



Observer automatiquement l'évocation des
lieux de Paris dans la chanson francophone :
adaptabilité de la reconnaissance d'entités
nommées et True Casing

Mémoire
présenté par :
DE CARVALHO BENE Tiago André
17003974
« Langue et Informatique »

Table des matières

1 Présentation du sujet	2
1.1 Les lieux de Paris dans la chanson francophone	2
1.2 Reconnaissance d'entité nommée	3
1.3 Plan du Mémoire	4
2 Bibliothèques et lexiques utilisées	5
2.1 SpaCy	5
2.2 GLÀFF	7
3 Paroles des chansons et corpus	10
3.1 Les paroles	10
3.2 Corpus échantillon 1	14
3.3 Corpus complet	16
3.4 Lexique de lieux de Paris	16
3.5 Verrous gênant la Reconnaissance d'Entités Nommées	16
4 Étude de spaCy, amélioration et résultats	20
4.1 Premières observations sur les résultats de spacy sur le corpus d'échantillon de Jacques Dutronc	20
4.1.1 Les différentes proportions calculées sur les modèles de spacy	21
4.1.2 Diagramme de Venn sur les 3 modèles de spacy	30
4.2 Pré-traitement du corpus	31
4.3 Observation des résultats sur le corpus échantillon de Jacques Dutronc pré-traité.	35
4.4 L'éloignement des classes populaires de Paris.	36
4.5 Enrichissement du lexique des lieux de Paris	50
5 Conclusion	52

Chapitre 1

Présentation du sujet

1.1 Les lieux de Paris dans la chanson francophone

Aujourd’hui, la chanson est omniprésente dans la culture populaire. Nous la retrouvons au quotidien dans les transports, dans la rue, ou encore à la télé, sur nos téléphones. C'est avec le partage sans limites à travers des applications ou des réseaux sociaux comme Youtube, Instagram, Spotify, sans oublier Apple Music, que la chanson réussit en grande majorité, à transmettre des émotions à ceux qui la consomment. Les artistes ne cessent d'innover dans ce domaine et l'évolution des mentalités au sein d'une société ou d'une communauté peut ainsi être mesurée à l'un des thèmes qui sont abordés selon les périodes temporelles.

C'est avec ce flux de données constamment renouvelé que l'on peut extraire des informations pertinentes, linguistiquement parlant dans les paroles des chanteurs qui composent notre monde.

L'objectif du mémoire, orienté linguistique de corpus, est d'étudier dans quelle mesure les noms de lieux évoqués dans la chanson française ont évolué de la moitié du XXème siècle jusqu'à aujourd'hui. Les données étant multilingues, la volonté de vouloir travailler sur les chanteurs francophones est d'une part une tâche conséquente, mais aussi présentant de nombreuses variations des formes, ainsi le/la chanteur/chanteuse francophone peut aussi bien s'exprimer en français comme en anglais dans le même couplet. Nous n'étudierons pas tous les noms de lieux ; le sujet va se restreindre aux lieux Parisiens : monuments, bâtiments, rues . . . afin d'observer leur importance à travers le temps dans cette partie de la culture francophone que constitue la chanson.

Paris reste en effet le lieu le plus mentionné dans la chanson française

[Leclanche, 1998]. Étant la ville lumière, celle de la mode, ou encore celle de l'amour, Paris est, si l'on peut dire, la ville de la réussite, des rêves et de l'avenir. Étant internationalement connue, on peut s'attendre alors à la mention de lieu touristique ou bourgeois dans les chansons populaires francophones. Afin d'identifier l'importance des lieux de Paris, nous allons vouloir détecter les différentes mentions de ces lieux et en connaître la fréquence d'apparition dans les paroles. On pourra ainsi d'une pierre de coup déduire si un chanteur aime parler spécifiquement d'un quartier, d'un lieu ou d'une adresse en particulier. Pour ce faire, nous disposons d'un outil de Reconnaissance d'Entités Nommées (REN/NER) reconnu, nous possédons aussi une ressource listant les noms des lieux de Paris classés par catégories.

Une des hypothèses est que l'évolution de Paris vers une ville globale [Sassen, 1996] accompagnée d'un éloignement des classes populaires vers des banlieues de plus en plus éloignées [Chevalier, 1958, Clerval, 2013]) se traduit dans la culture populaire et notamment dans les occurrences de noms de lieux dont il est question dans la chanson.

Par exemple, les chanteurs réalistes d'avant-guerre (Fréhel, Aristide Bruant, Edith Piaf) habitaient dans un Paris populaire qu'ils relataient dans leurs chansons.[Wikipédia, 2022b] Ils le font par exemple par des références à des lieux spécifiques : cafés, rues, squares. Ceci s'expliquerait notamment par le fait que leur public est avant tout Parisien. Au contraire, pour des interprètes plus contemporains (Renaud, Mano Solo, Bénabar ...), les représentations sont différentes ; il s'agit plus de lieux beaucoup plus emblématiques de Paris et donc identifiables par des non-Parisiens. Qu'en est-il pour un registre de la chanson populaire tel que le rap ? C'est une question qui se pose également. De façon globale, les objectifs de ce mémoire sont :

- compléter les métadonnées d'un corpus existant en ajoutant les dates d'écriture des chansons
- établir une cartographie en diachronie des lieux évoqués
- combiner différentes approches de REN pour améliorer les résultats

1.2 Reconnaissance d'entité nommée

La Named Entity Recognition ou Reconnaissance d'entité nommée est une des tâches du TAL¹ (ou NLP²). Le principe de cette tâche est d'identifier, récupérer et classer les noms des lieux, les noms d'organisations ou

1. Traitement Automatique des langues
2. Natural Language Processing

encore les noms de personnes apparaissant dans un document texte, quel qu'il soit. En effet, le but du TAL demande des connaissances dans les disciplines qui rassemblent l'informatique et la linguistique, car le langage possède une structure complexe et ambiguë. Il est difficile de conclure à des résultats seulement par la grammaire générative que sont les algorithmes.

La tâche de reconnaissance d'entité nommée n'est pas une tâche aussi simple qu'il y paraît. Car outre le fait qu'une entité dans un texte se définit normalement par la présence de majuscules, l'entité se définit aussi dans son contexte, mais aussi dans son sème : une personne peut posséder le nom d'un objet, d'une entreprise ou d'un lieu. spacy, construit et entraîné sur des articles Wikipédia, va user de pipeline, de vecteur multi-dimensionnel et d'I.A. afin de réaliser la difficile tâche de détection d'entités nommées.

1.3 Plan du Mémoire

En premier lieu, dans le chapitre 2, je vais présenter les outils qui m'ont permis de réaliser la tâche de détection d'entités nommées, ce qu'il est possible de réaliser avec ces outils ainsi que leurs avantages et leurs inconvénients. En second lieu, je présente, dans le chapitre 3, les corpus sur lesquels j'ai travaillé, expliquant en même temps la structure, la complexité de la tâche, ce qui peut manquer ou être amélioré.

Ensuite, dans le chapitre 4, je présenterai les tâches effectuées pour réaliser et valider mon hypothèse. Enfin, dans le chapitre 5, je ferai des analyses linguistiques sur mes résultats et je déboucherai sur une conclusion de ce qu'il faut retenir et apporterai des perspectives d'amélioration.

Chapitre 2

Bibliothèques et lexiques utilisées

2.1 SpaCy

SpaCy, développé par Ines Montani et Matt Honniba [Honnibal and Montani, 2017], est une bibliothèque Python sous licence MIT¹. Disponible sur le site officiel de spaCy², il possède un guide pratique sur son installation et son utilisation sous différents systèmes d'exploitation (Windows, Linux, macOS/ OSX). Actuellement dans sa version 3.3, il fonctionne sur la version 3 de python.

SpaCy est aujourd’hui capable de traiter plus de 66 langues et possède 73 pipelines³ entraînées pour 22 langues. Il possède aussi des modèles pré-entraînés et un apprentissage en multitâches en utilisant des transformers comme BERT⁴.

SpaCy propose aussi un *pipeline* personnalisable, si celui proposé n'est pas assez performant pour un problème donné. Il est aussi possible d'accéder à des cours interactifs en ligne.

Finalement, la bibliothèque possède une communauté active sur Github avec plus de 3 800 forks⁵ et 23 500 étoiles. Ils sont notamment présents sur twitter et possèdent une chaîne youtube où ils publient encore des vidéos.

1. Copyright (C) 2016-2022 ExplosionAI GmbH, 2016 spacy GmbH, 2015 Matthew Honnibal

2. <https://spacy.io/>

3. Les pipelines sont le traitement d'un texte avant de l'utilisation à la tâche.

4. Bidirectional Encoder Representations from Transformers est un projet de recherche permettant de comprendre le langage naturel

5. sur github un forks correspond est la copie d'un travail pour pouvoir travailler dessus

Token	Lemme	POS	Tag	Dep	Shape	Alpha	Stopword
Bonjour	Bonjour	PROPN	PROPN	advmod	Xxxxx	True	False
,	,	PUNCT	PUNCT	punct	,	False	False
je	je	PRON	PRON	nsubj	xx	True	True
suis	être	AUX	AUX	cop	xxxx	True	True
Tiago	Tiago	PROPN	PROPN	ROOT	Xxxxx	True	False
Etudiant	étudiant	ADJ	ADJ	acl	Xxxxx	True	False
en	en	ADP	ADP	case	xx	True	True
M1	m1	NOUN	NOUN	obl:mod	xd	False	False
langue	langue	NOUN	NOUN	amod	xxxx	True	False
et	et	CCONJ	CCONJ	cc	xx	True	True
informatique	informatique	ADD	ADD	conj	xxxx	True	False

FIGURE 2.1 – Exemple des résultats de spaCy analysant les mots de la phrase "Bonjour, je suis Tiago Étudiant en M1 langue et informatique

Qu'est ce que SpaCy ?

C'est une bibliothèque permettant plusieurs usages. Sa principale utilisation est celle utilisée dans ma recherche : la détection d'entité nommées. Il possède notamment d'autres fonctionnalités comme une analyse sur les mots/tokens comme on peut le voir sur la figure 2.1 qui illustre la manière dont spaCy analyse le token. Dans ce tableau, "token" correspond à l'unité de sens trouvée par spaCy ; le "lemme" correspond à sa forme canonique "suis" forme canonique → "être", "POS"⁶ sont les parties du discours que spaCy associe à la forme en contexte. Dans le domaine informatique, en syntaxe, les étiquettes POS peuvent être propres à un logiciel ou universelles⁷. Il peut aussi identifier si la phrase est une phrase nominale ou non. SpaCy peut afficher les dépendances syntaxiques comme on peut le remarquer sur la figure 2.2 - image qui illustre les dépendances syntaxiques d'une phrase - ou encore une fonctionnalité pour effectuer des calculs de similarité⁸.

Simple d'utilisation/installation et possédant 3 modèles d'analyse (small, médium, large), spaCy permet d'obtenir des résultats sans être un expert dans le domaine de la programmation.

Pour la NER, les trois modèles de détection de spaCy possèdent des pipelines déjà implémentées. Ce qui différencie les modèles dans leur construction sont

6. Part of Speech

7. récupérable sur le site <https://universaldependencies.org/u/pos/>

8. Quel mot est proche de quel mot, on peut faire le calcul sur une phrase, un texte, voire un corpus

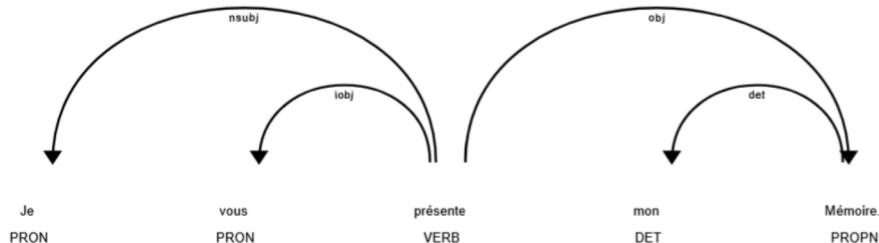


FIGURE 2.2 – Image de la construction des dépendances syntaxiques de spaCy sur la phrase "Je vous présente mon Mémoire"

les "word vector"⁹. Pour le modèle small, il fonctionne sans "word vecteur" mais les modèles medium et large possèdent respectivement 20 000 et 500 000 vecteurs.

C'est au cœur du domaine du NLP (ou TAL en français) qu'un outil, tel que spaCy avec toutes ses fonctionnalités, permet aujourd'hui de faire traiter du langage humain à l'ordinateur permettant ainsi de développer ou d'améliorer des applications en lien avec le domaine.

SpaCy n'est pas la seule bibliothèque permettant la détection d'entités nommées et il est souvent mis en comparaison et évalué par rapport à d'autres outils comme casEN [Nouvel et al., 2010] (utilisant des transducteurs), NLTK¹⁰, openNLP, Gate [Schmitt et al., 2019]...

L'utilisation de la bibliothèque spaCy dans ce mémoire permettra sûrement de tirer des conclusions sur le fonctionnement de celui-ci dans un corpus musical, où les paroles n'ont pas une structure de phrase, ou bien l'on peut voir apparaître des termes d'autres langues - point développé dans la partie 3.1.

2.2 GLÀFF

GLÀFF, acronyme pour "Gros Lexique À tout Faire du Français" [Sajous et al., 2013], est une ressource disponible en ligne construite à partir du wikitionnaire¹¹ contenant plus de deux millions d'entrées. Il est actuellement à sa Version 1.2.2 (22/12/2017) et permet de recenser les mots existant dans le

9. ou vecteur de mot, dans un vecteur, on place les mots pour en faire des statistiques et ainsi comprendre le sens du mot grâce à sa place dans le vecteur.

10. <https://www.nltk.org/>

11. wikitionary version française

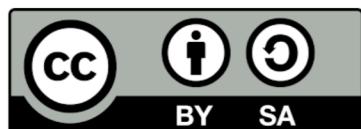


FIGURE 2.3 – Représentation de la licence Creative Commons By-SA

vocabulaire français. Pour chaque entrée, le GLÀFF indique une description morphosyntaxique et une transcription phonémique.

Il se distingue notamment par la taille du lexique, mais aussi par les indications de la forme phonétique d'un mot, de la présence des flexions, des synonymes, antonymes, hyponymes, hyperonymes et possède même des traductions.

Les lexiques contenant des gros volumes de données sont souvent payants ou non partageables car c'est la propriété intellectuelle d'une personne. Il faut donc une autorisation. Le GLÀFF, lui, est une ressource en accès libre sous licence Creative Commons By-SA comme le montre la figure 2.3, représentant les logos à utiliser lorsque l'on souhaite afficher la licence. C'est-à-dire que c'est une ressource non gouvernementale à but non-lucratif permettant aux auteurs de mettre leurs œuvres à disposition du public selon des conditions prédéfinies. Pour le GLÀFF, la licence BY-SA¹² requiert la citation de l'auteur. Il est possible d'y faire des modifications de l'œuvre. La licence doit toujours être citée dans chaque dérivé de l'œuvre originale et le partage est permis. Ainsi, le GLÀFF a été principalement conçu pour les communautés linguistiques, psycholinguistiques et TAL.

Comparaison avec différents lexiques français.

D'autres lexiques du français sont disponibles en ligne comme ABU¹³ qui contient une liste de mots communs (+300 000 mots), une liste de prénoms (12 437 prénoms), une liste de nom de cités française (39 076 noms), une liste de nom de pays (170 pays), une liste de difficultés de la langue (1 500 mots) ; le Lefff¹⁴ qui contient 404 634 formes et 105 595 lemmes ; Morphalou3 [ATILF, 2015] qui comprend 159 271 lemmes et 954 690 formes fléchies, du français moderne.

J'ai choisi de travailler sur le GLÀFF car c'est d'une part un lexique disponible en ligne sur lequel j'ai pu travailler en cours et, d'autre part un outil avec plusieurs millions d'entrées de formes fléchies qui, comparé aux

12. Share Alike

13. <http://http://abu.cnam.fr/>

14. <https://www.labri.fr/perso/clement/lefff/>

autres lexiques disponible, possède un volume de données beaucoup plus considérable. Le GLÀFF me permettra d'affiner mes recherches lors de mes travaux.

Chapitre 3

Paroles des chansons et corpus

3.1 Les paroles

Les paroles dans une chanson ont une construction particulière et diffèrent en fonction des époques, mais aussi des chanteurs. Il me semble alors intéressant de comprendre du point de vue du temps, mais aussi de la structure des paroles, les comportements des paroles de chansons qui peuvent peupler mon corpus.

Le temps d'une chanson

De nos jours, avec le développement du numérique, la sur-information est quelque chose de presque inévitable. Que ce soit dans la rue, les transports ou encore chez soi, à la maison, nous sommes pratiquement tous soumis à des informations auditives et/ou visuelles. Ce phénomène de sur-information est telle que les personnes ne possédant pas d'ordinateurs, de téléphones, de télévisions ou même de radios seront facilement qualifiées de marginaux - c'est-à-dire de personnes vivant en marge de la société.

Parmi tout ce flux d'informations quotidiennes, l'audio occupe une place importante ; au restaurant, ou encore dans les ascenseurs d'hôtels, il est désormais courant d'entendre des tubes des années quatre-vingt à nos jours. En vue de tout cela, il serait intéressant, voire judicieux de se demander comment s'y prennent les chanteurs visant la classe populaire pour faire accrocher leur auditoire dans ce monde qui tend vers une sur-information abondante.

Dans ce monde rempli de publicités, l'information se doit d'être rapide et efficace. Les publicités télévisées sont fabriquées de façon à attirer le potentiel

client en un temps restreint ; c'est-à-dire environ 21 secondes (en moyenne)¹. De plus, nous observons que le format des chansons populaires, lui, tourne autour de ce qui équivaut à 150 secondes. Ce format d'une durée limitée est dû au format de l'époque où les disques phonographiques étaient limités à 78 tours par minute. Avec le progrès technologique d'aujourd'hui, nous pouvons stocker des masses de données assez importantes. Le format des 150 secondes persiste, ainsi il est ancré dans les moeurs. Cela démontre que la chanson populaire souhaite se consommer de manière assez rapide. De surcroît, nous pouvons également prendre l'exemple des radios numéro 1 en France. En effet, des radios comme Skyrock² ou Generation 88.2³ ont un format bien précis et peuvent se permettre de ne pas ajouter une chanson dans leur playlist en raison de sa longueur excessive. Avec cet exemple, nous observons donc que la structure a une place importante du point de vue du temps.

Poésie VS paroles de chansons

Quand on pense paroles de chansons souvent on pense aussi aux poésies - c'est une sorte d'assimilation à laquelle on insère poésie et parole de chanson dans le même panier. Cela est sûrement dû au fait que l'on peut retrouver une structure similaire comme des vers à chaque ligne où l'on inscrit des rimes plates, embrassées ou encore croisées. Mais les paroles de chanson sont malgré tout ce que l'on peut dire différentes de la poésie.

En effet, la poésie est structurée, fonctionne en ensemble de strophes structurées en vers. Mais admettant la prose, elle est régie par une métrique, servant à amener la sonorité et l'agencement dans le choix des mots. "La poésie est un outil artistique qui utilise le langage pour créer une composition"⁴. Elle peut se réciter comme se chanter et est considérée comme œuvre littéraire, au contraire de la chanson. Ne connaissant pas entièrement le domaine de la poésie, hormis les périodes du classicisme et humanisme, il n'était pas courant de reprendre des poèmes afin d'honorer et de célébrer les poètes anciens. On ne retrouve peut-être pas des reprises du Corbeau et du Renard dans les poésies modernes, comme on peut retrouver une reprise

1. © Statista 2022, Information donnée sur le site Statista, sur une période d'enquête des années 2009 à 2020. <https://fr.statista.com/statistiques/694581/duree-moyenne-des-spots-publicitaires-television-france/>

2. <https://www.skyrock.com>

3. <https://generations.fr>

4. D'après le site <https://fr.sawakinome.com/articles/language/difference-between-poetry-and-song.html>

de "BELLE" de "Luc Plamondon" chanter par "GAROU"⁵ et reprise aujourd'hui par "GIMS", "DADJU" et "SLIMANE"⁶ (3 chanteurs populaires aujourd'hui).

Les paroles des chansons sont divisées en couplet et refrain et non régies par des règles de structure (peut exister sans refrain ou couplet). Les paroles se construisent par le rythme de la musique, mais cette dernière peut exister sans elles. Elle a comme objectif de s'ancrer dans la mémoire en jouant sur le rythme et les paroles qui peuvent être répétitifs.

La chanson populaire fait en sorte d'être renouvelée à chaque nouvelle sortie, tout en reprenant les codes des anciennes chansons. Elle peut même être le témoin direct de l'évolution de la parole avec de nouveaux termes propres aux langages des chanteurs qui peuvent être issus des lieux où ils vivaient ou vivent encore. Ce que nous pouvons relever de commun entre la chanson populaire du XXème siècle et celle de nos jours, ce sont les lieux qui d'une certaine manière reste inaltérés. Nous pouvons retrouver par exemple dans les paroles de "Serge Gainsbourg" dans "Le Poinçonneur des Lilas" le lieu lilas ou dans cette chanson, il parle du métro, tout comme le rappeur "Sefyu" dans "Invaincu" qui parle lui aussi de ce lieu.

structure des paroles

Les paroles de chansons peuvent respecter une certaine structure. Cela peut dépendre de l'artiste, de s'il est accompagné ou non, et du type de la chanson. Malgré tout, on retrouve beaucoup de chansons qui respectent une certaine structure. La plus connue, il me semble, est intro/couplet suivi du refrain suivi d'un ou plusieurs couplets, puis à nouveau refrain et finalement outro/couplet (ABABA). D'autres chanteurs préfèrent ne mettre uniquement le refrain car la musique entraînante suffit, voire des paroles sans refrain - le tout sous un temps limité. Chaque partie de cette structure musicale sert d'indication à l'artiste s'il souhaite respecter celle-ci.

L'introduction (Intro) Elle décrit dès les premières secondes, le rythme et le thème de la chanson. Il peut parfois, à la place de l'intro, y avoir un couplet.

Le refrain est la partie répétée dans la chanson. Il revient au moins 2 fois et doit donner à l'auditeur une vive émotion.

5. <https://www.youtube.com/watch?v=MEODTN06mJE>

6. <https://www.youtube.com/watch?v=BNWM-cZhqLU>

Les couplets Ce sont les parties qui présentent du contenu nouveau, faisant évoluer le récit et ajoutant de la profondeur dans la chanson, si les paroles possède 2 chanteurs. En général, les chanteurs alternent chacun leur tour couplet 1 pour le chanteur 1 puis couplet 2 pour le chanteur 2.

L'outro Elle conclut la chanson. Opposée à l'intro, elle prend plusieurs formes comme un "fade-out"⁷. Cette structure est construite en général sur papier par une indication ou si elle manque par un double saut de ligne pour marquer la séparation.

D'autres éléments ajoutent du contenu, mais n'étant pas forcément essentiel à la structure, comme les ad-libs⁸, mais plutôt au remplissage de la mélodie sur laquelle le chanteur va jouer avec cela pour améliorer la sonorité, ou encore laisser sa patte d'artiste comme une sorte de signature vocale.

D'après l'article Wikipédia sur la segmentation d'une musique [Wikipédia, 2022h], le retour à la ligne, à l'instar de la poésie, n'est pas l'indication d'une fin de vers proprement dite, mais elle est liée à la fin d'une mesure. La mesure se segmente en temps⁹. Ils sont eux-mêmes structurés par une accentuation (forts, faibles) ou pulsation - composant la structure rythmique qui se répète de manière cyclique.

Mais certaines fois, le texte musical ne suit pas le rythme, par effet de style ou par simple volonté de l'artiste. Nous constatons ainsi que la sonorité et les paroles ont une place importante au niveau de la structure qui est classée alors comme artistique.

Ceci soulève un point important et est de plus en plus évoqué, que ce soit dans le monde de la chanson, ou dans tout autre domaine qui produit du contenu - problèmes qui sont les droits d'auteurs. Tout créateur a le droit à la reconnaissance de son travail. Dans le domaine de la musique, qui est une forte source de revenus, chacun aimeraient tirer un profit et les reprises, qui ne sont pas nécessairement constitutives de plagiat dans ce domaine, sont monnaie courante. C'est grâce aux droits d'auteurs qui, aujourd'hui, sur-protègent ses musiciens empêchant par exemple, un "influenceur"¹⁰ sur Youtube de mettre une vidéo avec une chanson sans se faire bannir de la plate-forme, car ne possédant pas les droits. La chanson étant omniprésente, il est un comble de devoir demander à chaque créateur de la diffuser (ceci reste cependant mon point de vue). On peut donc se poser la question, comment

7. baisse du volume de la chanson de manière progressive.

8. les ad-libs sont des sons ou des mots prononcé entre les couplets ou fin de ligne par le chanteur pour compléter la mélodie de la musique qu'il peut juger nécessaire.

9. les temps étant des unités pour mesurer la durée.

10. Appelés ainsi du fait de leur métier qui consiste à poster des vidéos qui ont le pouvoir d'influer sur la consommation/information/... des gens.

récupérer les paroles de chansons à des fins d'étude sans être dans l'illégalité ?

Ainsi, grâce à mon directeur de mémoire, nous avons pu récupérer, non sans mal, sur des sites à données restreintes, des paroles de la chanson française grâce à l'API Genius¹¹ et me les fournir. En définitive, je possède 2 corpus. Un échantillon sur les paroles de chanson de Jacques Dutronc qui m'a permis d'étudier le fonctionnement de spaCy et de créer des tâches qui s'exécuteront sur le corpus complet. Il m'a notamment permis d'observer les phénomènes et les problématiques qu'offre la lecture des données et ainsi éviter d'avoir des erreurs sur le corpus complet qui possède un volume d'informations conséquent ralentissant ainsi le traitement et un corpus complet comprenant plusieurs chanteurs francophones.

3.2 Corpus échantillon 1

Mon premier corpus est une liste des paroles de chansons de Jacques Dutronc. Il contient 92 chansons de Jacques Dutronc dont 42 qui possèdent des paroles et 50 qui n'en possèdent pas. Tout n'est pas disponible et dépend des contributeurs. L'API Genius n'a donc pas permis de récupérer toutes les paroles de chansons, pour certaines nous n'avons que les métadonnées.

C'est un fichier json¹² qui contient en l'occurrence une liste de dictionnaire contenant les informations suivantes sous forme de clef-valeurs du chanteur Jacques Dutronc.

- **full_title** : Titre de la chanson.
- **id** : identifiant.
- **release_date** : date à laquelle la chanson a été réalisée.
- **album** : contenant toutes les informations sur l'album.
- **artiste** : contenant quelques informations sur l'artiste, notamment son nom et son identifiant.
- **featured_artists** : s'il a chanté la chanson avec quelqu'un.
- **lyrics** : les paroles de la chanson.

Dans cet échantillon, quelques données, autres que les paroles, manquent aussi à l'appel. Par exemple, les dates marquées d'une valeur `Null` pour signifier qu'il n'y a pas de date. En tout, seulement deux paroles de chansons possèdent une date. Et certaines fois, l'artiste n'est pas nommé. Cet échantillon est récupérable sur avec l'API Genius sous licence © 2022 Genius Media Group Inc.

11. <https://docs.genius.com/>

12. JavaScript Object Notation ; format qui permet de présenter les données de manière structurée et lisible par l'ordinateur et l'humain, alternative à XML.

Ce corpus comporte 10 236 tokens sur toutes les paroles de chanson d'après mon expression régulière et 11 467 d'après celle de spaCy. En comparant ces résultats, je peux déduire que spaCy segmente plus que mon expression régulière. On peut en voir un exemple dans la phrase :

```
"bonjour, je \t m'appelle Tiago\n \n C.N.T.L., R.A.T.P".
```

SpaCy va compter 11 tokens, alors que mon expression régulière va en compter 9. Cela est dû aux caractères spéciaux que sont "\t" et "\n". Présent dans mon corpus pour le retour à la ligne, spaCy va les compter alors que mon expression régulière non. Le fonctionnement de mon expression régulière sépare les mots de 7 manières différentes.

```
r"(\w[\w\.]{1,}|\w+-\w+|\w+\$|\w+|\$|\?|\!)"
```

Le premier cas "\w[\w\.]{1,}" s'il s'agit d'une siglaison, il sépare à la fin de celui-ci.

Le deuxième cas "\w-\w", si deux chaînes de caractère qui se succèdent possèdent un "-" alors il sépare après la deuxième chaîne de caractère.

Le troisième cas "\w\\$", si une chaîne de caractère est suivie d'une ponctuation il sépare après la ponctuation.

Le quatrième cas "\w", il sépare après avoir détecté autre chose qu'une suite de chaîne de caractères.

Le cinquième cas "\\$", il sépare après avoir détecté n'importe quel caractère qui n'est pas un blanc.

le sixième et le septième cas "\?|\!" , il sépare après avoir vu un "?" ou un "!".

Nous pouvons aussi observer la clef `featured_artists` indiquant si l'artiste a chanté seul la chanson ou non. C'est le cas pour une chanson appelée "Les Roses Fanées" by Jacques Dutronc (Ft. Jane Birkin)".

Et enfin nous pouvons observer des chansons dites `remastered`, il s'agit de chansons ayant déjà été chantées par un premier artiste, mais reprises par la suite par un autre artiste - ce qui est courant dans le monde la musique.

En ce qui concerne la structure des paroles des chansons, il est important de préciser qu'il ne s'agit pas d'un texte. Ce n'est ni un article, ni un texte encyclopédique, ni un texte romanesque, mais des paroles de chansons. Ces paroles sont structurées sous la forme de couplets, refrains, mais aussi de retours à la ligne fréquents suivi d'une majuscule - problème qui sera développer dans la partie (paragraphe 3.5).

3.3 Corpus complet

Mon corpus complet possède la même structure que le corpus échantillon, mais il est beaucoup plus conséquent. C'est en effet un corpus qui compte 1 251 artistes. Un artiste possède en moyenne 350.45 mots par chansons, pour un total de 5 953 242 tokens. Ce corpus contient 34 819 chansons dont 16 992 sans les paroles (donc non utilisables pour notre tâche) et 17 827 utilisables.

Au niveau de la structure, certaines paroles de chansons sont structurées avec une annotation explicite de l'intro, du refrain ou encore du couplet directement dans les valeurs de la clé `lyrics`. Ces indications peuvent être du bruit au niveau de la NER car peuvent détecter "Refrain" ou n'importe quelle entité nommée. Il serait dans les verrous (Section 3.5) intéressant de le développer.

3.4 Lexique de lieux de Paris

Enfin, je possède un lexique listant les lieux de Paris, il s'agit de plusieurs fichiers `json` contenant une liste de lieux de Paris comme Avenue Jean Jaurès ou Place de Clichy ...

Ce lexique me permettra de faire une multitude de tâches comme vérifier si les lieux se situent bien sur Paris, ou proche banlieue, ainsi qu'à faire des évaluations.

Ce lexique contient 8 507 lieux Parisiens que j'ai classés automatiquement selon certaines catégories. Par exemple Porte de la Chapelle se classe dans la catégorie "Porte", ou autrement "Saint" tout comme les autres catégories, ici désigne le fait qu'un lieu de Paris qui commence par "Saint" (Saint-Michel) soit classé dans la catégorie Saint. Les résultats de cette classification sont présentés dans le tableau 3.1.

3.5 Verrous gênant la Reconnaissance d'Entités Nommées

SpaCy permet la reconnaissance d'entité nommée (NER) dans un texte. Travaillant sur un corpus de chanson, le problème peut se trouver au niveau de la structure du texte qui ne ressemble pas aux articles de presse sur lesquels les outils sont habituellement entraînés. Lorsque j'ai simplement appliqué les textes musicaux avec spaCy, beaucoup d'erreurs me sont survenues. J'ai quelques hypothèses des causes de ces erreurs, mais certaines semblent

Catégorie	Effectif	Part d'entités nommées
Non classé	3517	0.4134
Rue	3464	0.4071
Place	515	0.0605
Avenue	350	0.0411
Eglise	132	0.0155
Saint*	120	0.0141
Boulevard	117	0.0137
Cite	114	0.0134
Porte	99	0.0116
Pont	37	0.0043
Château	22	0.0025
Arrondissement	20	0.0023

TABLE 3.1 – Répartition par catégorie des Entités Nommées du lexique des lieux de Paris.

plus complexes. Voici quelque hypothèses.

Les néologismes

Chaque artiste possède un langage particulier, venant de tout horizon ou milieu sociale, l'artiste possède sa manière de parler et même de chanter. C'est là que le problème des néologismes peut survenir.

En prenant l'exemple d'une artiste devenue célèbre ces dernières années "Aya Nakamura" ayant su se démarquer et étonné un très bon nombre d'auditeurs avec des paroles qui ont fait "polémique" car incompréhensibles pour certain mais courant pour d'autres. Je pense notamment à certains termes tel que "pookie"¹³, "djadjja"¹⁴. Si l'artiste emploie des termes aussi "nouveau", cela peut induire en erreur et mettre à mal le système de reconnaissance d'entités nommées qui peut potentiellement confondre avec un lieu.

Les retours à la ligne

Les retours à la ligne sont très présents dans mon corpus. Ils indiquent le lieu où la rime doit se produire et donne l'information au chanteur de s'arrêter pour obtenir une bonne sonorité. Comme cela indique la fin d'une sonorité, il faut que le chanteur ait pu exprimer ce qu'il avait à chanter dans

13. Le fait de raconter aux autres ce qui n'est pas censé être raconté.

14. terme pour indiquer qu'un homme raconte n'importe quoi sur telle personne.

cet enchaînement . Ce retour à la ligne est suivi pour chaque parole d'une majuscule. Ces majuscules sont un problème pour la NER. En effet, une entité nommée dans un texte en français se définit le plus souvent par la présence d'une majuscule, comme c'est un retour à la ligne et non un point, la détection peut ainsi se tromper pensant que le mot "Djadja" en début de ligne (si je reste dans le thème de la chanteuse Aya Nakamura) est une entité nommée au lieu d'être juste un mot. Rapidement, un mot que spaCy ne connaît pas peut être analysé comme entité nommée s'il possède une majuscule. Ce qui va donner en sortie beaucoup de Faux Positifs (des mots qui ne sont pas des entités nommées, mais qui sont détectés comme tels). On peut se demander à quel niveau d'imperfection un outil NER peut s'arrêter.

Les désignations pour un lieu, Métonymie

Un autre problème est le fait qu'un lieu puisse être désigné sous plusieurs noms. Pour prendre un exemple familier, la "maison de la recherche" peut posséder aussi le nom de "Serpente" car l'on comprend qu'elle se situe à la rue Serpente, figure de style appelé Métonymie, "il utilise un mot pour signifier une idée distincte, mais qui lui est associée."¹⁵ Cela peut nous faire manquer certaines entités nommées qui en fonction de l'artiste peuvent nommer un lieu connu de sa communauté, de son milieu social, ou autre. Si je prend l'exemple d'un artiste qui dit "rendez-vous au Champs", ici on peut soit comprendre les champs où les agriculteurs travaillent leurs plantations ou les Champs-Elysées si les locuteurs et les allocutaires se sont compris et partagent la même connaissance préalable du lieu.

Détails dans les paroles de chanson

La séparation en mots est propre à spaCy, et ne peut peut-être pas convenir à un corpus des paroles de chanson. Comme dans ce corpus on écrit ce qu'a chanté l'artiste, il est possible de rencontrer des mots ou suite de mots atypiques comme "j'm'inquiète", mot apparaissant dans la chanson "Truc de fou by 113 (Ft. Doudou Masta)". D'autres informations peuvent être indiquées comme dans la liste des paroles de chanson de 113 "[Instrumental]", "À compléter", "Extrait du film Scarface", "Interlude musical", qui ne sont pas les paroles de l'artiste, mais doivent avoir une indication particulière pour celui qui a construit ce corpus. Ainsi, dans le corpus, nous pouvons trouver une multitude de détails qui peuvent avoir un impact sur les données textuelles.

15. <https://fr.wikipedia.org/wiki/M%C3%A9tonymie>

La diversité du lexique pour chaque chanteur peut énormément varier, passant d'un langage soutenu à un langage courant, voire familier. Outre cette difficulté, mais encore il y a certaines chansons possédant sous forme de caractère incorporé dans les paroles, entre crochet, le refrain, le couplet, et l'intro qui indiquent le changement d'artiste, mais surtout les parties qui constituent la chanson ou encore les "ad-libs"¹⁶. Cela peut possiblement influer sur les résultats de NER qui pourraient détecter cela comme un lieu ou un autre type d'entité nommée. La question serait alors si l'on supprime ou non ces informations du texte qui pourraient être un élément d'information importante, notamment si l'on veut avoir des statistiques comme à quel moment les lieux sont le plus cités, refrain ou couplet ?

16. Ad libitum est une expression latine qui sont dans le domaine de la musique, des sons que les artistes prononcent entre des couplets et fin de ligne.

Chapitre 4

Étude de spaCy, amélioration et résultats

4.1 Premières observations sur les résultats de spaCy sur le corpus d'échantillon de Jacques Dutronc

Pour comprendre au mieux le fonctionnement de spacy, les erreurs et les exactitudes faites par les modèles, j'ai commencé par simplement appliquer spacy sur mon premier corpus afin d'observer et d'y faire des statistiques sur les différents modèles proposés. Cela m'a permis rapidement d'identifier les verrous que ce soit au niveau de la compréhension de la structure de mes échantillons, des résultats obtenus ainsi que la manière dont je dois récupérer, analyser et traiter mes résultats.

Lors des observations, j'ai pu constater que spacy détectait beaucoup d'entités nommées et qu'une partie d'entre elles étaient des faux positifs (résultats présenté dans la figure 4.1 qui liste les entités nommées de lieu). Ces mots détectés ont un point commun : ils possèdent une majuscule et la plupart se trouvent en début de ligne. Cela m'a amené à faire des différentes comparaisons avec la structure des mots et de les présenter sous forme de tableau pour mettre en valeur les résultats.

4.1.1 Les différentes proportions calculées sur les modèles de spacy

Ainsi, j'ai décidé de calculer les proportions des entités nommées découvertes par SpaCy et de les comparer aux "vraies entités nommées". En l'absence de vérité de terrain annoté, ce que j'appelle vraies entités nommées, sont les entités nommées présentes dans mon lexique des lieux de Paris et qui ont été détectées par SpaCy dans mon corpus de paroles de chansons.

Je me suis aussi intéressé à la proportion des majuscules afin de déterminer dans quelle mesure les problèmes liés à la "casse"¹ pourrait avoir une influence sur l'efficacité des modèles proposés par SpaCy ainsi que la rencontre de difficultés sur certaines paroles de chansons particulières , ou encore, sur l'évaluation de spacy et le rapport entre les Vrais Positif (donc un lieu détecté par spacy qui est bien un lieu) et les Faux Positif (donc un lieu détecté par spacy mais qui n'en ait pas un). Cela pourra ainsi potentiellement montrer que les majuscules ont bien un impact sur les résultats.

Je vais donc, ci-dessous, présenter des tableaux qui présentent le titre de la chanson, la proportion d'entités nommées détectées par SpaCy, la proportion d'entités nommées qui sont des lieux, la proportion des vraies entités nommées, ainsi que la proportion de majuscules présentes dans les paroles de la chanson.

J'accompagnerai ces tableaux d'observations de résultats pour expliquer ces chiffres.

Modèle Small de spacy : Dans le tableau 4.1, figurent les résultats des proportions calculées sur le modèle small de spacy.

On peut observer que ce modèle détecte beaucoup d'entités nommées à chaque chanson. C'est le cas de la chanson "Mini, Mini, Mini" où spacy a détecté sur le total des paroles de la chanson 5% de lieux et 11% d'entités nommées (organisation, personne, lieu). Dans cette chanson, j'ai relevé à la main 4 entités nommées (Docteur Schweitzer 3 fois, et Ministère 1 fois) sur les 111 tokens qui composent les paroles de la chanson (sans compter les retours chariot les caractères spéciaux \n, \t), c'est-à-dire 3,6% d'entités nommées (organisation, personne et lieu confondu) et 0,9% de lieux à proprement parler (ambiguïté sur Ministère, qui dans la chanson est soit un jeu de mot soit le lieu Ministère). Pour d'autres observations, au niveau des résultats sur les entités nommées qui sont des lieux, la bibliothèque a détecté : 'Minimum', 'Mini-jupe', 'Maxistère', 'Miniature', 'Mini-moke' qui ne sont pas des lieux,

1. le fait de savoir si un mot a la bonne capitalisation lorsque cette information n'est pas disponible.

CHAPITRE 4. ÉTUDE DE SPACY, AMÉLIORATION ET RÉSULTATS22

Titre chanson	Prop lieux	Prop EN	Prop match	Prop Maj
L'Aventurier by Jacq...	11.0672	15.8103	0.3953	26.087
Mini, Mini, Mini by ...	5.5556	11.8056	0.0	23.6111
Mini, mini, mini - r...	5.5556	11.8056	0.0	23.6111
Transat En Solitaire...	5.5556	11.5741	0.0	31.9444
Hippie Hippie Hourra...	4.9451	12.0879	0.0	21.978
Le Testamour by Jacq...	4.8485	18.1818	0.0	37.5758
Les Cactus by Jacque...	4.2735	15.812	0.0	26.0684
La Publicité by Jacq...	4.2146	8.8123	0.0	18.0077
J'aime les filles by...	3.6036	9.009	0.0	19.3694
Fais Pas Ci, Fais Pa...	3.4392	9.5238	0.0	16.6667
Le Courier Du Cœur ...	3.271	7.0093	0.0	19.6262
L'Idole (Je N'En Peu...)	3.1496	5.5118	0.0	17.7165
Le Bras Mécanique by...	3.0702	16.2281	0.0	35.0877
Elle m'a rien dit m'...	2.6012	5.2023	0.0	17.052
J'ai déjà donné by J...	2.4096	5.4217	0.0	9.6386
Le Petit Jardin by J...	2.3438	5.0781	0.0	18.75
Entrez M'Sieur Dans ...	2.1186	10.5932	0.0	15.678
Brèves Rencontres by...	1.9481	11.039	0.0	29.8701
Les Rois De La Réfor...	1.9231	7.2115	0.0	19.2308
Le Roi De La Fête by...	1.9231	6.5934	0.2747	26.9231

TABLE 4.1 – Proportions sur les entités nommées détectées comme lieu, sur toutes les entités nommées (organisation, lieu, personne), sur les entités nommées détectées par spacy et qui sont dans mon lexique de lieux de Paris et sur la proportion des majuscules par mot, sur le modèle "small" de spacy sur les titres de chansons de Jacques Dutronc.

mais je suppose des jeux de mots faits par l'auteur. Ainsi, on remarque que pour beaucoup de paroles de chansons trop de mots sont détectés comme lieu, cela est probablement dû au fait qu'il y ait beaucoup de majuscules (24% pour la chanson "Mini, Mini, Mini"). Enfin, lorsque je compare les entités nommées avec celles de mon lexique de lieux de Paris, le tableau m'indique que seules deux paroles de chansons possèdent des lieux correspondants à la fois à des lieux de mon corpus et à la fois à des lieux détectés par spacy.

Par conséquent, cela montre non seulement que le modèle small de spacy sur-interprète (c'est-à-dire que dès qu'un mot sort de l'ordinaire, il le classe en entité nommée), ou que quelque chose peut l'induire en erreur. Cependant, pour les paroles des chansons, soit l'artiste ne parle que peu, voire pas du tout des lieux de Paris, soit il utilise d'autres noms pour parler de ces lieux ; et donc que mon lexique manque de lieux.

Dans la figure 4.1, je présente les résultats sous forme de liste d'entités nommées qui sont des lieux trouvés par spacy. Quelques commentaires sur ce qu'il a pu trouver.

1. D'abord, il détecte des verbes, des pronoms : 'Viens', 'Mange', 'Touche', 'Tu', ce qui est étonnant sachant que spacy possède un morphologizer² et un lemmatizer qui sont censés remettre à la forme canonique le token.
2. On peut aussi noter des difficultés à sélectionner les tokens puisque certains tokens en fin de ligne sont collés au premier mot de la ligne suivante et spacy les détecte en tant qu'entités nommées.
(exemple :**Demi-jour\nD'un hotel**)
3. Le modèle small nous propose aussi des phrases comme "J'ai fait des games" suivi d'une entité nommée classée comme lieu, au lieu de nous proposer le lieu directement qui vient après la phrase. Dans son contexte entier, la ligne est "L'Aventurier " Jacques Dutronc x "J'ai fait des games à Binningham".
4. Ou encore des mots suivis d'une apostrophe "J'", des interjections comme "Ouilles" et des prépositions "A".

Néanmoins, spacy présente aussi des résultats positifs inattendus : il a pu repérer des lieux de Paris qui ne sont pas dans mon lexique de lieux, comme "Chaussée-d'Antin", "Boulevard Poissonnière", ou encore "Trocadéro". On peut se demander alors si spacy modèle small va détecter une entité nommée juste grâce à ses majuscules ? Ou bien va-t-il aussi utiliser son contexte ?

2. c'est-à-dire qui prédit les caractéristiques morphologiques

CHAPITRE 4. ÉTUDE DE SPACY, AMÉLIORATION ET RÉSULTATS 24

```
[['Mini-moke', 'Mini-jupe', 'Miniature', 'Ministère', 'Minimum', 'Mini-moke', 'Maxistère', 'Maxistère'],
['Canigou', "J'", 'A'], ['Ouille', 'Ouille', 'Ouille', 'A', 'Ouille', 'Ouille', 'Y', 'Ouille',
'Ouille'], ['Viens', 'Mange', 'Touche', 'Mange', 'rue\nSinon', 'Viens', 'Réponds', 'Sois', 'Faut', 'Dis',
'Viens', 'Prends', 'Tu'], [J'', J'', 'Mégève', 'Saint-Tropez', "", "Rochelle\nJ\"", J'', J'"], ['Hippie',
'Vive', 'Hippie', 'Vive', 'Ca', 'la France'], ['Toute', 'Placés',
'Fatigué', 'A'], ["J'ai des tracas j'ai des tourments\nJ'ai pas l'moral", "J'ai pas de chance j'ai pas d'amis\n
\nJ'ai pas de pot j'ai des impôts", "Parc"], ['Celles', 'Aidez', 'Aidez'], ['Dont'], [J'', 'Celles', 'Celles',
"J'ai maigrí je suis", 'Aidez', 'A', 'Aidez', "J'me tape la tête contre les murs\nn"], ['Les', 'Les', 'Les',
'Les'], ['Chaussée-d'Antin\nn', 'jardin\nPassa', 'Chaussée-d'Antin\nn', 'jardin\nPassa', 'Chaussée',
'Chaussée'], [], ['A', 'A', 'Prendre', 'Val-de-Grâce'], [J'aimerais], ['Ca', 'Vieux'], ['Lavabo', 'Merde in
France', 'Eh watching', 'Lavabo', 'Merde in France', 'Lavabo', 'Merde in France'], [J'', 'Varsovie',
'Baltimore', 'Yaoundé', 'Amsterdam', "J'ai fait des games", 'Bornéo', 'Papeete', "J'ai bu de l'eau", 'Bordeaux',
'Tampico', 'Calcutta', 'A', 'Téhéran', 'Saana', 'Kinshassa', 'Dôle', 'Lourdes', J'', J'', 'Créteil', "J'ai eu
la berlue", J'', 'Port-Gentil', 'Tripoli', 'Varsovie', 'Baltimore', 'Maintenant'], ['Amakawogo', 'Amakawogo',
'Gourougourou', 'Bonsoir'], ['la France', 'Instantané', 'Moi', "M'", 'Moi', "M'", 'Venez', 'A', 'Moi', "M",
"J'fonds"], ['Ca', 'Dois', 'Aidez', "J'ai du plaisir", 'Aidez', 'Marcelle', 'Venez'], ['Réserve', 'Voulez',
'Moi', "J'ai Maginot", "J'ai dit comme tes bas : mi-long\nnElle", "J'ai ouï\nn", 'Dire', 'Amiens', 'Yvette'], [],
["J'ai faim de toi", "J'ai confiance", 'Est', 'matin\nEst', 'Croire'], [], ['Mini-moke', 'Mini-jupe',
'Miniature', 'Ministère', 'Minimum', 'Mini-moke', 'Maxistère', 'Maxistère'], ['Presque', 'Toujours'], ['Jamais'],
[Gagnez', 'Hâtez', 'Portez', 'Sortez', 'Courez'], ['Paris'], J'', 'Euh!Euh', 'He!He', J'', 'rue Le Regrantier',
"J'ai des talents\nn"], ['Roméo', 'Pénélope', 'Disigny', 'Goldwyn', 'Métro\nMa', 'Bernhart\nArroy', 'Ménard',
'Venise'], [], [], ['Électronique', 'Lux', 'Pyrex', 'Gilet', 'Braqué', "", 'N'], ['être', 'Avoir', 'Repousse',
'Engagez'], ['Dou Doucement', 'Dou Doucement'], ['Trocadéro\nPris'], ['Placés', 'Fatigué'], ["L'eau de mer",
'Saint Omer', 'Houlgate', 'Jamais', "Qu'Agathe", 'Jamais', "Qu'Agathe", 'Boulevard Poissonnière', 'Jamais',
"Qu'Agathe", 'Jamais', "Qu'Agathe"], ['De', 'Demi-jour\nD'un hotel', 'Demi-tour'], ['Canigou', "J'"]]]
```

FIGURE 4.1 – Liste des entités nommées qui sont des lieux détecté par spacy modèle small.

CHAPITRE 4. ÉTUDE DE SPACY, AMÉLIORATION ET RÉSULTATS 25

Titre chanson	Prop lieux	Prop EN	Prop match	Prop Maj
L'Aventurier by Jacq...	8.3004	19.7628	0.3953	26.087
Mini, Mini, Mini by ...	2.0833	15.9722	0.0	23.6111
Mini, mini, mini - r...	2.0833	15.9722	0.0	23.6111
Le Petit Jardin by J...	1.5625	8.9844	0.0	18.75
Les Roses Fanées by ...	1.4815	13.3333	0.0	20.7407
J'aime les filles by...	1.3514	12.1622	0.0	19.3694
Le Testamour by Jacq...	1.2121	21.8182	0.0	37.5758
Hippie Hippie Hourra...	1.0989	14.8352	0.0	21.978
Les Rois De La Réfor...	0.9615	10.0962	0.0	19.2308
Transat En Solitaire...	0.9259	18.0556	0.0	31.9444
Le Bras Mécanique by...	0.8772	18.4211	0.0	35.0877
L'Idole (Je N'En Peu...)	0.7874	9.4488	0.0	17.7165
Amour Toujours, Tend...	0.7692	11.5385	0.0	16.1538
Brèves Rencontres by...	0.6494	16.2338	0.0	29.8701
Sur Une Nappe De Res...	0.5319	9.0426	0.0	15.9574
Entrez M'Sieur Dans ...	0.4237	8.8983	0.0	15.678
La Publicité by Jacq...	0.3831	9.9617	0.0	18.0077
On Nous Cache Tout, ...	0.3817	8.0153	0.0	16.0305
L'idole by Jacques D...	0.3333	4.3333	0.0	9.6667
Proverbes by Jacques...	0.3257	6.5147	0.0	14.0065
Elle m'a rien dit m'...	0.289	7.5145	0.0	17.052
Le Roi De La Fête by...	0.2747	11.5385	0.2747	26.9231

TABLE 4.2 – Proportions sur les entités nommées détectées comme lieu, sur toutes les entités nommées (organisation, lieu, personne), sur les entités nommées détectées par spacy et qui sont dans mon lexique de lieux de Paris et sur la proportion des majuscules par mot, sur le modèle medium de spacy et sur les titres de chansons de Jacques Dutronc.

CHAPITRE 4. ÉTUDE DE SPACY, AMÉLIORATION ET RÉSULTATS26

Modèle medium de spacy : Pour le modèle médium sur le tableau 4.2, ce qu'on observe en premier, c'est la proportion d'entités nommées - qui sont des lieux détectés par spacy - qui a fortement chutée : on passe ainsi de 5% de lieux détectés à 2% pour la chanson "Mini, Mini, Mini". On observe également qu'une chanson comme "Le Petit jardin" se trouve en cinquième position selon ce modèle, qui était 16ème dans le classement du tableau 4.1 présenté précédemment, avec 2.34% de lieux détectés par spacy, changement dû au fait que le modèle médium a détecté moins d'entités nommées que le modèle small. On peut déduire que cette chanson possède réellement quelques lieux, car toutes les paroles ont baissé au niveau des proportions. On peut aussi démontrer une volonté du modèle à étudier d'autres paramètres ; peut-être le contexte des mots³ pour préciser les résultats. En revanche, on passe de 11% à 15% pour la totalité des entités nommées - cela peut insinuer que spacy médium détecte plus d'entité nommées d'autres types.

Au niveau des résultats donnés par le modèle médium de spacy, figure 4.2 listant les lieux détectés, le modèle médium a détecté beaucoup moins d'entités nommées que le modèle small.

1. Les paroles de la chanson "Le Petit Jardin" possèdent bien des lieux.
2. Les verbes qui représentaient un problème pour le modèle small ont disparu. Cela est sûrement dû au modèle medium de spacy, qui propose une entrée vectorielle beaucoup plus grande 20k⁴
3. Côté séparation des mots, on peut noter une amélioration et une détérioration. La Rochelle a bien été séparée, en revanche "Chaussée-d'Antin" qui était correctement détecté sur le small a été détecté uniquement sous la forme "Chaussée".
4. Pour ce qui concerne les phrases, on en retrouve encore quelques-unes, par exemple : "J'ai écrit", "J'ai fait l'soldat".
5. Puis, on retrouve encore des prépositions comme "A" et des démonstratifs comme "Celles". C'est encore une détection pas cohérente puisque spacy possède un lexique sur *les stop words*⁵ du français.

Modèle large de spacy : Finalement, pour le modèle large, dont les résultats sont dans le tableau 4.3, certaines valeurs pour la proportion des lieux ont légèrement augmenté et d'autres ont diminué. Pour l'augmentation, cela

3. quand on parle de contexte d'un mot, on parle des mots qui se trouvent avant et après celui-ci

4. d'après une réponse formulé sur le site <https://github.com/explosion/spacy/discussions/3517>

5. Mots outils en français

```
[['Mini-moche', 'Mini-moche', 'Maxit re'], [], [], [], ['Castel\n', 'la Rochelle', 'France'], ['San Francisco', 'Ca'], [], [], ['Celles'], ['Ob lisque\nA'], ['Celles', 'A'], [], ['Chauss e', 'Chauss e', 'Chauss e', 'Chauss e'], [], ['A', 'Val-de-Gr ce'], [], ['Ca', 'Refrain']], ['France'], ['Varsovie', 'Baltimore', 'Yaound ', 'Amsterdam', 'Birmingham', 'Born o', 'Papeete', 'Bordeaux', 'Tampico', "J'ai fait l'soldat", 'Calcutta', 'T ran', 'Saana', 'Cotonou', 'D le', 'Lourdes', 'Cr teil', 'Port-Gentil', 'Tripoli', 'Varsovie', 'Baltimore'], [], ['A'], [], ["Bouquin"], ["J'ai cris"], ['Ventre'], [], ['Mini-moche', 'Mini-moche', 'Maxit re'], [], ['Amour'], ['H tez'], ['Paris'], ['Millau', 'Venise'], [], [], ['Essie', 'Nickel s\nCale ons'], [], [], ['Trocad ro\nPris de remords il se mord'], [], ['Houlgate', 'Boulevard Poissonni re'], ['Demi-ton'], []]
```

FIGURE 4.2 – Liste des entit s nomm es qui sont des lieux d tect s par spacy mod le medium.

Titre chanson	Prop lieux	Prop EN	Prop match	Prop Maj
L'Aventurier by Jacq...	7.5099	20.5761	0.3953	26.087
Merde In France (Cac...	3.3079	16.2602	0	19.084
Le Petit Jardin by J...	2.3438	5.3942	0	18.75
Mini, Mini, Mini by ...	2.0833	22.5225	0	23.6111
Mini, mini, mini - r...	2.0833	22.5225	0	23.6111
Br�ves Rencontres by...	1.9481	14.7287	0	29.8701
Le Courrier Du C�ur ...	1.8692	8.867	0	19.6262
Le Testamour by Jacq...	1.8182	20.1258	0	37.5758
Le Roi De La F�te by...	1.6484	9.4801	0.2747	26.9231

TABLE 4.3 – Proportions sur les entit s nomm es d tect es comme lieu, sur toutes les entit s nomm es (organisation, lieu, personne), sur les entit s nomm es d tect es par spacy et qui sont dans mon lexique de lieux de Paris et sur la proportion des majuscules par mot, sur le mod le large de spacy et sur les titres de chansons de Jacques Dutronc.

CHAPITRE 4. ÉTUDE DE SPACY, AMÉLIORATION ET RÉSULTATS28

```
[['Ministère', 'Minimum', 'Maxitaire'], ['Canigou'], ['A'], [], ['Mégève', 'Saint-Tropez', "Camaret\nJ'aime les
filles intellectuelles"], ['San Francisco\nJe suis hippie', 'Ca'], [], ["Parc"], ['Celles'], [], ['Celles',
'Grâce'], [], ['Ca', 'Vieux'], ['Merde', 'France', 'France', 'Merde', 'France', 'France', 'Merde',
'France', 'France', 'Merde', 'France'], ['Varsovie', 'Baltimore', 'Yaoundé', 'Amsterdam', 'Bornéo',
'Papeete', 'Bordeaux', 'Tampico', 'Calcutta', 'Téhéran', 'Saana', 'Cotonou', 'Dôle', 'Lourdes', 'Créteil', 'Port-
Gentil', 'Tripoli', 'Varsovie', 'Baltimore'], ['Akéké', 'Akékékéké'], ['la France'], ['Ca', 'Aidez', 'Aidez',
'Venez'], ["J'ai Maginot", "Bouquin"], ['Amiens'], ['Nagent'], [], [], ['Ministère', 'Minimum', 'Maxitaire'],
['Déformante'], [], ['Hâtez'], ['Paris', 'Prrri!Prrr', 'He!Hee!Ils', 'He!He', 'Oh!Ooh!Quand', "J'habite rue Le
Regrantier"], ['Vénus', 'Millau', 'Venise'], [], [], ['Harpic', 'Antiatomique', 'Braqué'], [], [],
['Trocadéro\nPris'], [], ['Houlgate', 'ile de la Jatte', 'Boulevard Poissonnière'], ['Demi-sourires', 'Demi-
soupir', 'Demi-tour'], ['Canigou']]
```

FIGURE 4.3 – Liste des entités nommées qui sont des lieux détectés par spacy modèle large.

s'explique par le fait que certaines entités ont été oubliées par le modèle médium, ou par le fait que spacy modèle large ait détecté un lieu qui n'en est pas un. Et la diminution s'explique par le fait que trop de lieux qui n'en sont pas dans le modèle medium ont été détectés, ou par le fait que des lieux détectés par le modèle medium sont oubliés par le modèle large. De plus, la proportion des EN est encore en hausse. Effectivement, le modèle large possède plus d'entrées vectorielles (500k), et cela peut être normal. Par exemple, si je prends la chanson "Merde In France" de Jacques Dutronc, "France" revient plusieurs fois et cela a bien été détecté par spacy modèle large, alors que le modèle médium de spacy n'a détecté "France" qu'une seule fois. Cela montre que le modèle médium ne détecte pas forcément l'ensemble des entités nommées présent dans une chanson.

Au niveau de la figure 4.3, de nouveaux et différents résultats du médium vont s'ajouter et d'autres encore vont manquer.

- Concernant les paroles de la chanson "Merde in France", le modèle large détecte bien l'entité nommée "France" qui apparaît plusieurs fois et n'était pas dans le modèle médium, ainsi que la suppression des phrases "J'ai fait l'soldat", "J'ai écrit" qui apparaissaient dans le modèle médium et qui disparaissent dans le modèle large.
- Les prépositions persistent ainsi que les démonstratifs, et nous retrouvons beaucoup d'interjections suivies d'un mot (prononc "il" ou la conjonction "quand"). 'Prrri!Prrr', 'He!Hee!Ils', 'He!He', 'Oh!Ooh!Quand'. Cela montre que quand spacy détecte une forme anormale, c'est-à-dire plusieurs mots suivis avec un "!" entre chaque mot comme dans les résultats, il le classe comme un lieu.
- On retrouve encore des problèmes au niveau de la séparation en mots.

"jardin\nPassa"

CHAPITRE 4. ÉTUDE DE SPACY, AMÉLIORATION ET RÉSULTATS29

4. Chaussée-d'Antins ayant été détecté correctement dans le modèle small, les modèles plus larges n'arrivent pas à garder le nom du lieu en entier.
5. Au niveau des entités nommées, il se fait plus précis notamment dans la chanson "L'aventurier", car il détecte des lieux que les autres modèles n'ont pas détecté.

Dans ces analyses sur le premier échantillon, on remarque que la chanson "L'aventurier" est toujours première, possédant un grand nombre de lieux différents dans la chanson. SpaCy l'a bien remarqué et si l'on prend le modèle supérieur, spacy va être plus précis comme l'ont montré les figures 4.2 et 4.3 illustrant les listes des entités nommées détectés par spacy médium et large. Dû aux entrées vectorielles qui augmentent. Malgré beaucoup de problèmes de détection et de découpage des mots qui ont pu réduire nos résultats ou passer à côté des lieux, spacy semble bien fonctionner. Afin d'observer l'efficacité de spacy sur ce corpus, j'ai décidé de présenter mes résultats deux manières : une avec un graphique et une avec des pourcentages. L'évaluation de spacy se porte sur trois paroles de chanson de mon échantillon de Jacques Dutronc, possédant des lieux ("L'aventurier", "Paris s'éveille..." et "J'aime les filles"). Deux métriques d'évaluation sont présentées dans la figure 4.4, la micro-moyenne et la macro-moyenne.

Au niveau de la micro-moyenne, elle représente pour chaque lieu détecté un calcul de sa métrique : le calcul se fera sur chaque occurrence du lieu, autrement dit un lieu plus fréquent aura plus d'impact sur les résultats qu'un lieu rare, la micro-moyenne traite les tokens. Au niveau de la macro-moyenne, elle représente le calcul général des lieux détectés : le calcul prendrait en compte un lieu, s'il apparaît plusieurs fois, alors sa métrique est calculée une fois ; si on le revoit, alors on ne le compte plus, car déjà vu. Cela évite donc de sur-évaluer le modèle, car un lieu apparaissant trop de fois donnerait trop de bons résultats. Par exemple, dans la chanson "Paris, s'éveille...", Paris apparaît 11 fois, il sera donc compté 1 fois. La macro-moyenne traite les types écartant les doublons.

En sommes, spacy modèle large nous donne - au niveau de la micro-moyenne - une précision de 63.79, un rappel de 77.08% et une f_mesure de 69.81%. Au niveau de la macro-moyenne, nous avons une précision de 63.16%, un rappel 68.57% de et une f_mesure de 65.75%.

Sous forme de proportion, la précision ou accuracy calcule le taux de résultats pertinents parmi les éléments. Il se calcule avec la formule suivante :

$$\text{vrais Positifs(VP)}/\text{vrais Positifs(VP)} + \text{faux Positifs(FP)}.$$

Le rappel calcule le taux de résultats pertinents parmi tous les éléments

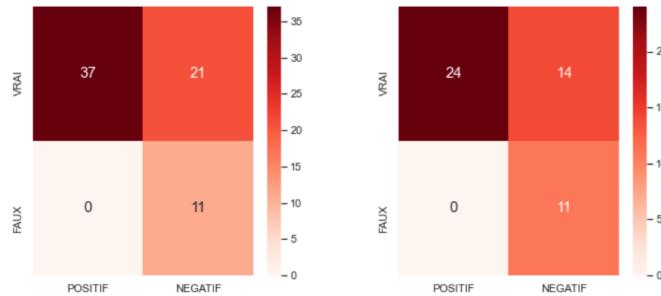


FIGURE 4.4 – Matrice de confusion sur les métriques d'évaluation micro puis macro, sur spacy modèle large, sur 3 chansons des paroles de Jacques Dutronc.

pertinents. Il se calcule avec la formule suivante :

$$\text{vrais Positifs(VP)}/\text{vrais Positifs(VP)} + \text{faux Négatifs(FN)}.$$

La $f_$ mesure est la moyenne harmonique du rappel et de la précision. Elle se calcule avec la formule suivante :

$$2 \times (\text{précision} \times \text{rappel}) / (\text{précision} + \text{rappel}).$$

4.1.2 Diagramme de Venn sur les 3 modèles de spacy

Un diagramme de Venn est une représentation graphique des instances peuplant différents ensembles finie, montrant toutes les relations logiques possibles. Par exemple, dans deux ensembles données A = Paris, Marseille, Dijon, B = Lyon, Marseille, Bordeaux, les instances peuplant les deux ensembles (Union) sont Paris, Marseille, Dijon, Lyon, Bordeaux. Les instances peuplant à la fois l'ensemble A et l'ensemble B (intersection) sont Marseille.

Le diagramme présenté dans la figure 4.5 est un diagramme sur 3 ensembles. Il représente les entités nommées trouvées par les trois modèles de spacy sur l'échantillon des paroles de chanson de Jacques Dutronc. Ce graphique montre pour chaque modèle le nombre d'entités nommées qui sont des lieux trouvés qui sont en commun ou non : le modèle small ayant trouvé 138 "lieux" dont 95 sont trouvés uniquement par ce modèle, 17 trouvés par le modèle small et le modèle large et 26 trouvés par les trois modèles en même temps.

Ce diagramme montre en premier lieu la sur-interprétation du modèle small qui a détecté plus d'entités nommées qu'il n'aurait dû. En second lieu, il montre que chaque modèle trouve des entités que l'autre modèle n'a pas trouvées - les modèles entre eux peuvent ne pas détecter des lieux, ou alors le modèle a bien détecté le lieu, mais a récupéré le lieu avec une syntaxe différente. Par exemple, "la France" pour le modèle medium, "France" pour le modèle small et large ou encore "San Francisco" qui n'a pas été détecté par le modèle small et large. Cela me fait donc dire que pour des résultats complets, il faudrait rendre les 3 modèles complémentaires. Et finalement, les 26 lieux trouvés par les 3 modèles sont ceux pour lesquels les modèles sont d'accord pour dire qu'il s'agit bien d'entités nommées classées comme lieux.

Si l'on regarde au niveau des résultats, on obtient bien des lieux, mais quelques anomalie comme "Ca", "A" et "Celles" et les prépositions que j'avais cité précédemment, ainsi que le verbe "hâter" ("Celles" dans son contexte est bien le démonstratif pluriel féminin, et non le village dans l'Hérault). Voici la liste des lieux en commun :

- 'Port-Gentil', 'Yaoundé', 'Tampico', 'Ca', 'Papeete', 'Saana', 'Chausée', 'Bornéo', 'Celles', 'Venise', 'Boulevard Poissonnière', 'Amsterdam', 'Baltimore', 'Calcutta', 'Téhéran', 'Paris', 'Varsovie', 'A', 'Créteil', 'Hâtez', 'Houlgate', 'Val-de-Grâce', 'Dôle', 'Tripoli', 'Bordeaux', 'Lourdes'

En vue de ces quelques erreurs de détection, et peut-être la non-détection d'autres lieux, une proposition d'amélioration des corpus pourrait faciliter ainsi la détection à spacy, ce que j'explique dans la section suivante.

4.2 Pré-traitement du corpus

Lors de mes premières observations, j'ai pu remarquer que le modèle small de spacy ne fonctionnait pas très bien, il détectait énormément d'entités nommées alors que ce n'est souvent pas le cas. De surcroît, spacy avait quelques petites difficultés à séparer des mots, notamment en fin de ligne ou ne détectait pas le même lieu. C'est donc dans ce sens que pour améliorer la détection dans spacy, j'ai décidé de pré-traiter le corpus afin qu'on obtienne moins de FP⁶, c'est-à-dire un mot détecté comme un lieu, mais qui n'en est pas un.

True Casing Premièrement, nous avons vu que les majuscules ont un possible impact sur les résultats. Si l'on devait ainsi mettre tout le corpus en

6. Faux Positif

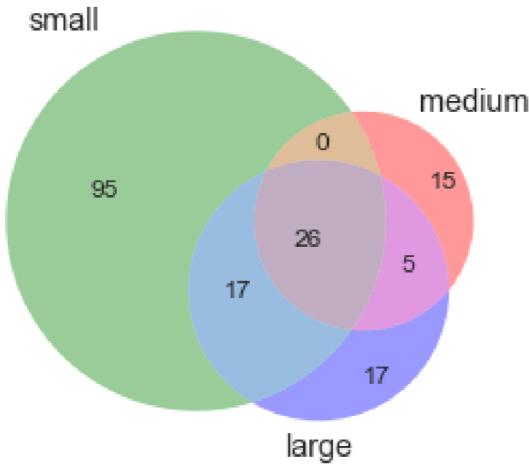


FIGURE 4.5 – Diagramme sur les résultats de spacy sur les entités nommées, sur le corpus d'échantillon de Jacques Dutronc.

minuscule, nous perdrons ainsi des données et des résultats qui peuvent avoir du sens linguistiquement. En effet, à l'écrit, le français, pour des raisons distinctives, va spécifier un lieu, une organisation ou une personne avec les majuscules [Shelar et al., 2020]. Cela est intéressant, car la bibliothèque spacy, si l'on applique un texte sans majuscule, sera beaucoup moins efficace. Possédant beaucoup de majuscules dans ce corpus notamment en début de ligne, une proposition serait alors de modifier le mot en début de ligne s'il n'est pas pertinent pour la détection (True casing⁷).

C'est là qu'intervient GLÀFF. Possédant un lexique de plusieurs millions d'entrées, il pourra me permettre d'identifier un mot courant (comme les prépositions vues dans mes analyses) qui peut potentiellement être détecté comme entité nommée et ainsi renforcer la détection. En admettant que cela fonctionne, cela évitera au modèle une sur-interprétation des résultats (l'erreur sur la NER). C'est ainsi qu'a été construit la méthode d'évaluation suivante figure 4.6 Illustrant la construction du pré-traitement sur 4 mots.

Pour un mot en début de ligne, s'il est dans GLÀFF et qu'il est un lieu dans mon lexique de lieux de Paris, alors c'est un FP (on n'enlève pas la majuscule et on a raison).

7. vraie casse, qui a pour but de remettre la majuscule au bon endroit

Mot en début de ligne	Dans GLÀFF	Dans mon corpus lieu de Paris	Modification
Je	OUI	NON	OUI
Paris	OUI	OUI	NON
Créteil	NON	OUI	NON
Maxistère	NON	NON	NON

FIGURE 4.6 – Tableau illustrant la méthode de construction du Pré-traitement du corpus sur 4 mots servant d'exemples.

S'il est dans GLÀFF mais qu'il n'est pas dans la ressource des lieux de Paris, alors c'est un VP⁸ (on enlève la majuscule et on a raison d'enlever).

S'il n'est pas dans GLÀFF et ni dans la ressource lieux de Paris, c'est un FN⁹ (on n'enlève pas la majuscule et on a eu tort).

Et s'il n'est pas dans GLÀFF mais dans la ressource lieux de Paris alors c'est un VN¹⁰ (on n'enlève pas la majuscule et on a raison).

Ce pré-traitement évaluable me donne les résultats suivants figure 4.7 : Où j'obtiens 1 262 VP résultats correspondant à la modification de la majuscule et le fait qu'il fallait bien la modifier. Cela correspondrait à des mots simples. 307 FN résultats correspondant à la non-modification de la majuscule, alors qu'il fallait la modifier, cela correspondrait aux mots n'existant pas dans le lexique GLÀFF. On obtient ainsi une précision, un rappel et une f_mesure relevable.

- une précision de 100% (1.0)
- un rappel de 80% (0.8043)
- une f_mesure de 89% (0.8915)

Nettoyage des paroles Deuxièmement, on remarque que spacy présente quelques difficultés à séparer certains mots en fin de ligne, puisque le retour à la ligne est désigné par convention du \n. On peut observer quelques formes comme `Trocadéro\nPris., Camaret\nJ'aime les filles intellectuelles` apparaissant dans la figure 4.3 où le retour à la ligne est collé au mot. La tâche serait de modifier dans un nouveau corpus les paroles de chansons en

8. Vraie Positif

9. Faux Négatif

10. Vrai Négatif

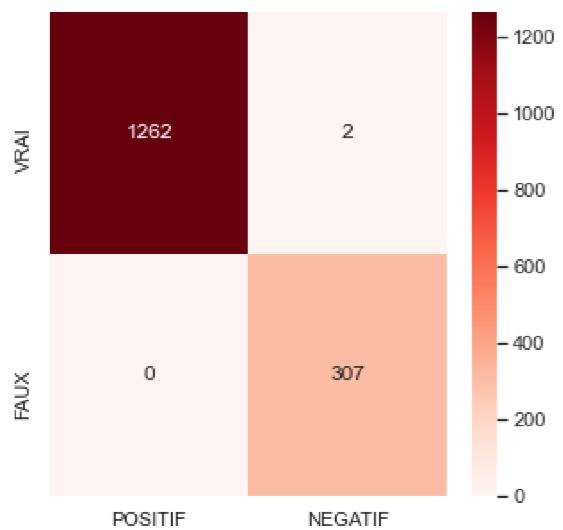


FIGURE 4.7 – Évaluation du pré-traitement étape 1 sur les paroles de chansons de Jacques Dutronc, VP, VN, FP, FN

```

    "petit , petit , petit\n"
    tout est mini dans notre vie\n
    Mini-moke et mini-jupe\n
    Mini-moque et lilliput\n
    il est mini Docteur Schweitzer\n
    mini mini ça manque d'air\n
    Mini-jupe et mini-moque\n
    miniature de quoi j'me moque\n
    ministère et terminus\n
    minimum et minibus\n
    petit , petit , petit\n
    tout est mini dans notre vie\n
  
```



```

    "petit , petit , petit \n"
    tout est mini dans notre vie \n
    Mini-moke et mini-jupe \n
    Mini-moque et lilliput \n
    il est mini Docteur Schweitzer \n
    mini mini ça manque d'air \n
    Mini-jupe et mini-moque \n
    miniature de quoi j'me moque \n
    ministère et terminus \n
    minimum et minibus \n
    petit , petit , petit \n
    tout est mini dans notre vie \n
  
```

FIGURE 4.8 – Modification des paroles de chansons étape 2. Du retour à la ligne à espace puis retour à la ligne sur le corpus d'échantillon des paroles de Jacques Dutronc.

y ajoutant un espace avant le retour à la ligne, là où l'espace n'y est pas, comme illustré en figure 4.8.

Effectuer ceci permettrait d'avoir un corpus prêt à une nouvelle analyse de spacy qui pourrait potentiellement voir une amélioration des résultats.

4.3 Observation des résultats sur le corpus échantillon de Jacques Dutronc pré-traité.

Grâce aux modifications apportées au corpus d'échantillon des paroles de chansons de Jacques Dutronc, on observe une baisse significative des résultats dans le modèle small comme l'illustre le diagramme 4.9, au point qu'il détecte moins d'entités nommées que les autres modèles. Cela montre que le pré-traitement a eu un réel impact sur le traitement des informations, et que spacy détecte beaucoup moins de fausses entités nommées. Ainsi, avec la baisse des résultats, on observe 68 entités nommées trouvées par le modèle small par rapport à 138 précédemment trouvées. Sur les 68, 35 sont à la fois dans les résultats du modèle medium et dans les résultats du modèle large, ce qui peut permettre de dire qu'il s'agit bien de lieu. Il reste néanmoins 41 entités nommées dont 26 détectées uniquement par le modèle medium et 15 uniquement par le modèle large - résultat qu'il faudrait filtrer. J'y reviendrai plus tard dans la section 4.5.

Ainsi, appliquer ce pré-traitement à mon corpus complet des paroles de chansons serait une tâche efficace de *True casing* donnant des résultats plus

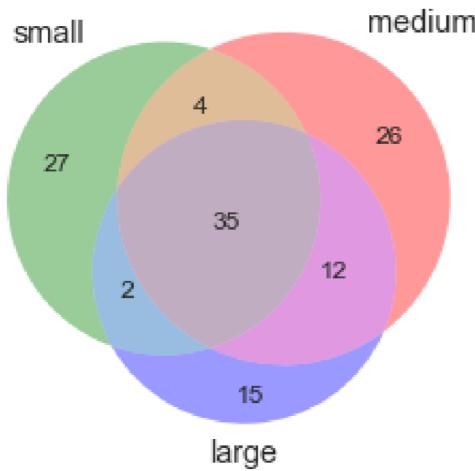


FIGURE 4.9 – Diagramme sur les résultats de spacy sur les entités nommées après le pré-traitement du corpus échantillon de Jacques Dutronc.

performants pour spacy.

Résultats sur mon corpus complet

En somme, après avoir effectué la tâche sur le corpus complet, sans surprise dans la figure 4.10 Illustrant l'évaluation du pré-traitement sur le corpus entiers, nous retrouvons des résultats à peu près similaires, mais sur une échelle plus grande. Ceci donne une évaluation :

- d'une précision de 99% (0.9979)
- d'un rappel de 76% (0.7627)
- d'une f_mesure de 86% (0.8646)

4.4 L'éloignement des classes populaires de Paris.

Le but ici est de vérifier l'hypothèse initialement énoncée dans l'introduction de mon mémoire, selon laquelle les paroles de la chanson populaire française auraient été influencées par le phénomène appelé "gentrification". En fait, l'éloignement progressif des classes populaires vers des banlieues de

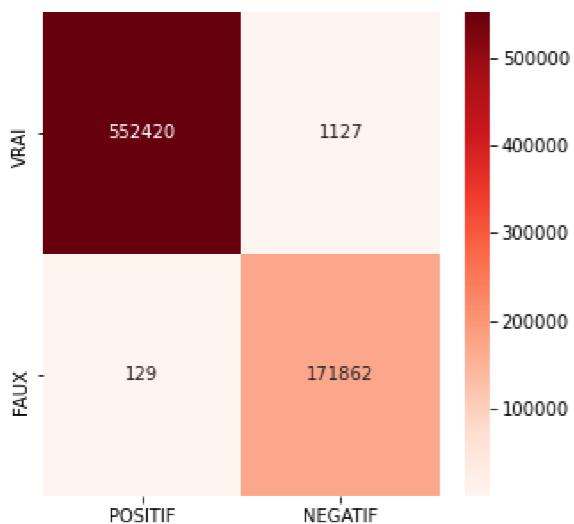


FIGURE 4.10 – Évaluation du pré-traitement étape 1 sur les paroles des chansons du corpus complet, VP, VN, FP, FN

CHAPITRE 4. ÉTUDE DE SPACY, AMÉLIORATION ET RÉSULTATS38

plus en plus distantes des lieux symboliques du centre-ville aurait changé le lexique employé par les chanteurs pop, puisque cette partie de la ville aurait été de moins en moins habitée par des personnes à faibles ou moyens revenus.

Je vais présenter différents graphiques qui peuvent montrer ce phénomène en partant d'un lieu. À quelle fréquence apparaît-il, qui en parle dans ses chansons et à quelle intensité ? Quels sont les lieux les plus cités par un chanteur et pourquoi ? Pour chaque date, quel est le lieu le plus évoqué ? Enfin, je présenterai le tout sur une carte où l'on apercevra les lieux et leurs localisations, avec une couleur qui tend vers le bleu si le lieu a été cité il y a longtemps (au XXème siècle par exemple) et une couleur qui tend vers le rouge si le lieu a été cité récemment.

Histogramme sur les chanteurs et les lieux de Paris

Ce corpus contient des chanteurs venant de différents mondes musicaux (R.N.B, rap, variété française, ...) ou différents milieux sociaux. Cependant, ils ont tous un point commun : leurs chansons sont écoutées par tous, notamment dans le milieu populaire. Les chanteurs francophones font souvent référence à des lieux de Paris pour parler de la beauté du lieu, célébrer celui-ci, raconter une histoire, parler du lieu qui les a vu naître ou tout simplement faire plaisir à leurs fans. Les références et leurs raisons sont multiples. Par exemple, dans les paroles de chanson de la musique "Trouvez-la moi" du chanteur "Dadju" il parle de la Belgique pour citer quelqu'un qui vient de ce pays : "Rudy le locksé de Belgique, to monani".

J'ai donc récupéré les résultats de spacy après avoir fait le pré-traitement et structuré de la manière suivante dans un fichier json : pour chaque nom d'artiste, je place dans une liste le titre de la chanson, sa date de parution et les lieux détectés par spacy. Cela me permet ainsi de chercher plus rapidement les lieux cités et leurs informations. Ainsi, si je prend Paris dans le corpus complet, j'aurais d'après le graphique 4.11 illustrant un histogramme des 5 premiers chanteurs qui parlent le plus de Paris ainsi que le nombre de fois où il sera cité (cité 47 fois par 'A2H', 40 fois par 'Rouda', 36 fois par 'JoeDassin', 31 fois par 'CharlesAznavour', 30 fois par 'LéoFerré'). Au total, il sera cité 1 639 fois dans tout le corpus.

Je peux aussi récupérer pour un artiste le lieu le plus fréquent apparaissant dans ses paroles de chanson. Si je prends le chanteur "Joe Dassin" (1938-1980) 4.12 illustrant un histogramme des fréquences d'apparition d'un lieu, qui d'après mon corpus parlerait ainsi trois fois de 'Pologne', deux fois de 'Clignancourt', deux fois de 'Opéra', deux fois de 'Clichy'. Il est étonnant

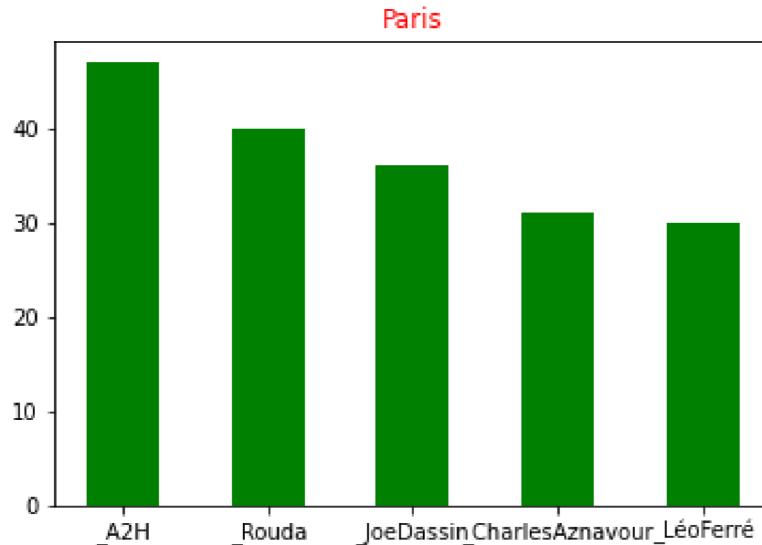


FIGURE 4.11 – Histogramme sur les chanteurs et la fréquence d'apparition du lieu "Paris", dans mon corpus complet.

de ne pas voir "Champs-Élysées" dans l'histogramme - lieu apparaissant plusieurs fois dans les paroles de la chanson "Aux Champs-Elysées" par Joe Dassin (1969), étant une chanson très populaire dans la francophonie et même à l'étranger. Malheureusement, le corpus complet ne possède pas toutes les paroles des chansons de chaque chanteur. Étant construit automatiquement, il est possible que certaines paroles de chansons connues n'apparaissent pas dans celui-ci.

Voici tous les lieux cités par Joe Dassin détecté par spacy et étant dans mon lexique de lieux de Paris : {'Paris' : 36, 'Pologne' : 3, 'Clignancourt' : 2, 'Opéra' : 2, 'Clichy' : 2, 'Tour Eiffel' : 1, 'Amérique' : 1, 'Luxembourg' : 1}

Dans ces résultats, on peut lire "Amérique" et "Pologne" : étant né et vivant actuellement à Paris, je connais beaucoup de recoins de celui-ci, j'ai pu comprendre le quartier Amérique qui est un autre nom pour une partie du 19ème arrondissement de Paris, sauf que ce terme apparaît dans les paroles de la chanson "Moi j'ai dit non" (1975), le chanteur ne parle pas du quartier de Paris mais bien des États-Unis comme la ligne "Un oncle d'Amérique, millionnaire en dollars" l'indique, tâche compréhensible si l'on regarde dans le contexte.

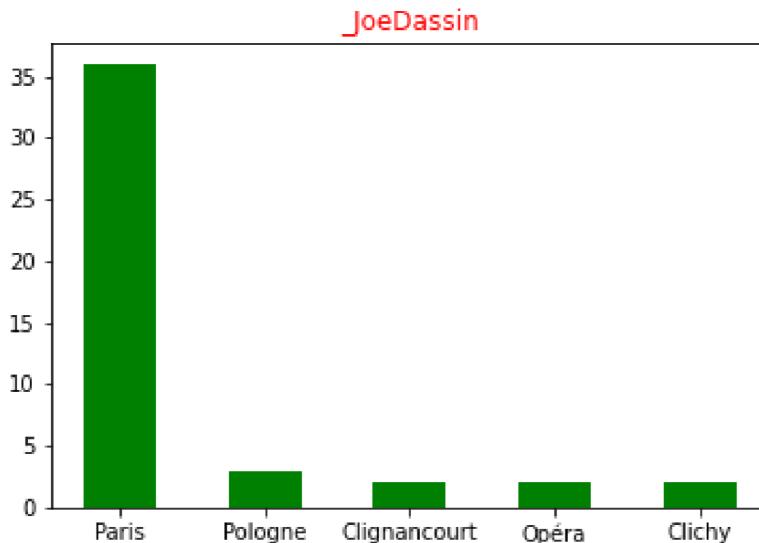


FIGURE 4.12 – Histogramme sur le chanteur Joe Dassin du lexique complet des lieux de Paris et la fréquence d'apparition de ces lieux les plus cités.

Pour le terme "Pologne" apparaissant dans la chanson "Chanson triste" (1965) : Personnellement, je ne connais pas le quartier ou la rue du nom de "Pologne" et dans ces paroles de chanson lui aussi semble désigner le pays.

"Il chante pour lui-même
Les notes qui lui viennent
Comme un vent d'automne
C'est pas toujours gai, la Pologne"

C'est un extrait de la chanson de Joe Dassin "Chanson triste" où figure le token Pologne. On pourrait alors se retrouver avec des lieux dans le filtrage qui sont présents dans le lexique, mais qui ne font pas référence à des lieux de Paris. Un problème résoluble si les résultats sont filtrés par un humain, mais automatiquement elle serait une tâche beaucoup plus complexe.

Une autre tâche qui serait envisageable, à condition que les dates soient présentes dans la chanson, serait de chercher pour une date les lieux les plus évoqués, ainsi que le nombre de fois où ils apparaissent. Cela pourrait ainsi indiquer le quartier, la rue, le boulevard ou le pont à la mode en cette période, s'il y en avait un et en partie valider mon hypothèse. Lors d'un séminaire de l'équipe STIH proposé par mon directeur de mémoire Monsieur Gaël Lejeune, sur le thème "Problématiques d'élaboration d'un corpus de

CHAPITRE 4. ÉTUDE DE SPACY, AMÉLIORATION ET RÉSULTATS 41

chansons situées temporellement"¹¹, il a été évoqué le fait que les paroles de chansons récupérées dans le corpus datant du XXième siècle possédaient beaucoup moins de dates que les chansons du XXIème. C'est le cas, dans les graphiques 4.13 et 4.14 : beaucoup de dates entre 1950 et 2000 manquent, et donc beaucoup de chansons possédant des lieux ne sont pas récupérables à cause des dates manquantes. Je présente donc dans ce qui suit les dates avec lesquelles j'ai obtenu des résultats pertinents.

Dans les graphiques 4.13 et 4.14, sont présentés les 5 lieux ayant la plus grande fréquence d'apparition dans les années 1975, 1976 et 2000. On observe les lieux détectés par spacy modèle large, outre Paris, nous avons en 1975 des lieux de Paris qui se trouvent soit bel et bien dans Paris (intra-muros) soit proche de Paris (départements de la petite couronne).

Villette, apparaît dans les paroles de la chanson "Gueule d'aminche" de Renaud (1975-04-03), est un ancien abattoir de Paris, situé dans le 19ème arrondissement, construit en 1867 puis détruit en 1974. C'est en 1987 qu'il devint le parc de la Villette, que l'on connaît actuellement (il est le plus grand parc de Paris, qui abrite également un musée scientifique). Il raconte dans la chanson l'histoire triste d'une personne qui traîne dans les quartiers du 19ème/18ème arrondissement de Paris et qui tombe amoureux d'une fille. [Wikipédia, 2022k]

Tour Eiffel, apparaît dans les paroles de la chanson "Amoureux de Paname" de Renaud ('1975-04-03'), du surnom "la dame de fer" et construit par Gustave Eiffel, tour de 312 mètres de haut. Elle est située dans le 7ème arrondissement de Paris, proche du bord de la Seine et située entre le Jardin du Trocadéro et les Champs de Mars. Elle fut construite en 1887. L'artiste évoquant Paris.

"Écoutez-moi, vous les ringards
Écologistes du samedi soir"

Il semble répondre aux gens qui se disent "écolo" et qui critiquent Paris en les traitant de "ringards". Il raconte à quel point il aime Paris et son "bitume", à défaut des écologistes qui préfèrent la verdure. Citant la Tour Eiffel qui pour lui est un lieu où l'amour règne.[Wikipédia, 2022o]

Pantin, apparaît dans les paroles de la chanson "Amoureux de Paname" de Renaud ('1975-04-03'), ville de la région Île-de-France et du département

11. http://www.lejeunegael.fr/ressources_seminaire-LC/2022_Seminaire-Lejeune_chansons.pdf

CHAPITRE 4. ÉTUDE DE SPACY, AMÉLIORATION ET RÉSULTATS 42

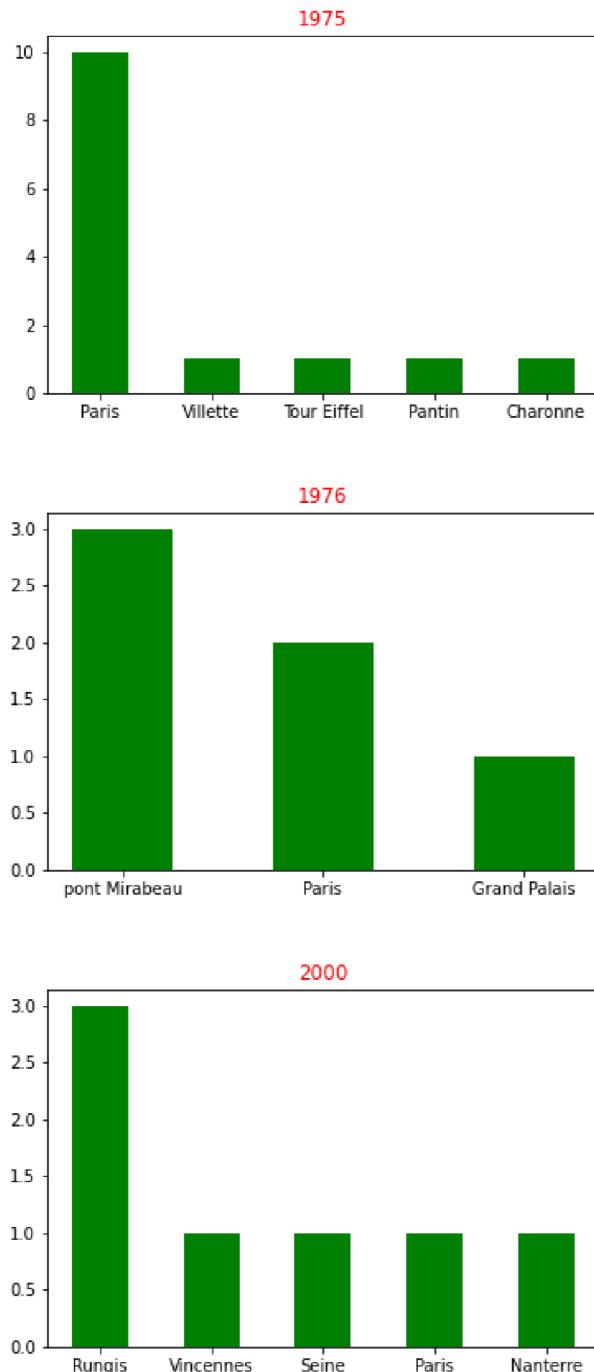


FIGURE 4.13 – 5 lieux de Paris (ou de proche banlieue) les plus fréquemment détectés par spacy modèle large dans les paroles de chansons du corpus complet, avant le XXième siècle (1975, 1976, 2000)

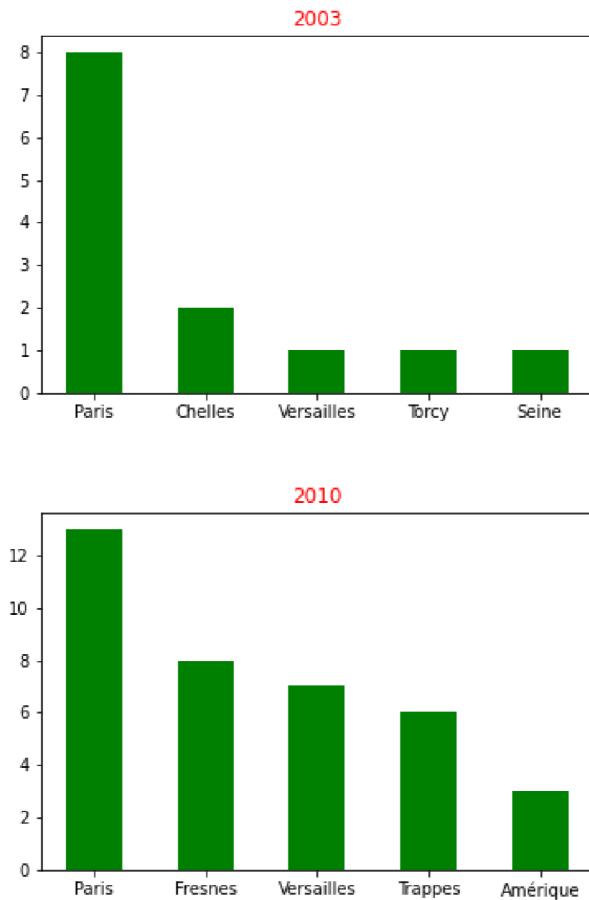


FIGURE 4.14 – 5 lieux de Paris (ou de proche banlieue) les plus fréquemment détectés par spacy modèle large dans les paroles de chansons du corpus complet, après le XXI^e siècle (2003, 2010)

CHAPITRE 4. ÉTUDE DE SPACY, AMÉLIORATION ET RÉSULTATS44

Seine-Saint-Denis. Cette municipalité se situe proche de la Villette, et surtout à la frontière de Paris Nord. [Wikipédia, 2022j]

Charonne, apparaît dans les paroles de la chanson "Hexagone" Renaud (1975-04-03). Charonne est une station de métro se situant sous Paris [Wikipédia, 2022c]. Le contexte est très ambigu, même pour moi. Ne connaissant pas la chanson, on peut supposer que l'artiste a voulu faire une allégorie. On peut se demander s'il parle de la station de métro Charonne ou bien d'une femme dénommée Charonne, puisque dans ces paroles, il fait référence à une affaire de violence policière ayant eu lieu le 8 février 1962, dans la station de métro "Charonne" [Wikipédia, 2022a].

Pour l'année 1976, deux nouveaux lieux détectés toujours présents dans la capitale Paris :

Pont Mirabeau, apparaît 3 fois dans les paroles de la chanson "Les Ricochets" de Georges Brassens (1976-12-01). Construit de 1893 à 1896, situé sur la Seine reliant les 15 et 16ème arrondissement, ce pont intéresse de nombreuses personnes par son architecture. Il sera cité dans de nombreux poèmes, des pièces de théâtre feront une référence au pont et apparaîtra dans quelques films. Georges Brassens raconte son arrivée à Paris à ses 18 ans, et fera référence à d'autres lieux de Paris[Wikipédia, 2022l]/

Grand Palais, apparaît dans les paroles de la chanson "Michèle" de Gérard Lenorman (1976-01-01). Situé dans le 8ème arrondissement de Paris, proche des Champs-Elysées, ce monument sert de lieu d'exposition en tout genre. Sur son côté est accessible le "Palais de la découverte", musée passionnant où j'aimais aller étant jeune. Dans sa chanson, il parle d'une personne (homme/femme) qui habitait près du Grand Palais et dont il semble être amoureux, mais qui finalement n'a jamais pu lui dire. [Wikipédia, 2022f]

Pour l'année 2000, on commence déjà à sortir de Paris avec l'apparition des lieux suivant (toujours sans compter le token Paris lui même) :

Rungis, apparaît dans les paroles de la chanson "Les comiques tripiers" de Éric Toulis (2000-01-01). Rungis est une ville située au Sud de Paris, dans le département du Val-de-Marne, proche d'Orly et de Thiais. Rungis est connu pour son marché international. Il évoque dans sa chanson les bouchers de Rungis travaillant et couverts de sang animal. [Wikipédia, 2022m]

Vincennes, apparaît dans les paroles de la chanson "Mauvais œil" de Lunatic (2000-09-28). Vincennes est une commune située dans l'Est de Paris

CHAPITRE 4. ÉTUDE DE SPACY, AMÉLIORATION ET RÉSULTATS45

dans le département du Val-de-Marne. Vincennes comporte le château et son bois du même nom (Château de Vincennes, Bois de Vincennes). Dans la chanson, Lunatic semble évoquer la difficulté de son quartier : il fait références à Vincennes pour y faire une figure de style.

"Ici c'est la merde, la jungle urbaine
Plein de fauves évadés de Vincennes"

La commune possédant un zoo, il joue sur cette information pour faire une comparaison avec les personnes de son quartier qui se comporteraient comme des "fauves" faisant références aux fauves du Zoo de Vincennes. [Wikipédia, 2022r]

Seine, apparaît dans les paroles de la chanson "Intro (Mauvais œil)" de Lunatic (2000-09-28). Seine dans son contexte est Hauts-de-Seine, il semblerait que spacy ait mal détecté ; il a sûrement compté Hauts-de-Seine comme trois mots au lieu d'un seul. Situé à l'Ouest de Paris en Île-de-France, les Hauts-de-Seine est un département. Dans son contexte, Lunatic chante et exprime la haine des quartiers pauvres des Hauts-de-Seine. Il utilise "Mon frère ça vient des Hauts-de-Seine" comme beaucoup de rappeurs qui font référence au lieu qui les a vu grandir (en général). [Wikipédia, 2022g]

Nanterre, apparaît dans les paroles de la chanson "La lettre" de Lunatic (2000-09-28). Nanterre, commune des Hauts-de-Seine située à l'Ouest de Paris, est connue comme étant une commune difficile d'Île-de-France. Dans cette chanson, il veut dénoncer les injustices d'une peine de prison qu'il a sûrement reçu et il évoque notamment Nanterre pour souhaiter qu'une connaissance s'en aille.[Wikipédia, 2022i]

Pour le XXI^e siècle, j'ai récupéré 2 dates aléatoirement : on remarque la forte présence de Paris à nouveau dans les chansons, preuve d'une volonté des artistes francophones de mentionner cette ville, ainsi que des nouveaux lieux qui cette fois-ci sont plus centrés sur les banlieues de Paris. Je commence par l'année 2003.

Chelles, apparaît dans les paroles de la chanson "Panam All Starz" de Sniper (Ft.0113, Diam's, G-Kill, Haroun, L'Skadrille, Mano Kid Mesa, Salif, Sinik, Tandem & Zoxea) (2003-09-15). Située en Seine-et-Marne 77 en Île-de-France, Chelles se situe à l'Est de Paris. Dans les paroles de la chanson, Chelles est citée pour faire une sorte de dédicace pour représenter et appeler les habitants la ville qui les soutient sans avoir à les citer un par un, "Du Mée à Savigny en passant par Chelles, Noisiel et Torcy". [Wikipédia, 2022d]

CHAPITRE 4. ÉTUDE DE SPACY, AMÉLIORATION ET RÉSULTATS46

Versailles, apparaît dans les paroles de la chanson "Bonne vibration" de Daddy Mory (Ft. Fefé & Sir Samuel) ('2003-06-03'). Commune d'Île-de-France dans le département des Yvelines, Versailles est connue pour son célèbre château (internationalement connu) - Château de Versailles - d'où le nom de la commune. Dans sa chanson, Daddy Mory parle de Versailles pour exprimer le fait que le type de sa chanson (reggae) est écouté par ce qu'il appelle "les Massives" et étant écouté de Montpellier jusqu'à Versailles, sûrement pour jouer avec la distance Montpellier-Versailles et peut-être la classe sociale. [Wikipédia, 2022q]

Torcy, apparaît dans les paroles de la chanson "Panam All Starz" de Sniper (Ft.0113, Diam's, G-Kill, Haroun, L'Skadrille, Mano Kid Mesa, Salif, Sinik, Tandem & Zoxea) (2003-09-15). Situé en Seine-et-Marne au Nord-Est de Paris, c'est une commune avec beaucoup de lacs où l'on peut s'y baigner. Torcy est cité pour les mêmes raisons que pour Chelles. [Wikipédia, 2022n]

Seine, apparaît dans les paroles de la chanson "Panam All Starz" by Sniper (Ft.0113, Diam's, G-Kill, Haroun, L'Skadrille, Mano Kid Mesa, Salif, Sinik, Tandem & Zoxea) ('2003-09-15'). Encore une fois le mot "Seine" a été détecté au lieu des Hauts-de-Seine. Il a été cité pour les mêmes raisons que Chelles et Torcy.

Pour ce qui est de 2010, j'ai choisi de m'intéresser à une autre série de lieux.

Fresnes, apparaît dans les paroles de la chanson "Dans nos quartiers" de La Fouine (Ft. Alonzo & Teddy Corona) ('2010-07-12') mais aussi dans "Veni, vidi, vici" de La Fouine ('2010-12-10'), Fresnes est située dans le Val-de-Marne, au Sud de Paris. La ville est connue pour son centre pénitentiaire étant le plus important de France. Fresnes est citée uniquement dans le refrain, pour marquer le fait qu'elle soit connue du Nord au Sud d'Île-de-France. Il essaie de dénoncer dans ses paroles les fortes interpellations policières de sa communauté, mais aussi pour montrer comment il se débrouille pour s'en sortir.

Dans les paroles de la chanson "Veni, vidi, vici", il cite Fresnes comme un lieu dont il doit se rendre s'il n'a pas le choix.

"Je traîne de hall en hall quitte à finir à Fresnes"

Il insinue ainsi que les jeunes des quartiers sensibles traînent souvent dans les halls des bâtiments, notamment à Fresnes, où il semblerait (à cette date) que cela est courant. [Wikipédia, 2022e]

CHAPITRE 4. ÉTUDE DE SPACY, AMÉLIORATION ET RÉSULTATS47

Versailles, apparaît dans les paroles de la chanson "Procureur de Versailles de La Fouine (Ft.A2P, Chabodo & Vincenzo)" ('2010-01-18') mais aussi "Les Balances Sont Respectées" de La Fouine', ('2010-01-18'). Dans les paroles de la chanson, Versailles est citée comme complément du nom pour désigner le procureur exerçant dans la commune de Versailles et le comparer à plusieurs personnes - une sorte d'allégorie, "Bande de Procureur de Versailles, (Yeah de Versailles)". Dans celle de "Les Balances sont respectées"

Trappes, apparaît dans les paroles de la chanson "Veni, vidi, vici" de La Fouine, ('2010-12-10') mais aussi "Dans nos quartiers" de La Fouine (Ft. Alonzo & Teddy Corona)', ('2010-07-12') ainsi que dans "Procureur de Versailles" de La Fouine (Ft.A2P, Chabodo & Vincenzo)', ('2010-01-18'). Trappes est une commune située dans les Yvelines en Île-de-France, Sud-Ouest de Paris. Il désigne Trappes comme un lieu dangereux et sale comme le montre les lignes sorties des 3 paroles de chanson.

"J'ai grandi à Trappes la merde sous mes semelles"(Dans nos quartiers),

"A Trappes quand le chat n'est pas là, les souris se tirent dessus!"(Veni, vidi, vici),

"Cousin avec une équipe man du cent pour-cent Trappes et Du Vice," (Procureur de Versailles),

le terme "vice" accentue le péché, la malhonnêteté, la dangerosité lié par les mots "se tirent dessus" expliquant le port d'arme qui n'est pas légal en France, et "la merde sous mes semelles" qui cherche à exprimer que Trappes n'est pas une commune pour grandir sûrement pour faire référence à l'état insalubre des bâtiments, ou de l'état des rues. [Wikipédia, 2022p]

Amérique, apparaît dans les paroles de la chanson "On a tous une Lula" de Damien Saez', ('2010-03-29') ainsi que dans "Pilule" de Damien Saez', ('2010-03-29'). Il cite Amérique pour les États-Unis et non pour le quartier français "Amérique". Ici, nous sommes à nouveau dans un cas où l'ambiguïté de la correspondance entre les lieux se produit et donc peut influer sur les résultats.

Visualisation des Lieux

Pour bien illustrer l'évolution des lieux dans la chanson francophone, l'idéal serait de le représenter sur une carte de l'Île-de-France, comprenant les lieux du corpus. Cela pourrait ainsi montrer, grâce à des points situés

CHAPITRE 4. ÉTUDE DE SPACY, AMÉLIORATION ET RÉSULTATS48

sur la carte et des formes correspondant à des dates, l'hypothèse évoquée au début du mémoire.

C'est ainsi qu'à l'aide de Google my maps¹² et la création d'un fichier .csv qui récupère les lieux, leurs latitudes, leurs longitudes, les dates où ils ont été cité, et les titres dans lesquels ils apparaissent, que j'ai pu dans la figure 4.15 qui représente Paris et les lieux entre 1950 et 2020, mettre automatiquement et manuellement une partie des lieux détecté par spacy modèle large et présent dans mon lexique lieu de Paris.

C'est donc Présenté avec un système de forme pour les différencier que j'ai classé mes lieux comme suit.

- Pour les formes avec un petit carré, elles correspondent aux lieux étant cités entre les années 1940 à 2000.
- Pour les formes possédant un carré plus grand, Elles correspondent aux lieux étant cités entre les années 2001 à 2010.
- Pour les formes possédant un losange. Elles correspondent aux lieux étant cités entre les années 2011 à 2019.
- Pour les formes possédant une étoile. Elles correspondent aux lieux étant cités entre les années 2020 à plus.
- Pour les formes avec une croix et un fond gris. Elles représentent les lieux cité dans les paroles de chanson dont je n'ai pas pu retrouver la date.

En sommes, dans ce premier plan, on remarque pour les années allant de 1940 à 2000, les lieux précédemment cité "Charonne", "Pont Mirabeau", "Villette", "Pantin", mais aussi quelques lieux détecté par spacy comme "Arsenal" dans la chanson "Bercy Madeleine" de Pierre Perret qui en réalité dit dans ses chansons "Elle avait tout un Arsenal". On retrouve aussi "Pont Marie" du titre "Au café de la paix" de Thomas Fersen et "Pont Neuf" du titre "Pour Me Rendre À Mon Bureau" de Georges Brassens. Dans les autres fourchettes de date, "clichy" pour la chanson "Place de Clichy" de Julien Clerc ainsi que "rue Rivoli" pour la chanson "Décollement pulmonaire" de Octobre Rouge, ou encore un lieu pas forcément visible sur la carte, "Pont d'Arcole" du titre "Le bateau mouche" d'Alain Souchon pour les années allant de 2001 à 2010. Et pour des dates plus récentes, "Rue de la Paix" dans "Cantare" par Soprano (Ft. Soolking), "Auteuil" par Kalash pour le titre "Insta Twitter".

Dans ses résultats, ce qui s'observe en premier, c'est la forte présence de lieu Parisien pour les années datant d'avant 2000. Suivi de quelques lieux

12. <https://www.google.com/maps/d/u/0/viewer?mid=10e0Pw0tbMBe9h-x0FDWbB9eZZAip2hs&ll=48.853069698940544%2C2.344381392980326&z=12>

CHAPITRE 4. ÉTUDE DE SPACY, AMÉLIORATION ET RÉSULTATS 49

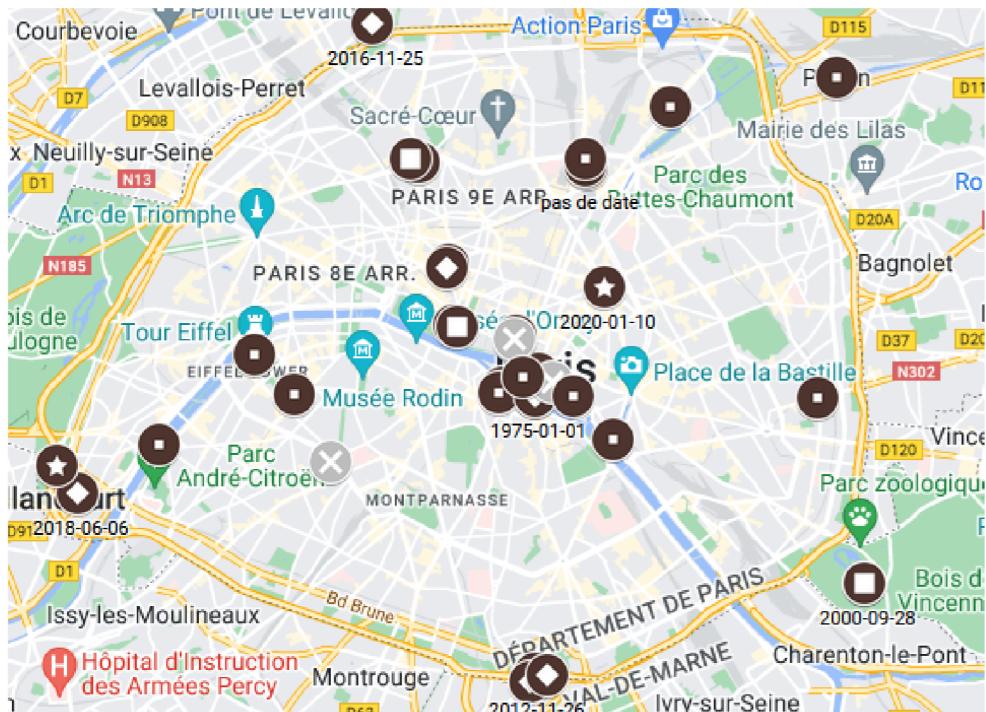


FIGURE 4.15 – Illustration d'une carte de Paris localisant des lieux de Paris détecté par spacy modèle large sur le corpus complet Entre les années 1940 et 2020. Le petit carré → 1940 à 2000, le grand carré → 2001 2010, le losange → 2011 à 2019, l'étoile → 2020 à plus.

plus récent (2011-2019), soit situé en plein Paris, soit au niveau des Portes de Paris.

C'est en prenant un plan à échelle plus grande, comme sur ce second plan figure 4.16, qui illustre Paris et sa banlieue, où l'on observe des occurrences de lieu d'Évry pour le Sud jusqu'à Pontoise pour le Nord-Ouest. Ce sont notamment des lieux cité pour des dates entre 2001 et 2019. Mais l'on retrouve quelques lieux cités avant 2000 vers Versailles et Nanterre.

C'est à échelle plus grande qu'on peut observer plus d'occurrences des lieux datant du XXI^e siècle. Cela nous montre donc un léger décalage vers les banlieues Parisiennes entre les années 2001 à 2010 et 2011 à 2019. C'est notamment pour les années de 2001 à 2010 que les occurrences des lieux se font plus présente. Cela peut se justifier par une volonté de parler de ses banlieues par les chanteurs. Banlieue sûrement oublié qui représente une grosse partie des ménages Parisien migrant ainsi vers sa périphérie pour des

loyers moins chers ou des espaces plus grands.¹³.

4.5 Enrichissement du lexique des lieux de Paris

Afin d'améliorer mon lexique de lieux de Paris, et surtout pour désambiguïser¹⁴ un lieu qui ne possède pas le même nom que dans mon lexique, j'ai décidé de créer une sorte de filtre qui demande à un humain si les lieux trouvés par spacy sont bien des lieux de Paris qui ne sont pas dans mon lexique, afin de les ajouter. Alors que dans le cas où ces lieux existent sous un autre nom, l'humain peut - s'il le souhaite - les mentionner.

Le lexique d'un locuteur étant propre à lui-même, il peut utiliser des noms de lieux connus par sa communauté ou ses proches a minima. Une machine ne peut le savoir sauf si on lui a expliqué au préalable le nom du lieu. En revanche, un humain peut se douter de l'ambiguïté potentielle d'un nom de lieu. Par exemple, un ami ne connaissant pas BNF¹⁵ comme expliqué dans la partie "verrous" et la sous-section "les désignations pour un lieu" 3.5, si je lui dis dans le contexte que je vais aller étudier à la BNF, il va comprendre un institut, une bibliothèque, ou un lieu pour étudier et ensuite s'il le souhaite me poser la question où se situe le lieu plus précisément. Un autre exemple s'appuyant sur mon corpus, dans les paroles de la chanson "Possédé" du chanteur Vald (2018-02-02), le chanteur cite "QG" dans la ligne suivante "Dans la rue ou au QG, le sourire est lié au budget". Il fait référence à un lieu uniquement connu par lui et ceux qui le côtoient. Souvent utilisé par les jeunes qui aiment rester dehors, pour parler du lieu où l'on se donne rendez-vous.

Cette tâche étant chronophage, et n'ayant pas trouvé de moyen de "désambiguïser" un lieu, ceci est un supplément à mon mémoire qui pourrait aider à la tâche de désambiguïsation d'un lieu.

13. <https://www.insee.fr/fr/statistiques/2011101?geo=DEP-75#consulter-sommaire>, https://50ans.apur.org/data/b4s3_home/fiche/100/01_evolution_population_Paris_dynamique_perspectives_a6718.pdf

14. j'utilise ce terme pour décrire le fait qu'un lieu peu posséder plusieurs noms

15. Bibliothèque National de France

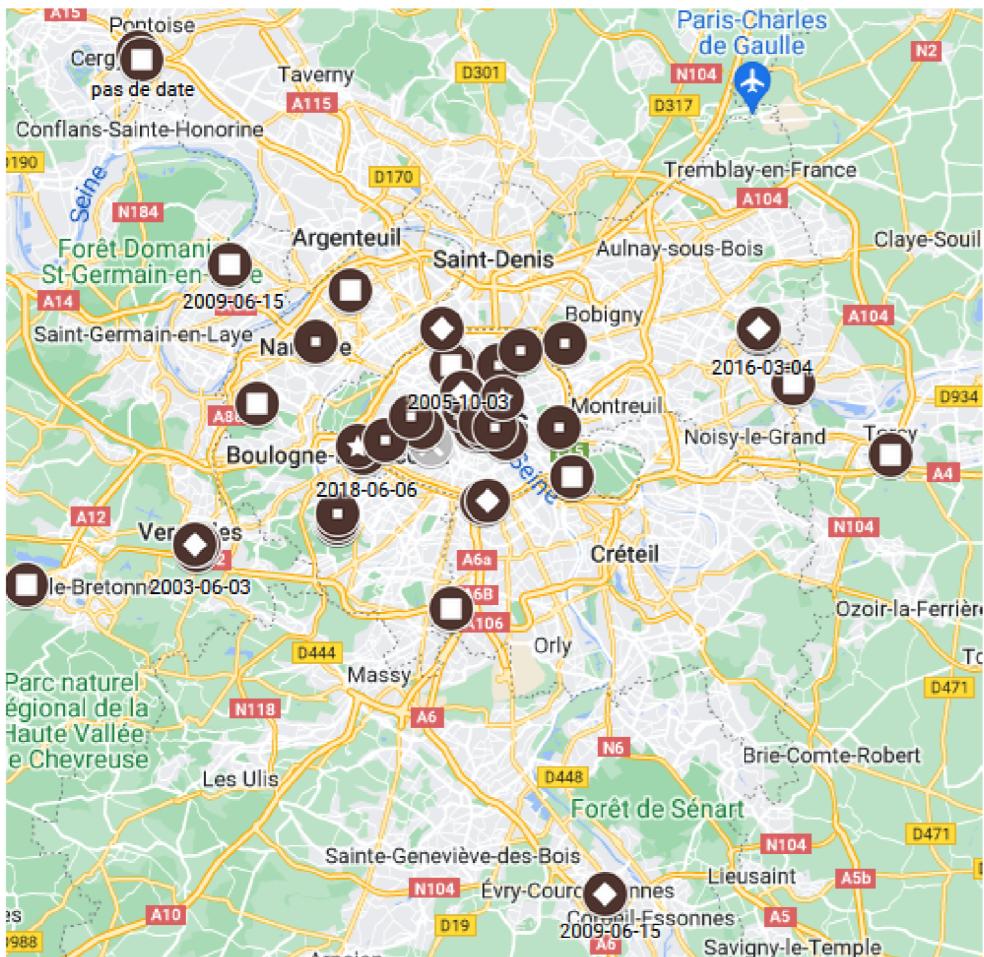


FIGURE 4.16 – Illustration d'une carte de Paris et de sa banlieue localisant des lieux détectés par spacy modèle large sur le corpus complet Entre les années 1940 et 2020. Le petit carré → 1940 à 2000, le grand carré → 2001 à 2010, le losange → 2011 à 2019, l'étoile → 2020 à plus.

Chapitre 5

Conclusion

En somme, le travail fourni au cours de ce semestre pour ce mémoire m'a permis de comprendre l'une des tâches dans le domaine du TAL et les problématiques qui entourent la NER : des paramètres non seulement textuels, mais aussi au niveau des données pouvant être difficiles à récupérer, de par leur limitation. En effet, aujourd'hui, les droits d'auteur peuvent avoir un impact sur la récupération de données du fait que des personnes potentiellement malveillantes peuvent s'en servir pour s'enrichir, ou uniquement partager mais en dépit du créateur, qui peut souhaiter rester dans l'anonymat, puisse se sentir volé ou une quelconque raison. Impactant ainsi le domaine de la recherche, les corpus recueillant les paroles de chanteur se limitent de plus en plus, voire se payent : ce qui est normal pour le code de la propriété intellectuelle, mais à des fins de recherche cela peut être facteur d'abandon. Heureusement, une réglementation¹ voit le jour sur la fouille des textes à des fins de recherche qui dans le droit français autorise les personnes à consulter et étudier les données telles que les images fixes ou animées, sons, musiques, logiciels si cela est à but scientifique.

Ayant tout de même pu récupérer un corpus avec des chanteurs du XXième et XXIème siècle, ces derniers proposent des variétés du français tant bien au niveau diachronique, qu'au niveau synchronique, observant ainsi beaucoup de paroles de chansons. J'ai pu non seulement constater des phénomènes vus en cours, comme écouter des accents régionaux de différents chanteurs ou la façon de chanter de certains, mais aussi des problèmes sociaux abordés dans les chansons du XXième siècle qui sont encore d'actualité. Mon travail étant concentré sur un but précis, l'étude de ce qui l'entoure est bénéfique au plan personnel, point que je souhaitais relever.

Pour revenir à la tâche principale qui était l'éloignement des classes po-

1. <https://www.ouvrirlascience.fr/la-fouille-de-textes-et-de-donnees-a-des-fins-de-recherche>

pulaires illustré par des lieux dans les paroles de la chanson francophone, j'ai utilisé la bibliothèque spacy. Souvent nommé dans les articles traitant de tâche, spacy est populaire et offre de bons résultats dans le domaine du TAL. Un travail de recherche a été précédemment fait sur cette bibliothèque [Koudoro-Parfait et al., 2021], il était donc normal de continuer ainsi dessus et observer les résultats proposés par celui-ci qui sur un texte de chanson obtient un score plutôt bon pour la détection des entités nommées.

Afin d'améliorer mes résultats de recherche sur la NER qui dans ce corpus contenait plus de 1 000 artistes, et ainsi éviter un maximum de Faux Positif, (des entités qui seraient classées comme lieux alors qu'en réalité ce ne sont pas des lieux), j'ai tout d'abord essayé de comprendre quel était le facteur qui induisait en erreur spacy. J'ai donc pu illustrer à l'aide de tableaux que les majuscules avaient un impact sur la détection d'entité nommée et faussaient énormément les résultats. Ceci m'a directement amené à une tâche de "True casing" qui était de remettre à la bonne casse. C'est-à-dire qu'il s'agit de dire si le mot est bien en majuscule ou non, pour les éléments en début de ligne car la structure des paroles de chansons force le mot en début de ligne à être en majuscule. Grâce à la ressource du GLAFF, qui était disponible en ligne, cette tâche, en définitive, s'est révélée pertinente et efficace dans mes résultats. Par exemple, les résultats de détection de spacy sur mon échantillon des paroles de Jacques Dutronc se sont considérablement précisés. Appliqué à l'échelle du corpus, cela m'a permis de conserver le sens de chaque token qui peuvent être de précieuses informations.

Par la suite, grâce aux entités nommées détectées par spacy dans mon corpus, j'ai pu créer un dictionnaire d'interrogation qui regroupait pour chaque artiste, le titre de la chanson, la date et une liste des lieux détectés en me fournissant les données nécessaires pour la validation de mon hypothèse. Malgré l'absence de paroles chez certains, ce qui a été un réel frein était les dates manquantes pour en grande majorité les chanteurs du XXI^e siècle, élément nécessaire pour prouver qu'un chanteur parle de Paris à cette date. Il faut aussi noter qu'un artiste, dans ses paroles de chanson, cherche à passer un message, et cela influe indirectement sur mon travail. J'ai pu constater que certains artistes jouent des mots et des figures de style soit pour le bon fonctionnement de la chanson soit par pur plaisir de jouer sur les mots ou encore pour faire passer un message. Il peut donc utiliser des noms de lieux : comme pour "Hexagone" de Renaud (1975-04-03) qui souhaitait dénoncer les violences policières ayant eu lieu à cette époque et parler de "Charonne" qui est une station de métro. Sans contexte, sans connaissance de la chanson et sans adjectif épithète antéposé, on peut facilement se méprendre comme ce fut mon cas.

En définitive, c'est avec l'obtention, la transformation et l'observation des

données de mon corpus que j'ai pu construire, à l'aide de Google, une carte de Paris et de sa banlieue 4.16, me permettant d'illustrer un léger phénomène d'éloignement des lieux au cours des années. L'on retrouve sur cette carte un certain nombre de lieux de Paris intra-muros, évoqués au XXième siècle, en revanche en dehors de Paris ce qui saute aux yeux sont les occurrences des lieux du XXIème siècle. Malheureusement par manque de méta-données de date dans le corpus, mais aussi du fait d'un manque d'hétérogénéité des lieux présents sur la carte, il est difficile de bien illustrer ce phénomène et voir si l'hypothèse est concluante. Puis, nous pouvons retrouver des occurrences de lieux situés en banlieue pendant le XXième siècle qui sont à nouveaux cités plus tard dans le XXIème siècle, comme "Nanterre" dans "La vie intime est maritime" d'Alain Souchon (1985), qui a été cité dans "Cap Carnaval" du chanteur Naps parut en (2019-06-28), date assez récente.

Si je devais expliquer la raison du phénomène des citations des lieux qui tend à être au dehors de Paris, serait que les chanteurs souhaitent ainsi dépeindre Paris qui n'est pas uniquement des infrastructures et des monuments, mais c'est avant tout ces gens qui l'habitent et qui s'en éloignent de plus en plus pour des espaces et des prix bas que n'offre plus Paris, j'ai pu relever dans quelques paroles de chanson pris aléatoirement, mais aussi pendant l'étude, des chanteurs qui dénoncent des faits comme l'état de leur ville/quartier, comme d'autres qui festoient pour partager avec ceux qui les ont vues grandir.

Propositions d'amélioration et perspectives

Pour finir, ce mémoire m'a beaucoup inspiré et souvent, de nouvelles idées que je n'ai pas pu développer ont traversé mon esprit, notamment au niveau statistique. Je souhaitais savoir au niveau du contexte la fréquence des mots qui entourait les noms des lieux, si l'on rencontrait en majorité les mots "rue", "boulevard", ..., on pourrait ainsi si l'on avait plus de dates regarder l'évolution des mots qui figurent dans le contexte du token "Paris". Ou encore à quel moment les lieux sont-il le plus cités, couplet ou refrain. J'avais une petite idée avec le logiciel "TXM" [Heiden et al., 2010], qui propose une fonctionnalité sur un texte, où l'on observe pour l'occurrence d'un mot à quel moment il apparaît dans un texte. Cela me permettrait d'utiliser d'autres "outils" informatiques que le logiciel python sur lequel j'ai travaillé tout du long.

Je souhaiterais aussi mieux illustrer ce phénomène d'éloignement des lieux vers la banlieue, par un gain d'occurrences des lieux, mais encore en complétant les méta-données. Difficilement complétable via des ressources en ligne tel que "Wikidata", ces méta-données verraien une implémentation auto-

matique des dates pour un titre qui n'en possède pas. Ce serait une avancée intéressante au niveau personnel et au niveau des résultats.

Une autre idée serait d'utiliser sur ce même corpus, une autre bibliothèque ou logiciel de détection d'entité nommée et ainsi évaluer et comparer l'efficacité de spacy.

Je voulais encore développer et trouver une solution sur le problème de désambiguïsation d'un lieu et sur le plan automatique qu'offre les ordinateurs. En effet, expliqué dans la partie 4.5, j'exprimais que parfois un artiste utilisait un vocabulaire propre pour désigner un lieu - point que je ne trouve pas assez abordé dans mon mémoire. J'avais proposé dans celui-ci une sorte de questionnaire à un humain qui enregistre le lieu détecté par spacy et qui n'est pas dans mon lexique des lieux de Paris. Cela permettait pour chaque consultant d'enrichir le lexique, en étant quasi certain de ne pas se tromper, mais cela ne relevait pas d'un automatisme.

Bibliographie

- [ATILF, 2015] ATILF (2015). Morphalou. ORTOLANG (Open Resources and TOols for LANGuage) –www.ortolang.fr.
- [Chevalier, 1958] Chevalier, L. (1958). *Classes laborieuses et classes dangereuses*. Plon, Paris.
- [Clerval, 2013] Clerval, A. (2013). *Paris sans le peuple. La gentrification de la capitale*. La Découverte, coll. Hors collection Sciences Humaines, Paris.
- [Heiden et al., 2010] Heiden, S., Magué, J.-P., and Pincemin, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. In Bolasco, S., Chiari, I., and Giuliano, L., editors, *10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, volume 2, pages 1021–1032, Rome, Italy. Edizioni Universitarie di Lettere Economia Diritto.
- [Honnibal and Montani, 2017] Honnibal, M. and Montani, I. (2017). Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *Unpublished software application*. <https://spacy.io>.
- [Koudoro-Parfait et al., 2021] Koudoro-Parfait, C., Lejeune, G., and Roe, G. (2021). Spatial named entity recognition in literary texts : What is the influence of ocr noise ? In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities*, pages 13–21.
- [Leclanche, 1998] Leclanche, M. (1998). *Le refrain dans la chanson française de Bruant à Renaud*. Pulim.
- [Nouvel et al., 2010] Nouvel, D., Antoine, J.-Y., Friburger, N., and Maurel, D. (2010). An analysis of the performances of the casen named entities recognition system in the ester2 evaluation campaign.
- [Sajous et al., 2013] Sajous, F., Hathout, N., and Calderone, B. (2013). GLAFF, un Gros Lexique À tout Faire du Français. In *Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2013)*, pages 285–298, Les Sables d'Olonne, France.

- [Sassen, 1996] Sassen, S. (1996). *La ville globale*. New York, Londres, Tokyo. Descartes & Cie, Paris.
- [Schmitt et al., 2019] Schmitt, X., Kubler, S., Robert, J., Papadakis, M., and LeTraon, Y. (2019). A replicable comparison study of ner software : Stanfordnlp, nltk, opennlp, spacy, gate. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 338–343.
- [Shelar et al., 2020] Shelar, H., Kaur, G., Heda, N., and Agrawal, P. (2020). Named entity recognition approaches and their comparison for custom ner model. *Science & Technology Libraries*, 39(3) :324–337.
- [Wikipédia, 2022a] Wikipédia (2022a). Affaire de la station de métro charonne — wikipédia l'encyclopédie libre. En ligne ; Page disponible le 8-février-2022.
- [Wikipédia, 2022b] Wikipédia (2022b). Chanson réaliste — wikipédia, l'encyclopédie libre. [En ligne ; Page disponible le 14-juin-2022].
- [Wikipédia, 2022c] Wikipédia (2022c). Charonne (métro de paris) — wikipédia, l'encyclopédie libre. [En ligne ; Page disponible le 16-juin-2022].
- [Wikipédia, 2022d] Wikipédia (2022d). Chelles — wikipédia l'encyclopédie libre. En ligne ; Page disponible le 30-mai-2022.
- [Wikipédia, 2022e] Wikipédia (2022e). Fresnes (val-de-marne) — wikipédia l'encyclopédie libre. En ligne ; Page disponible le 5-mai-2022.
- [Wikipédia, 2022f] Wikipédia (2022f). Grand palais (paris) — wikipédia l'encyclopédie libre. En ligne ; Page disponible le 24-mai-2022.
- [Wikipédia, 2022g] Wikipédia (2022g). Hauts-de-seine — wikipédia l'encyclopédie libre. En ligne ; Page disponible le 30-mai-2022.
- [Wikipédia, 2022h] Wikipédia (2022h). Mesure (musique) — wikipédia l'encyclopédie libre. En ligne ; Page disponible le 17-mars-2022.
- [Wikipédia, 2022i] Wikipédia (2022i). Nanterre — wikipédia l'encyclopédie libre. En ligne ; Page disponible le 30-mai-2022.
- [Wikipédia, 2022j] Wikipédia (2022j). Pantin — wikipédia l'encyclopédie libre. En ligne ; Page disponible le 10-mai-2022.
- [Wikipédia, 2022k] Wikipédia (2022k). Parc de la villette — wikipédia l'encyclopédie libre. En ligne ; Page disponible le 8-mai-2022.
- [Wikipédia, 2022l] Wikipédia (2022l). Pont mirabeau — wikipédia l'encyclopédie libre. En ligne ; Page disponible le 11-mars-2022.
- [Wikipédia, 2022m] Wikipédia (2022m). Rungis — wikipédia l'encyclopédie libre. En ligne ; Page disponible le 3-mai-2022.

- [Wikipédia, 2022n] Wikipédia (2022n). Torcy (seine-et-marne) — wikipédia, l'encyclopédie libre. [En ligne ; Page disponible le 3-juin-2022].
- [Wikipédia, 2022o] Wikipédia (2022o). Tour eiffel — wikipédia, l'encyclopédie libre. [En ligne ; Page disponible le 14-juin-2022].
- [Wikipédia, 2022p] Wikipédia (2022p). Trappes — wikipédia l'encyclopédie libre. En ligne ; Page disponible le 6-juin-2022.
- [Wikipédia, 2022q] Wikipédia (2022q). Versailles — wikipédia l'encyclopédie libre. En ligne ; Page disponible le 24-mai-2022.
- [Wikipédia, 2022r] Wikipédia (2022r). Vincennes — wikipédia l'encyclopédie libre. En ligne ; Page disponible le 5-juin-2022.