendante qui est désormais au centre de différentes campagnes d'évaluation d'outils dédiés à l'extraction d'informations. Plusieurs campagnes comme ESTER ou ETAPE évaluent justement 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.

L'objectif du travail est d'annoter l'ensemble des mentions de lieux présentes dans l'oral transcrit. Les lieux correspondent ici à la définition de Lesbeguerrie (2007) des entités spatiales absolues (ESA) qui représentent les informations spatiales les plus « primitives » 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 qualifiés d'entités spatiales relatives (ESR). A la différence des projets de REN mentionnés précédemment, notre analyse se fonde sur un corpus oral spontané. 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 au contexte oral. Dans l'optique de faire l'analyse de la perception des entités spatiales identifiées dans notre corpus, nous privilégions la constitution d'un nouveau système de reconnaissance des lieux qui préparera directement l'analyse de la perception de ces mêmes lieux.

[2] http://www-nlpir.nist.gov/related_projects/muc/
[3] http://nlp.stanford.edu/software/CRF-NER.html
[4] http://polyglot.readthedocs.io/en/latest/NamedEntityRecognition.html

3 Présentation des données exploitées

3.1 Le projet ESLO

Le corpus traité est le corpus ESLO (Eshkol-Taravella et al. 2012). Ce corpus comprend différentes situations d'enregistrement (entretiens face à face, interviews de personnalités, enregistrements dans des cours de récréation, pendant des repas, etc.).

Chaque enregistrement est transcrit orthographiquement avec une distinction entre les tours de parole sans signe de ponctuation, les points d'interrogation et les majuscules pour les noms propres étant les seules exceptions admises.

3.2 Modules exploités

Deux modules du corpus ESLO2 ont été sélectionnés : Entretiens et Itinéraires.

Lors de l'entretien, le locuteur, un habitant d'Orléans ou de son agglomération, est amené à faire état de son histoire personnelle, à partager ses habitudes de vie, etc. Les locuteurs témoins expriment ainsi à travers les entretiens leur perception de la ville d'Orléans. Au total, le module Entretiens d'ESLO2 comprend 84 transcriptions pour un total de 150h et environ 1 166 660 mots.

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. La collecte a été effectuée dans divers endroits de la ville afin d'interroger des locuteurs représentatifs de la diversité sociologique de la ville. De par leur constitution, ces courts enregistrements forment un matériel très riche en mentions de lieux relatives à la ville d'Orléans. Au total, le module Itinéraires d'ESLO2 comprend 91 transcriptions qui représentent 5h d'enregistrements et environ 69 330 mots.

Ces deux modules constituent une ressource riche en mentions de lieux relatives à Orléans. Bien qu'elle servira à entraîner notre système de détection automatique des mentions de lieux, notre méthode reste généraliste dans le sens où elle doit être applicable sur l'ensemble du corpus ESLO.

4 Méthode

Dans l'objectif de repérer les lieux présents dans l'oral transcrit, notre démarche (cf. Figure 1) comprend 4 étapes suivantes : (1) Constitution d'un corpus d'entraînement pour l'exploration manuelle des données, (2) Création de convention d'annotations, (3) Constitution d'un corpus de référence et (4) Expérience d'annotation automatique des lieux.

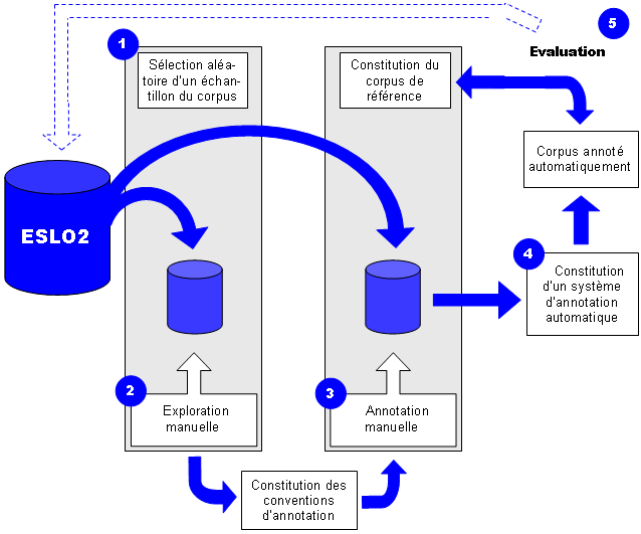

FIGURE 1 : Démarche globale pour le repérage des lieux

4.1 Exploration manuelle du corpus

Une dizaine de transcriptions issues des modules Entretiens et Itinéraires ont été sélectionnées pour l'examen manuel des types de noms de lieux présents dans le corpus ce qui a permis d'établir une première typologie des lieux (Eshkol-Taravella et Flamein, à paraître) et de constituer des conventions d'annotation en adéquation avec notre corpus.

Les noms de lieux peuvent eux-mêmes être désignés de manières très variées et parfois personnelles. Ils peuvent être abrégés, tronqués, transformés par le locuteur. Un même lieu peut aussi être mentionné sous différents noms sans que cela n'en gêne le sens comme dans :

1. *ce bout de d'île euh qui est euh pas très loin du pont euh euh du* ***pont George Cinq*** *oui c'est* ***le pont Royal*** (ESLO2_ENT_1034_C)

Pont George Cinq est le nom officiel actuel de l'un des ponts d'Orléans. Il a été inauguré en 1763 sous le nom de *Pont Royal*. Même s'il a changé plusieurs fois de noms entre temps, il est très régulier que les Orléanais se réfèrent à celui-ci en utilisant son ancien nom. L'enjeu de notre travail est de détecter toutes les mentions de lieux présentes dans le corpus tout en prenant en compte la variation dans leur dénomination. Le système de l'annotation automatique doit être capable de lier une entité nommée, nom officiel du lieu, à ses possibles variantes pour permettre de rendre géolocalisable sur une carte chacun des lieux identifiés, qu'ils soient mentionnés via leurs noms officiels ou via une variante de celui-ci. La prise en compte de ses éléments se traduit directement dans la convention d'annotation présentée dans la partie suivante.

4.2 Conventions d'annotation des mentions de lieux

Pour annoter l'ensemble des mentions de lieux présentes dans le corpus, nous utilisons la balise XML : <loc> … </loc> à laquelle nous ajoutons trois attributs. Ces attributs préciseront la nature des lieux identifiés et serviront de guide à l'analyse future de la perception.

4.2.1 Typologie des lieux

La convention d'annotation proposée s'inspire de celle des campagnes ESTER2[5] et ETAPE[6] (Rosset, Grouin et Zweigenbaum, 2011). ETAPE a pour objectif d'évaluer les performances des technologies vocales appliquées à l'analyse de flux télévisés en langue française tandis qu'ESTER2 est un projet antérieur à ETAPE avec des objectifs similaires de mesure de performances de systèmes de transcriptions d'émissions radiophoniques.

En vue de l'analyse de l'image d'une ville, il nous est nécessaire de conserver une typologie fine des lieux urbains. Une finesse que nous ne conservons pas pour la description des lieux naturels, extra-urbain. Notre convention reprend les codes de la convention Quaero en ce qui concerne les lieux que l'on peut découper administrativement (quartiers, villes, pays…). Les lieux géographiques seront considérés dans leur ensemble comme dans la typologie de la convention ESTER2.

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

TABLE 3 : Nouvelle typologie des lieux (1/2)

Les lieux considérés comme des constructions humaines et les organisations sont des entités souvent difficiles à distinguer. Le contexte d'emploi influence la catégorie dans laquelle classer l'entité observée. Ainsi dans les exemples [2] et [3] :

[5] http://www.afcp-parole.org/camp_eval_systemes_transcription/
[6] http://www.afcp-parole.org/etape.html

2. *alors euh je suis étudiante à l'**université euh d'Orléans La Source***

3. *comme je vais à la à l'**université** à La Source ça me fait quand même de la route* (ESLO2_ENTJEUN_03_C)

Le même locuteur mentionne à deux reprises *l'université d'Orléans* et le quartier *La Source*. Dans l'exemple [2], l'université est clairement désignée en tant qu'organisation alors que dans l'exemple [3], il s'agit d'un lieu. Que ce soit dans les conventions proposées par ESTER2 ou Quaero, l'*université d'Orléans* serait catégorisée comme une construction humaine dans l'exemple [3]. Aucune distinction n'est prévue au niveau du type de construction. L'université, la mairie ou toute autre structure commerciale sont catégorisées de la même façon. Pour une analyse plus fine de la perception des lieux de la ville d'Orléans, nous souhaitons préciser la nature des constructions humaines identifiées. Pour ce faire, nous nous référons à la typologie des entités nommées de type organisation définie dans les conventions ESTER2 et Quaero.

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

TABLE 5 : Typologie des entités nommées de type organisation selon ESTER2

A partir du contexte d'emploi, le système développé devra déterminer s'il a affaire à un lieu ou à une organisation. Il pourra s'appuyer sur la présence de verbes de déplacements ou sur l'association de certaines prépositions comme *à* ou *vers* et de certains verbes intrinsèquement liés à la notion de lieux : *habiter, naître, vivre…* Dans les cas où il détectera une organisation, aucune annotation ne sera faite. Si une construction humaine est identifiée, celle-ci sera annotée et considérée comme un lieu auquel on attribut une fonction. La typologie des fonctions possibles des lieux reprend celle des organisations dans la convention ESTER2 (cf. table [5]).

<loc type=" ">	
Lieux à dimension 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,…	

TABLE 6 : Nouvelles typologie des lieux (2/2)

Notre typologie conserve les catégories principales proposées par ESTER2 et les constructions humaines sont typées de façon similaire aux organisations (cf. table [6]). Toutefois, selon les conventions d'ESTER2, le type « politique » représente les organisations à caractères politiques 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. Nous ne conservons pas ce type puisque nous considérons que les organisations avec une fonction politique. Si des lieux à fonction politique sont évoqués, ils seront plutôt inclus avec le type « admin » de notre convention.

En conclusion, si l'on reprend l'exemple [3], le quartier de la Source et l'université d'Orléans seront annotés selon notre convention de la façon suivante :

3. *comme je vais à la à l'***<loc type="educatif" zone="2" label="université d'Orléans">*université*</loc>** *à* **<loc type="ville" zone="2" label="Orléans-la-Source">***La Source***</loc>** *ça me fait quand même de la route* (ESLO2_ENTJEUN_03_C)

4.2.2 Zone géographique

Trois zones géographiques sont distinguées dans l'annotation. Elles correspondent aux découpages administratifs de la ville d'Orléans et de son agglomération.

<loc type=" " zone=" ">	
zone ="0"	lieux hors agglomération orléanaise
Paris, Tours, Indre, Bretagne, Rhône, Seine …	
zone ="1"	lieux hors Orléans mais inclus dans l'agglomération
Saint Jean de la Ruelle, Saran, Auchan…	
zone ="2"	lieux situé à Orléans
Orléans, rue de Bourgogne, Key-West…	

TABLE 7 : Zone géographique

L'information de la zone géographique permet des traitements différents entre les annotations. Par exemple, un lieu considéré hors agglomération orléanaise n'aura pas à être géoréférencé sur la carte finale.

4. *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_C)

4.2.3 Label officiel

L'attribut du label officiel correspond au nom officiel du lieu identifié. Cet attribut répond à la capacité à varier du nom du lieu. Lorsqu'un locuteur mentionne un lieu, il peut se l'approprier, le personnifier en opérant des modifications sur son nom (troncations, utilisation de surnom…). Le nom officiel du lieu correspond à sa forme complète, sans aucune modification du locuteur.

5. *ah ben si tu peux redescendre tu prends la tu prends la rue qui est là et tu vas tout au bout jusqu'à la* **<loc type="voie" zone="2" label="rue de la République">*rue de la Rép-*</loc>** *tu vois où elle est ? la* **<loc type="voie" label=" rue de la République">*rue de la République*</loc>** *? (ESLO2_iti_06_11_C)*

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_C)*

Dans les exemples [3] et [4] les lieux *rue de la République* et *La Ferté-Saint-Aubin* ont été tronqués par le locuteur (*rue de la Rép-* et *La Ferté*). L'attribut label prend pour valeur la forme complète du nom de ces lieux.

L'intérêt de faire figurer le nom officiel dans la balise du lieu réside dans notre volonté de géoréférencer les lieux identifiés. Cette information servira à rechercher dans une base de données les coordonnées géographiques du lieu pour le placer sur la carte finale.

4.3 Annotation manuelle

Les conventions d'annotation définies, nous procédons à l'annotation maneulle du corpus pour constituer un premier corpus de référence. Le corpus de référence est constitué de 4 transcriptions aléatoirement sélectionnées parmi les 91 transcriptions du module Itinéraires différentes de celles analysées dans la première étape d'observation. Ce module a été privilégié puisque plus concentré en mentions de lieux proportionnellement à sa taille par rapport aux transcriptions du module Entretien. Cette concentration offre un gain de temps et d'efficacité au moment de l'annotation manuelle.

Transcriptions	Durée	Nombre de mots
ESLO_iti_02_01	02:39	570
ESLO_iti_02_03	03:20	704
ESLO_iti_02_05	04 :01	893
ESLO_iti_02_06	12:23	2815
Totaux	22:23	4982

TABLE 8 : Volume de données par transcriptions du corpus de référence

Afin de tester l'efficacité de la convention d'annotation, nous avons procédé à une évaluation de l'accord inter-annotateur avec la mesure du Kappa de Cohen. Cette évaluation concerne en particulier le calcul d'accord sur le type de lieu choisi par les annotateurs lorsqu'ils ont détecté le même lieu. Dans le cadre d'un cours de master, notre propre version annotée (A1) du corpus présenté en table [8] a été comparé à celle d'une étudiante (A2). Le Kappa de Cohen permettra de montrer l'accord des deux juges en ce qui concerne le choix des types de lieux.

Pour procéder au calcul, nous utilisons l'application mise à disposition librement en ligne par Philippe Bonnardel[7]. En suivant la procédure, nous obtenons un Kappa égal à 0.810. Si l'on se rapporte à la grille de lecture proposée par (Landis et Koch, 1977), l'accord est considéré comme

[7] http://kappa.chez-alice.fr/

excellent. Cette première évaluation montre un accord élevé entre les annotateurs mais laisse paraître quelques désaccords. L'évaluation de l'annotation s'est ensuivie d'une phase d'adjudication durant laquelle chaque désaccord a été analysé. Par exemple, dans :

7. 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*

A1 et A2 sont en désaccord sur la façon de catégoriser le lieu *mairie d'Orléans*. Même si le bâtiment abritant la mairie peut avoir un intérêt touristique, nous considérons que la fonction administrative prévaut sur l'attrait touristique de la structure. Le type "admin" sera donc privilégié.

Après comparaison des deux annotations et résolution des désaccords, nous avons pu aboutir à une version consensuelle. Cette troisième version annotée constitue notre corpus de référence. Ce corpus annoté sera utile pour l'évaluation de notre système.

4.4 Expérience d'annotation automatique

4.4.1 Méthodologie

Pour repérer des lieux, nous utilisons les méthodes heuristiques prenant en compte leur contexte d'emploi dans le discours. Le système d'annotation est en cours d'élaboration.

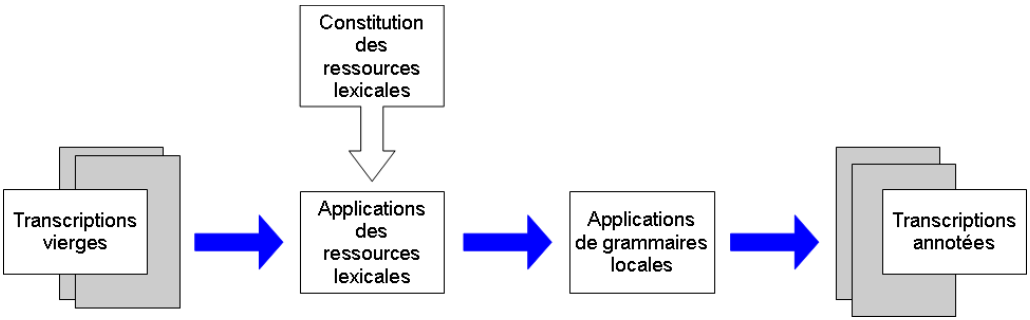

FIGURE 2 : Chaîne de traitement automatique du corpus

L'annotation des lieux dans le corpus repose dans un premier temps sur l'utilisation de ressources lexicales dédiées à l'information géographique. En complément de ces ressources, différentes règles sont appliquées afin d'identifier toutes les mentions de lieux absentes des listes.

4.4.1.1 Ressources utilisées

Les ressources utilisées sont constituées à partir de bases de données disponibles en ligne et dédiées à l'information géographique comme GEOFLA®[8] qui contient la liste de tous les noms de voies de circulation, de communes, de départements, et de régions en France. Ces bases de données géographiques permettront l'identification de tous les lieux mentionnés par leur nom officiel, sans modification apportées par le locuteur. De plus, elles pourront renseigner l'attribut de la zone géographique à renseigner dans l'annotation.. Ces bases de données mettent aussi à disposition des coordonnées géographiques de chaque lieu répertorié. Grâce à cette information, nous serons capables de placer sur une carte tous les lieux identifiés dans les transcriptions analysées.

4.4.1.2 Variations des noms de lieux

Le nom d'un lieu peut varier et peut être mentionné sous un nom complètement différent du son nom d'origine.

Pour gérer ces variations dans la désignation des lieux, nous utilisons notamment la distance de Levenshtein, dans la lignée de Zesnani et al. (2005), qui l'avaient utilisée pour l'enrichissement d'un dictionnaire d'entités spatiales. Cette distance mathématique permet de mesurer la similarité entre deux chaînes de caractères et est exploité dans le projet pour gérer les cas d'abréviation des noms de lieux comme dans les exemples (3) et (4). C'est la distance de Levenshtein qui doit permettre d'identifier *La Ferté* comme la version abrégée de *La Ferté-Saint-Aubin*, nom renseigné dans la base.

Cette méthode connaît toutefois des limites. Pour décider si les deux chaînes comparées renvoient à la même entité, il faut définir un seuil afin de déterminer si la distance obtenue est trop grande ou non. Ce seuil est généralement défini en fonction de la longueur de la chaîne observée. En l'état, la distance amène autant de bonnes détections que de bruits dans l'annotation. Si l'on reprend l'exemple de la chaîne de caractères *La Ferté* comparée à la chaîne *La Ferté-Saint-Aubin*, la distance de Levenshtein vaut 12. La différence entre les deux chaînes est supérieure à la taille de la chaîne *La Ferté* (8 caractères). Le seuil définit dans le système demandera automatiquement le rejet de *La Ferté* comme une variante de *La Ferté-Saint-Aubin* alors que celle-ci est pertinente.

Pour contrer cette difficulté, il est nécessaire d'ajouter d'autres règles pour compléter l'analyse. L'une des premières pistes pour pallier à ce problème est l'adaptation même de la distance pour qu'elle ne raisonne plus en nombre de caractères différents mais plutôt en nombre de mots. La combinaison de la distance de Levenshtein classique et de sa version adaptée pourrait permettre d'améliorer la qualité de l'annotation. Dans le cas de *La Ferté* et *La Ferté-Saint-Aubin*, la distance en mots serait de 2 et l'utilisation classique de la distance pourra déterminer que deux mots sont identiques dans les deux chaînes. Ces deux informations pourront permettre de valider *La Ferté* comme une variante de *La Ferté-Saint-Aubin*.

Cependant, le seuil à définir doit rester suffisamment strict afin de limiter, voir éliminer le bruit. Les conclusions de la distance devront être confirmés ou infirmés par d'autres règles basées sur le contexte ou sur la nature des mots constituant l'entité observée.

[8] Base de données mise à disposition librement par l'Institut National de l'Information Géographique et Forestière (IGN) - http://professionnels.ign.fr/geofla

4.4.1.3 Evaluation du système

L'évaluation du système n'a pas encore été réalisée puisque le système d'annotation est encore en cours d'élaboration. Si l'annotation des noms de lieux simples (c'est-à-dire les entités nommées sans variations) sont déjà identifiées grâce aux ressources lexicales, beaucoup d'autres mentions comme les noms de lieux génériques ou les noms soumis à variations ne le sont pas encore. Il est nécessaire de continuer le développement de l'outil pour proposer une évaluation pertinente de celui-ci.

5 Perspectives

5.1 Poursuite de l'élaboration du système

Comme évoqué précédemment, il est nécessaire d'ajouter d'autres règles dans le système afin d'améliorer notamment l'efficacité de la distance de Levenshtein mais surtout pour permettre l'identification des les lieux absents des bases de données utilisées. Ce problème concerne en particulier les lieux avec une activité commerciale ou les lieux publics pour lesquels nous ne disposons pas de listes similaires aux bases de GEOFLA®. Ces lieux font partie intégrante du quotidien des locuteurs interrogés. On peut alors supposer que les locuteurs vont plus naturellement agir sur le nom du lieu auquel ils font référence. Les phénomènes les plus courants sont l'abréviation ou l'utilisation de variantes du nom officiel. L'enjeu est ici de reconnaître le lieu sans distinction du procédé utilisé pour le mentionner. La difficulté principale dans le repérage est de tenir compte de toutes les variations possibles et de tenir compte de la coréférence entre plusieurs mentions du même lieu. Il sera nécessaire de constituer des lexiques dédiés à l'information géographique en listant par exemple tous les noms communs représentants des lieux. Ces listes ou dictionnaires pourront être combinés à des règles syntaxiques prenant en compte le contexte d'emploi des lieux.

Il faudra aussi être capable de résoudre les cas d'homonymie provoquée par l'abréviation d'un nom de lieu. Par exemple, l'agglomération orléanaise englobe notamment les villes Saint-Jean-de-la-Ruelle, Saint-Jean-de-Braye et Saint-Jean-le-Blanc. commeDans l'exemple :

8. *EW15: #1 oui c'était un à **Saint Jean** #*
   *ch_AC7: #2 c'était où à ? # à **Saint Jean-le-Blanc** ?*
   *EW15: **Saint Jean-le-Blanc** #1 oui #*
   ch_AC7: #2 ah oui
   (ESLO2_ENT_1015_C)

à laquelle de ces trois villes renvoie la mention *Saint Jean* ? Il sera nécessaire de prendre en compte le contexte d'énonciation et les tours de parole précédents afin de désambigüiser ce type de cas. On peut supposer qu'avant d'utiliser la forme abrégée du nom de la ville, le locuteur aura utilisé la forme complète afin de s'assurer que son interlocuteur le comprenne. En interrogeant le reste de la transcription et en associant la recherche à la distance de Levenshtein, le système pourrait identifier le nom de ville le plus proche en termes de position dans la transcription et d'orthographe. Dans l'exemple [7], *Saint Jean-le-Blanc* serait alors identifié comme la forme complète de *Saint Jean*.

5.2 Module de géoréférencement

L'objectif final du travail est de représenter les lieux et les avis qui les concernent sur une carte. Dans cette optique, nous allons intégrer à notre système un module de géoréférencement. Nous pourrons nous appuyer sur les ressources lexicales constituées et les bases de données dédiées à l'information géographique dans lesquelles sont renseignées ces coordonnées. Les coordonnées seront incluses dans un Système d'Information Géographique et chacun des lieux d'Orléans ou de son agglomération pourront être placés les uns par rapport aux autres sur la carte. Il ne sera pas pertinent de représenter les lieux en dehors de ces zones, d'où la nécessité d'obtenir cette information au moment du repérage.

5.3 Analyse de la perception des lieux

Le repérage des lieux est la première étape de notre travail. A partir de ce repérage, nous serons en mesure de faire l'analyse de la perception de ces lieux. La dénomination d'un lieu est un processus social réapproprié subjectivement et est déterminée par la personnalité, l'histoire du locuteur. Un lieu peut être apprécié, ou non, par le locuteur. Au niveau de l'analyse des variations dans les noms de lieux, nous pouvons nous rapprocher de la notion des lieux subjectifs développée par Dominguès et Eshkol (2013, 2015). Mentionner un lieu n'est pas un processus neutre et les moyens pour exprimer ses sensations à son sujet sont multiples comme cela a été démontré dans Eshkol-Taravella et Flamein (à paraître).

6 Conclusion

Nous proposons une démarche pour la création d'un module de reconnaissance automatique des lieux qui se veut généraliste puisqu'il est applicable sur toutes les transcriptions du corpus ESLO. Les perspectives du travail présenté est d'offrir une visualisation cartographique de la perception des Orléanais de leur lieu de vie. Pour ce faire, nous avons défini une convention d'annotation ce qui nous a permis de constituer un corpus de référence dédié à l'évaluation de notre système d'annotation automatique.

A terme, les expressions subjectives identifiées dans les transcriptions analysées formeront la vision que des Orléanais ont de leur propre ville. Cette vision sera comparée avec celle proposée par les données issues des bases de données Linked Open Data comme Wikipédia. Le fait de prendre en compte les témoignages d'ESLO et d'y associer l'ensemble des informations neutres et objectives des bases de données donnera une dimension anthropologique et sociologique à la carte produite et permettra de constituer un véritable portrait de la ville d'Orléans.

Références

ABEILLE A., CLEMENT L. & TOUSSENEL F. (2003). Building a treebank for French. In *Anne Abeillé*, éditeur : Treebanks, pages 165 –187. Kluwer, Dordrecht.

BRANDO C., DOMINGUES C., CAPEYRON M. (2016). Evaluation of NER systems for the recognition of place mentions in French thematic corpora, In: *Proceedings of the 10th Workshop on Geographic Information Retrieval (GIR '16)*. ACM, New York, NY, USA, article 7, 10 p ages DOI: 10.1145/3003464.3003471

COHEN J. (1960). A coefficient of agreement for nominal scales., *Educ. Psychol. Meas*., 20, 27-46.

DOMINGUES C., ESHKOL-TARAVELLA I. (2015). Toponym recognition in custom-made map titles. *International Journal of Cartography*, Volume 1, Taylor & Francis.

DOMINGUES C., ESHKOL-TARAVELLA I. (2013). Repérer des toponymes dans les titres de cartes topographiques. *TALN2013*, Jun 2013, Les Sables d'Olonne, France. 2013,

EHRMANN M. (2008). *Les entités nommées, de la linguistique au TAL : statut théorique et méthode de désambiguïsation.* PhD thesis, Université Paris 7

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C. & TELLIER I. (2012). Un grand corpus oral « disponible » : le corpus d'Orléans 1968-2012. in *Ressources linguistiques libres*, *TAL*. (vol. 52, n° 3, p. 17-46).

ESHKOL-TARAVELLA I., FLAMEIN H. (à paraître)

FORT K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : vers une méthodologie de l'annotation manuelle de corpus. Traitement du texte et du document*. Université Paris-Nord - Paris XIII, 2012. Français.

KERGOSIEN E., MAUREL P., ROCHE M., TEISSEIRE M. (2015). Senterritoire pour la détection d'opinions liées à l'aménagement d'un territoire. *Revue Internationale de Géomatique*, Hermes, 25 (1), pp.11-34.

LANDIS J.R., KOCH G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, (33) :159–174, 1977.

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

MAUREL D., FRIBURGER N., ANTOINE J.-Y., ESHKOL-TARAVELLA I. & NOUVEL D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *TAL*, (vol. 52, n° 1, pp. 69-96).

NADEAU N., SEKINE S. (2009). A survey of named entity recognition and classification. In *S. Sekine & E. Ranchhod (eds.), John Benjamins publishing company*, Amsterdam, (pp. 3-28).

FINKEL J. R., GRENAGER T. & MANNING C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

AL-RAFOU R., KULKARNI V., PEROZZI B. & SKIENA S. (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.

ROSSET S., GROUIN C., ZWEIGENBAUM P. (2011). *Entités Nommées Structurées : guide d'annotation Quaero*. Technical report.

ZENASNI S., KERGOSIEN E., ROCHE M, TESSEIRE M. (2016). Extracting new Spatial Entities and Relations from Short Messages, In the *8th International ACM Conference on Management of Digital EcoSystems (MEDES'2015)*, pp. 8, Hendaye (France).