LES NOMS DE LIEUX MENTIONNÉS DANS DES RÉCITS DE VIE DE RÉPUBLICAINS ESPAGNOLS : DISTRIBUTION GÉOGRAPHIQUE ET PERCEPTIONS ASSOCIÉES

par Laurence Jolivet , Catherine Domingués
LASTIG, Univ Gustave Eiffel, IGN-ENSG,
laurence.jolivet@ign.fr; catherine.domingues@ign.fr

et Carmen Brando
École des hautes études en sciences sociales (EHESS)
carmen.brando@ehess.fr

Les lieux peuvent être définis par différentes caractéristiques telles que les coordonnées géographiques, les éléments du paysage, mais également des informations subjectives relatives à la perception. Le corpus constitué par le Réseau des acteurs de l'histoire et la mémoire de l'immigration (RAHMI) concerne des entretiens oraux décrivant des parcours de Républicains espagnols pendant la guerre civile d'Espagne et la Seconde Guerre mondiale. Ces récits de vie contiennent des noms de lieux, mentionnés par les personnes, correspondant à des noms propres (Npr) ou des noms communs (Nc). Ces noms de lieux peuvent être enrichis par des informations perceptives. Des méthodes linguistiques et statistiques ont été déployées afin d'extraire automatiquement et de caractériser certaines informations dans le texte par le biais d'annotations. L'objectif du travail présenté est d'analyser les distributions géographiques des lieux et des caractéristiques associées, ainsi que de proposer des cartographies illustrant les récits de vie et accompagnant les analyses. Premièrement, la distribution des Npr de lieux a été étudiée, permettant de cibler des échelles spatiales pertinentes et de cartographier une partie ou la totalité du corpus sous l'angle de ces lieux. Deuxièmement, les caractéristiques des Npr de lieux ont été synthétisées : leur type (administratif, topographique), leur étendue spatiale à partir de leur combinaison avec des Nc. Enfin des perceptions, résumées en polarité positive ou négative, ont été identifiées et dans certains cas associées aux Npr de lieux. Les motifs spatiaux de chaque polarité, en plus des Npr sans polarité et donc neutres, ont été cartographiés. La synthèse des annotations, les résultats d'analyses et les cartes proposées constituent des points d'entrée dans le corpus par des lieux mentionnés. Ces témoignages individuels, ainsi que leur valorisation, contribuent à la connaissance de cette période de l'histoire et aux travaux de recherche dédiés.

***Mots-clés** : récit de vie, nom de lieu, cartographie, analyse spatiale, indice de Moran local*

Introduction

Les noms de lieux constituent des jalons dans des récits de vie. Leur localisation permet de mettre en lumière la géographie de trajectoires individuelles, c'est-à-dire de comprendre les modalités et les causes de l'utilisation et du parcours de l'espace. Un lieu est situé, grâce à des coordonnées et à une emprise géographique. Il est également caractérisé (Purves *et al.*, 2019) par la présence d'éléments du paysage (ex. configurations, matériaux de construction, couleurs) (Tuan, 1977), une temporalité (ex. date de visite du lieu) (Huyghe, 2012), les conditions météorologiques, les personnes, les sons, etc. ; les perceptions individuelles, incluant les opinions, les émotions, les sentiments, sont aussi pertinentes (Boccagni *et al.*, 2015 ; Zeile *et al.*, 2015). Un lieu est ainsi défini comme un objet d'étude à plusieurs dimensions. En outre, et notamment dans des récits de vie, les lieux peuvent être désignés par des dénominations de différents types et avec diverses caractéristiques.

Le traitement automatique des langues (TAL) a pour but de concevoir et d'implémenter des méthodes et outils afin d'annoter automatiquement des textes écrits dans un format numérique. Une méthode d'annotation s'appuie sur un schéma d'annotation qui est un modèle décrivant des catégories sémantiques et syntaxiques (attributs, ou propriétés) à attribuer à un segment du texte (une séquence de mots) ainsi que les règles précisant ces catégories. Afin de

mesurer la pertinence des annotations posées par les outils, des annotations manuelles considérées comme une référence, sont fabriquées et comparées aux annotations posées automatiquement en vue d'évaluer le modèle d'annotation automatique. Les indicateurs communément calculés pour cette évaluation sont la précision, le rappel et la F-mesure[1]. La précision (P) est égale au nombre de vrais positifs divisé par le nombre de vrais positifs ajouté à celui des faux positifs. Elle indique ainsi le pourcentage d'annotations automatiques effectivement pertinentes par rapport à la totalité des annotations automatiques. Le rappel (R) est égal au nombre de vrais positifs divisé par le nombre des vrais positifs ajouté à celui des faux négatifs. Il indique le pourcentage d'annotations automatiques effectivement pertinentes par rapport aux annotations de référence. La F-mesure est une valeur synthétique à partir de P et R.

En linguistique, des recherches se consacrent aux noms de lieux en tant qu'une catégorie des noms propres (Leroy, 2004). Les lieux peuvent être désignés par un nom propre (Npr), par exemple *Catalogne*, par un ou des noms communs (Nc), par exemple *le front*, incluant éventuellement un Npr, par exemple *le front de Catalogne*. Les noms de lieux (Npr ou Nc) peuvent être complétés par des mots qui les modifient, comme un adjectif dans le Npr modifié *Paris étendu*, ou comme un groupe prépositionnel dans le Nc modifié *le quartier de Barcelone de mon enfance*. Un enjeu important est d'identifier les frontières de la séquence.

Le projet MATRICIEL (PEPS 2016-2017, *Université Paris-Est*[2] et *Centre national de la recherche scientifique*) avait pour objectif d'étudier les lieux et les perceptions associées au cours de parcours migratoires, en mobilisant les mots exprimés par les personnes ayant réalisé ces parcours. Le corpus contenait des récits de vie de Républicains espagnols exilés en France pendant la guerre civile espagnole. Ces récits ont été collectés par le *Réseau des acteurs de l'histoire et la mémoire de l'immigration*[3] (RAHMI). Le projet a associé des méthodes de traitement de l'information linguistique et géographique pour identifier et interpréter des éléments d'histoires individuelles à la lumière de l'histoire et de la mémoire collectives. L'article présente la partie du projet et les perspectives concernant l'analyse spatiale et la cartographie des noms de lieux et des informations associées dont les perceptions, extraits des récits de vies. Les lieux ont été pris en compte comme un ensemble de plusieurs caractéristiques, celles-ci étant mentionnées dans les récits ou déduites, et ajoutées par annotation. L'analyse spatiale avait pour but de mettre en évidence les distributions géographiques des lieux et de leurs caractéristiques. Notre intérêt s'est en particulier fixé sur les perceptions exprimées par rapport aux lieux, et la détermination de possibles patrons spatiaux de ces perceptions. Les cartes visaient à rendre compte des annotations ou bien à accompagner les analyses (Caquard, 2017). La cartographie était réservée aux informations géolocalisées, même si les informations subjectives telles que la perception sont délicates à représenter dans une carte car elles n'ont généralement pas de contour géographique, diffèrent selon les acteurs des récits et peuvent changer au cours du ou des récits (Griffin *et al.*, 2012 ; Olmedo, 2016). La partie 2 de l'article décrit le corpus des récits de vie. La partie 3 concerne les résultats d'analyse et les cartes éditées. La partie 4 contient des améliorations envisagées et des perspectives du travail.

Corpus et méthode

Le corpus des récits de vie des Républicains espagnols

Le corpus a été constitué en 2009 par le RAHMI. Il contient 17 récits de vie oraux, produits par les Républicains eux-mêmes ou par leurs proches. La période de vie concernée s'étend de l'enfance des Républicains au début du XXe siècle à l'époque contemporaine jusqu'en 2009. Toutefois, la période ciblée est la fin de la guerre civile espagnole et la Seconde Guerre mondiale. Chaque récit correspond à un enregistrement en format mp4, entre 11 minutes (pour le plus court) et 3 heures et 25 minutes (le plus long).

La transcription des enregistrements a été réalisée manuellement (Dominguès *et al.*, 2019), à partir d'une transcription initiale fournie par le RAHMI, corrigée et enrichie successivement par plusieurs transcripteurs et annotateurs. En plus des mots formulés par les témoins interviewés, ont été transcrites des informations traduisant les conditions de l'oral : les tours de parole, les bruits ambiants, les pauses ; lorsque l'énonciateur utilise des mots d'une autre langue (principalement espagnole et allemande), la traduction est, si possible, associée à la transcription. Le nombre de tokens, c'est-à-dire mots et signes de ponctuation, est compris entre 2 300 et 52 000 par récit.

1 https://fr.wikipedia.org/wiki/Pr%C3%A9cision_et_rappel
2 Université Gustave Eiffel depuis 2020 : https://www.univ-gustave-eiffel.fr/
3 https://www.rahmi.fr/

Annotation des noms de lieux et des caractéristiques associées

Le corpus transcrit des récits de vie constitue un matériau à partir duquel des informations ciblées ont été annotées. Les annotations réalisées concernent des particularités de formulations orales, des noms de lieux et des mots de sentiment, en détail :

- des marqueurs de l'oral : disfluences (répétitions, révisions), accent, rire, volume de la voix, etc. ;
- les mots dans une langue autre que le français ;
- les lieux désignés par des noms propres (Npr) ;
- les lieux désignés par des noms communs (Nc) ;
- des mots de sentiment.

Les récits de vie ont été abordés comme un ensemble de lieux désignés par leurs noms et enrichis par des informations complémentaires. Nous nous sommes particulièrement intéressées aux lieux désignés par un Npr car davantage localisables avec des coordonnées géographiques et donc cartographiables. Les lieux désignés par un Nc et les mots de sentiment sont pris en compte dans cet article comme des informations complémentaires permettant de préciser l'emprise géographique des lieux désignés par un Npr et de déterminer la perception de ces lieux.

L'accord inter-annotateurs sur chaque type d'annotations (AIA), c'est-à-dire entre deux ou plusieurs personnes ayant annoté une partie du corpus manuellement, a été calculé. Cette mesure permet d'estimer la pertinence de concevoir des traitements automatiques pour annoter les informations prévues. Un AIA faible implique que des annotateurs humains ne peuvent établir une annotation de référence consensuelle. Définir un processus d'annotation automatique ne serait donc pas pertinent car les résultats seraient difficiles à évaluer. L'AIA a été calculé pour deux récits de vie annotés par deux annotateurs. Concernant les mots de sentiments qui sont a priori les plus difficiles à annoter automatiquement, les valeurs R, P et F-mesure sont respectivement 0,72, 0,73 et 0,73 pour un récit et 0,88, 0,89 et 0,89 pour le second récit. Les variations sont principalement dues aux différentes interprétations, par exemple le mot *hiver* a été annoté avec une polarité négative par une personne, et n'a pas été annoté par une deuxième personne. Les valeurs des AIA ont toutefois été satisfaisantes, ce qui a permis d'envisager de mettre en place des méthodes d'annotation automatique.

Les annotations ont été réalisées semi-automatiquement, en utilisant des traitements automatiques et des traitements manuels (Dominguès *et al.*, 2019). Les Npr de lieux ont été annotés à l'aide d'un gazetier et les mots de sentiment avec un lexique. Les Nc de lieux ont été annotés par une méthode d'apprentissage supervisé pour la reconnaissance d'entités nommées s'appuyant sur l'outil Stanford-NER (*Named Entity Recognition*) (Finkel *et al.*, 2005). Cette méthode implémente des modèles de séquence de chaîne linéaire CRF (*Conditional Random Field*) et un modèle pour reconnaître des Nc de lieux a été spécifiquement entraîné à partir des exemples annotés et appliqué sur l'ensemble des récits (Brando *et al.*, 2016). Les annotations produites à l'aide du gazetier, du lexique ainsi que du modèle d'apprentissage ont été évaluées par comparaison avec des annotations manuelles considérées comme des références. Les R, P et F-mesure ont été calculés pour deux récits. Des corrections ponctuelles ont été apportées manuellement. Les traitements et les corrections ont été effectués dans la plateforme GATE[4] (Cunningham *et al.*, 2013) qui possède une interface utilisateur et qui peut intégrer des plugins développés en Java.

Le gazetier de Npr de lieux a été construit à partir de gazetiers existants : GeoNames[5] produit de manière collaborative et BDNyme produit par l'Institut national de l'information géographique et forestière (IGN). Il contient des Npr de lieux et leurs coordonnées géographiques. Les formes étrangères des Npr (ex. *Francia* pour la France) ont été intégrées au gazetier. La couverture géographique du gazetier construit est mondiale : le niveau de détail est plus important en France et en Espagne, que pour le reste du monde où seuls les lieux les plus peuplés (pays, montagnes, etc.) ont été sélectionnés. L'annotation automatique des Npr de lieux a été suivie par une désambiguïsation manuelle dans les cas où plusieurs entrées identiques existent (ex. Paris en France et Paris au Texas). Pour chaque entrée du gazetier, le type de lieu, unité administrative ou élément topographique, a aussi été renseigné.

Les exemples des Nc qui ont servi à l'entraînement du modèle pour Stanford-NER ont été annotés manuellement avec des étiquettes BIO (*Begin, Inside, Outside*) (Ramshaw *et al.*, 1995). Pour chaque occurrence d'un Nc, le début est étiqueté par une balise B-LOC et le reste par une balise I-LOC. La balise O est attribuée à chaque mot extérieur au Npr de lieu identifié. Par exemple, le Nc de lieu *la ligne de démarcation* est annoté ainsi : *la* par O, *ligne* par B, *de*

4 https://gate.ac.uk/
5 https://www.geonames.org/

par I et *démarcation* par I. Un Nc peut correspondre à un nom de rue, de place, etc., parfois précisé par un numéro d'adresse, par exemple *23 rue Naujac à Bordeaux*. Lorsqu'un Nc contient (ex. *banlieue de Barcelone*) un Npr, il a été géolocalisé par la localisation de ce Npr. Les Npr modifiés, les Nc n'incluant pas de Npr ne sont pas étudiés dans cet article.

La perception des témoins a été étudiée à l'aide des mots énoncés exprimant des sentiments. Elle a été définie selon une polarité positive, négative ou neutre. Le lexique comprend une liste de mots de sentiments, chaque mot étant associé à une polarité positive ou négative. Les récits transcrits ont été annotés automatiquement à partir de ce lexique, puis les Npr de lieux ont été associés à une des trois polarités définies en fonction de l'influence des mots de sentiment.

Analyse spatiale et représentation cartographique des noms de lieux

Les analyses spatiales et textuelles avaient pour but de caractériser les noms de lieux présents dans le corpus, notamment grâce aux informations complémentaires. Les résultats d'analyse ont ensuite été cartographiés afin de mettre en évidence des distributions spatiales. Analyses et cartographies ont été réalisées pour l'ensemble du corpus et pour chaque récit de vie. Elles s'appuient sur les annotations décrites en 2.2.

Dans un premier temps, les analyses ont porté sur la distribution spatiale des Npr de lieux, qui étaient donc renseignés en coordonnées géographiques. La représentation cartographique des Npr a utilisé ces coordonnées correspondant à des points. Des cartes de chaleur, mobilisant les distances entre les points, ont été réalisées afin d'étudier leur densité spatiale. Les lieux sont de différents types et ainsi de différentes étendues et échelles spatiales. Par exemple, l'étendue d'une commune n'est pas à la même échelle que l'étendue d'un pays ou d'une région administrative ou géographique. Afin de conserver une cohérence d'échelle, les cartes de chaleur ont été réalisées en fonction des types de lieux. Les nombres d'occurrences ont été calculés relativement aux Npr cités et aux récits contenant les Npr. Les unités spatiales et les éléments topographiques sont décrits dans des bases de données géographiques. Nous avons utilisé Natural Earth[6] pour les contours des pays et de régions, et les bases de données téléchargeables sur le site des géoservices de l'IGN[7] pour les unités administratives françaises ; les réseaux de transport, les bâtiments, l'hydrographie et la végétation ont été rajoutés pour faciliter l'interprétation

Dans un deuxième temps, la description des noms de lieux extraits a été enrichie afin de préciser l'interprétation des distributions spatiales. En plus du type de lieu administratif ou topographique, l'emprise géographique a été estimée. Celle-ci est liée aux annotations avec le modèle BIO des Nc de lieux incluant un Npr de lieu. L'emprise géographique d'un lieu correspondant à un Nc peut être égale à l'emprise géographique du lieu Npr ou la modifier. Les modifications peuvent être une diminution ou un élargissement de l'emprise géographique, ou encore un déplacement.

Une troisième thématique est consacrée à la perception des Npr de lieu. Une occurrence de Npr est associée à une unique polarité, donc un même Npr est associé à autant de valeurs de polarité que son nombre d'occurrences. Les distributions spatiales des polarités ont été cartographiées puis étudiées. Des similarités spatiales potentielles ont été analysées en calculant les indices de Moran concernant l'autocorrélation spatiale locale (LISA). Pour un Npr localisé, une valeur de LISA positive signifie une similarité des polarités entre les Npr proches géographiquement. Une valeur de LISA négative signifie des polarités contrastées dans le voisinage spatial du Npr. Les valeurs de LISA peuvent être confrontées aux distributions spatiales des localisations des Npr et des valeurs de polarités.

La distribution spatiale des noms de lieux, des informations complémentaires et des perceptions

Cette partie décrit les analyses spatiales et les cartographies réalisées du corpus des récits de vie des Républicains espagnols. Ces récits ont été annotés en noms de lieux et en informations complémentaires. Les résultats présentés concernent les géolocalisations et les nombres d'occurrences, puis les types de lieux et les emprises géographiques, et enfin les perceptions.

La distribution spatiale des noms de lieux et les occurrences

Le corpus contient 351 Npr distincts de lieux, qui correspondent à 2 787 occurrences. L'annotation automatique des Npr de lieux a été évaluée sur deux récits. Les valeurs R sont égales respectivement pour chacun des récits à 0,96 et 0,92. La précision

6 https://www.naturalearthdata.com/

7 https://www.geoportail.gouv.fr/, https://geoservices.ign.fr/

et la F-mesure n'ont pas été calculées car la désambiguïsation a été effectuée manuellement. Les distributions géographiques et les décomptes ont été réalisés pour le corpus complet puis par récit.

Distribution spatiale des noms de lieux

Le nombre minimum de Npr distincts pour un récit est égal à 15, et le nombre maximum à 73. Pour l'ensemble du corpus, la représentation cartographique des Npr de lieux, considérés distinctement sans leur nombre d'occurrences, a permis de mettre en évidence trois échelles spatiales du corpus, correspondant à trois étendues géographiques significatives : mondiale, régionale et locale (figure 1). Elle permet de visualiser les localisations absolues et relatives des lieux cités, à interpréter au regard de connaissances historiques.

Le corpus possède une échelle mondiale car les lieux mentionnés sont répartis dans 22 pays en Europe, en Amérique et en Afrique, en plus de l'océan Atlantique et de la mer Méditerranée. Toutefois 95 % des lieux sont situés en Europe continentale et 85 % des lieux en France (n = 177) et en Espagne (119).

L'échelle régionale la plus significative du corpus correspond au sud-ouest de la France et au nord-est de l'Espagne. 41 % des Npr de lieux sont situés dans les régions administratives françaises contemporaines Nouvelle-Aquitaine et Occitanie, créées en 2016 par fusion des précédentes régions. Les témoins se sont déplacés et se sont installés dans cette zone qui émerge donc logiquement.

Les échelles locales significatives ont été identifiées grâce à des cartes de chaleur. Les étendues sont restreintes, correspondant à des lieux proches spatialement. La carte de chaleur des noms de communes (256 noms distincts dans le corpus) fait ressortir des densités spatiales locales. Les zones à forte densité sont la ville de Bordeaux et ses alentours, dans et autour du département du Lot-et-Garonne et sur les côtes méditerranéenne et atlantique au niveau de la frontière franco-espagnole. Ces densités s'expliquent car Bordeaux est un centre économique important et les entretiens ont été menés à Bordeaux et ses environs où se sont installés les témoins ; le Lot-et-Garonne correspond à un endroit où les témoins et leurs proches ont vécu lorsqu'ils sont arrivés en France ainsi qu'un haut lieu de la Résistance ; la frontière franco-espagnole était jalonnée de points de passage et de camps de regroupement ou d'internement. Les zones à densité moyennement élevée sont situées dans les environs de Barcelone, Madrid et Cordoue, dans les Pyrénées françaises et dans la grande région parisienne. Les lieux en Espagne sont ceux où les témoins sont nés et ont grandi, avant de venir en France. Il s'agit également de fronts de bataille pendant la guerre civile espagnole (*front de l'Èbre*, *bataille de l'Alto*). Les Pyrénées françaises et la grande région parisienne contenaient des camps de regroupement et des manufactures où ont travaillé les témoins.

Nombre d'occurrences des noms de lieux

Dans tout le corpus, les Npr de lieux le plus cités sont *Espagne* et *France* dont les nombres d'occurrences sont respectivement 378 et 224. Ce sont les deux seuls Npr mentionnés dans chacun des 17 récits. Ensuite, les Npr les plus cités désignent des villes : *Bordeaux* (174), *Barcelone* (77), *Argelès-sur-Mer* (65), *Madrid* (63), *Perpignan* (58) et *Villeneuve-sur-Lot* (58). Les nombres d'occurrences des Npr et les nombres de récits dans lesquels les Npr sont cités, sont positivement corrélés (coefficient de Pearson égal à 0,8, p-valeur < 0.001). Quelques exceptions existent telles que *Villeneuve-sur-Lot* correspondant à 58 mentions mais cité dans uniquement 3 récits, et *Allemagne* correspondant à 34 mentions et pourtant cité dans 14 récits. La figure 2 illustre le nombre de récits dans lesquels les Npr de lieux distincts sont cités. 22 Npr distincts (6 %) figurent dans 5 récits et plus ; 244 Npr (70 %) dans un seul récit.

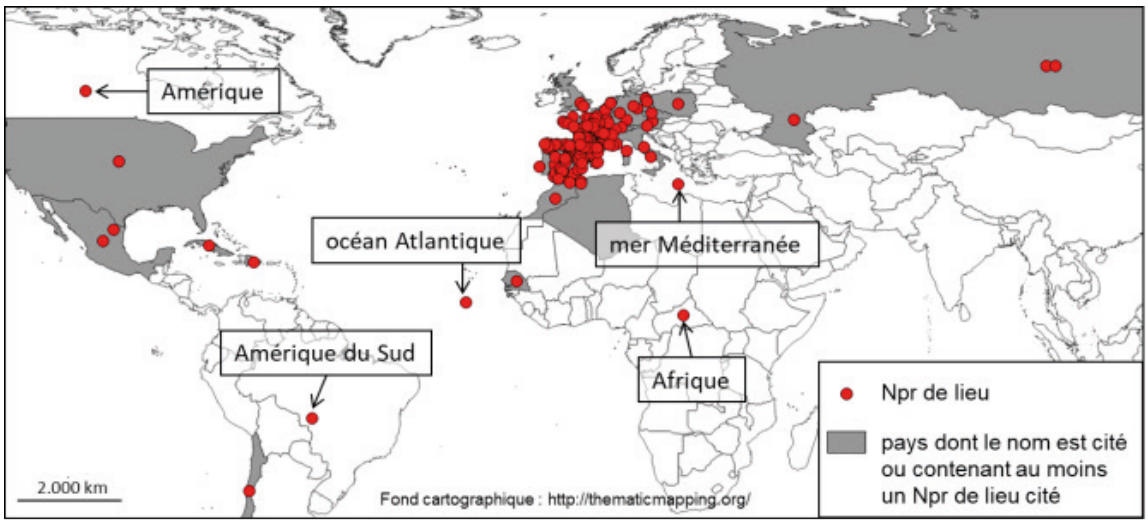
Par récit, le nombre minimum d'occurrences, tous Npr de lieux confondus, est égal à 48, et le nombre maximum à 450. Ces valeurs sont significativement corrélées avec les nombres de tokens (coefficient de Pearson égal à 0,9 avec une p-valeur < 0.001).

Informations associées aux noms de lieux

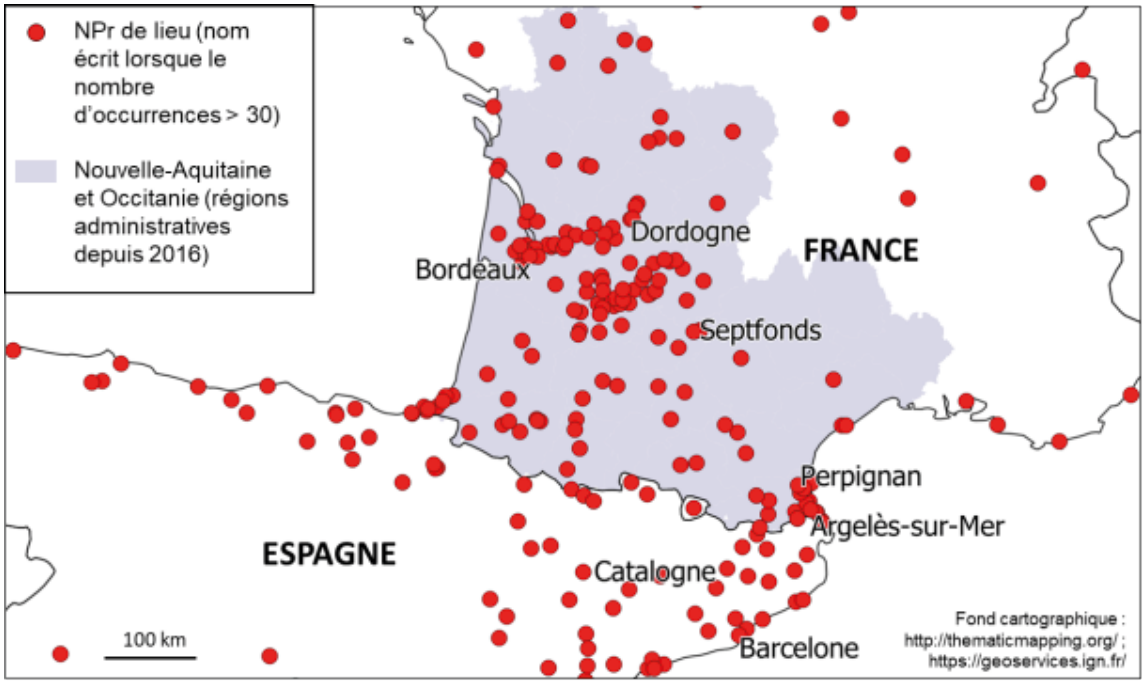
Types de lieu

Les lieux sont de deux grands types : unité administrative et élément topographique. Les unités administratives correspondent à des continents (*Afrique*), des parties de continent (*Amérique du Sud*), des pays (*Italie*), des communes (*Figueras*) et des entités intermédiaires comme des régions ou des départements (*Catalogne*, *Alpes-Maritimes*). Des entités comme des aires culturelles ou des régions historiques (*Castilla*, *Médoc*, *Touraine*) ont également été incluses dans les unités administratives. Les éléments topographiques comprennent des caractéristiques du relief (*Pointe de Grave*, *Sierra Morena*), l'hydrographie (*Guadalquivir*, *Lac Saint-André*) et certains sites notables, c'est-à-dire d'emprise limitée mais de grande notoriété, (*fort de San Cristobal*, *chantiers de l'Atlantique*). Le décompte par type de lieu est indiqué en tableau 1.

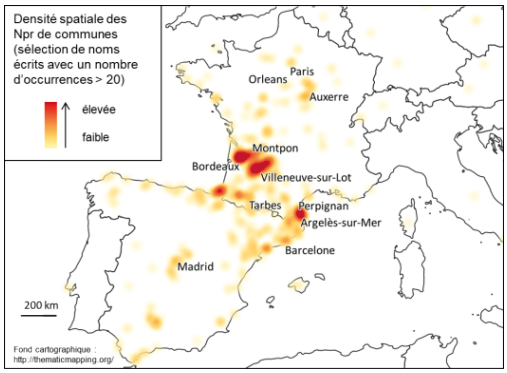
a

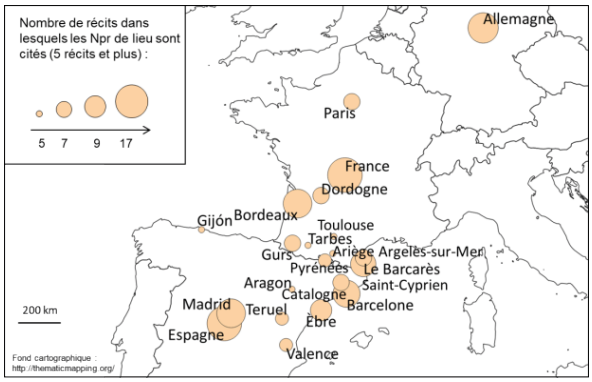


b



c





a. *b. c.*

Figure 1 : Les échelles spatiales du corpus déterminées à partir des localisations des noms de lieux cités : a. mondiale, b. régionale, en particulier dans le sud-ouest de la France et le nord-est de l'Espagne, c. locale, notamment à Bordeaux et ses environs, dans le département du Lot-et-Garonne et à la frontière franco-espagnole.

Figure 2 : Le nombre de récits de vie dans lesquels les Npr de lieux distincts sont cités, à partir de 5 récits et plus.

Type de lieu cité par un Npr	Décompte
Continent	3
Partie de continent	2
Pays	19
Entité intermédiaire	47
Commune	256
Relief	8
Hydrographie	10
Site notable	6

Table 1. Décompte des occurrences par type détaillé des lieux cités par un Npr.

Le type de lieu cité par un Npr le plus présent dans le corpus est la commune. Comparées à des unités de plus haut niveau, les communes sont des repères souvent mobilisés par les témoins afin de situer les événements lors de leurs récits. Les communes partitionnent en effet le territoire à une échelle en adéquation et pertinente avec le parcours et l'utilisation de l'espace par les témoins. Elles structurent d'une certaine manière l'espace vécu et perçu. Certains noms de communes cités ont changé depuis la période des récits. Cependant des continuités dans les noms permettent d'identifier les communes et de les localiser. C'est par exemple le cas pour *Cosne* qui est intégré depuis 1973 à la commune de *Cosne-Cours-sur-Loire*.

Emprise géographique des Nc de lieux incluant un Npr

Entre 50 et 400 Nc de lieux ont été annotés par récit de vie. L'évaluation des annotations des Nc pour un récit correspond à des valeurs R, P et F-mesure respectivement égales à 0,53, 0,85 et 0,66 pour la balise B-LOC (le caractère initial de l'annotation) et 0,28, 0,67 et 0,39 pour la balise I-LOC (les caractères suivants dans l'annotation). 250 annotations distinctes sont des Nc de lieux contenant un Npr. L'emprise géographique du Npr de lieu contenu peut être modifiée selon les quatre situations suivantes : le Nc égale, agrandit, réduit ou déplace le Npr inclus. Dans le corpus, ces situations ont été identifiées et déclinées en cinq catégories :

- la localisation exacte (*ville de Bordeaux*) ;
- une inclusion dans (*quartier entièrement rénové à Gijón*, *plages du Barcarès*) ;
- à l'extérieur proche (*banlieue de Barcelone*, *front de l'Èbre*) ;
- un espace englobant (*région de Perpignan*) ;

… inconnu (*direction Bordeaux*).

Peu de Nc contenant un Npr correspondent à la localisation exacte du lieu Npr (environ 10 %) ; dans ce cas, le Nc (ex. *la ville de*) renforce le Npr en précisant le type de lieu. Dans la catégorie la plus fréquente (environ 50 %), le lieu Nc est inclus dans le lieu Npr, mais n'a pas de nom dédié (ex. le front de mer d'une commune) ou ce nom est ignoré du locuteur (ex. le nom d'un quartier d'une commune). Le cas où le lieu Nc est à l'extérieur proche d'un lieu Npr est également fréquente (environ 30 %). De même que pour l'inclusion, il s'agit des lieux non nommés par un Npr dédié, comme dans le cas où les limites ne sont pas claires (ex. un territoire couvrant plusieurs communes proches d'une ville importante), et des lieux dont les noms ne sont pas connus par les témoins. La catégorie de l'espace englobant n'est pas évidente à discriminer. En effet, le gazetier construit contient des entrées de Npr de province, ex. *province de Cordoue*. Il s'agit néanmoins principalement d'unités administratives dans le gazetier, qui peuvent ne pas correspondre aux lieux désignés par les témoins (ex. le mot *région* peut faire référence à l'unité administrative ou à l'espace englobant le lieu désigné par un Npr sans contour défini ou partagé).

Parmi les 351 Npr de lieux du corpus, 96 Npr distincts (27 %) ont au moins une mention dans un Nc. Les Npr les plus cités sont aussi ceux les plus inclus dans les Nc (*Bordeaux*, *Espagne*, *Barcelone*, *France*). La représentation cartographique en figure 3 contient les extraits de corpus annotés correspondant aux Npr et aux Nc incluant un Npr dans les environs de Bordeaux. Les 255 Npr de lieux restant (73 %) n'ont pas été annotés comme inclus dans un Nc de lieu.

Une sémiologie a été définie afin de symboliser chaque catégorie d'emprise spatiale des lieux Nc par rapport aux lieux Npr (figure 3). Des combinaisons de symboles ont été définies pour les lieux Nc répondant à différentes catégories selon les annotations. La carte obtenue aide, par la visualisation, à se rendre compte des possibles influences des Nc sur les Npr qu'ils contiennent.

Perceptions

Distribution spatiale des perceptions

Entre 5 et 200 mots de sentiments ont été annotés par récit. L'évaluation de l'annotation pour un récit a abouti à des valeurs R, P et F-mesure respectivement égales à 0,30, 0,80 et 0,4. 40 % des annotations en Npr de lieux ont été associées à un ou plusieurs mots de sentiment. Parmi ces annotations, 30 % correspondent à une polarité positive et 70 % à une polarité négative. Les 60 % des Npr non associés à des mots de sentiments ont été qualifiés de neutre.

La perception d'un même lieu peut varier au cours d'une vie, conséquence d'événements, ainsi qu'entre

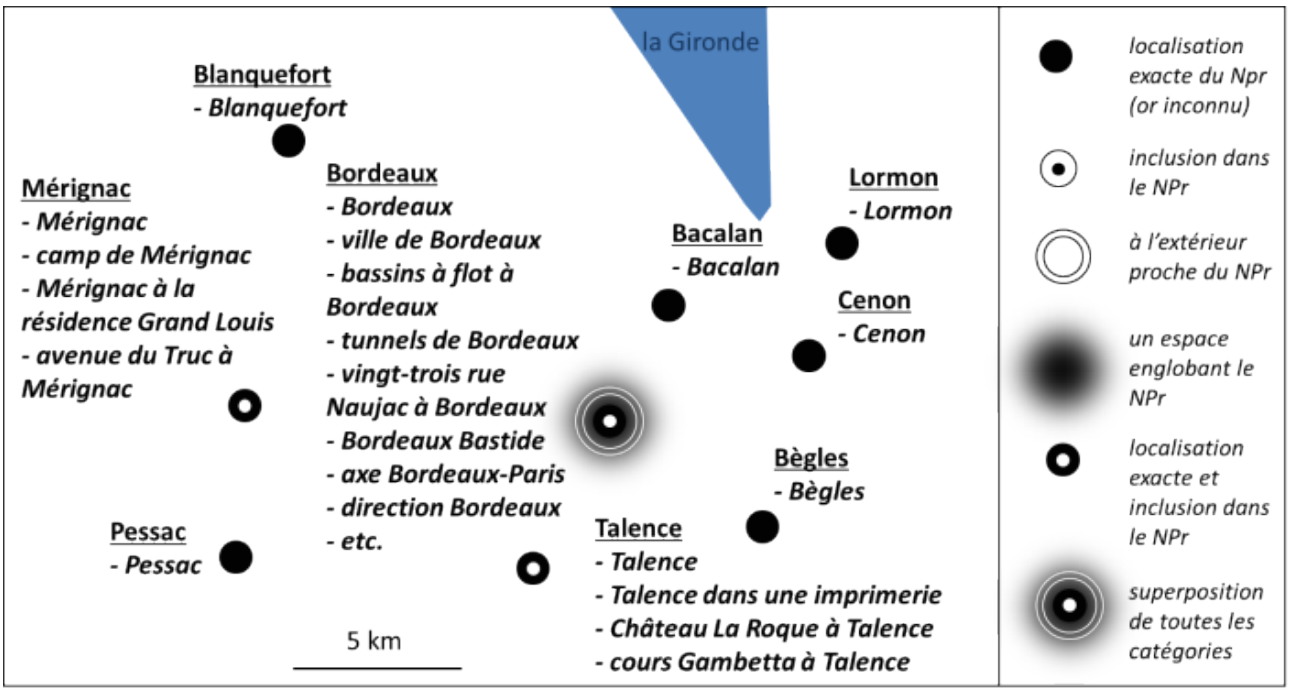


Figure 3 : Représentation cartographique de l'emprise géographique des Npr de lieux : les Npr (géolocalisés) et les Nc incluant un Npr dans la région de Bordeaux. Le symbole qui localise le Npr décrit le nombre de catégories où il est impliqué. La catégorie "inconnu" est ici assimilée à une localisation exacte.

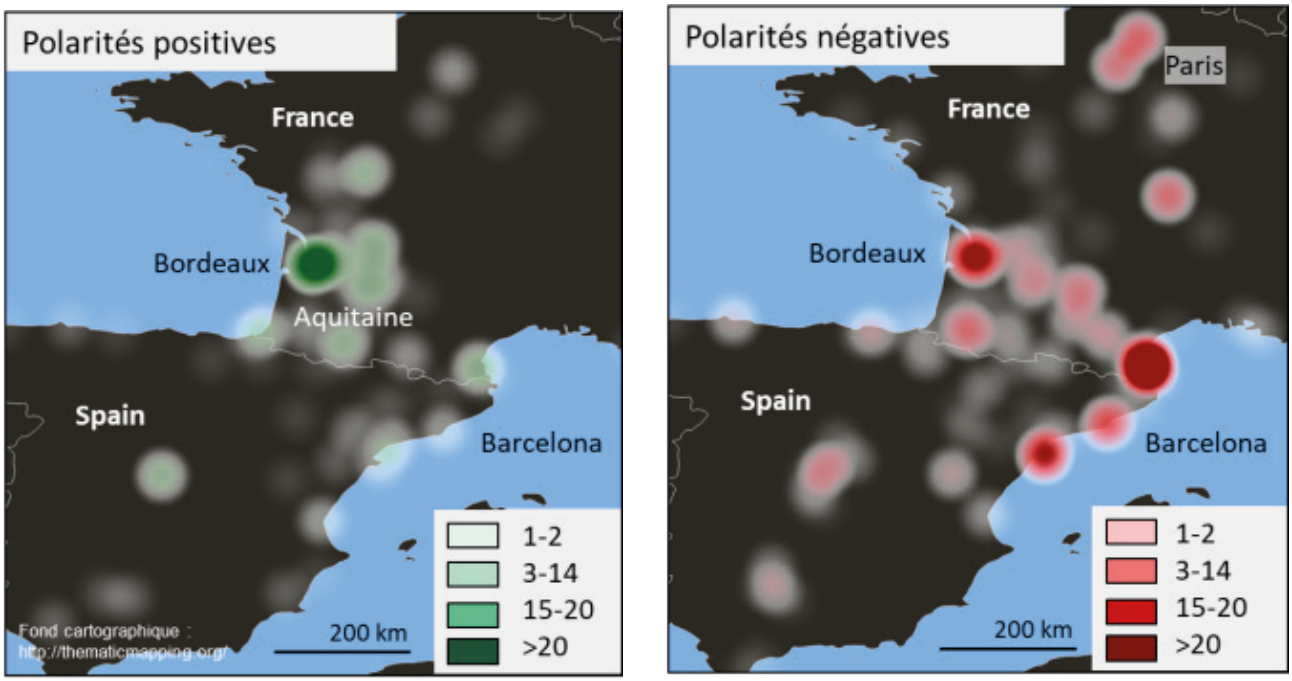


Figure 4 : Distribution spatiale des polarités associées aux Npr de lieux : cartes de chaleur centrées sur l'Espagne et la France des polarités a. positives et b. négatives. Les nombres d'occurrences sont additionnés par proximité spatiale en fonction de l'échelle spatiale des cartes.

les récits, et donc les témoins. Les résultats par Npr de lieux distincts sont les suivants : 7 % des Npr ont été annotés avec uniquement des polarités positives, 11 % avec uniquement des polarités négatives, 40 % avec deux ou trois polarités (positive, négative, neutre), et enfin 42 % avec uniquement des polarités neutres. Les distributions spatiales ont été cartographiées pour les Npr de communes. Polarités positives et polarités négatives ont été comparées grâce à des cartes de chaleur à une échelle régionale (figure 4) et à des cartes de décomptes d'occurrences à des échelles locales (figure 5).

À une échelle régionale, la région autour de Bordeaux ressort logiquement avec de fortes densités de lieux associés à la fois à des polarités positives et négatives. Concernant les polarités positives en particulier, la région autour de Barcelone et le sud de l'Aquitaine ressortent également, même si les nombres d'occurrences sont moins élevés que dans la région autour de Bordeaux. Concernant les polarités négatives, la côte méditerranéenne à la frontière espagnole (camps de regroupement) et des zones proches de Paris (manufactures) apparaissent.

À des échelles locales, les décomptes entre polarités positives et négatives peuvent être détaillés et cartographiés. En figure 5 par exemple, un équilibre entre polarités positives et négatives est observé pour la ville d'*Agen*, alors qu'une différence existe en faveur des polarités positives pour *La Sauvetat-du-Dropt*.

Similarités spatiales étudiées avec l'indice de Moran local (ou LISA).

Les LISA ont été calculés pour chaque Npr de lieux. Les valeurs sont cartographiées en figure 6, ici pour tous les types de Npr. Concernant les polarités positives (figure 6.a), une similarité spatiale élevée est observée dans la région autour de Bordeaux. Les dissimilarités spatiales apparaissent comme peu prononcées, sauf pour la localisation du Npr *France*, ce qui est peu concluant car la France est ici résumée par un point au centre de la surface du pays. Concernant les polarités négatives (figure 6.b), les similarités spatiales élevées sont localisées dans la région autour de Barcelone. Des dissimilarités élevées sont observées autour de Tolède (conséquence de la proximité de lieux d'enfance de témoins, et de fronts de bataille), la côte méditerranéenne française proche de l'Espagne (par ex. *Saint-Cyprien*, *Collioure*), l'Aquitaine (*Tarbes*, *Agen*), la région autour de Bordeaux (*Libourne*) et le long du fleuve Loire (*Nevers*). Les polarités positives et négatives considérées en même temps (figure 6.c) aboutissent à des similarités spatiales relativement équilibrées parmi les polarités pour des lieux proches. Même si les similarités et dissimilarités sont moins fortes que pour les polarités négatives, les mêmes zones ressortent, sauf l'Aquitaine et le long de la Loire.

Les valeurs de LISA qui sont significatives permettent de mettre en évidence des Npr qui correspondent également à des nombres d'occurrences importants (par ex. *Saint-Cyprien*). Dans ces cas, des similarités et dissimilarités locales sont associées à de nombreux mots de sentiments formulés par les témoins, décrivant ainsi des perceptions variées. Des perceptions contrastées existent aussi entre récits, portant sur des parcours de vie différents, mais dans des lieux relativement proches dans le corpus.

Discussion

Cette partie revient sur la démarche et les perspectives envisagées concernant les annotations des lieux et des informations associées, ainsi que leur analyse spatiale et leur cartographie.

Perspectives d'amélioration des annotations des Npr et des Nc de lieux et des mots de sentiment

Les approches proposées pour l'annotation automatique appellent des améliorations en fonction des informations ciblées. Concernant le Npr de lieux, une désambiguïsation entre les entrées identiques du gazetier reste une tâche à implémenter, même si le corpus est limité et que celle-ci peut être menée manuellement comme cela a été le cas. Plusieurs approches existent, par exemple dans Hoffart *et al.* (2011) et Nguyen *et al*. (2016), et devraient être adaptées au corpus et son contexte.

L'annotation des Nc de lieux par apprentissage donne de bons résultats. Toutefois des erreurs subsistent, dues à un corpus d'entraînement restreint, à la rareté de certaines formes de Nc dans le corpus d'entraînement, et à certains mots polysémiques pouvant désigner à la fois un lieu géographique et autre chose. Des perspectives d'amélioration concernent les confusions entre les balises B-LOC et I-LOC. Dans l'exemple *usine d'armements*, *usine* a bien été annoté en B-LOC (vrai positif) mais *d'armements* n'a pas été annoté alors qu'il devrait normalement l'être en I-LOC (faux négatif). En plus de la confusion entre balises, des faux négatifs et des faux positifs pourraient être corrigés, par exemple dans les résultats d'annotation présentés : *arènes de Madrid* qui n'a pas été annoté, et équipe d'Agonac qui a été annoté en

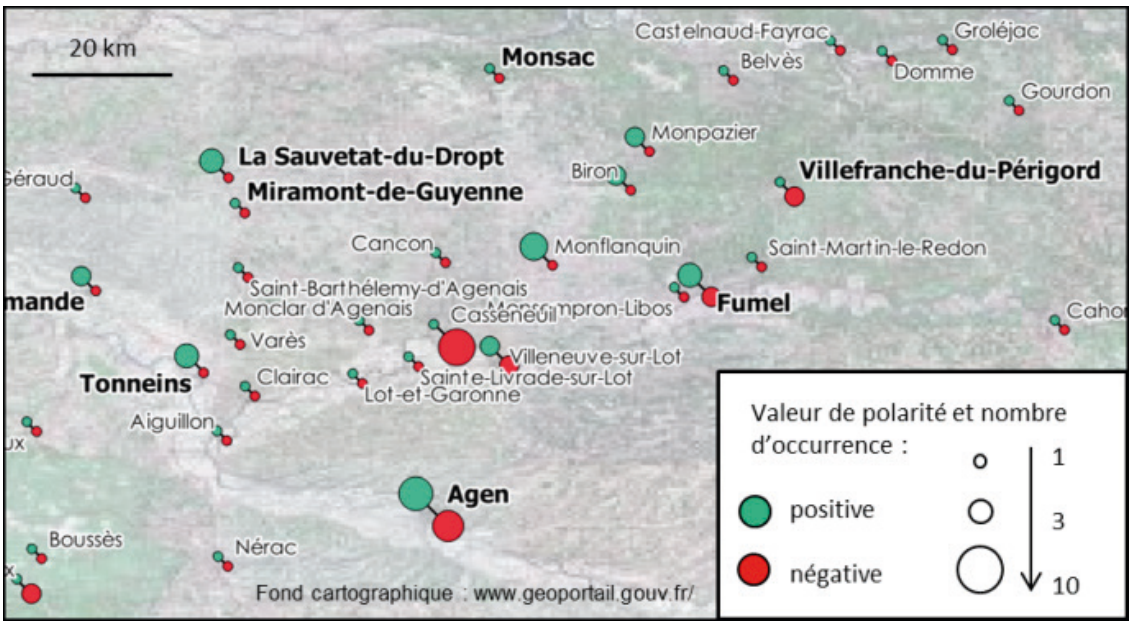


Figure 5 : Décompte des polarités positives et négatives attribuées par Npr de lieu. La carte est centrée dans le sud-ouest de la France.

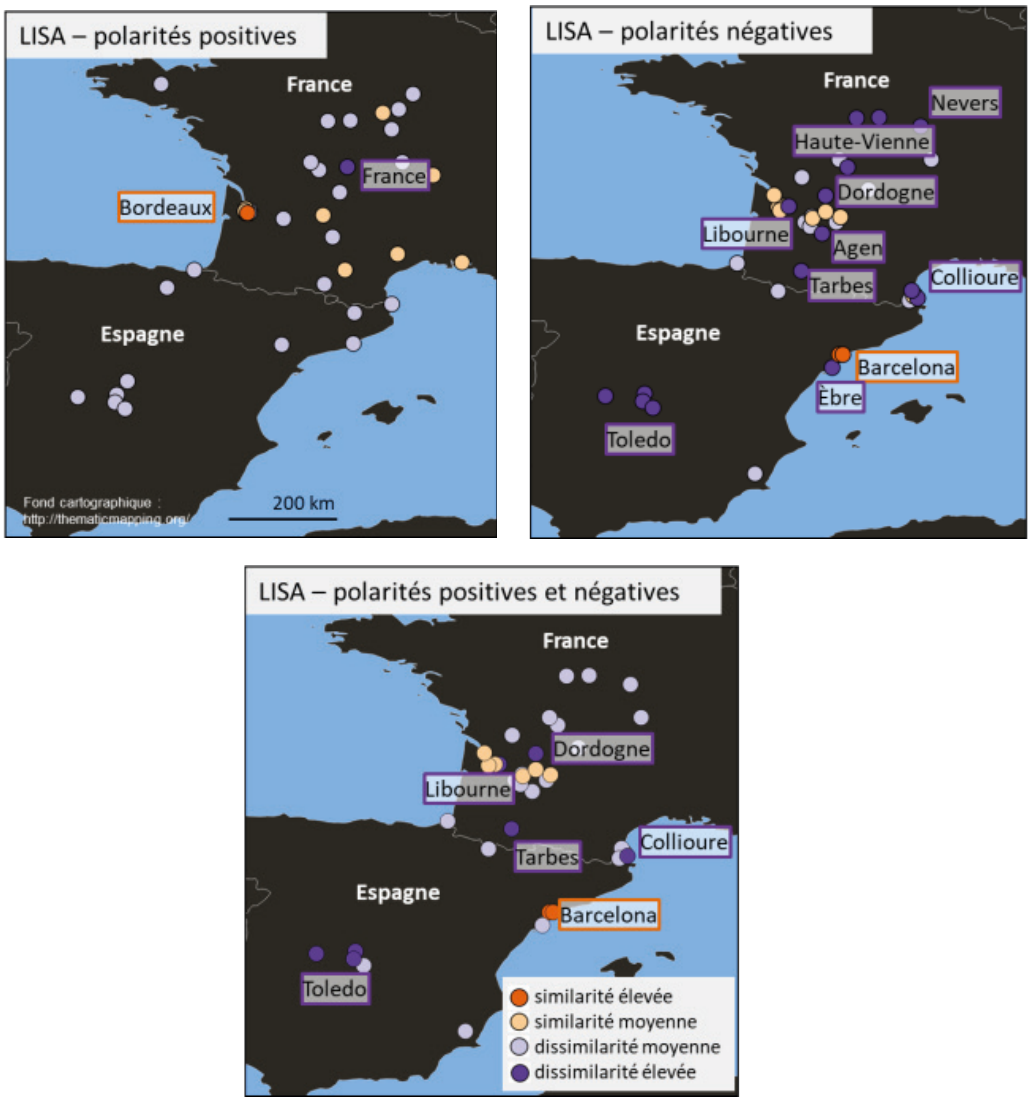


Figure 6 : Similarité spatiale locale entre Npr de lieux relativement à leur polarité : ont été représentés les Npr avec des LISA significatifs calculés pour a. les polarités positives, b. les polarités négatives, c. les positives et négatives.

Nc de lieu alors qu'il s'agit d'une équipe de sport. Des informations ciblées pourraient être ajoutées au schéma d'annotation, notamment des mots indiquant des relations spatiales entre les lieux telles que *près de* ou *au nord de* (Clementini *et al.,* 1993 ; Egenhofer *et al.,* 1991). Lorsqu' un Nc de lieu inclut un Npr localisé, la localisation du Npr est propagée au Nc, même si cette localisation doit être affinée. La propagation entre Npr et Nc est à définir lorsqu'il n'y a pas d'inclusion, par exemple en utilisant la proximité dans le texte ; un tel algorithme reste un défi.

Des perspectives de recherche portent également sur l'enrichissement du lexique des mots de sentiment et sur leur propagation aux noms de lieu. Le lexique devrait être complété par du vocabulaire spécifique au contexte de la guerre et en particulier de la Seconde Guerre mondiale et de la guerre civile espagnole. La propagation des polarités de perception à partir des mots de sentiments est manuelle dans le travail présenté. La définition d'une approche automatique de propagation bénéficierait de la prise en compte des négations qui peuvent inverser les polarités, et d'éléments de contexte tels que des faits énoncés, par ex. *on avait faim*.

Cartographie *ad hoc* de récits de vie

La création de cartes à échelle locale

Les trois échelles spatiales identifiées dans le corpus permettent de visualiser des distributions spatiales pertinentes des lieux mentionnés par les témoins. Elles permettent d'alternativement détailler et contextualiser les distributions, par exemple les zones où les densités spatiales sont importantes. Des cartes ad hoc pourraient donc être produites, notamment à l'échelle locale. Une expérimentation a été menée sur la ville de Bordeaux. Des représentations cartographiques ont été créées en utilisant des cartes de l'IGN des années 1950 et des images aériennes historiques. Ont été ajoutés les Npr de lieux annotés avec les gazetiers, ainsi que les lieux correspondant à certains bâtiments remarquables (ex. *bassins à flot à Bordeaux*) et à des adresses localisées grâce au service en ligne de géocodage de l'IGN (ex. *8 rue Constantin à Bordeaux*).

Enrichissement des cartes par les emprises géographiques des lieux

Les Npr de lieux et les Nc incluant un Npr de lieu ont été représentés par des géométries ponctuelles correspondant à celles du gazetier, parfois adaptées selon l'emprise spatiale. Ces propositions doivent être évaluées (Harrie *et al.,* 2015 ; Fabrikant *et al.,* 2012). Une perspective est de symboliser un même lieu en fonction des différentes annotations associées, Npr et Nc. Les cartes des lieux Npr et Nc, pourraient être enrichies par des emprises spatiales plus représentatives de l'emprise réelle de ces lieux que celles ponctuelles utilisées. Les bases de données contenant des géométries surfaciques (communes, régions) et linéaires (rivières, routes) pourraient être mobilisées, telles que celles produites par l'IGN et Natural Earth.

Une perspective de travail concerne aussi l'identification et la représentation cartographique des Npr modifiés ainsi que des mots exprimant des relations spatiales, ce qui complèterait les catégories des emprises géographiques arrêtées dans cet article. La localisation des Nc de lieux n'incluant pas de Npr augmenterait la connaissance de la répartition spatiale des lieux cités par les témoins. Les localisations et les implantations géographiques pourraient être imprécises, et une sémiologie adaptée devra être définie. La représentation cartographique de l'incertitude est un enjeu car elle cumule toutes les imprécisions dues aux traitements automatiques et aux choix de modélisation, de construction des ressources, etc. Elle constitue une information complémentaire de la représentation des lieux, qui éclairerait l'interprétation des résultats et la lecture des cartes (MacEachren *et al.,* 2005).

L'ajout de la temporalité attachée aux lieux

Les lieux et les informations associées ont été abordés et cartographiés en rapport avec les nombres d'occurrences et la géolocalisation, permettant d'obtenir une connaissance synthétique de la géographie du corpus. Un travail similaire a été réalisé pour chacun des récits de vie. Des informations de temporalité pourraient s'ajouter aux informations décrivant les lieux mentionnés, par exemple les dates et les durées de présence, dans chacun des lieux, des témoins ou des personnes citées. Des cartes de trajectoires correspondant à une sélection de lieux, ont déjà été produites par le RAHMI et diffusées sur leur site[8]. Cette approche par identification des parcours chronologiques entre les lieux rend compte des déplacements dans l'espace (Morcrette, 2016) et s'ajoute aux clés d'interprétation des changements de perceptions dans le temps vis-à-vis d'un lieu.

Conclusion

La cartographie et l'étude des distributions spatiales des noms de lieux dans un corpus de récits de vie permettent un accès à la géographie du corpus, dans son ensemble ou par récit de vie. Des points communs et des divergences ont pu être identifiés entre les lieux cités par les témoins et les manières dont ils sont cités. Il s'agit d'une démarche d'abord exploratoire concernant la caractérisation des lieux, qui inclut leurs perceptions, en vue de la formulation d'hypothèses de recherche concernant le possible rôle déterminant de certains lieux dans ces récits et les informations additionnelles à envisager pour définir les lieux. Les informations annotées, notamment la perception, peuvent être situées par rapport à un voisinage spatial similaire ou contrasté. Elles peuvent évoluer dans le temps, pendant la durée des événements racontés, puis être altérées par le souvenir (plusieurs dizaines d'années séparant les faits du récit qui en témoigne). Les synthèses des analyses et les cartes ont vocation à être utilisées en support de travaux, notamment en géographie et en histoire, et contribuent à la valorisation des récits de vie oraux des Républicains espagnols.

Bibliographie

Boccagni, P. Baldassar, L. (2015). "Emotions on the move: Mapping the emergent field of emotion and migration". *Emotion, Space and Society*, Volume 16, p. 73-80.

Caquard, S. (2017). "Cartography I: Mapping narrative cartography". *Progress in Human Geography*, Volume 37 (1), p. 135144.

Clementini, E., Di Felice, P., Van Oosterom, P. (1993). "A small set of formal topological relationships suitable for end-user interaction". *Advances in Spatial databases*, Volume 692, p. 277-295.

Cunningham, H. *et al* (2013). "Getting more out of biomedical documents with GATE's full lifecycle open source text analytics". *PLoS computational biology*, Volume 9 (2), en ligne.

Dominguès, C., Jolivet, L., Brando, C., Cargill, M. (2019). « Place and Sentiment-based Life story Analysis". *Revue française des sciences de l'information et de la communication*, Volume 17, en ligne.

Egenhofer, M., Franzosa, R. (1991). "Point-Set Topological Spatial Relations". *International Journal of Geographical Information Systems*, Volume 5 (2), p. 161-174.

Fabrikant, S. I., Christophe, S., Papastefanou, G., Maggi, S. (2012). Emotional response to map design aesthetics. Proceedings of GIScience, 18-2 septembre 2012, Columbus, Ohio, USA.

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),* p. 363-370.

Griffin, A.L., McQuoid, J. (2012). « At the Intersection of Maps and Emotion: The Challenge of Spatially Representing Experience". *Thematische Kartographie*, Volume 62 (6), p. 291-299.

Harrie, L., Stigmar, H., Djordjevic, M. (2015). "Analytical Estimation of Map Readability". *ISPRS International Journal of Geo-Information*, Volume 4 (2), p. 418446.

Hoffart, J., Yosef M. A., Bordino, I., *et al* . (2011). Robust Disambiguation of Named Entities in Text. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, p. 782-792.

Huyghe, R. (2012). « Evénements et localisation spatiale : quand le lieu est un événement ». In *Langue, espace, cognition*, Fagard, B. and Stosic, D. (eds), *Corela* numéro thématique.

Leroy, S. (2004). *Le nom propre en français*. (Coll. L'essentiel Français), Ophrys.

MacEachren, A. M., Robinson, A., Hopper, S., *et al.* (2005). "Visualizing Geospatial Information Uncertainty: What We Know and What We Need to Know". *Cartography and Geographic Information Science*, Volume 32 (3), p. 139-160.

Morcrette, Q. (2016). « L'espace récité de la carte d'itinéraire: exemples français et américains ». *Mappemonde*, n° 118.

Nguyen, D. B., Theobald, M., Weikum, G. (2016). "J-NERD: Joint Named Entity Recognition and Disambiguation with Rich Linguistic Features." *Transactions of the Association for Computational Linguistics*, Volume 4, p. 215–229.

Olmedo, É (2016). “Pour une cartographie affective des récits des femmes de Sidi Youssef Ben Ali (Marrakech, Maroc) ». In Fournier, M. (dir.), *Cartographier les récits*, Clermont-Ferrand, Presses universitaires Blaise Pascal, Collection CERAMAC, p. 179-196.

Purves, R. S., Winter, S., Kuhn, W. (2019). “Places in Information Science”. *Journal of the association for information science and technology*, Volume 70 (11), p. 1173-1182.

Ramshaw, L. A., Marcus, M. P. (1995). “Text Chunking using Transformation-Based Learning”. In *Natural language processing using very large corpora*, Springer, Dordrecht, 1999, p 157-176.

Tuan, Y.-F. (1977). *Space and place: The perspective of experience.* University of Minnesota Press, Minneapolis, Minnesota, USA.

Zeile, P., Resch, B., Exner, J.-P., Sag, G. (2015). “Urban Emotions: Benefits and Risks in Using Human Sensory Assessment for the Extraction of Contextual Emotion Information in Urban Planning”. In *Planning Support Systems and Smart Cities*, Geertman, S. *et al*. (eds), Springer International, p. 209-225.