Humanités numériques

3 | 2021
Humanités numériques spatialisées

Exploitation du corpus *Enquêtes sociolinguistiques à Orléans* (*ESLO*) par les outils du traitement automatique des langues et de la géomatique

Exploiting the Enquêtes sociolinguistiques à Orléans (ESLO) *Corpus through Natural Language Processing and Geomatics Tools*

Hélène Flamein et Iris Eshkol-Taravella

Édition électronique
URL : https://journals.openedition.org/revuehn/1911
DOI : 10.4000/revuehn.1911
ISSN : 2736-2337

Éditeur
Humanistica

Référence électronique
Hélène Flamein et Iris Eshkol-Taravella, « Exploitation du corpus *Enquêtes sociolinguistiques à Orléans* (*ESLO*) par les outils du traitement automatique des langues et de la géomatique », *Humanités numériques* [En ligne], 3 | 2021, mis en ligne le 01 mai 2021, consulté le 16 juillet 2021. URL : http://journals.openedition.org/revuehn/1911 ; DOI : https://doi.org/10.4000/revuehn.1911

humanités
numériques

3 | 2021
Humanités numériques spatialisées
INFORMATION SPATIALE ET CORPUS TEXTUELS

Exploitation du corpus *Enquêtes sociolinguistiques à Orléans* (*ESLO*) par les outils du traitement automatique des langues et de la géomatique

Exploiting the Enquêtes sociolinguistiques à Orléans (ESLO) Corpus through Natural Language Processing and Geomatics Tools

Hélène Flamein et Iris Eshkol-Taravella

Résumés

Les travaux présentés dans cet article s'intéressent à l'exploitation du corpus oral *Enquêtes sociolinguistiques à Orléans* (*ESLO*), qui offre un intérêt particulier pour les humanités numériques spatiales car il part d'une enquête menée par des linguistes orléanais auprès des habitants de leur ville. Le corpus *ESLO* est traité avec des outils du traitement automatique des langues (TAL) et de la géomatique, qui permettent d'en extraire des renseignements sur l'espace et sa perception afin de les représenter sur une carte de la ville.

The work presented here focuses on the exploitation of the *ESLO* oral corpus (*Enquêtes sociolinguistiques à Orléans*, "Sociolinguistic Surveys in Orleans"). This corpus is especially relevant for the digital spatial humanities because it is conducted by linguists from Orléans among the inhabitants of their city. The corpus is processed using Natural Language Processing (NLP) and geomatics tools to extract information about space and its perception in order to represent it on a map of the city.

Entrées d'index

MOTS-CLÉS : humanités numériques spatialisées, linguistique et sciences du langage, corpus oral, annotation, traitement automatique des langues, système d'information géographique

KEYWORDS: spatial digital humanities, linguistics, oral corpus, annotation, natural language processing, geographic information system

Introduction

1 L'évolution constante des nouvelles technologies et des outils à notre disposition multiplie et diversifie les usages et les attentes des utilisateurs vis-à-vis des données auxquelles ils sont confrontés. 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 des données disponibles aujourd'hui. Le traitement de ces données multimodales ainsi que la gestion de leur quantité incommensurable sont des enjeux dans différents domaines dont le traitement automatique des langues (TAL). Il existe de véritables besoins en matière de traitement automatique des masses de données : gagner en rapidité de traitement tout en gérant la quantité et la variété de la nature des données.

2 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. Quel que soit le domaine d'application ou le but recherché, il est nécessaire d'adopter une démarche pluridisciplinaire pour répondre à ces enjeux. Cette idée se retrouve dans la définition même des humanités numériques donnée par Piotrowski et Xanthos (2020) qui considèrent l'objet de recherche comme « la construction de modèles pouvant être manipulés par l'ordinateur, c'est-à-dire des modèles formels, dans le but de soutenir et faire progresser la recherche en SHS ».

3 Si les outils d'aide à l'analyse de corpus atteignent aujourd'hui une certaine maturité, ils 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. C'est ce que propose le travail décrit dans cet article fondé sur l'exploitation du corpus linguistique *Enquêtes sociolinguistiques à Orléans* (*ESLO*), une ressource authentique apportant des connaissances nouvelles liées aux représentations et à la perception de l'espace de vie par des locuteurs francophones. Pour pouvoir exploiter une telle ressource, nous avons eu recours aux outils du TAL et de la géomatique permettant de modéliser, d'extraire et d'analyser des informations spatiales avec le contenu subjectif qui leur est relatif.

Données exploitées

4 Les données analysées dans cet article proviennent du corpus *Enquêtes sociolinguistiques à Orléans* (*ESLO*[1]) [Abouda et Baude 2006 ; Eshkol-Taravella *et al.* 2011 ; Baude et Dugua 2011], programme du Laboratoire ligérien de linguistique (LLL, UMR 7270) de l'université d'Orléans, qui met au cœur de son investigation les pratiques langagières dans la ville d'Orléans. 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. Le corpus *ESLO* est une ressource riche tant sur le plan qualitatif que quantitatif. Ce corpus linguistique, composé d'enregistrements sonores et de leurs transcriptions réalisés à Orléans entre 1968 et 1974 (*ESLO*1) et à partir de 2008 (*ESLO*2), cumule plus de 700 heures d'enregistrement. Ce travail s'appuie particulièrement sur *ESLO*2 dont les enregistrements se répartissent entre 18 modules représentant chacun une situation d'enregistrement différente : interviews d'habitants et de personnalités de la ville, captations de paroles dans la rue, les transports publics, les commerces, les lieux de travail, etc. Chaque enregistrement du corpus est transcrit orthographiquement selon les recommandations du *Guide du transcripteur et du relecteur des ESLO*[2].

5 Par son mode de constitution, Orléans est le fil conducteur du corpus *ESLO*. Chacune des situations captées met en avant différents aspects de la ville. Les locuteurs parlent de ce lieu et d'une manière plus générale, de leur vie dans ce lieu. L'enjeu principal de notre travail est d'offrir aux utilisateurs une nouvelle manière d'accéder à ce corpus, et plus particulièrement à la perception qu'ont les Orléanais de leur ville. 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.

6 Pour constituer le corpus, deux modules d'*ESLO*2 ont été privilégiés : Entretiens et Itinéraires (notés « ENT » et « ITI », respectivement, dans les exemples). Le module Entretien est une collection de discussions en face-à-face entre un enquêteur et un locuteur témoin, habitant à Orléans ou dans sa région proche. L'enquêteur mène la discussion et invite son interlocuteur à retracer son histoire personnelle, à partager ses habitudes de vie, etc. Le module Itinéraires regroupe, quant à lui, des enregistrements réalisés dans les rues d'Orléans. Les enquêteurs vont à la rencontre de piétons pour demander leur chemin vers certains endroits connus de la ville comme la gare, la mairie ou encore la piscine. La collecte a été effectuée dans divers points de la ville afin d'interroger des locuteurs en tenant compte 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.

Problématique de la dénomination d'un lieu

7 La notion de lieu est centrale dans notre réflexion. Le *Trésor de la langue française informatisé* (*TLFi* [3]) 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 lieu se définit en tant qu'espace qui occupe une certaine étendue et que l'on peut mesurer et situer par rapport à ce qui l'entoure. Cette propriété offre la possibilité d'allouer des coordonnées géographiques au lieu afin de le placer sur un plan, de le localiser dans un espace.

8 Le lieu participe à la définition de l'identité de chacun, idée centrale du domaine de la géographie humaine. Selon Clerc (2004), l'homme se définit par son environnement, par son appartenance spatiale et, à l'inverse il donne une identité, « et même plus fondamentalement une existence, au lieu ». Ainsi, le lieu est caractérisé par les interactions de l'homme avec son environnement. Frémont (1980) introduit une notion de l'espace vécu, 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 appropriation transparaît dans le nom du lieu qui n'est pas une simple étiquette sur un espace : ce nom symbolise l'histoire du lieu, les enjeux qui lui sont propres. Un même lieu peut donc être nommé de différentes façons en fonction du locuteur, de son histoire et de sa perception.

9 Nommer un lieu est d'abord un processus institutionnel qui dépend des procédures décrites par Bouvier (1999) :

- La dénomination résultant 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'Église, place du Marché, etc. Ce sont les toponymes d'usage, dominants, sinon exclusifs, en dehors des espaces agglomérés.
- La dénomination provenant de 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. Cette dénomination s'impose en principe à l'usage des habitants et de tous les utilisateurs du lieu : rue Victor-Hugo, rue Richelieu, place de la Concorde, etc. Il s'agit des toponymes créés, très rares, sinon inexistants en dehors des agglomérations.

10 Cependant, ces dénominations officielles de lieux ne sont pas toujours utilisées par les locuteurs. La variation de la norme dans la dénomination des lieux dépend de la situation dans laquelle se produit le discours. Dans l'exemple 1,

> [Exemple 1 : ESLO2_ENT_1034_C] « ce bout de d'île euh qui est euh pas très loin du pont euh euh du pont George V oui c'est le pont Royal »,

11 « Pont George V » 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 Orléa-

nais se réfèrent au pont en utilisant son ancien nom.

12 Une modalité (oral/écrit) peut aussi influencer la dénomination d'un lieu. Si l'on prend le cas du corpus *ESLO*, où les locuteurs s'expriment librement à l'oral, plusieurs phénomènes apparaissent.

13 Les noms de villes peuvent être abrégés comme Saint-Pierre-des-Corps qui devient Saint-Pierre. La troncation est un autre phénomène qui peut intervenir sur différents types de noms de lieux comme les noms de ville (Saint-Étienne peut devenir Saint-É) ou comme dans l'exemple 2 dans lequel le quartier Saint-Euverte devient Saint-Eu.

> [Exemple 2 : ESLO2_ITI_10_03] « ça va em- vous emmène jusqu'au coin de la centre de conférence ce quartier Saint-Eu »

14 Les noms de lieux peuvent aussi être amorcés comme « école d'éducateurs spécialisés » dans l'exemple 3 :

> [Exemple 3 : ESLO2_ENT_1003] « à Orléans y a p- et en plus c'est dommage parce qu'à Olivet y a une école de d'éduca- euh d'éducateurs spécialisés de euh je sais plus moniteur euh je sais plus comment ça s'appelle »

15 Le locuteur hésite au moment de nommer l'école à laquelle il se réfère, ce qui provoque la troncation du terme éducateur et fait apparaître l'onomatopée « euh ». Il se reprend ensuite pour donner la forme complète « école d'éducateurs spécialisés ».

> [Exemple 4 : ESLO2_ITI_10_04] « y a oui si c'est Sainte-Croix quoi la cathédrale »

16 L'instantanéité du discours oral a une autre influence sur la façon dont un locuteur mentionne un lieu. Puisque l'oral s'élabore au fil de la pensée, l'ordre syntaxique peut facilement se trouver perturbé. Le lieu « la cathédrale Sainte-Croix » est mentionné dans l'exemple 4 par le locuteur dans l'ordre inverse « Sainte-Croix la cathédrale », ce qui se produit rarement dans le discours écrit. De plus, l'expression est entrecoupée par le marqueur discursif « quoi ». Le locuteur utilise les éléments du nom complet « la cathédrale Sainte-Croix » mais les donne dans le désordre. Il y fait allusion d'abord par le nom propre et précise ensuite le type du monument.

> [Exemple 5 : ESLO2_ITI_07_01] « on regardait notre cathédrale tous les deux »

17 Les noms de lieux peuvent être désignés de manière personnelle. Cette idée est abordée dans les travaux de Dominguès et Eshkol-Taravella (2013, 2015) par le biais du concept de lieu subjectif. Dans l'exemple 5, le locuteur emploie 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 lui appartient. Le possessif est un indice de l'appropriation d'un lieu par le locuteur, et plus généralement, cet indice montre la perception que cet habitant a de la ville d'Orléans.

18 Le traitement des noms de lieux pose donc certaines difficultés. À chaque maillon de la chaîne de traitement proposée, les spécificités décrites sont prises en compte dans l'élaboration des modules d'annotation et d'extraction.

Modélisation et traitements pour le repérage automatique des lieux dans l'oral transcrit

Lieux et difficulté de leur détection

19 Les lieux sont abordés dans ce travail du point de vue du TAL où ils font partie des entités nommées, « des éléments informationnels pertinents dont on parle et qui jouent un rôle dans la description d'un événement, d'un fait » (Nouvel, Ehrmann et Rosset 2015 : 13) ou des entités spatiales (Lesbegueries 2007 ; Zenasni *et al.* 2016).

20 Que ce soit à l'écrit ou à l'oral, les traitements automatiques rencontrent plusieurs difficultés pour détecter les noms de lieux.

21 L'un des problèmes principaux réside dans leur polysémie. Un lieu peut renvoyer vers un autre référent comme dans le cas de la ville Orange qui peut aussi être une entreprise et un fruit. Il est donc nécessaire d'avoir recours au contexte pour pouvoir le désambiguïser : présence d'une majuscule, d'un verbe de mouvement, d'une préposition locative, etc. Les noms de lieux peuvent aussi être employés en tant que personne comme dans « La Région annonce des mesures exceptionnelles [4] ».

22 Au-delà des ambiguïtés sémantiques pouvant exister, différentes opérations concourent à faire varier le nom du lieu de la norme comme il a été montré dans la section précédente. Les usagers du lieu peuvent être tentés d'utiliser un nom différent de la norme pour désigner un lieu en fonction de leurs pratiques, de leurs habitudes ou d'une caractéristique propre à l'endroit. Que ce soit à l'oral ou à l'écrit, la nature morphologique des mots désignant un lieu varie. Les noms de lieux peuvent être composés d'un ou de plusieurs mots (place du Général-de-Gaulle, université d'Orléans…), être liés ou non par un trait d'union (le quartier Orléans-La Source ou Orléans La Source, Saint-Jean-de-la-Ruelle…), etc. Ces variations dans la dénomination d'un lieu ne sont pas toutes répertoriées dans les ressources toponymiques, ce qui complique leur détection.

23 La majorité des travaux en TAL reposent sur des données textuelles comme des tweets (Zenasni *et al.* 2016), des récits de voyages (Loustau, Gaio et Nodenot 2008), des œuvres littéraires (Moncla *et al.* 2016), des écrits du Web (Dominguès et Eshkol-Taravella 2013, 2015) ou des récits de vie (Dominguès *et al.* 2018). Le travail présenté ici est effectué sur un corpus oral composé de conversations spontanées.

24 L'objectif du traitement automatique des noms de lieux mentionnés dans le corpus *ESLO* est de pouvoir les extraire afin de les situer sur la carte de la ville [5].

Modélisation de l'information spatiale et son annotation

25 Avant de procéder au traitement des données, il est nécessaire de s'intéresser à la modélisation de l'information spatiale à identifier. Une étape d'observation manuelle du corpus a permis d'établir des conventions d'annotation. Ces conventions renseignent les informations suivantes :

- Le type du lieu afin d'expliciter son périmètre et de donner un premier renseignement sur le rôle qu'il peut revêtir. Par exemple, une entité identifiée comme une ville ne sera pas considérée de la même façon qu'une rue, un commerce ou une forêt. La typologie proposée est inspirée des conventions d'Ester [6] (Bonastre *et al.* 2008) et de Quaero [7] (Rosset, Grouin et Zweigenbaum 2011) : villes (Orléans, La Source), région (Loiret, Beauce), pays (France, Russie), supranational (Europe, Asie du Sud), voies (rue de la République, pont Royal), naturel (forêt d'Orléans, Loire), monument (cathédrale Sainte-Croix, hôtel Groslot), lieux administratifs (mairie d'Orléans, office du tourisme, CAF), lieux éducatifs (lycée Pothier, université d'Orléans), lieux commerciaux (Carrefour, H&M, Memphis Coffee) et non commerciaux (hôpital de la Source).
- La zone géographique, une information relative à la localisation géographique des lieux, pour différencier les lieux situés à Orléans (zone 2), dans son agglomération (zone 1) ou à l'extérieur (zone 0).
- Le label officiel qui permet de renseigner le nom normalisé d'un lieu lorsque le nom identifié diffère de la norme. Cette information permet d'anticiper 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.

26 Cette convention d'annotation permet d'annoter l'exemple 2 de la manière suivante :

> [Exemple 2 : ESLO2_ITI_10_03] « ça va em- vous emmène jusqu'au coin de la centre de conférence ce <loc type="voie" zone="2" label="quartier Saint-Euverte">quartier Saint-Eu-</loc> »

27 La forme tronquée « quartier Saint-Eu- » est annotée comme une voie, située en zone 2, soit à Orléans, et son label officiel est : quartier Saint-Euverte.

28 Un échantillon du corpus (15 transcriptions distinctes de celles utilisées lors de l'observation manuelle du corpus) a été annoté manuellement en accord avec des conventions d'annotation préétablies. Ce corpus représente un corpus de référence essentiel pour l'évaluation finale du système élaboré. L'enjeu est que le système d'annotation automatique parvienne à produire la même annotation.

29 L'annotation manuelle a été effectuée par deux annotatrices (i.e. deux étudiantes en linguistique). L'évaluation de cette annotation a été réalisée grâce au calcul de l'accord interannotateurs qui montre dans quelle mesure les annotatrices sont en accord ou en désaccord dans la tâche d'annotation des lieux en fonction de conventions d'annotation préétа-

blies. Pour mesurer cet accord, un kappa de Cohen (1960) a été utilisé. On obtient un score de 0,81 jugé excellent selon la grille de Landis et Koch (1977).

Détection automatique de l'information spatiale

30 Pour développer le système de la détection automatique des lieux, nous utilisons une approche symbolique fondée sur des lexiques et des règles élaborées grâce à une analyse approfondie des lieux et de leur dénomination. Comme évoqué précédemment, les noms de lieux peuvent être abrégés, tronqués, substitués par de nouvelles expressions, nommés d'une manière conventionnelle ou pas, etc. Toutes ces spécificités propres à l'oral ou au registre familier sont prises en compte dans le développement du système automatique.

31 Le module de détection automatique mis en place s'appuie sur des lexiques extraits de bases de données dédiées à l'information spatiale comme GEOFLA[8] et data.gouv.fr[9] qui référencent des noms de lieux normalisés, associés à des coordonnées géographiques permettant leurs géolocalisations sur une carte. En l'état, ces ressources ne suffisent donc pas pour l'annotation exhaustive des lieux. Pour détecter les noms de lieux non normalisés, les ressources lexicales sont enrichies en variantes de noms de lieux grâce aux règles construites spécifiquement pour cette tâche. Dans le cas des noms de lieux composés :

- Pour des noms de voie, les locuteurs ne conservent que le type de voie et le dernier mot composant le nom officiel : « place de Gaulle » au lieu de « place du Général-de-Gaulle », « rue Bannier » au lieu de « rue du faubourg-Bannier ».
- Pour des noms de villes, c'est plutôt le premier terme qui est conservé, alors que les derniers sont supprimés : « Saint-Cyr » pour « Saint-Cyr-en-Val » ou « La Chapelle » pour « La Chapelle Saint-Mesmin ».

32 De nouvelles entrées sont générées et ajoutées au lexique sur la base de ces observations.

33 D'autres noms de lieux, composés de noms communs comme rue de la gare, boulangerie du coin, etc. sont détectés en utilisant des règles et des patrons décrivant leur structure et les contextes de leur apparition (par exemple, un nom désignant un type de lieu « bar », « boulangerie », suivi d'un groupe prépositionnel).

34 Les lieux sont ainsi identifiés tels qu'ils sont nommés, c'est-à-dire sous la forme utilisée quotidiennement par les gens.

35 Finalement, le module développé est évalué à partir du corpus de référence et obtient une F-mesure de 0,91 avec un rappel de 0,90 et une précision de 0,93. Cette évaluation prend en compte l'annotation automatique des caractéristiques des lieux identifiés (type de lieu, zone géographique et label officiel) et leur délimitation.

36 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 6 dans lequel trois lieux sont cités (Orléans, la rue de la République et la rue Royale).

[Exemple 6 : ESLO2_ITI_03_01] « 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 »

37 Néanmoins, certains lieux restent non détectés comme le bateau restaurant L'Inexplosible mentionné dans l'exemple 7. Ces cas peuvent s'expliquer principalement par l'absence d'exhaustivité des ressources lexicales employées que l'on utilise pour leur détection.

[Exemple 7 : ESLO2_ENT_1042] « 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 »

38 Dans l'exemple 8, il est question du Campo Santo, un cloître accueillant régulièrement des événements organisés par la mairie d'Orléans. Ce nom n'est pas répertorié dans des ressources lexicales exploitées et il n'est pas relevé par le module développé.

[Exemple 8 : ESLO2_ENT_1042] « 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 »

39 La précision, qui représente le degré de pertinence de l'annotation, est jugée aussi satisfaisante (0,93). Le module produit une annotation de qualité dans laquelle la grande majorité des détections de lieux 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.

40 Cependant, certaines détections sont erronées. Dans l'exemple 8, si le système n'a pas identifié le Campo Santo, il a annoté Campo comme une ville en renvoyant à la commune corse Campo qui est référencée dans la base de données GEOFLA.

41 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.

42 Les résultats montrent des performances satisfaisantes quant à l'efficacité du module de détection de lieux et la gestion des difficultés liées aux variations des noms de lieux complexifiées par les spécificités de l'oral. L'étape de détection des lieux sert d'ancrage pour l'analyse de leur perception.

Détection de la perception

43 Partager sa perception d'un objet est un processus subjectif. La notion de perception est liée donc avec celle de subjectivité. On définit communément la subjectivité en opposition à l'objectivité. La subjectivité est définie dans le *TLFi* [10] 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 (1969), repris dans le *Grand Dictionnaire* de l'Office québécois de la langue française [11], 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 qui correspondrait à tout ce qui n'est pas objectif : 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.

44 Kerbrat-Orecchioni (2009, 125-130) définit plusieurs types de subjectivité :

- La subjectivité affective, qui considère les expressions qui montrent que le sujet d'énonciation est émotionnellement impliqué dans ce qu'il énonce.
- La subjectivité interprétative, qui s'intéresse au processus de dénomination d'un objet, soit le fait de choisir entre plusieurs étiquettes pour identifier son environnement. Une même expression ne sera pas interprétée de la même façon selon son contexte d'énonciation.
- 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 à 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 ».

45 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. La subjectivité interprétative, quant à elle, peut se rapprocher de l'opinion.

46 La notion centrale de cette étude est la perception. C'est un processus subjectif dans lequel chacune des notions citées joue un rôle plus ou moins important. La perception peut donc apparaître à travers l'avis, le jugement, le sentiment, la sensation exprimée par un locuteur sur ce lieu. Ces expressions subjectives se retrouvent dans le contexte d'emploi des lieux comme dans l'exemple suivant :

> [Exemple 9 : ESLO2_ENT_1070_C] « bon puis les bords de Loire sont magnifiques maintenant »

47 Le jugement que porte le locuteur sur les bords de la Loire est exprimé à travers le lexique évaluatif avec l'adjectif « magnifique ».

48 La caractérisation des différentes stratégies de dénomination des lieux amorce l'analyse de la perception qui leur est relative. Barbaras (2009) présente 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 entre observation de la réalité et interprétation des observations réalisées. 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. Percevoir un objet passe d'abord par son expérimentation physique et sensorielle. La perception est une expérience interne 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.

49 L'analyse de la perception relative à Orléans s'appuie sur les techniques d'apprentissage automatique supervisé qui reposent sur des données préalablement annotées. Pour que ces méthodes soient efficaces, il faut disposer d'un grand nombre de données annotées manuellement pour former un corpus de référence sur lequel le système développé pourra apprendre à réaliser la tâche qu'on lui a attribuée.

50 Partager la perception d'un objet est un processus subjectif, c'est pourquoi l'analyse automatique de la perception débute par la reconnaissance des énoncés porteurs de la subjectivité.

51 Dans les conversations analysées, les déclarations subjectives émises par les locuteurs ne portent pas toutes sur la ville d'Orléans. Pour isoler les énoncés pertinents, le corpus est segmenté en fonction des lieux identifiés : un tour de parole est sélectionné, avec son contexte proche (deux tours précédents et deux tours suivants), s'il contient un lieu annoté. Les techniques d'apprentissage supervisé sont appliquées à ce nouveau corpus uniquement constitué d'énoncés mentionnant des lieux.

52 La détection de la perception consiste d'abord à distinguer les segments subjectifs des segments objectifs, puis à déterminer le caractère positif, neutre ou négatif des segments jugés subjectifs.

53 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'entraînement, le corpus de validation pour tester l'efficacité du modèle entraîné sur un nouveau jeu de données et le corpus de test pour évaluer définitivement le système. Dans cette perspective, trente transcriptions ont été retenues et annotées manuellement par deux linguistes. Elles ont classé manuellement les tours de paroles sélectionnés en fonction de trois étiquettes. Un accord interannotateurs a été mesuré en utilisant un kappa de Cohen (1960), il obtient 0,85 et est considéré comme excellent selon la grille d'interprétation proposée par Landis et Koch (1977).

54 Différentes expériences ont été menées afin d'entraîner un modèle. Pour améliorer les résultats de l'entraînement, certains traits linguistiques ont été intégrés au modèle :

- La lemmatisation, soit une proposition de la forme de base de chaque mot du corpus (infinitif pour le verbe, masculin singulier pour les formes variables comme les noms, adjectifs, pronoms, etc.).
- L'étiquetage morphosyntaxique, consistant à attribuer une étiquette comprenant des informations de nature morphologique pour chaque mot (partie de discours : nom, verbe, adjectif, etc. ; genre : masculin, féminin ; nombre : singulier ou pluriel ; temps, mode pour un verbe, etc.). Ces deux traitements ont été réalisés avec Treetaggerwrapper [12], soit l'implémentation de l'outil Treetagger [13] (Schmid 1994) dans une bibliothèque Python par Pointal en 2004.
- Le calcul d'un score de polarité et d'un score d'émotion obtenus à partir de l'exploitation du dictionnaire de termes annotés en polarités et en émotions (French Expanded Emotion Lexicon [FEEL], Abdaoui *et al.* 2017 [14]).

55 Observons le tableau 1 qui montre cette distinction. Les deux segments a et b sont considérés comme subjectifs puisqu'ils sont porteurs de l'opinion d'un locuteur à propos du « quartier de la cathédrale » et de la « rue du Martroi ». Ils se différencient l'un de l'autre par leur polarité respectivement positive (« préféré », « joli ») et négative (« défaut »). Cette polarité est déterminée grâce au lexique polarisé FEEL. Le dernier segment, quant à lui, est considéré comme objectif puisque aucune information subjective n'est associée aux Alpes du Nord dont il est question. Le locuteur ne fait que mentionner ce lieu dans son discours. Aucune polarité n'est donc associée à ce segment. Dans l'éventualité où des termes positifs et négatifs sont présents dans la même phrase, le système tranche et penche pour la polarité majoritaire.

TABLEAU 1. EXEMPLE D'ANNOTATION DE SEGMENTS DE TRANSCRIPTION EN SUBJECTIVITÉ ET EN POLARITÉ

	Segments de transcription	Subjectivité	Polarité
a	[ESLO2_ITI_10_03] « mon lieu préféré dans la ville ça serait quand même du quartier cathédrale euh/voilà justement oui c'est très joli »	Subjectif	Positif
b	[ESLO2_ENT_1047] « rue du Martroi ouais le seul petit défaut ouais/c'est les places de parking c'est vrai que là euh »	Subjectif	Négatif
c	[ESLO2_ENT_1047] « c'est dans les Alpes du Nord/d'accord »	Objectif	X

Tableau produit par les auteurs

56 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, qui consiste en sa transformation en liste de nombres (ou vecteurs) afin de faciliter la manipulation des données par les algorithmes de classification. Différentes représentations vectorielles des documents ont été comparées et évaluées. Les meilleurs résultats ont été obtenus en utilisant une mesure de TF-IDF [15] qui s'intéresse au nombre d'occurrences de chaque terme dans le corpus. Plus un terme est fréquent, moins il sera considéré comme distinctif. Au contraire, s'il est rare, alors sa valeur discriminante sera plus grande.

57 Après avoir testé différents modèles utilisés en TAL [16] pour des tâches similaires, comme une fouille d'opinion, nous nous sommes arrêtées sur les machines à vecteurs supports (ou SVM pour *Support Vector Machines*) qui obtiennent les meilleurs résultats sur nos données. Toutes ces expériences combinant différents traits, représentations vectorielles et classifieurs ont permis de définir le modèle obtenant les meilleurs résultats, présenté dans le tableau 2. En l'occurrence, le modèle associant le TF-IDF et l'ensemble des traits disponibles dans un SVM, présente les meilleures performances pour la détection de la subjectivité (macro-average de 0,77) et de la polarité (macro-average de 0,76).

TABLEAU 2. MODÈLE RETENU POUR LA DÉTECTION DE LA SUBJECTIVITÉ ET DE LA POLARITÉ

Vectorisation	Features	Classifieur	Cible	Macro--average	Précision	Rappel	F-mesure
TF-IDF	Lemmatisation + Score polarité + Score émotions + POS	SVM	Subj.	0,77	0,69	0,56	0,63
			Obj.		0,77	0,89	0,83
			Pos.	0,76	0,78	0,91	0,82
			Nég.		0,67	0,46	0,54

Tableau produit par les auteurs

58 Les dynamiques observées lors de l'entraînement du modèle de détection de la subjectivité se retrouvent dans celui de la détection de la polarité. Ainsi, les mesures de rappel pour la détection des segments subjectifs et négatifs sont les plus faibles. Cela signifie que très peu de détections sont réalisées : seuls 56 % des segments subjectifs et 46 % des segments négatifs sont identifiés. Les scores de précision, qui représente la part de détections correctement réalisées parmi l'ensemble des détections réalisées, montrent que la qualité des identifications est meilleure (respectivement 69 % et 67 %) mais reste moyenne. Néanmoins, la détection des segments positifs est plus satisfaisante : la grande majorité des segments objectifs et des segments positifs sont détectés et, en moyenne, 78 % des détections faites sont pertinentes.

59 Ces résultats s'expliquent très probablement par la composition du corpus de référence. Les classes objectives et positives étant majoritaires, le classifieur dispose de plus d'exemples pour apprendre à les détecter au détriment des autres classes. On peut faire l'hypothèse que l'équilibrage de la composition du corpus de référence, et donc du corpus d'entraînement, par l'ajout de segments subjectifs, en particuliers négatifs, permettrait l'amélioration du rappel. Une nouvelle phase d'annotation manuelle doit donc être réalisée.

60 Une autre explication pour les non-détections de segments négatifs provient du lexique FEEL employé pour le calcul des scores de polarité et d'émotion. Dans l'exemple 10,

> [Exemple 10 : ESLO2_ENT_1016] « si y avait vraiment la crue de la Loire ça serait une catastrophe »,

61 le locuteur parle de la 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 aussi éventuellement retrouver dans le terme « crue ». Pourtant, aucune émotion n'a été identifiée dans cet ex-

trait pour la raison que les termes « crue » et « catastrophe » ne sont pas référencés dans le lexique. Le score de subjectivité n'est donc pas représentatif du segment. Les seules mesures obtenues à partir du lexique FEEL ne sont pas suffisantes pour analyser la polarité et la subjectivité, mais leur intégration dans le processus d'apprentissage automatique peut *a minima* donner des indices intéressants à l'algorithme à propos du caractère objectif ou subjectif de chaque segment. Ainsi, il reste nécessaire d'ajouter d'autres *features*, comme le nombre de mots composant les segments, la position des termes porteurs dans le segment, afin d'améliorer l'apprentissage du modèle.

Perspectives typologiques de la perception

62 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.

Typologie des cibles de la perception

63 En premier lieu, nous proposons d'élaborer une typologie de cibles sur lesquelles porte la perception émise, c'est-à-dire les thématiques abordées. En effet, il serait intéressant de pouvoir détecter quels sont les éléments mis en avant par les locuteurs pour partager leur perception de l'environnement. 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. À 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 qu'il se fait de son environnement. Ces choix spontanés, sans réflexion préalable, ne sont pas anodins, ils illustrent sa perception.

64 L'exploration du corpus révèle que l'expression de la perception à propos d'une ville se fait en fonction de quatre thématiques précises :

- L'« urbanisme » qui regroupe les déclarations subjectives sur la situation au sens géographique des espaces les uns par rapport aux autres. Le lieu est perçu sous l'angle de son accessibilité générale et de l'organisation des espaces qui le composent. Dans l'exemple 11, le locuteur compare le cinéma Pathé et le cinéma des Carmes d'Orléans en accordant plus d'importance à leur accessibilité qu'à leur offre cinématographique.

  [Exemple 11 : ESLO2_ENT_1020] « 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 »

- La thématique « économique » dans laquelle les lieux sont perçus par le biais de leurs fonctions, et donc par l'usage qu'en ont ou non les personnes, comme on peut l'observer dans l'exemple 12. 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.

[Exemple 12 : ESLO2_ENT_1050] « maintenant c'est Carrefour city oui c'est vrai que c'est plus sympa et c'est moins cher »

- Le « social » pour les cas où les lieux sont décrits sous l'angle des personnes qui y vivent, des relations qui s'y nouent, de l'animation sociale. Dans l'exemple 13, il est question de l'ambiance dans une ville et de son animation. Il ne s'agit pas dans cette catégorie de s'intéresser aux activités qui poussent les personnes à se réunir ou non dans le lieu mais plutôt de garder un regard général sur l'importance de l'occupation humaine de l'endroit.

[Exemple 13 : ESLO2_ENT_1050] « 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 »

- La thématique « historique » à travers laquelle l'individu se projette dans son environnement en considérant les activités qui s'y sont déroulées, se déroulent ou se dérouleront. Dans l'exemple 14, le locuteur parle de la rumeur d'Orléans, 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.

[Exemple 14 : ESLO2_ENT_1004] « 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 »

Typologie de la nature de la perception

65 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. Au-delà de cette typologie, on peut aussi essayer de décrire 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 qui considère le positionnement d'un individu par rapport à son environnement via le prisme de ses sens (le quartier est plutôt bruyant, calme, etc.) à partir desquels apparaissent ses émotions et ses sentiments (un beau fleuve, une ville triste et grise, etc.).
- L'expérience personnelle lorsqu'un individu perçoit le lieu par le biais de ses propres expériences ou activités (c'est là que je vais, je n'y vais jamais, etc.). Le fait d'énoncer quelles activités une personne

peut ou ne peut pas faire dans un certain endroit contribue à en dresser le portrait. De plus, l'individu s'implique personnellement dans l'expression de sa perception, ce qui en renforce la valeur.

- L'expérience collective – ou en tout cas la représentation qu'il en a – qu'un individu partage avec une ou plusieurs autres personnes (il paraît que, d'autres ont dit que, les gens disent, etc.). Même s'il peut partager leur avis, il prend de la distance avec les personnes dont il rapporte les propos.

Annotation manuelle de la cible et de la nature de la perception

66 Afin de contrôler la pertinence de ces nouvelles typologies, le corpus de référence utilisé pour l'évaluation de la subjectivité et de la polarité a été annoté manuellement.

67 La répartition des quatre cibles de la perception est montrée dans la figure 1. Elle est influencée sans doute par les thématiques abordées lors des enregistrements. Trois étiquettes sont réparties d'une manière équivalente : urbanistique, sociale et économique, ce qui confirme leur pertinence car les locuteurs parlent de leur vie de tous les jours en mentionnant les lieux de commerce, de loisirs, etc. ; ils décrivent aussi leur travail et leurs relations sociales. Cependant, l'étiquette historique (6 %) est faiblement représentée, ce qui peut être expliqué par la nature de conversations menées. Parmi les segments faisant partie de cette catégorie, on trouve un surnom historique des habitants, les chiens d'Orléans, des événements comme la reconstruction de la ville après la Seconde Guerre mondiale, la rumeur d'Orléans déjà mentionnée ci-dessus, etc.

FIGURE 1. RÉPARTITION DES ANNOTATIONS EN FONCTION DE LA CIBLE DE LA PERCEPTION

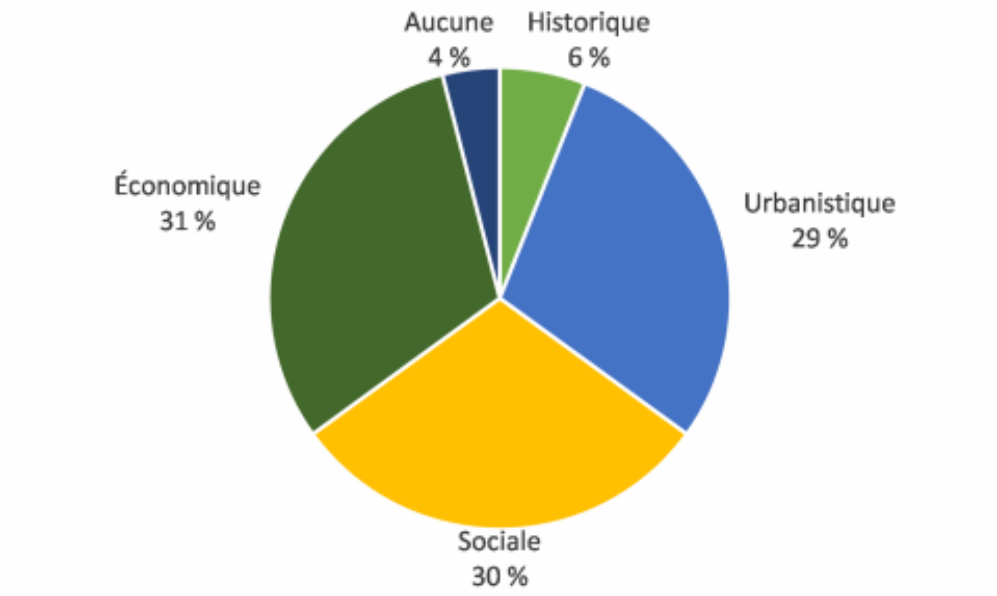

Image produite par les auteurs

68 Quelques segments n'ont pas été catégorisés (4 %) car ils n'appartiennent à aucune des catégories définies. Il s'agit de cas particuliers :

> [Exemple 15 : ESLO2_ENT_1016] « je ne retiens pas les noms de rues à Orléans »

69 La figure 2 montre la répartition des catégories décrivant la nature de perception. Si les étiquettes « expérience individuelle » et « expérience collective » sont représentées en nombre significatif, l'étiquette sensorielle est très fortement sous-représentée (4 %). Ce constat montre que la perception ne doit pas être abordée que du point de vue des sens. Un lieu peut être perçu ainsi différemment à travers les expériences qui y sont vécues.

FIGURE 2. RÉPARTITION DES ANNOTATIONS EN FONCTION DE LA NATURE DE LA PERCEPTION

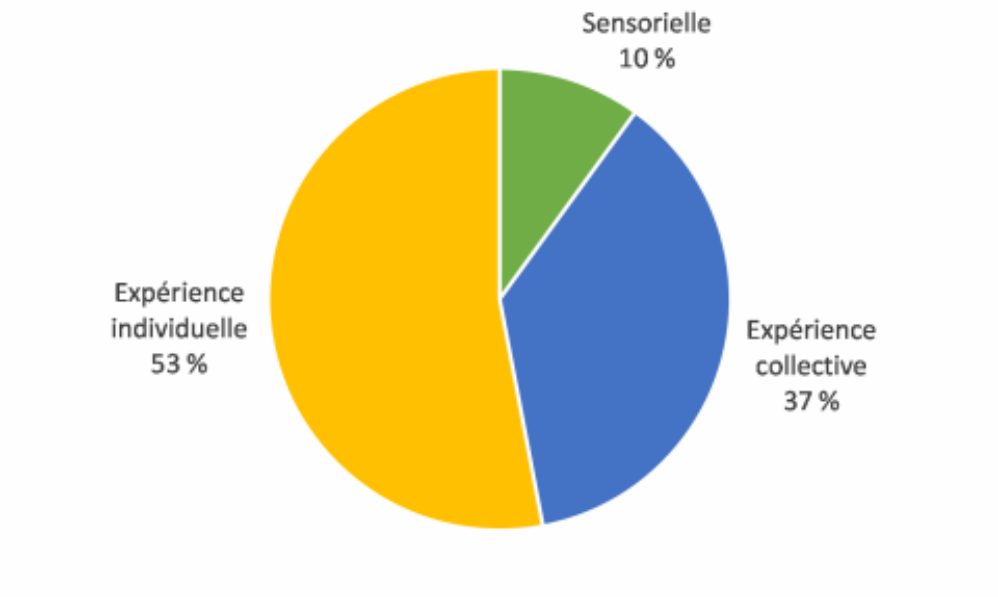

Image produite par les auteurs

Analyse et visualisation de l'information spatiale et de sa perception

70 L'objectif de la dernière étape du travail est d'analyser et de représenter visuellement les résultats obtenus grâce aux traitements décrits ci-dessus.

71 La visualisation des informations extraites participe à l'émergence de nouvelles connaissances à propos des objets étudiés. Elle 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 et Beaudouin-Lafon (2001, 3), son but 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 permet une « exploration rapide d'ensembles d'informations inconnues », une mise en évidence de « relations et de structures dans les informations » et de « chemins d'accès à des informations pertinentes » ainsi que la « classification interactive des informations ».

72 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.

73 À la suite des différents traitements appliqués pour détecter les noms de lieux ainsi que la perception qui leur est associée, un extrait du corpus *ESLO* a été annoté. Après avoir montré la répartition des annotations réalisées automatiquement dans le corpus, nous le projetterons dans un système d'information géographique afin de matérialiser la perception des locuteurs au sujet des lieux évoqués.

Répartition des annotations automatiques

74 Trente transcriptions ont donc été sélectionnées dans les modules Entretiens et Itinéraires du corpus *ESLO*. Parmi elles, 3 178 segments ont été établis à partir de la détection des lieux. Les figures 3 et 4 présentent la répartition des lieux parmi les 3 178 entrées du corpus observé en fonction du type de lieux et de leur zone géographique.

FIGURE 3. RÉPARTITION DES SEGMENTS EN FONCTION DES TYPES DE LIEUX IDENTIFIÉS

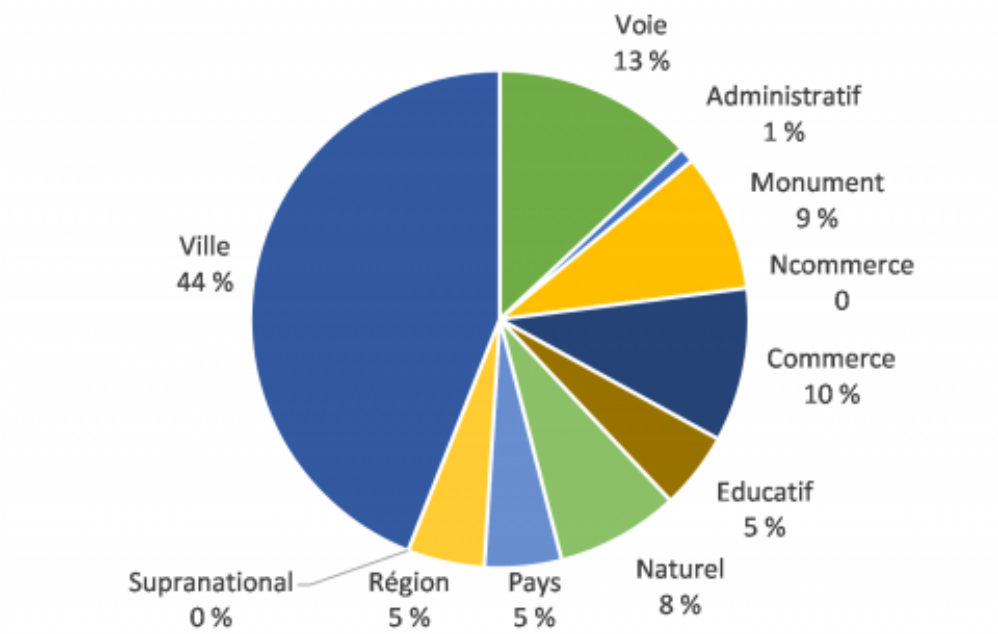

Image produite par les auteurs

75 La figure 3 montre que l'étiquette ville (44 %) est la plus représentée, ce qui reflète la nature des données traitées. Les voies représentent 12 % des lieux mentionnés, les monuments 12 %, les commerces 9 %, les lieux naturels 7 %, et les régions, les lieux éducatifs et les pays représentent chacun 5 %. Les lieux administratifs sont rares et n'apparaissent que 15 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 ont tendance à se référer au bâtiment pour son attrait touristique plutôt que pour sa fonction administrative. Enfin, seulement 10 lieux à fonction non commerciale (« Ncommerce ») et un seul lieu étiqueté comme supranational sont présents dans le corpus.

76 La figure 4 montre, quant à elle, une répartition plutôt équilibrée entre les trois zones géographiques définies. La majorité des segments mentionnent des lieux situés à Orléans (37 %, soit 1 175 segments). Le corpus contient aussi 34 % de lieux situés dans l'agglomération d'Orléans

et 29 % en dehors de l'agglomération. Au total, 71 % des lieux mentionnés dans ce corpus ont un lien direct ou très proche avec la ville d'Orléans placée au cœur de notre étude.

FIGURE 4. RÉPARTITION DES SEGMENTS EN FONCTION DE LA ZONE GÉOGRAPHIQUE DES LIEUX IDENTIFIÉS

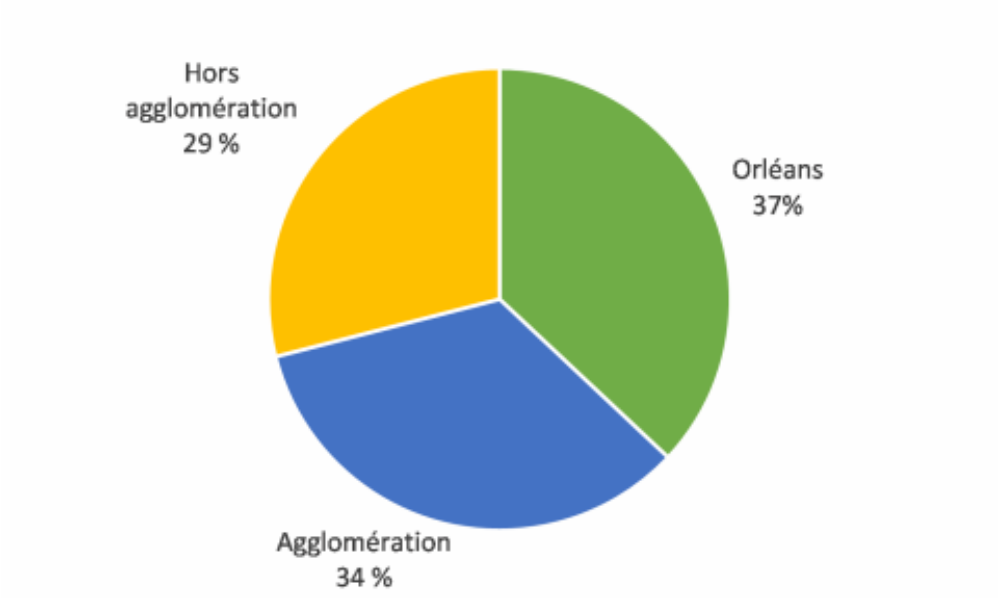

Image produite par les auteurs

77 En ce qui concerne la subjectivité, la figure 5 montre 1 112 expressions subjectives qui représentent 35 % des segments du corpus annoté. Sur le total des segments, 24 % ont une connotation positive et 11 % ont une connotation négative.

FIGURE 5. RÉPARTITION DE LA SUBJECTIVITÉ ET DE LA POLARITÉ DANS LE CORPUS DE RÉFÉRENCE

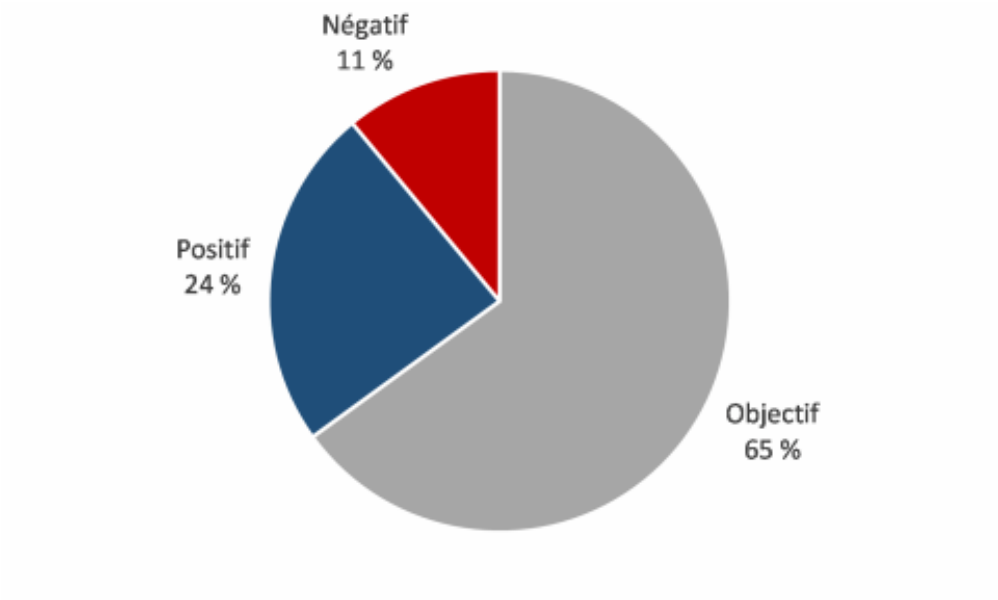

Image produite par les auteurs

78 Cette répartition est conditionnée par le contenu du corpus : les entretiens portent sur la ville d'Orléans appréhendée par le prisme 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.

Systèmes d'information géographique

79 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 (SIG) est un enjeu majeur pour la représentation de l'information spatiale. Ces problématiques s'intègrent dans une logique pluridisciplinaire en cherchant à mettre en évidence les liens qu'entretient l'être humain avec l'espace.

80 Pour permettre la visualisation de la perception, les SIG semblent être les plus adaptés. La visualisation des lieux et des annotations réalisées contribue à faire émerger les relations qui unissent les différentes informations détectées.

81 Pour répondre à cet objectif, l'ensemble des données collectées lors des traitements décrits précédemment a été projeté dans un SIG disponible en ligne [17] grâce à l'outil ArcGIS Online [18], développé par Esri [19]. Plusieurs couches d'informations sont projetées dans le système et permettent de figurer l'ensemble des lieux détectés en fonction des informations annotées au cours du traitement.

82 La figure 6 montre l'ensemble des lieux identifiés dans le corpus de référence pour lesquels il existe des déclarations subjectives. 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. Les lieux sont associés aux informations qui les décrivent (zone, type, polarité, etc.) et aux déclarations faites à leur sujet comme l'avenue Jean-Zay que l'on voit catégorisée comme une voie. Une fenêtre contextuelle au milieu de la figure indique la déclaration, le code du locuteur qui l'a prononcée, le nom de l'enregistrement dont elle est extraite ainsi que le lien vers sept autres déclarations à propos de ce lieu.

FIGURE 6. PROJECTION DES LIEUX EN FONCTION DE LEUR TYPE

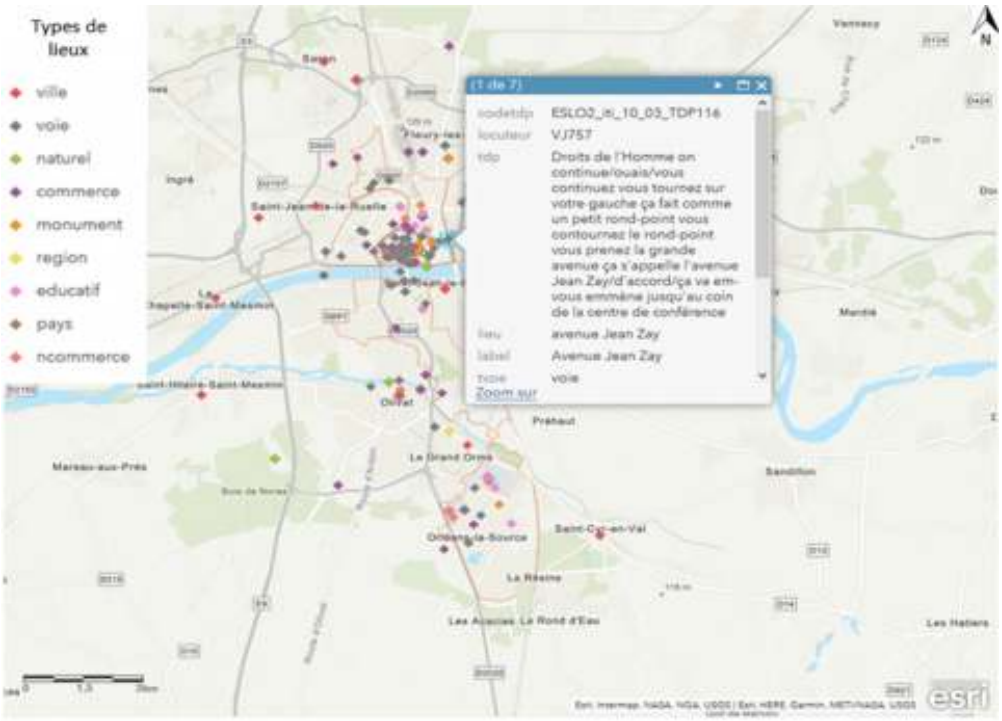

Image produite par les auteurs

83 La figure 7 présente, quant à elle, les lieux situés dans le centre-ville d'Orléans en fonction de la polarité qui leur a été associée. Un ratio correspondant à la part de déclarations positives et négatives est calculé et corrélé avec le nombre total de déclarations subjectives exprimées à propos du lieu. Plus le nombre de déclarations est important, plus le ratio de la polarité est significatif. Ainsi, les points les plus foncés sur la figure 7 montrent les avis les plus positifs à propos d'Orléans, ce qui permet de constater que le centre de la ville, par exemple est plus apprécié que le quartier de la gare.

FIGURE 7. RÉPARTITION DE LA POLARITÉ À PROPOS DES LIEUX DU CENTRE-VILLE D'ORLÉANS

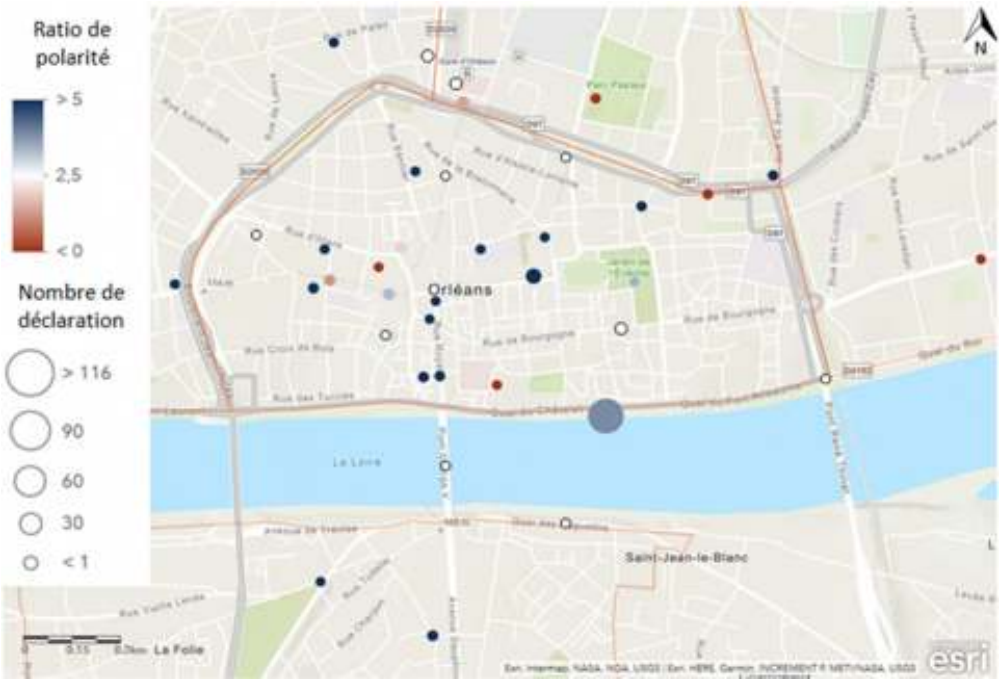

Image produite par les auteurs

Conclusion

84 Les objectifs de ce travail sont aussi applicatifs que théoriques. En premier lieu, il s'agit 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 ici est composé de transcriptions de conversations orales. Le traitement de l'oral diffère de celui de l'écrit et pose certaines difficultés. Dans chaque maillon de la chaîne de traitement présentée, les spécificités propres de l'oral sont analysées et prises en compte dans l'élaboration des différents modules d'annotation et d'extraction.

85 Par ailleurs, la notion de lieu est complexe. Sa définition pose question, que ce soit du point de vue de la linguistique, du TAL ou même de la géographie. La détection automatique des mentions de lieux, le deuxième objectif visé, n'est donc pas une tâche simple. Elle est rendue encore plus difficile sur les données transcrites de l'oral.

86 Ce travail aborde également une autre problématique, celle de l'analyse de la perception que peut avoir un locuteur de son environnement et comment cette perception transparaît à l'oral. L'analyse de la perception d'un lieu commence par l'étude des opinions émises à son sujet, ce qui suppose une réflexion sur la façon dont quelqu'un donne son opinion, en particulier à propos de son environnement. Cependant, pour être complète, l'analyse de la perception d'un lieu doit aller plus loin que la caractérisation de l'opinion en s'intéressant aussi aux habitudes et aux actions rapportées par les locuteurs.

87 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.

88 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 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. 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. Ils pourront estimer dans quel endroit ils préféreront vivre en fonction de ce que les Orléanais auront partagé à propos de leurs habitudes : les commerces dans lesquels ils vont faire leurs courses, l'école dans laquelle sont inscrits leurs enfants, les endroits dans lesquels ils vont se détendre, ceux dans lesquels ils ne vont pas, etc.

89 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 sé-

jour dans la ville : restaurants, shopping, musées, manifestations culturelles, etc. L'analyse de la carte pourrait même répondre à des problématiques d'aménagement du territoire en mettant par exemple en évidence certaines améliorations demandées par les administrés ou en permettant d'évaluer l'impact que des travaux ont pu avoir sur les habitants.

90 Plusieurs perspectives peuvent être envisagées.
En premier lieu, la représentation cartographique des informations extraites peut être améliorée. On peut imaginer une distinction de représentation visuelle entre les différentes dimensions spatiales évoquées : les places, les voies, les zones imbriquées ou non, les villes, les quartiers, etc. 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, en s'appuyant sur les annotations relatives à la cible ou à la nature de la perception, on pourrait observer quel type de lieu est le plus souvent perçu par le biais des sens ou par les actions réalisées. On pourrait aussi se demander si certaines thématiques sont plus ou moins susceptibles d'être polarisées et dans quelle mesure. Des liens pourraient également être établis avec d'autres ressources comme Wikipédia, par exemple.

91 En deuxième lieu, un travail sur la détection automatique des cibles et de la nature de la perception à partir de la typologie proposée pourrait être mis en œuvre et ainsi faire progresser les recherches sur cet objet souvent abordé sous l'angle de la perception sensorielle (Philippot 2007 ; Cance et Dubois 2016). Le travail réalisé a bien montré qu'un lieu peut aussi être perçu à travers les expériences qui y sont vécues et que ces expériences définissent le rapport qu'un être humain établit avec son environnement.

92 Les travaux présentés s'inscrivent dans un cadre pluridisciplinaire en exploitant 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. Ils s'interrogent sur l'exploitation nouvelle de données linguistiques dans un corpus à dimension sociolinguistique avec l'objectif d'en extraire un contenu subjectif. La démarche décrite est généralisable à d'autres données et peut être appliquée à des enregistrements portant sur d'autres villes.

Bibliographie

Abdaoui, Amine, Jérôme Azé, Sandra Bringay et Pascal Poncelet. 2017. « FEEL : A French Expanded Emotion Lexicon ». *Language Resources and Evaluation* 51 (3) : 833-855. https://doi.org/10.1007/s10579-016-9364-5.

Abouda, Lotfi et Olivier Baude. 2006. « Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des *ESLO* ». Dans *Corpus en lettres et sciences sociales. Des documents numériques à l'interprétation. Actes du XXVII^e^ colloque d'Albi « Langages et signification », Albi, 10-14 juillet*, édité par François Rastier et Michel Ballabriga. https://halshs.archives-ouvertes.fr/halshs-01162506.

Barbaras, Renaud. 2009. *La Perception. Essai sur le sensible*. Paris : Vrin.

Baude, Olivier et Céline Dugua. 2011. « (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? » *Corpus* 10 (novembre) : 99-118. https://doi.org/10.4000/corpus.2036.

Bonastre, Jean-François, Guillaume Gravier, Édouard Geoffrois, Khalid Choukri et Djamel Mostefa. 2008. « Évaluation des systèmes de transcription enrichie d'émissions radiophoniques ». Dans *L'évaluation des technologies de traitement de la langue*, édité par Stéphane

Chaudiron et Khalid Choukri, 165-182. Paris : Hermès, Lavoisier.

Bouvier, Jean-Claude. 1999. « Odonymes d'agglomération entre l'écrit et l'oral ». *Nouvelle revue d'onomastique* 33 (1) : 303-310. https://doi.org/10.3406/onoma.1999.1349.

Cance, Caroline et Danièle S. Dubois. 2016. « Diversité des ressources linguistiques et des discours dans différentes modalités sensorielles : de l'expression de l'expérience sensible à la construction de l'objectivité ». Communication présentée au *Colloque international « La perception en langue et en discours »*, Opole, Pologne, 21 avril. https://hal.archives-ouvertes.fr/hal-02343065.

Clerc, Pascal. 2004. « Lieu ». *Hypergéo*. http://www.hypergeo.eu/spip.php?article214.

Cohen, Jacob. 1960. « A Coefficient of Agreement for Nominal Scales ». *Educational and Psychological Measurement* 20 (1) : 37-46. https://doi.org/10.1177/001316446002000104.

De Landsheere, Gilbert. 1969. *Comment les maîtres enseignent. Analyse des interactions verbales en classe*. Bruxelles : ministère de l'Éducation nationale et de la culture française.

Dominguès, Catherine et Iris Eshkol-Taravella. 2013. « Repérer des toponymes dans les titres de cartes topographiques ». Communication présentée à *TALN2013*. Les Sables d'Olonne, 17-21 juin. https://hal.archives-ouvertes.fr/hal-01174713.

Dominguès, Catherine et Iris Eshkol-Taravella. 2015. « Toponym Recognition in Custom-Made Map Titles ». *International Journal of Cartography* 1 (1) : 109-120. https://doi.org/10.1080/23729333.2015.1055935.

Dominguès, Catherine, Serge Weber, Carmen Brando et Laurence Jolivet. 2018. « Projet PEPS MATRICIEL. Lieux des migrants à travers des récits de vie : mots, perceptions, émotions, cartes ». Poster présenté aux *Journées de la recherche IGN*, Champs-sur-Marne, 22 mars. https://halshs.archives-ouvertes.fr/halshs-02344538.

Eshkol-Taravella, Iris, Olivier Baude, Denis Maurel, Linda Hriba, Céline Dugua et Isabelle Tellier. 2011. « Un grand corpus oral "disponible" : le corpus d'Orléans 1968-2012 ». *Ressources linguistiques libres* 52 (3) : 17-46.

Flamein, Hélène. 2019. « Étude de la perception d'une ville. Repérage automatique, analyse et visualisation ». Thèse de doctorat, université d'Orléans.

Flamein, Hélène et Iris Eshkol-Taravella. 2020. « Noms de lieux dans le corpus de français parlé : une approche symbolique pour un traitement automatisé ». *Le Français moderne* 88 (1) : 320.

Frémont, Armand. 1980. « L'espace vécu et la notion de région ». *Travaux de l'Institut de géographie de Reims* 41 (1) : 47-58. https://doi.org/10.3406/tigr.1980.1081.

Hascoët, Mountaz et Michel Beaudouin-Lafon. 2001. « Visualisation interactive d'information ». *Revue I3* 1 (1) : 77-108.

Kerbrat-Orecchioni, Catherine. 2009. *L'Énonciation. De la subjectivité dans le langage*. Paris : Armand Colin.

Landis, J. Richard et Gary G. Koch. 1977. « The Measurement of Observer Agreement for Categorical Data ». *Biometrics* 33 (1) : 159-174. https://doi.org/10.2307/2529310.

Lesbegueries, Julien. 2007. « Plate-forme pour l'indexation spatiale multi-niveaux d'un corpus territorialisé ». Thèse de doctorat en informatique, université de Pau et des Pays de l'Adour. https://tel.archives-ouvertes.fr/tel-00258534/document.

Loustau, Pierre, Mauro Gaio et Thierry Nodenot. 2008. « Interprétation automatique d'itinéraires à partir d'un corpus de récits de voyages pilotée par un usage pédagogique ». *Revue des nouvelles technologies de l'information* E (13) : 177-206.

Moncla, Ludovic, Mauro Gaio, Javier Nogueras-Iso et Sébastien Mustière. 2016. « Reconstruction of Itineraries from Annotated Text with an Informed Spanning Tree Algorithm ». *International Journal of Geographical Information Science* 30 (6) : 1137-1160. https://doi.org/10.1080/13658816.2015.1108422.

Nouvel, Damien, Maud Ehrmann et Sophie Rosset. 2015. *Les Entités nommées pour le traitement automatique des langues*. Londres : ISTE Éditions.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Mathieu Brucher, Matthieu Perrot et Édouard Duchesnay, 2011. « Scikit-Learn : Machine Learning in Python ». *Journal of Machine Learning Research* 12 (85) : 2825-2830.

Philippot, Pierre. 2007. *Émotion et psychothérapie. L'Influence des émotions dans la société.* Wavre : Mardaga.

Piotrowski, Michael et Aris Xanthos. 2020. « Décomposer les humanités numériques ». *Humanités numériques* 1. https://doi.org/10.4000/revuehn.381.

Poibeau, Thierry. 2014. « Le traitement automatique des langues pour les sciences sociales. Quelques éléments de réflexion à partir d'expériences récentes ». *Réseaux* 188 (6) : 25-51. https://doi.org/10.3917/res.188.0025.

Rosset, Sophie, Cyril Grouin et Pierre Zweigenbaum. 2011. « Entités nommées structurées : guide d'annotation Quaero ». Notes et documents LIMSI n° 2011-04.

Schmid, Helmut. 1994. « Probabilistic Part-of-Speech Tagging Using Decision Trees ». Communication présentée à *International Conference of New Methods in Language Processing*, Manchester.

Zenasni, Sarah, Éric Kergosien, Mathieu Roche et Maguelonne Teisseire. 2016. « Extracting New Spatial Entities and Relations from Short Messages ». Dans *Proceedings of the 8th International Conference on Management of Digital EcoSystems – MEDES'16, Biarritz, November*, 189-196. New York : ACM Press. https://doi.org/10.1145/3012071.3012079.

Notes

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

2 http://eslo.huma-num.fr/index.php/pag6methodologie?id=71/.

3 http://stella.atilf.fr.

4 http://regions-france.org/actualites/en-direct-des-regions/bretagne-covid-19-region-annonce-mesures-exceptionnelles/.

5 Pour une description plus détaillée de la chaîne de traitement et des expériences menées, voir Flamein (2019), Flamein et Eshkol-Taravella (2020).

6 http://www.afcp-parole.org/campagne-devaluation-ester/.

7 http://www.afcp-parole.org/campagne-devaluation-etape/.

8 Ontologie des unités administratives de l'IGN. Révision : version 1.1 – 2019-02-12. http://data.ign.fr/def/geofla/20190212.htm.

9 https://www.data.gouv.fr.

10 http://stella.atilf.fr.

11 http://www.granddictionnaire.com/ficheOqlf.aspx?Id_Fiche=8462502/.

12 https://treetaggerwrapper.readthedocs.io/en/latest/.

13 https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

14 http://advanse.lirmm.fr/feel.php.

15 La vectorisation de notre corpus en fonction du TF-IDF est réalisée grâce à la bibliothèque Python Scikit-learn (Pedregosa *et al.* 2011) (https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html#sklearn.feature_extraction.text.TfidfVectorizer/).

16 La bibliothèque libre Python Scikit-learn (https://scikit-learn.org/stable/) 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.

17 http://arcg.is/Dymu8/.

18 https://www.arcgis.com/index.html.

19 https://www.esri.com/fr-fr/home/.

Auteurs

Hélène Flamein

UMR 7270 Laboratoire ligérien de linguistique, université d'Orléans, Orléans, France
Hélène Flamein est docteure en sciences du langage et s'intéresse à l'analyse et la valorisation de la subjectivité à travers l'exploitation de corpus oraux grâce aux apports de la linguistique et du traitement automatique des langues (TAL).
helene.flamein@univ-orleans.fr

Iris Eshkol-Taravella

UMR 7114 MoDyCo, université Paris-Nanterre, Nanterre, France
Professeure de sciences du langage à l'université Paris-Nanterre, Iris Eshkol-Taravella mène depuis vingt ans des recherches portant sur la constitution, le traitement et l'analyse des corpus avec les outils informatiques ; ses travaux s'inscrivent dans les trois domaines : traitement automatique des langues (TAL), linguistique de corpus et humanités numériques. Ses travaux portent plus particulièrement sur l'annotation de corpus, le développement des outils permettant d'avoir un accès plus facile au corpus et à son contenu, l'exploitation et l'étude des données annotées.
ieshkolt@parisnanterre.fr

Droits d'auteur