Master 2 - Traitement Automatique des Langues
Années 2020 – 2021
Université Paris Nanterre

Rapport de Stage

Du 15/02/2021 Au 13/08/2021

Université Paris Nanterre

Fondation

Chaire La Défense en Perspectives

Représentations de La Défense dans La Presse

Valentin-Gabriel SOUMAH

Enseignant-Référent :

Iris Eshkol-Taravella

Tutrice dans l'Organisme d'Accueil :

Ghislaine Glasson Deschaumes

1 Remerciements

Je tiens à remercier tous les membres du personnel enseignants du master TAL de L'université Paris-Nanterre, de L'Université Sorbonne-Nouvelle et de l'Institut National des Langues Orientales grâce à qui j'ai pu acquérir les compétences théoriques et techniques qui m'ont permis de mener ce stage à bien.

Je dois témoigner ma gratitude envers la Fondation « Chaire la Défense en Perspective » ainsi que tous ses membres et partenaires et en particulier madame Ghislaine Glasson Deschaumes qui a été ma tutrice de stage.

Je tiens à remercier Madame Iris-Taravella, responsable du Master 2 TAL à L'Université Paris-Nanterre grâce à qui j'ai pu faire ce stage et pour sa disponibilité et ses précieux conseils pendant tout le cours du stage.

Il est également très important de remercier madame Camille Debras pour sa participation active dans le projet.

Je remercie également Corentin Vialar qui a contribué à un échange productif au sein et hors des réunions pendant le stage ainsi que les étudiants de Master 1 et de Licence 3 qui m'ont grandement aidé au cours de ce stage.

Je remercie à nouveau Madame Iris Eshkol-Taravella qui a permis la présentation du travail réalisé à *l'Atelier Humanités Numériques de la Conférence SAGEO (Spatial Analysis and Geomatics Conference)* 2021 et au colloque *De l'espace à la langue* à Dijon en 2021 .

2 Synthèse

Ce rapport de stage explore les problématiques liées à la représentation d'un lieu dans mes media écrit. Notre étude se concentre plus spécifiquement sur la représentation du quartier d'affaires de la Défense dans la Presse française et dans la presse Anglaise. Le Projet utilise diverses techniques du traitement automatique des langues dont les approches symboliques et statistiques ainsi que l'analyse linguistique pour proposer une méthode capable d'analyser la perception du lieu dans le corpus.

2 Abstract

This internship report delves into the various topics surrounding the portrayal of a location in written media. This study namely focuses on the portrayal of The La Défense buisness district in in the french and english presses. Through the Projet, multiple Natural Language Processing techniques, including rule-based and statistics-based approaches as well as linguistics studies allows us to implement a method that lets us analyze how the location is portrayed in the corpus.

3 Table des matières

4 Introduction

Dans le cadre de mon master 2 en traitement automatique des langues, j'ai effectué un stage de 6 mois au sein de la fondation : La chaire la Défense en perspective. Ce programme de recherche transversal repose sur une pluralité de disciplines en sciences humaines et sociales et conduit une diversité de projets en lien avec le quartier d'affaires de La Défense. Il mobilise conjointement les savoirs et les savoir-faire des entreprises et acteurs publics dont notamment l'Université Paris-Nanterre.

Ce stage ayant commencé le 8 Février et devant se terminer le 8 Août, seuls 4 des 6 mois prévus de stage auront été effectués à la date de rendu de ce rapport. Aussi, ce document ne portera que sur le travail effectué sur ces 4 premiers mois même s'il inclura les perspectives pour le projet pour les mois restants.

Dans ce cadre, ce stage fait suite à un stage précédent et a pour objectif d'étudier la perception et représentation du quartier d'affaires de La Défense dans la Presse écrite en utilisant les diverses méthodes du Traitement Automatique des Langues.

Dans la suite de ce document nous décrirons donc dans un premier temps le cadre de travail et les membres de l'équipe avec lesquels nous avons collaboré. Nous évoquerons les travaux déjà effectués en lien avec notre problématique, à la fois dans le domaine et dans les domaines connexes. Puis nous détaillerons le travail effectué en insistant sur le cheminement de notre projet. Enfin nous conclurons en abordant les résultats et perspectives du projet.

5 Contexte de travail

5.1 La Chaire « La Défense en Perspective »

Penser le devenir des quartiers d'affaires mondiaux dans leur relation avec les villes, les territoires et leurs acteurs s'impose comme une urgence. Quelles transformations de ces quartiers pour quels usagers ? Pour quelles dynamiques sociales et économiques ? Pour quelle durabilité ? Ces préoccupations vont de pair avec un chantier stratégique de réflexion sur l'avenir du travail : les transformations technologiques, l'évolution des liens vie au travail/vie hors travail et celle des mobilités modifient le rapport au temps et à l'espace et, ce faisant, bousculent les modes de management.

Pour répondre à ces défis majeurs, de grandes entreprises, des acteurs territoriaux, le monde de la recherche au travers de l'université Paris Nanterre et sa fondation, se fédèrent autour d'une chaire : La Défense en perspectives. Elle vise à nourrir les orientations stratégiques permettant d'accroître l'attractivité des entreprises, des territoires, des acteurs publics et de faire face à la compétition mondiale des quartiers d'affaire en partageant les savoirs et les retours d'expérience.

Ce programme de recherche transversal repose sur une pluralité de disciplines en sciences humaines et sociales, d'outils méthodologiques innovants, de plateformes expérimentales proposées par l'université Paris Nanterre, avec les expertises qui en découlent. Il mobilise conjointement les savoir et les savoir-faire des entreprises et acteurs publics.

La Chaire est en partenariat avec BNP Paribas, l'Association des Utilisateurs de la Défense, Elior, La Fondation de l'Université Paris Nanterre, le Conseil Général des Hauts-De-Seine, L'Etablissement Paris La Défense, Visiativ, Samsic, EDF, Vinci, Paris Ouest La Défense.

5.2 L'Equipe Projet

Durant ces 4 premiers mois de stage le travail a été effectué en grande partie en Autonomie. Mes référents principaux ont été :

- Madame Iris ESHKOL-TARAVELLA, Responsable du Master 2 Traitement Automatique des Langues à Université Paris-Nanterre, Enseignant-Référent dans le cadre de mon Stage, Membre du Conseil Scientifique de la Fondation « Chaire La Défense en Perspective » avec la spécialité Sciences du Langage.
- Madame Camille DEBRAS, Membre du Conseil Scientifique de la Fondation « Chaire La Défense en Perspective » avec la spécialité Linguistique Anglaise.
- Madame Ghislaine GLASSON DESCHAUMES, directrice de la Fondation « Chaire La Défense en Perspective », représentante de l'organisme d'accueil et tutrice signataire dans le cadre de mon stage.

Le stage a été effectué en télétravail au vu de la situation sanitaire. Pour échanger concernant le stage, nous avons mis en place un groupe sur *Google Teams* à partir duquel nous pouvions partager des documents, échanger à l'écrit et organiser notre emploi du temps. Cette plateforme a ainsi pu être utilisée pour échanger facilement hors des réunions.

Google Teams fut un outil adapté pour pouvoir organiser des réunions en visio-conférence afin de présenter le travail effectué et discuter de l'avancement et des futures tâches à réaliser. Présent durant ces réunions, nous devons aussi mentionner :

- Corentin VIALAR, en Master 2 Traitement Automatique des Langues, un autre stagiaire de la chaire la Défense en Perspective travaillant sur un sujet différent.

Les réunions en visio-conférence ont eu lieu régulièrement, au rythme de deux à trois fois par mois selon les besoins immédiats du projet. En mai a eu lieu une réunion comptant, en plus des participants habituels, l'ensemble des membres du Conseil Scientifique de la Chaire La Défense en Perspective. Cette réunion avait pour objectif de présenter l'avancée des travaux de chacun ; nous avons pu à cette occasion faire la connaissance des membres et prendre mieux conscience de l'étendue et de la diversité du programme de recherche.

Durant les mois de Mars et Avril, une équipe d'étudiants en Master 1 de Traitement Automatique des Langues ont intégré l'équipe pour aider à réaliser, entre autres, des tâches d'annotation. Ceux-ci ont effectué ce travail dans le cadre d'un des cours de leur Master afin d'avoir un aperçu des tâches de leur futur stage/emploi. Ultérieurement, nous avons aussi supervisé l'annotation réalisée par des étudiants en Licence 3 de Sciences du Langage.

- Etudiants de Master 1 :
  - Zhuang QI, Master 1 Traitement Automatique des Langues, Sorbonne-Nouvelle
  - Santiago HERRERA YANEZ, Master 1 Traitement Automatique des Langues, Sorbonne-Nouvelle
  - Nixon SATURNE, Master 1 Sciences du Langage, Université Paris-Nanterre
- Etudiants de Licence 3 :
  - Julie HALBOUT, Licence 3, Sciences du Langage, Université Paris-Nanterre
  - Baptiste BERLIOZ, Licence 3, Sciences du Langage, Université Paris-Nanterre

6 Problématique et Etat de l'Art

Avec plus de 180 000 d'employés, le quartier d'affaires de la Défense est connu pour être le premier quartier d'affaires européen : son image est souvent associée à celle du centre économique important où se situent 3 600 entreprises, dont 15 des 50 premières entreprises mondiales. Mais la Défense ne peut pas être résumé à cet aspect, c'est aussi un ensemble architectural emblématique, un lieu de travail, d'habitat, de divertissement ; un hub par lequel 23 millions de Franciliens passent tous les jours.

Le quartier d'affaires de la Défense évoque diverses représentation dans l'imaginaire mais celles-ci demeurent floues, c'est pourquoi nous nous demandons :

Comment étudier les représentations du quartier d'affaires de la Défense dans un corpus de presse ?

Le Quartier d'affaires de la Défense entre dans la catégorie plus générale des lieux. Nous nous sommes donc renseignés sur les travaux précédents effectués pour étudier la perception ou la représentation d'un lieu. Ce sujet est a pu être abordé sous l'angle de diverses disciplines dont la géographie et plus proche de notre champ d'étude, la linguistique de corpus et le traitement automatique des langues.

Gouvert aborde la place de la toponymie dans l'histoire de la linguistique en France, depuis le XVIIe siècle jusqu'à nos jours. Il soulève les problèmes théoriques de la toponymie et s'intéresse particulièrement aux noms propres. Il définit la notion de lieu comme une catégorie linguistique et discute des buts et techniques de la toponymie. (GOUVERT X. , 2008)

Borillo s'intéresse aux prépositions spatiales et relève leurs principales caractéristiques distinctives au niveau morphologique et sémantique et tente d'expliquer celles- ci. (BORILLO, 2001) La thèse de Cadiot soutient qu'un repérage dans l'espace est le modèle de base de la signification des prépositions, l'identification de ce sens spatial – ou plus précisément de la référence au monde physique étant constitutif de ces signes linguistiques. (CADIOT, 2002)

Les travaux de C. Dominguès et I. Eshkol-Taravella présentent une méthode pour détecter automatiquement les toponymes dans les titres de cartes personnalisées. Cette méthode prend en compte à la fois le contexte extra-linguistique (ici géoréférencement) et le contexte linguistique. (DOMINGUES & ESHKOL-TARAVELLA, Toponym recognition in custom-made map titles, 2015)

Le Projet MATRICIEL réalisé par Dominguès, C., S. Weber, C. Brando, L. Jolivet et M.-D. Van Damme est fondé sur l'analyse et la cartographie d'un corpus de récits de vie de républicains espagnols ; il s'intéresse à la migration sous l'angle des lieux décrits par les migrants dans les récits de leur vie. La méthode proposée utilise l'apprentissage automatique supervisé pour la reconnaissance automatique des noms de lieux générique. Les sentiments associés aux lieux ont été inférés à partir d'un lexique de sentiments analysé en contexte. La représentation vise une cartographie des informations sensibles liées aux lieux. (DOMINGUES, WEBER, BRANDO, JOLIVET, & VAN DAMME, 2017)

La publication de Catherine Dominguès, Laurence Jolivet, Carmen Brando et Marion Cargill fait suite au projet MATRICIEL et présente la démarche mise en place pour une analyse pluridisciplinaire de ce corpus de récits de vie, qui combine des méthodes et outils du traitement automatique des langues et de la cartographie. L'identification puis la représentation cartographique des lieux noms propres ont mis en évidence la répartition spatiale des parcours de vie des témoins, déterminés par le contexte historique et les choix personnels. (DOMINGUES, JOLIVET, BRANDO, & CARGILL, 2019)

C., C. Dominguès and M. Capeyron ont réalisé une étude comparative des systèmes de reconnaissances d'entités nommées spécifiquement pour les entités de type "toponyme". Cet état de l'art se concentre sur les outils supervisés comme Stanford NER pour l'identification de lieux génériques dans des corpora thématiques. (BRANDO, DOMINGUES, & CAPEYRON, 2016)

« Dis-moi Orléans » par I. Eshkohl-Taravella et H. Flamein porte sur les lieux, leur perception et l'expression de celles-ci par les différents locuteurs du corpus ESLO2 (Enquête Sociolinguistique à Orléans). Les lieux sont repérés automatiquement à l'aide d'un outil de reconnaissance d'entités nommées avant une correction manuelle qui tient compte de la variation dans leur dénomination et du contexte de leur emploi. Les lieux annotés ont ensuite été étudiés quantitativement et qualitativement. Les résultats permettent de faire la première typologie des différents processus que le locuteur met en œuvre pour exprimer sa perception du lieu. (ESHKOL-TARAVELLA & FLAMEIN, 2017) Le corpus *ESLO* est traité avec des outils du traitement automatique des langues et de la géomatique, qui permettent d'en extraire des renseignements sur l'espace et sa perception afin de les représenter sur une carte de la ville. (FLAMEIN & ESHKOL-TARAVELLA, Exploitation et analyse du corpus ESLO par les outils du TAL et de la géomatique, 2021)

La thèse de Hélène Flamein s'interroge sur l'exploitation de données linguistiques dans un corpus d'oral à dimension sociolinguistique avec l'objectif d'en extraire automatiquement du contenu subjectif. Cette étude exploite le corpus ESLO pour modéliser, détecter et visualiser la perception qu'ont les locuteurs de la ville d'Orléans. Une première étape de reconnaissance des mentions de lieux en français parlé s'appuie sur une analyse linguistique de la variation des noms de lieux et prépare l'étude de la perception qui leur est associée. Des techniques de visualisation de l'information permettent de confronter les indices subjectifs extraits des transcriptions et aboutissent à une représentation graphique des lieux identifiés associée aux déclarations subjectives des locuteurs, matérialisant la perception de la ville d'Orléans par ses habitants. (FLAMEIN, « Étude de la perception d'une ville. Repérage automatique, analyse et visualisation »., 2019)

7 Le Corpus

Un stage précédent, effectué par Marina Baidina nous a fourni une base en terme de méthodologie et de corpus. Lors de son stage, Marina Baidina avait constitué un corpus composé de 353 articles de presse en français et de 75 articles de presse en anglais. La figure 1 présente la chaîne de traitement qu'elle a réalisée.

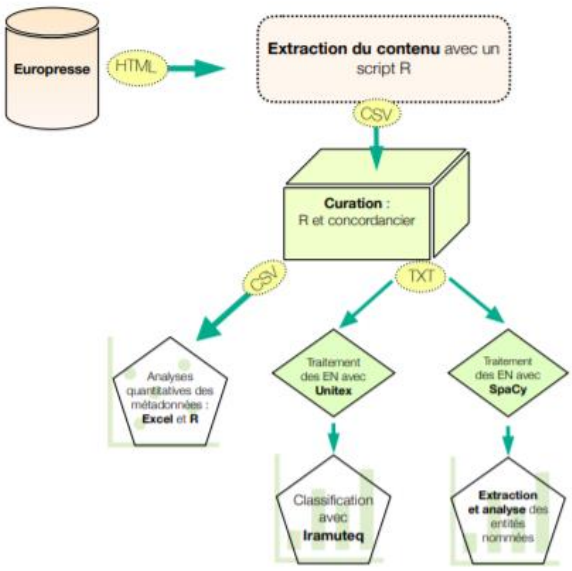

Figure 1 - Chaîne de Traitement par Marina Baidina

Le corpus a été récolté sur Europresse au format HTML à l'aide des mots-clefs « La Défense » et « quartier d'affaire ». Nous avons repris le même corpus en français que celui déjà récolté par Marina Baidina ; il nous a été fourni par cette dernière au format tabulaire (CSV). Il a ensuite été converti au format XML avec un script python de manière à le rendre plus facilement annotable pour la suite du projet.

Pour le corpus en anglais en revanche, en vue d'avoir deux corpus de tailles comparables, nous avons décidé d'étendre le nombre d'articles en langue anglaise. Nous avons également utilisé Europresse pour récolter les articles avant de les convertir au format XML en suivant la même structure que pour le corpus en français.

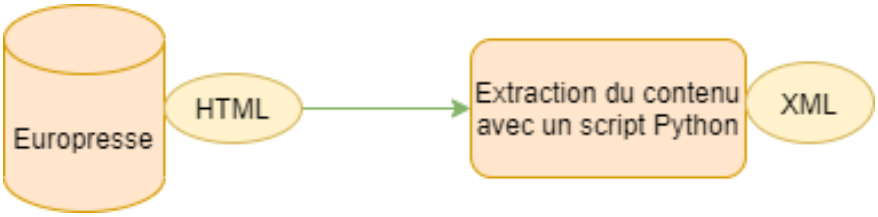

Figure 2 - Constitution du Corpus en Anglais

Les articles en français sont issus de 8 journaux différents dont la répartition est illustrée par la figure 4. Ce corpus comporte 10000 phrases comportant 229000 tokens, avec 16000 lemmes différents.

Langue	Tokens	Types	Lemmes	Phrases
Français	228819	21268	15951	10243

Figure 3 - Caractéristiques du Corpus en Français

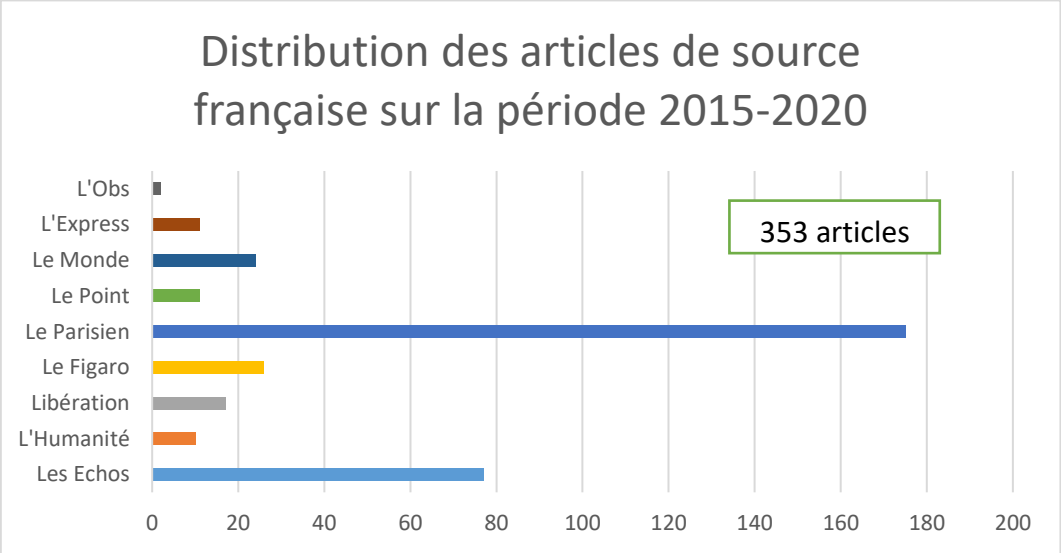

Figure 4 - Répartition des Journaux dans le Corpus en Français

Les articles en anglais eux, sont issus de 12 journaux différents dont la répartition est illustrée par la figure 6. Nous avons environ 5590000 tokens et 27000 phrases. Le corpus est plus large même si un nombre comparable d'article dû au fait que les articles sont plus longs en moyenne, en revanche comme nous le verrons les mentions de la Défense sont moins nombreuses.

Langue	Tokens	Types	Lemmes	Phrases
Anglais	558971	34577	29639	27784

Figure 5 - Caractéristiques du Corpus en Anglais

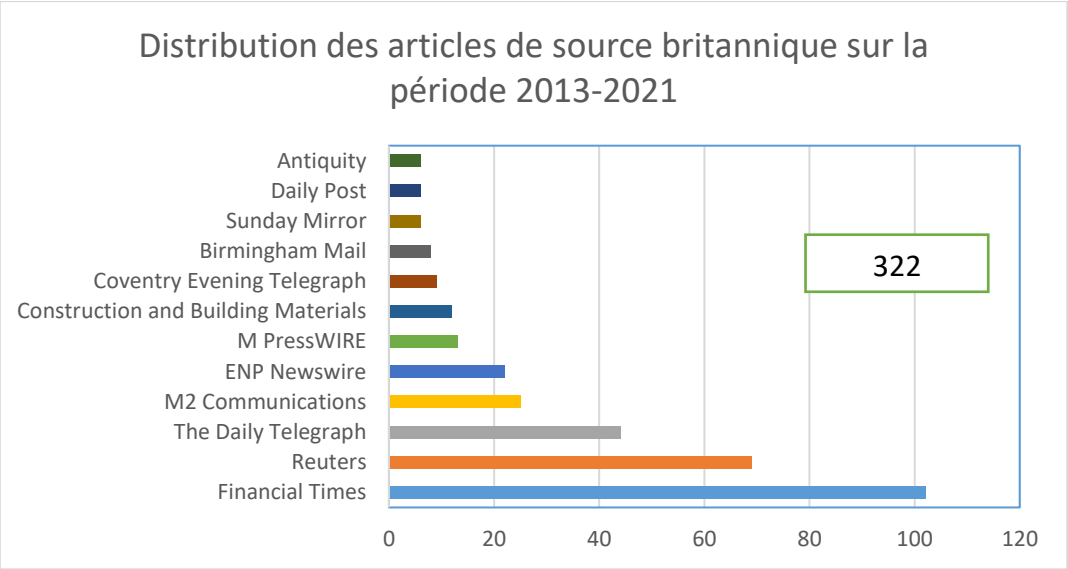

Figure 6 - Répartition des Journaux dans le Corpus en Anglais

8 La Méthodologie

Pour pouvoir réaliser une étude des désignations de la Défense et de leur contexte nous avons suivi la méthodologie illustrée par la figure 7 :

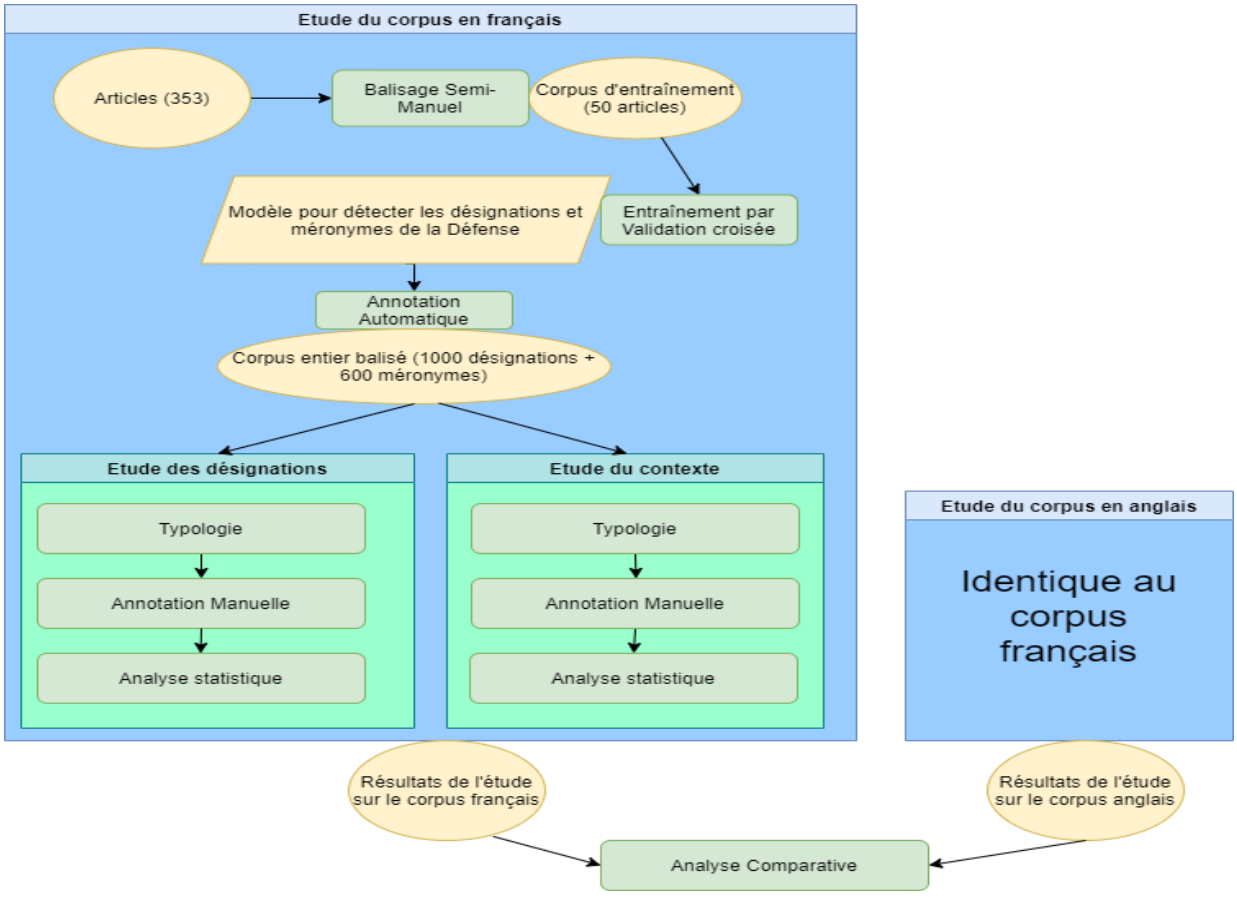

Figure 7 - Méthodologie se suivie pour étudier la représentation de la Défense dans la presse français et anglaise

En premier lieu, nous avons dû repérer l'ensemble des désignations de la Défense dans notre corpus ainsi que les désignations des lieux qui composent la Défense. Pour ce faire, nous avons

pré-annoté une partie du corpus (50 articles) en utilisant des méthodes symboliques conjuguées à une annotation manuelle. Ce corpus pré-annoté a servi de set d'entraînement pour un modèle de langage qui a ensuite permis de repérer les désignations dans l'ensemble du corpus.

Par la suite, une première analyse des désignations nous a permis de dégager une typologie pour celles-ci. Une seconde analyse effectuée à partir du contexte des désignations et du lexique présent dans celui-ci a nourri la conception d'une seconde typologie, cette fois adaptée au contexte.

Les deux typologies ont permis une annotation manuelle de chacune des désignations de la Défense et du contexte associé. Pour cela nous avons rédigé une Convention d'Annotation qui permet d'effectuer ce travail avec plusieurs annotateurs.

Enfin, nous avons conduit une analyse statistique sur les résultats de l'annotation afin d'étudier la distribution de chacune des catégories. Nous réitérons toutes les étapes précédentes avec le corpus anglais, ce qui nous permet de comparer les résultats obtenus pour les deux corpus.

9 Repérer les désignations et méronymes

Au cours de ce stage, nous avons adopté un angle d'étude qui ciblait précisément la représentation de la Défense et ce que l'on en dit. Notre première hypothèse est la suivante :

- **Les mots choisis pour nommer la Défense nous renseignent sur la représentation que l'on s'en fait.**

En effet, nommer est un acte linguistique qui n'est jamais neutre, c'est un acte qui mobilise nécessairement la subjectivité et reflète la manière dont on se représente un référent.

En explorant le corpus, nous avons également relevé la présence très nombreuses de désignations non pas du quartier de la Défense dans son ensemble mais de lieux faisant partie de la Défense. Ces lieux, étant des parties de la Défense, sont aussi très utiles pour étudier la représentation de la Défense dans la presse et réaliser une analyse plus fine en mettant en lumière les différences de perception entre les différentes parties géographiques de la Défense. **Ces lieux sont liés au quartier de la Défense par une relation de méronymie, c'est pourquoi dans la suite de l'étude nous les appelerons « les méronymes ».**

Pour pouvoir analyser ces désignations et méronymes, la première étape a été de repérer chacune des mentions de la Défense et des méronymes dans le corpus. Etant donné la taille du corpus, réaliser cette tâche manuellement aurait été très couteux en termes de temps. C'est pourquoi nous avons utilisé diverses approches pour pouvoir annoter l'ensemble du corpus avec des méthodes majoritairement automatisées.

Le format XML de notre corpus fut propice pour marquer ces désignations à l'aide de balisage. Chaque désignation a été entourée par une balise dédiée (<mention> ou <meronyme>). A chacune a été assigné un attribut « id » numérique unique qui a permis de retrouver chacune des désignations dans la suite de l'étude. Ce format a également permis, par l'intermédiaire des attributs, d'aisément enrichir l'annotation du corpus avec les nouvelles informations.

9.1 Pré-annoter le corpus semi-manuellement

9.1.1 Corpus Français

Avant d'user de méthodes de balisage automatique de notre corpus, nous avons effectué un balisage semi-manuel à des fins évaluatives. En effet, pour pouvoir évaluer la qualité de l'annotation automatique, il fallait nécessairement avoir une référence et donc annoter une portion du corpus au préalable. Nous avons décidé d'annoter 50 articles, ce qui correspond à environ 1/7 du corpus.

Sachant qu'annoter, même une portion du corpus, reste couteux, nous avons voulu alléger cette tâche en utilisant une méthode hybride qui conjugue l'annotation manuelle et les méthodes symboliques. L'annotation symbolique en premier lieu est illustrée par la figure 8 :

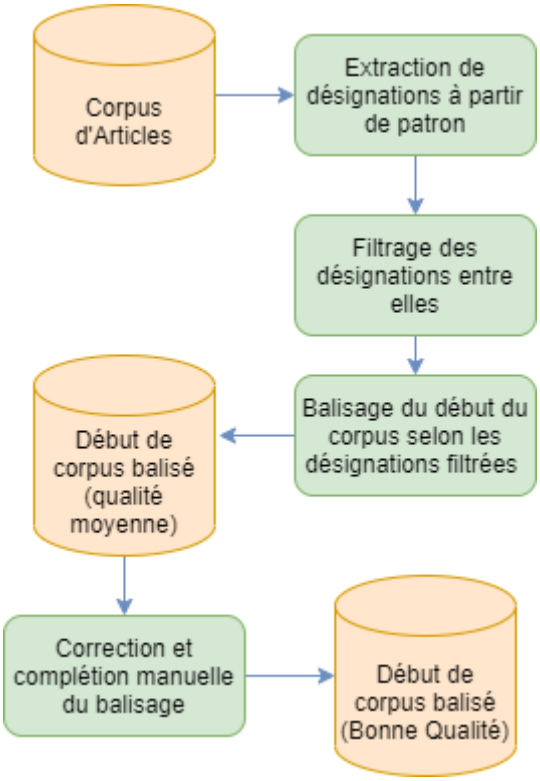

Figure 8 - Balisage Semi-Manuel

Nous avons utilisé un système de patrons à base d'expressions régulières pour repérer les mentions de la Défense, l'objectif étant de détecter automatiquement les formes les plus présentes qui réfèrent à la Défense. Celles-ci contenaient typiquement les expressions « La Défense » ou « Quartier d'affaire ». Elles pouvaient être encadrées ou entrecoupées par d'autres termes, d'où la nécessité d'utiliser des patrons à base d'expression régulières. De cette manière nous avons pu repérer des désignations telles que « Le quartier parisien de la Défense » ou encore « Le gigantesque quartier d'affaires ».

Pour les méronymes, nous avons constitué une liste de tous les monuments et lieux de la Défense. Celle-ci est obtenue à partir du site officiel du quartier d'affaire qui recense tous les lieux notables qui composent le quartier. A partir de cette liste nous avons construit d'autres patrons pour les méronymes.

Certaines correspondances de certains patrons se superposaient avec les correspondances d'autres patrons (par exemple « La Grande Arche de La Défense » englobe à la fois un

méronyme et une mention). Pour résoudre ce problème, une solution envisagée a été d'exploiter la possibilité d'avoir des balises enchâssées de manière à conserver la désignation des deux entités, et avoir un balisage de la sorte :

« <meronyme>La Grande Arche de <mention>La Défense</mention></meronyme> ».

Si cette solution était satisfaisante d'un point de vue sémantique, elle rendait très difficile l'exploitation de l'annotation. Elle rendait très laborieuse l'extraction les tokens pour la suite du travail pour des raisons que nous expliciterons dans la partie sur CamemBERT.

Aussi nous avons opté pour un système privilégiant la désignation la plus large : lorsqu'une désignation en contenait une autre, nous conversions celle-là. Lorsqu'en revanche il y a une superposition entre deux désignations sans que l'une n'englobe complètement l'autre, c'est la désignation qui compte le plus de tokens qui était conservée.

Après avoir pré-annoté 50 articles avec ces méthodes symboliques, nous avons effectué une vérification manuelle en corrigeant et complétant les balises. Les noms des balises attribués incorrectement (<meronyme> à la place de <mention>) ont été modifiés, les balises manquantes ont été ajoutés et les balises excédentaires retirées. Ainsi nous avons obtenu une annotation manuelle parfaite des désignations de ces 50 premiers articles qui nous ont servi comme étalon d'or pour la suite de l'annotation.

9.1.2 Corpus Anglais

La même méthode a été appliquée pour annoter les 50 premiers articles du corpus anglais en utilisant des patrons différents. Il faut noter que, même si le corpus entier est de taille comparable, le nombre total de mentions et de méronymes repéré dans le corpus anglais est bien moindre, car la presse française est bien plus encline à écrire des articles entiers portant sur la Défense ou ses méronymes (puisque c'est un quartier proche géographiquement des lecteurs français) alors que la presse anglaise n'évoque la plupart du temps La Défense que dans des brèves d' articles qui portent sur un sujet plus général ou connexe.

9.2 Balisage par résolution de Coréférence

9.2.1 Corpus Français

La première méthode envisagée pour annoter automatiquement le corpus employait un outil de résolution de coréférence. Il s'agissait d'une méthode entièrement non-supervisée (ou du moins qui ne nécessitait pas d'annotation supplémentaire de notre part) qui suivait la méthodologie suivante:

- Constitution de toutes les chaines de Coréférence dans le corpus.
- Parcours de toutes les mentions de toutes les chaines et comparaison avec un patron (« la Défense »)
- Annotation de toutes les mentions de la chaîne dont l'un des maillons a été identifié comme correspondant à « La Défense »
- Répétition des étapes précédente en remplaçant le patron par les éléments de la liste de lieux. Annotation comme « méronyme »

La figure 9 illustre le procédé pour une chaîne :

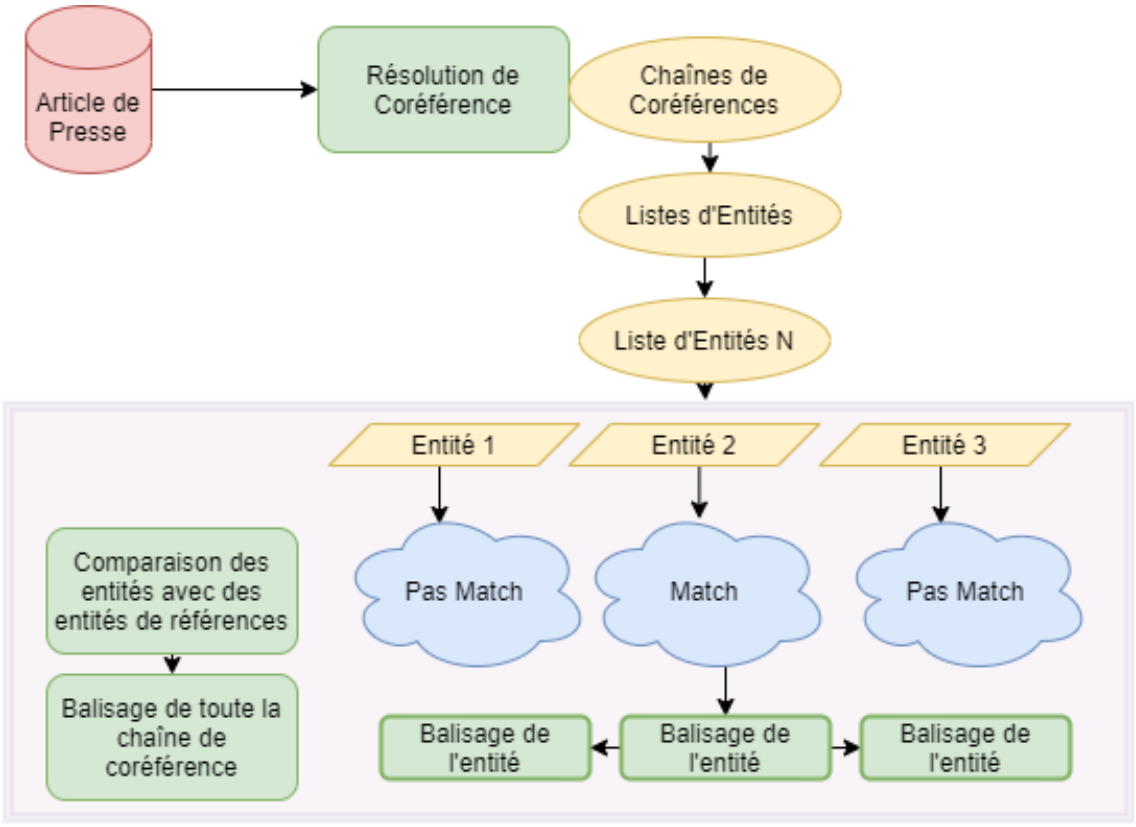

Figure 9 - Balisage Automatique Basé sur la Résolution de Coréférence

Pour mettre en œuvre cette méthode nous avons très vite dû faire face à la grande rareté des ressources et outils pour le français dans le domaine de la résolution de coréférence, la majorité de la recherche sur ce sujet portant sur l'anglais. Après avoir exploré l'état de l'art dans le domaine, nous avons retenu deux outils de coréférence disponibles et utilisables par le public :

- *COFR: COreference resolution tool For French* (WILKENS, OBERLE, LANDRAGIN, & TODIRASCU)
- *DeCoFre : Detecting Coreference For Oral* French (GROBOL, 2021)

Devant la grande difficulté d'utilisation du premier, nous avons opté pour *DeCofre* l'outil developpé par Loïc Grobol. Celui-ci fût très difficile à mettre en œuvre sur notre machine, entre autres à cause de la difficulté de mettre en phase les différentes versions requises de Python et des Bibliothèques associées mais aussi et surtout les besoins de mémoire et d'une puissante carte graphique Nvidia pour effectuer les calculs. Les résultats de l'évaluation de cette méthode apparaissent sur la figure 10.

	Précision	Rappel	F-Mesure	Support
O	0.98	0.98	0.98	15074
mention	0.23	0.44	0.30	146
meronyme	0.15	0.22	0.18	120
accuracy			**0.97**	**17340**
Macro Moyenne	**0.56**	**0.66**	**0.60**	**17340**
Moyenne Pondérée	**0.97**	**0.97**	**0.97**	**17340**

Figure 10 - Résultats pour le Balisage Basé sur la Coréférence avec DeCOFre (français)

Les résultats furent relativement mauvais, car *DeCofre*, comme indiqué par son nom est un modèle qui a été conçu avant tout pour la résolution de coréférence sur des corpus oraux (ou du moins transcrits depuis l'oral). Notre corpus d'articles de presse est un corpus purement écrit suivant assez rigoureusement les normes du français écrit. Aussi, ***Decofre* était très peu adapté à la résolution de coréférence dans ce corpus** et donne un rappel et une précision qui sont inférieurs à ceux de la pré-annotation avec des méthodes symboliques. Il ne présentait donc pas d'intérêt pour notre tâche en l'état.

9.2.2 Corpus Anglais

L'anglais étant bien mieux fourni que le français au niveau des ressources concernant la résolution de Coréférence, la recherche d'un outil pour le corpus anglais s'est révélée plus aisée. Nous avons choisi l'outil *neural-coref* (Huggingface) pour sa facilité d'utilisation, d'autant qu'il se marie parfaitement avec *Spacy* que nous utilisons par ailleurs pour notre projet. La modèle utilisé a été le modèle par défaut pré-entraîné pour l'anglais.

La méthodologie suivie a été la même que celle qui est développée plus haut pour le français. Il faut noter que le nombre de désignations est moins important Nous avons obtenu les résultats de la figure 11 :

	Précision	Rappel	F-Mesure	Support
O	0.99	0.99	0.99	11071
mention	0.47	0.84	0.60	119
meronyme	0.28	0.23	0.25	108
accuracy			**0.98**	**11298**
Macro Moyenne	**0.58**	**0.69**	**0.61**	**11298**
Moyenne Pondérée	**0.98**	**0.98**	**0.98**	**11298**

Figure 11 - Résultats pour le Balisage Basé sur la Coréférence avec neuralcoref (anglais)

Les résultats ont été nettement meilleurs qu'avec *Decofre* pour le français. En effet *neuralcoref* a été entraîné sur la partie en anglais de l'*OntoNotes Release 5.0 Dataset* qui contient divers genres pour la plupart appartenant à l'écrit. Cela explique la meilleure performance de ce modèle sur notre corpus. Le rappel est très bon pour les mentions de la Défense mais la précision peu élevée. Les méronymes quant à eux ont un score qui laisse grandement à désirer.

9.3 Balisage avec Un Modèle de Langage pré-entraîné (*CamemBERT* et *RoBERTa*)

9.3.1 Corpus Français

N'ayant pas obtenu des résultats satisfaisants avec l'approche non-supervisée, nous nous sommes tournés vers une approche supervisée se basant sur le modèle *CamemBERT* (MARTIN, et al., 2020). *CamemBERT* et un modèle de langage basé sur les transformers, c'est une adaptation de *BERT* entraîné sur un large corpus en langue française. Il est reconnu pour atteindre d'excellences performances pour diverses tâches en TAL dont l'étiquetage morphosyntaxique, l'analyse morphosyntaxique et la reconnaissance d'entités nommées. C'est cette dernière tâche qui nous intéressait le plus car *CamemBERT* a permis d'en améliorer l'état de l'art. Cherchant à repérer des séquences de tokens et à leur attribuer une étiquette particulière, **notre tâche revenait par conséquent à de la classification de tokens** (un cas plus général de la reconnaissance d'entités nommées). La méthodologie est indiquée par la figure 12.

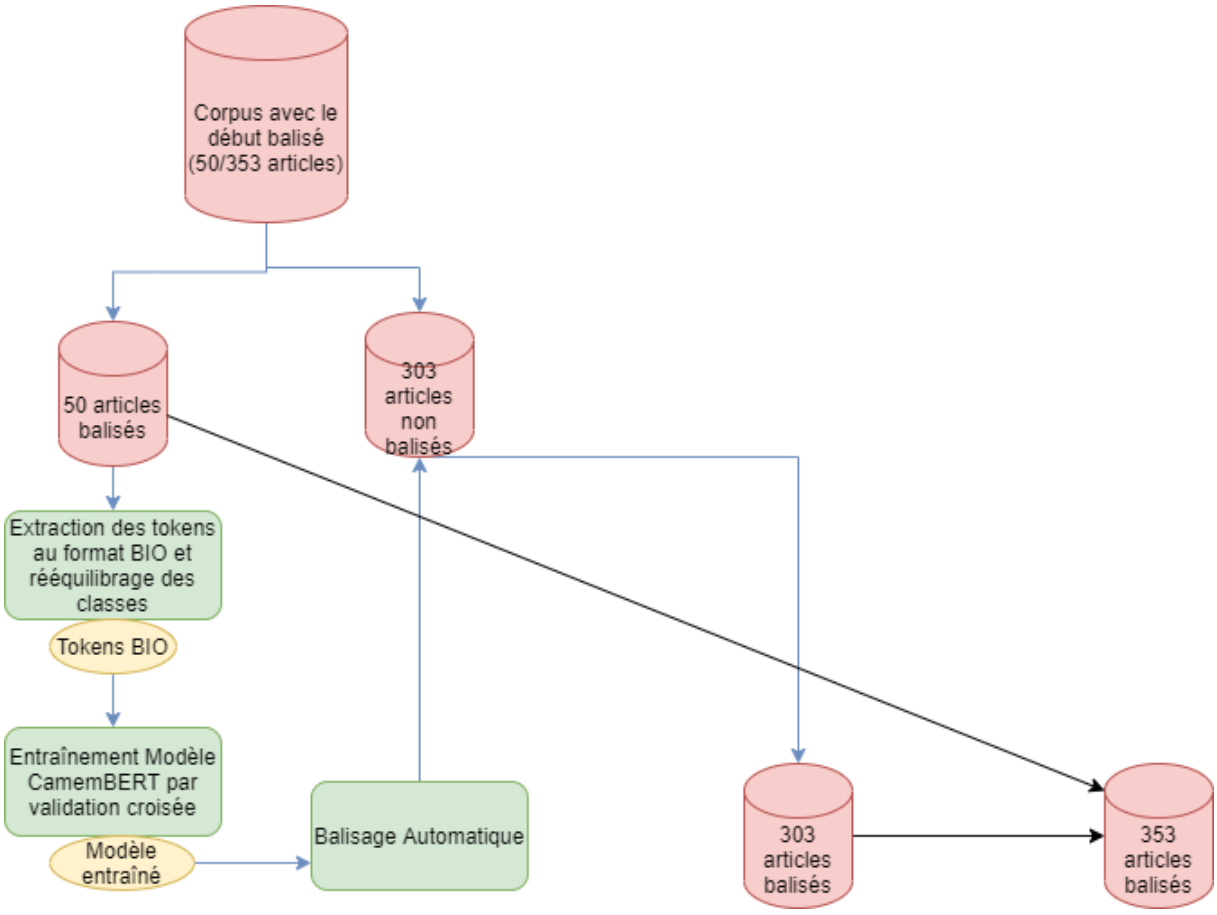

Figure 12 - Balisage du corpus avec un modèle de classification automatique de Tokens utilisant les Transformers

Cette méthode est supervisée par nécessitant une annotation manuelle pour l'entraînement en plus de celle effectuée pour l'évaluation. Nous avons utilisé la Bibliothèque *simple-transformers* qui nous a permis de paramétrer commodément *CamemBERT* pour une tâche de classification de tokens. Nous avons fourni au modèle un tableau contenant des séquences de token, chaque token étant associé à un numéro de phrase et à une étiquette. L'étiquette suivait le schema *B-I-O* (*Beginning-Inside-Outside*) exprimant la place du token dans l'entité :

- Beginning : Premier token de l'entité
- Inside : Deuxième token ou ultérieur
- Outside : en dehors de toutes les entités

Les tokens composant l'entité portaient donc à la fois une information indiquant leur place dans l'entité (B- ou I-) et à quel type d'entité ils correspondent (méronyme ou mention) : La figure 13 fournit un exemple. Le schéma BIO n'admettant qu'une seule étiquette pour chaque token, le corpus d'entraînement ne devait contenir aucune balise enchâssée.

<méronyme>La tour Hekla</méronyme> surplombe <mention>le quartier d'affaires</mention>.								
La	tour	Hekla	surplombe	le	quartier	d'	affaires	.
B-méronyme	I-méronyme	I-méronyme	O	B-mention	I-mention	I-mention	I-mention	O

Figure 13 - Exemple de Conversion d'une phrase balisée en XML en tokens classés selon le schéma BIO

Pour convertir notre corpus vers ce format, nous avons d'abord récupéré les indices de début et de fin des balises <mention> et <meronyme> dans la chaîne de caractère. Nous avons ensuite utilisé un Parseur de Spacy pour délimiter les Tokens et les phrases et déterminé les tokens de début et de fin pour chaque entité à l'aide des indice de chaînes. Enfin nous avons attribué un identifiant à chaque phrase et ajouté l'information « B » ou « I » (selon s'il s'agit du premier token de l'entité ou non) suivi du type d'entité. Les tokens restants se sont vu attribuer l'étiquette « O ».

De manière attendue, la très vaste majorité des tokens avaient une étiquette « O » ; **pour éviter un trop grand déséquilibre de classes, nous avons décidé de filtrer le corpus avant de commencer l'apprentissage** : Nous avons retiré tous les tokens des phrases ne contenant que des tokens de type « O ». De cette manière Les tokens de type O, même s'ils demeuraient majoritaires, n'occultaient pas les autres tokens qui sont ceux-qui nous cherchions à reconnaître.

Le rappel et la précision ont été optimisés en faisant varier les hyperparamètes du modèle ainsi que le nombre d'époques. *CamemBERT* nécessitant un excellent processeur graphique (que nous n'avions pas sur notre ordinateur personnel), nous avons utilisé les processeurs graphiques de *Google* à l'aide de la plateforme *Google Colab*.

Etant donné la taille limitée du corpus (d'autant plus en ne considérant que les phrases contenant au moins une mention ou un méronyme) nous avons opté pour la validation croisée pour évaluer le Modèle. Avec 15 époques et 4 plis, soit 80 % des phrases pour l'entraînement et 20 % pour le test à chaque pli, nous obtenons la Figure 14.

	Précision	Rappel	F-Mesure	Support
O	0.97	0.98	0.97	4863
mention	0.90	0.93	0.91	554
meronyme	0.75	0.65	0.70	437
accuracy			**0.95**	**5854**
Macro Moyenne	**0.87**	**0.85**	**0.86**	**5854**
Moyenne Pondérée	**0.95**	**0.95**	**0.95**	**5854**

Figure 14 - Résultats pour le Balisage avec CamemBERT (français)

Nous avons obtenus une précision et un rappel tous deux supérieurs à 0.9 pour les mentions de la Défense soit des résultats excellents. Les Méronymes qui représentent une tâche plus difficile comme nous l'a montré la méthode précédente sont eux reconnus de manière plus que satisfaisante. Nous avons remarqué une précision meilleure que le rappel pour les méronymes et l'inverse pour les mentions. Cela peut s'expliquer par le fait que les désignations de la Défense sont plus facilement reconnaissables de manière systématique (présence de « La Défense ») ce qui a pu augmenter la confiance du modèle alors que pour les méronymes la forme de l'expression est en moyenne bien plus variée, ce qui a pu encourager le modèle vers plus plus grande prudence (notamment grâce aux modules pour éviter le surapprentissage).

Une fois satisfaits de notre modèle nous avons pu le lancer sur tous les articles que nous n'avions pas encore annotés et ainsi baliser automatiquement l'intégralité du corpus (353 articles). Le corpus intégralement annoté contient **1039 mentions et 568 méronymes**.

9.3.2 Corpus Anglais

Pour le Corpus Anglais nous avons appliqué la même méthode mais cette fois en utilisant le modèle de langage *RoBERTa* (LIU, et al., 2019). *RoBERTA* est l'optimisation de BERT qui a inspiré le projet *CamemBERT*, ce dernier ayant été entraîné sur du français. *RoBERTa* est adapté pour traiter les données en anglais. Etant donné que l'outil de coréférence à disposition pour l'anglais était cette fois plus adapté à la tâche nous avons voulu comparer les deux méthodes :

- La méthode non-supervisée utilisant la coréférence
- La méthode supervisée de classification de tokens.

Les figures 15 et 16 comparent les résultats obtenus :

	Précision	Rappel	F-Mesure	Support
O	0.98	0.98	0.98	15074
mention	0.23	0.44	0.30	146
meronyme	0.15	0.22	0.18	120
accuracy			**0.97**	**17340**
Macro Moyenne	**0.56**	**0.66**	**0.60**	**17340**
Moyenne Pondérée	**0.97**	**0.97**	**0.97**	**17340**

Figure 15 - Résultats pour le Balisage Basé sur la Coréférence avec neuralcoref (anglais)

	Précision	Rappel	F-Mesure	Support
O	0.96	0.97	0.96	15074
mention	0.73	0.81	0.77	222
meronyme	0.82	0.60	0.69	190
accuracy			**0.93**	**2823**
Macro Moyenne	**0.84**	**0.79**	**0.81**	**2823**
Moyenne Pondérée	**0.93**	**0.93**	**0.93**	**2823**

Figure 16 - Résultats pour le Balisage avec RoBERTa (anglais)

Les résultats obtenus avec RoBERTa furent nettement meilleurs avec une f-mesure de 0.77 (mention) et 0.69 (meronyme) alors que neuralcoref n'obtient que 0.60 pour les mentions. **C'est pourquoi nous avons choisis d'utiliser aussi le modèle de token classification avec les transformers pour annoter le corpus en anglais**.

Il faut toutefois noter que RoBERTA sur l'anglais (F-mesures de 0.77 et 0.69) a donné de moins bonnes performances que CamemBERT sur le français (F-mesures de 0.91 et 0.70). Cette différence peut s'expliquer par un corpus d'entraînement plus large pour le français. Néanmoins cette différence concerne surtout sur les mentions (F-mesure de 0.77 contre 0.91) car les résultats sont comparables pour les méronymes (0.69 VS 0.70). Nous avons constaté un équilibre précision-rappel similaire entre les deux modèles (précision de mention < rappel de mention et précision de méronyme > rappel de méronyme).

10 L'Etude des désignations de la Défense et de leur Contexte

10.1 La typologie

10.1.1 La typologie pour désignations

Après avoir repéré toutes désignations de la Défense dans le corpus, une typologie de celles-ci en vue de pouvoir les étudier plus finement s'est révélée nécessaire. Une analyse manuelle immédiate étant très ardue au vu du très grand nombre de mésignations de la Défense, **une première étape de synthèse des désignations a été appliquée**.

Dans un premier temps, nous avons normalisé les désignations par conversion de majuscules en minuscule et retrait des mot-vides. Ensuite nous avons appliqué une étape de stemmatisation associant chaque mot du lexique sa racine lexicale. Nous avons ainsi pu regrouper les mots issus de la même racine (« France » avec « français »). Le nombre de désignation est descendu de 1039 à 60.

Nous avons ensuite extrait des vecteurs représentant sémantiquement les lemmes des désignations à l'aide de *Word2Vec*, desquels nous faisons la moyenne afin d'obtenir une représentation sémantique de chacune des 60 désignations. Ces vecteurs ont ensuite été donné en entrée à un algorithme de *clustering* (*K-means*) qui nous a permis de regrouper les désignations étant sémantiquement le plus proche.

Enfin, nous avons réalisé une analyse linguistique sur les clusters pour **établir une typologie**. Celle-ci définit les quatre catégories différentes :

- Le procédé de nommage

Il s'agit des procédés linguistiques sont mis en œuvre pour nommer la Défense. Il est possible d'utiliser plusieurs procédés différents à la fois pour nommer La Défense, c'est pourquoi cette catégorie admet une ou plusieurs étiquettes. La Norme correspond à « La Défense » qui, selon notre hypothèse, est la manière neutre de nommer La Défense.

- La cible

La Cible est l'aspect de la Défense qui est mis en avant par le procédé de nommage. Nous avons formulé l'hypothèse selon laquelle nommer permet de marquer et d'attirer l'attention vers un ou plusieurs aspects de La Défense. Chaque désignation peut admettre là aussi une ou plusieurs des 8 étiquettes. Si le Procédé de nommage correspond à la norme, il n'y a pas de cible visée, et cela est également valide pour les catégories suivantes.

- La Polarité

La Polarité est capitale pour étudier la perception de la Défense (ou d'un lieu en général) puisqu'elle nous renseigne sur la subjectivité du locuteur par rapport à ce lieu. La Polarité se base sur les dénotations et connotations des lemmes présents dans la désignation et indique si la désignation est plutôt positive, neutre, ou négative par rapport à l'aspect de la Défense visé. Elle ne peut prendre qu'une seule étiquette.

- Le positionnement

Le positionnement sert à indiquer si la manière de désigner la Défense est prise en charge par l'auteur de l'article. Certaines désignations peuvent être produites directement par l'auteur

et d'autres peuvent provenir de la parole rapportée. Cette catégorie sert donc avant tout à avoir du recul sur les autres catégories.

10.1.2 La typologie pour le contexte

Si la manière dont la Défense est nommée est riche en information concernant la représentation du quartier dans la presse, nous ne devions pas négliger une autre source d'information tout aussi riche : le contexte de ces désignations. Avec cette seconde typologie, que nous appelons « typologie du contexte » nous avons tiré parti du repérage des désignations déjà effectué pour analyser spécifiquement le discours tenu à propos de la Défense ; cela a permis une analyse bien plus précise et ciblée des thèmes et propos entourant le quartier d'affaires.

Nous avons d'abord établi une fenêtre contextuelle (la phrase) pour cadrer notre analyse. **Nous avons extrait les lemmes les plus fréquents afin de mettre en lumière les champs lexicaux. Par ailleurs, avec les étudiants de M1 TAL, nous avons analysé les premières désignations (les 50 premières) afin d'établir une typologie**. Celle-ci s'inspire de la première tout en ajoutant des spécificités pour le contexte.

- La cible

Il s'agit de du thème/sujet en rapport avec la Défense dont on parle dans la phrase. Nous nous intéressons avant tout au prédicat et nous demandons « De quoi on parle ? ». L'ampleur du cadre d'annotation (la phrase) présente un risque de désaccord entre les futurs annotateurs, c'est pourquoi nous avons choisis de nous limiter au thème principal de la phrase, et donc à seule étiquette pour chaque contexte désignation. Si la Défense n'est mentionnée que de manière périphérique sans que le sujet de la phrase porte réellement sur la Défense, alors le contexte n'est pas annoté.

- La temporalité

La temporalité associe à la cible une dimension temporelle. Il s'agit de déterminer si la phrase porte sur un passé révolu de la Défense, sur son présent ou bien sur futur. Cette catégorie nous a permis de placer les autres catégories sur un axe temporel.

- La polarité

Tout comme la polarité pour la désignation, cette catégorie essentielle aura été très utile pour étudier la perception. La polarité se doit d'être en phase avec La cible et la Temporalité.

- Le Positionnement

Enfin, il s'agit de la prise en charge ou non du prédicat.

Nous avons donc 8 catégories différentes chacune pouvant prendre un ou plusieurs étiquettes. Pour plus de détail sur la typologie, voir la Convention d'Annotation en Annexe.

10.2 L'annotation selon la Typologie

10.2.1 La Convention d'annotation

Une fois la typologie complétée, nous l'avons utilisée pour annoter l'ensemble du corpus. Etant donné l'étendue de la tâche, nous avons eu la chance de pouvoir faire appel à des étudiants de Master 1 et de Licence 3 qui nous aidé à réaliser l'annotation. Dans un premier temps, avec les M1, nous avons analysé les 50 premiers désignations afin d'affiner la typologie et de s'assurer qu'elle fonctionnait sur les exemples analysés. Nous avons pu réfléchir ensemble à l'amélioration de la typologie : Il s'agissait une tâche très itérative où chaque amélioration amenait une nouvelle question dont la réponse amenait une nouvelle interrogation.

Pour pouvoir réaliser une annotation de qualité avec un bon accord entre les annotateurs **nous avons réalisé une Convention d'annotation**. Celle-ci contient une méthodologie pour annoter ainsi qu'une description précise de chaque catégorie et étiquette de la typologie. Tous les cas d'annotation difficiles rencontrés ont été indiqués sur la Convention avec la règle et le raisonnement à suivre dans ces cas-là, en les généralisant le plus possible. La Convention a été réalisée en commun en répartissant les tâches.

C'était une expérience très enrichissante qui m'a permis de développer et améliorer mes compétences

- pédagogiques : Il m'a fallu expliquer précisément la typologie aux étudiants, les enjeux du projet et de l'annotation et répondre clairement à leurs question
- en animation et encadrement : Préparer une présentation, échanger avec les étudiants, coordonner et répartir les tâches, gérer les délais
- de contrôle qualité : Réviser efficacement les annotations faites par les étudiants, corriger et relever les points d'attention si nécessaire.

10.2.2 L'accord Inter-Annotateur

Une fois la Convention d'annotation réalisée, nous avons annoté 80 désignations communes afin de pouvoir calculer les accords inter-annotateur et évaluer la qualité de notre annotation. Nous avons calculé le Kappa de Cohen pour chaque pair d'annotateur (4 annotateurs donc 6 paires) ainsi que le Kappa de Fleiss pour tous les annotateurs, ceci pour chaque catégorie.

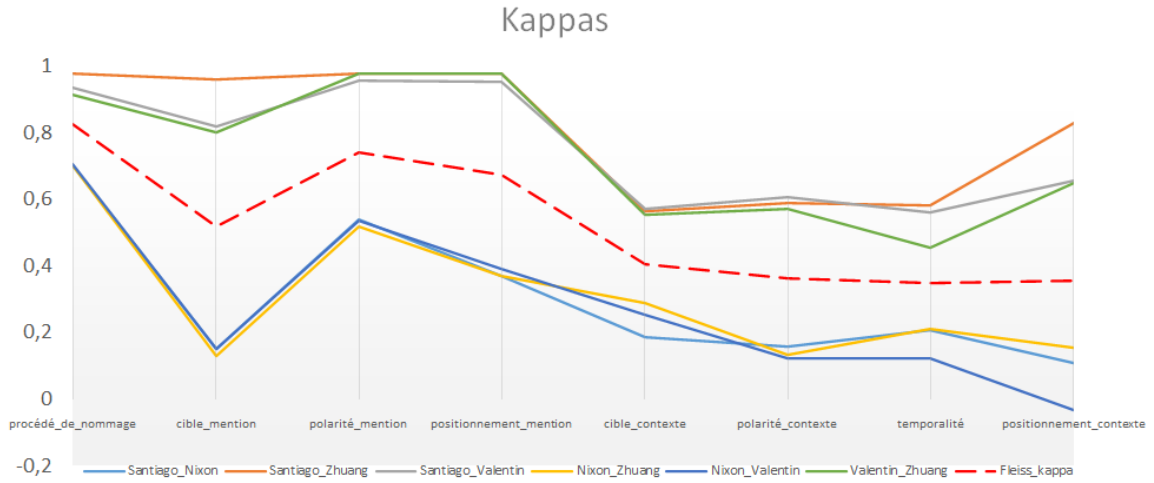

Figure 17 - Kappas de Cohen et de Fleiss entre les annotateurs

Chaque courbe du graphique de la figure 17 représente une mesure. Nous observons qu'à part pour les paires avec Nixon, **tous les accords concernant la désignation/mention sont excellents atteignant des valeurs supérieures à 0.8, parfois atteignant presque 1**. Le Kappa de Fleiss est plombé par l'accord avec un des annotateurs qui tire vers le bas, (les annotations différentes de cet annotateur sont dus au fait que sa maitrise du français n'est pas optimale). On peut le considérer comme une valeur aberrante et ne prendre en compte que les autres valeurs. Ces excellents résultats sont explicables à la fois par la qualité du guide d'annotation et par le nombre limité d'information à prendre en compte (un groupe nominal)

κ	Interprétation
< 0	Pauvre concordance
0.01 – 0.20	Faible concordance
0.21 – 0.40	Légère concordance
0.41 – 0.60	Concordance moyenne
0.61 – 0.80	Concordance importante
0.81 – 1.00	Concordance presque parfaite

Figure 18 - Interprétation des valeurs des Kappas

Pour Le contexte on les valeurs sont nettement inférieure mais elles restent au dessus de 0.6 globalement, ce qui demeure un bon accord. Cela est explicable par le bien plus grand nombre de token à interpréter et prendre en compte pour l'annotateur. Il y a bien plus de signifiants et donc une plus grande complexité potentielle des signifiés et plus d'information à traiter. Il est intéressant de remarquer que bien que la cible admette bien plus d'étiquettes que les autres catégories (qui n'ont que 3 étiquettes), l'accord demeure similaire.

Après s'être assuré de la qualité de la Convention d'annotation, celle-ci a facilité l'explication du travail aisément aux étudiants de Licence qui ont alors annoté le corpus en anglais en se basant sur la même Convention : les étudiants de Licence ont pu, sans confusion, réaliser une annotation de bonne qualité sur le corpus en anglais.

10.3 L'Analyse des résultats

10.3.1 Les résultats pour la presse française

10.3.1.1 Analyse des désignations

Une fois l'annotation réalisée pour le corpus français, nous avons effectué une analyse statistique des résultats. Pour ce faire, nous avons utilisé le langage *R*, ou plus précisément l'environnement *RStudio* pour, traiter les données, faire les calculs, et visualiser les résultats. Dans un premier temps nous avons cherché à mettre en lumière la distribution des étiquettes pour chacune des catégories.

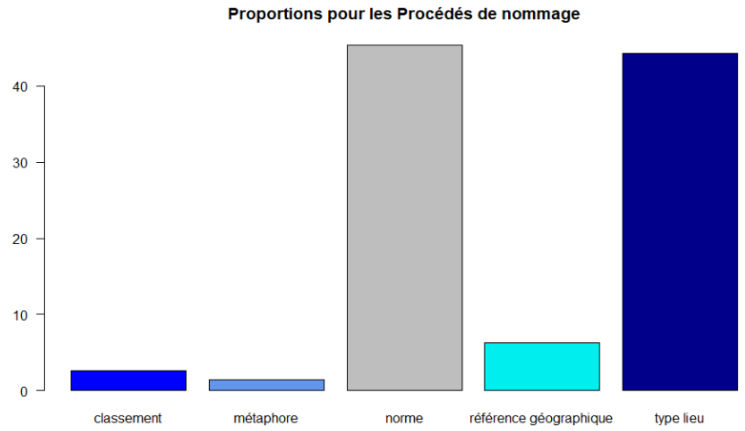

Figure 19 - Distribution des Procédés de Nommage dans le Corpus Français

La distribution des procédés de nommage révèle que **la *norme* est majoritaire** comme nous en avions fait l'hypothèse. Le procédé de nommage ***type de lieu* est presque aussi présent**. Cela s'explique par l'omniprésence de l'expression « quartier d'affaire » ou « quartier ». Les deux proportions étant presques identiques nous pourrions même conclure qu'il existe en réalité dans notre corpus deux normes de nommage. Utiliser le *type de lieu* est tellement courant qu'il est presque considéré comme la forme par défaut sur le même plan que « La Défense ». La Défense comme quartier est une représentation qui vient à l'esprit aussi natuellement que le nom officiel du lieu dans la presse. **La manière de nommer la moins fréquente est la *métaphore*** à moins de 2% , cette manière est donc l'opposée de la norme, c'est la manière de nommer la plus créative , signal d'une prise de position importante dans le nommage.

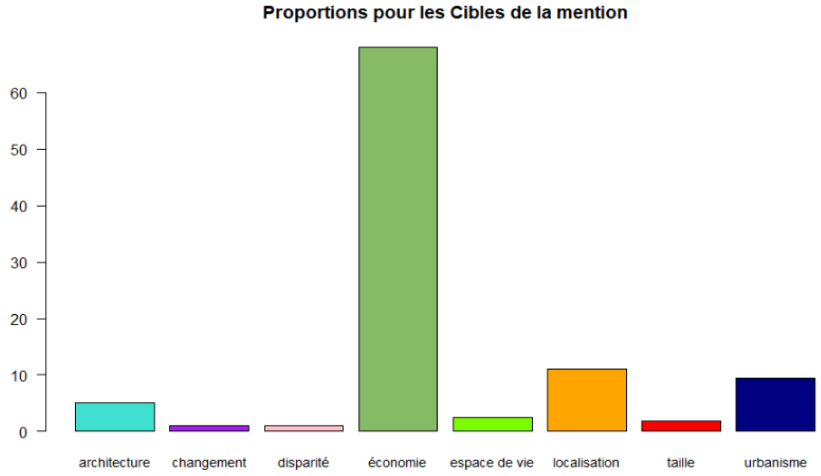

Figure 20 - Distribution des Cibles des Désignations dans le Corpus Français

En excluant la *norme* qui ne correspond à aucune cible, ***l'économie* se révèle être de très loin la cible principale des désignations de la Défense** avec une fréquence de 65%. Cela est dû, encore, à l'expression « quartier d'affaire » qui est notre « deuxième norme ». La *localisation*, l'*urbanisme* et l'*architecture* sont aussi des cibles significativement présentes dans le corpus (à plus de 5% chacune)

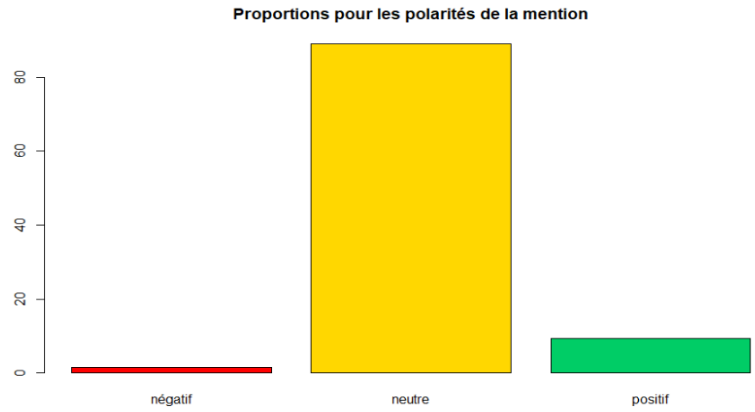

Figure 21 - Distribution de la polarité des désignations dans le Corpus Français

Les désignations sont assez peu polarisées avec 85% de désignations portant une polarité *neutre*. Les désignations sont plus fréquemment *positives* (10%) que *négatives* (2%).

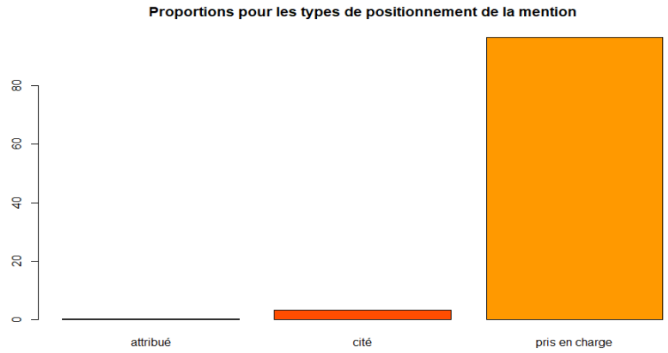

Figure 22 - Distribution du positionnement sur les désignations dans le Corpus Français

Enfin, sans surprise le positionnement est à plus de 90% *pris en charge*.

Pour pouvoir étudier plus en détail les liens entre les différentes catégories, nous avons effectué une analyse des correspondances multiples. Celle-ci a permis une exploration des données et une identification des points d'attention méritant une analyse plus fine : « quelles variables sont opposés ou groupées de manière significative ? ». La figure 23 illustre les oppositions pour les deux premières dimensions contenant respectivement 16% et 14.8% de la variance dans le corpus. Deux modalités (nos étiquettes) sont proches si elles se retrouvent souvent ensemble dans l'annotation. Nous avons aussi exploré les autres dimensions.

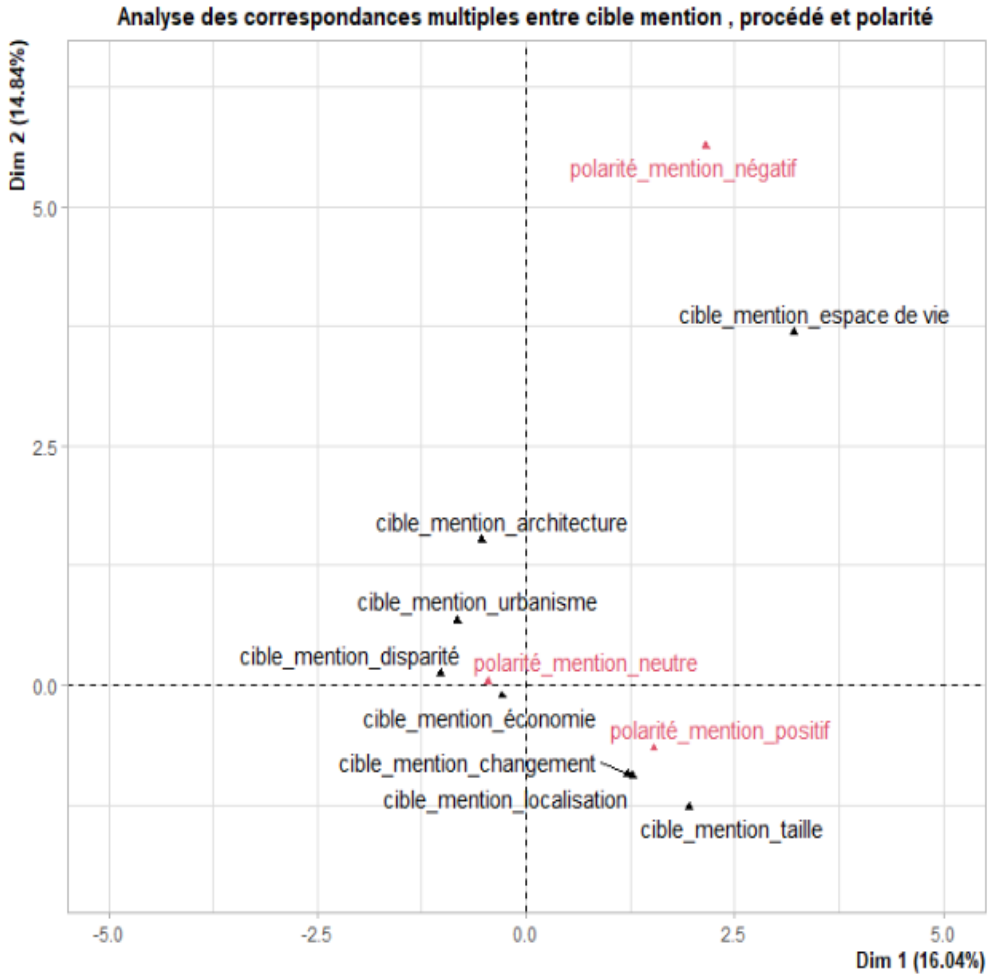

Figure 23 - Deux premières dimensions de l'Analyse des Correspondances Multiples entre les cibles des désignations, procédés de nommage et polarité des désignations dans le Corpus Français

Cette figure montre que polarité varie beaucoup selon les cibles : la cible *espace de vie* est proche de la polarité *négative*. A l'inverse, les cibles *localisation* et *changement* sont proches de la polarité *positive* et la cible *économie* est proche de la polarité *neutre*.

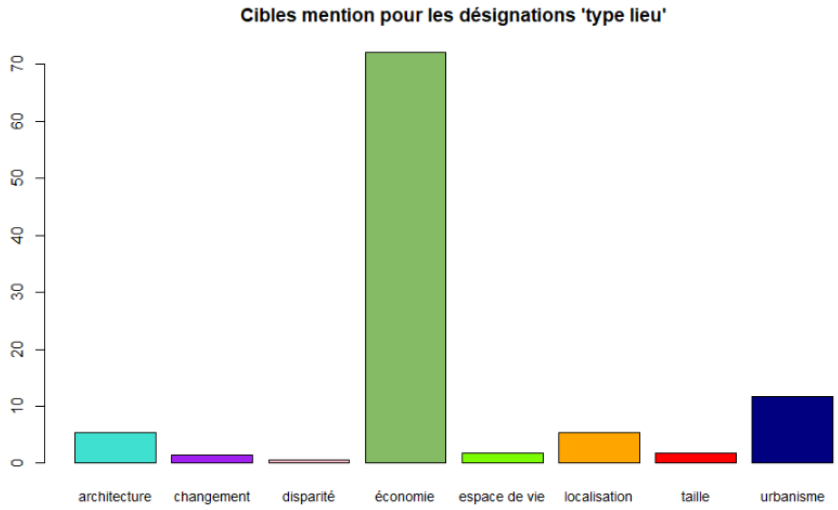

Figure 24 - Distribution des cibles des désignations pour le procédé de nommage "type de lieu" dans le Corpus Français

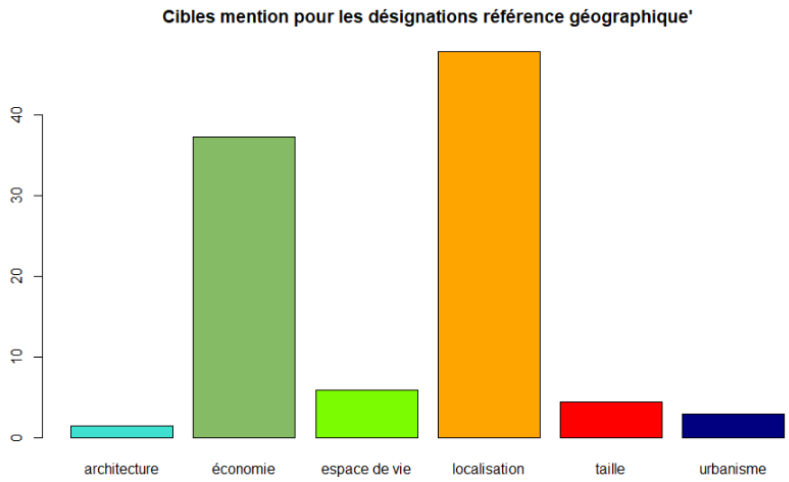

Figure 25 - Distribution des cibles des désignations pour le procédé de nommage "référence géographique" dans le Corpus Français

Nous avons analysé les proportions des cibles des désignations utilisant trois procédés de nommages différents qui ont attiré notre attention lors de l'analyse des correspondances multiples. Le procédé *Type de lieu* est très lié à l'aspect *économique*, dû à la deuxième norme « quartier d'affaire » susmentionnée. La distribution des *références géographiques* est très différente : elles attirent logiquement l'attention avant tout sur la *localisation* mais aussi sur l'aspect *économique*.

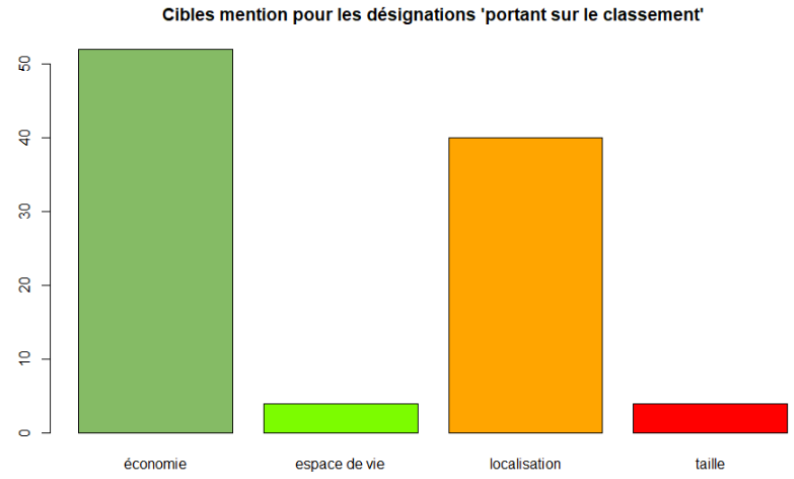

Figure 26 - Distribution des cibles des désignations pour le procédé de nommage "classement" dans le Corpus Français

La majorité des procédés reposant sur le *classement* est mis en lien avec l'*économie*. En effet, le corpus présente un nombre conséquent de désignations du type « premier quartier d'affaire ». Le *classement* est aussi mis en lien avec la *localisation*, il est très fréquemment d'ordre géographique comme par exemple « Le plus grand quartier d'affaire du pays ». Les notions de *taille* et d'*espace de vie* sont aussi liées au *classement*.

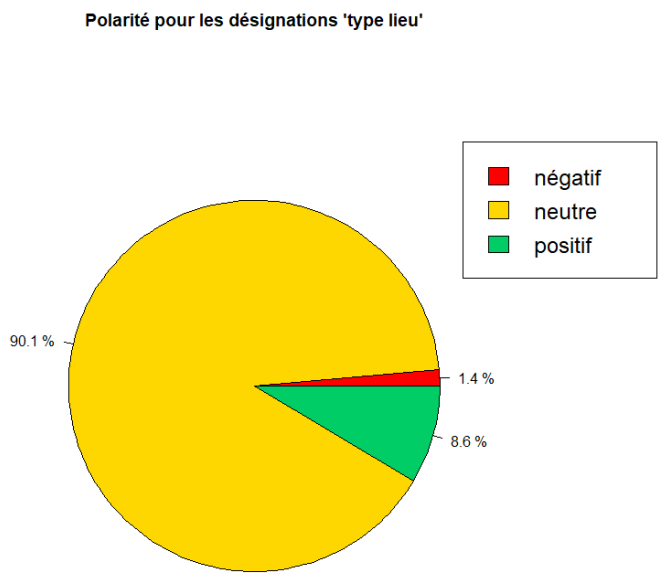

Figure 28 – Distribution de la Polarité pour le procédé de nommage « type de lieu dans le Corpus Français

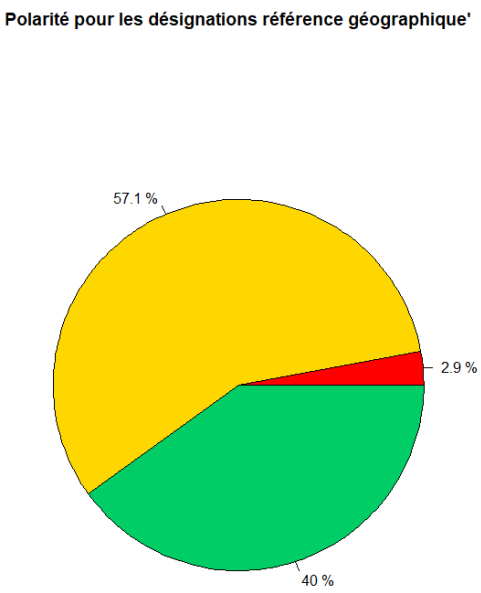

Figure 27 - Distribution de la Polarité pour le procédé de nommage "référence géographique" dans le Corpus Français

En observant la distribution de la polarité selon certains procédés de nommage, nous notons les appellations de type *type de lieu* (figure 28) sont très peu polarisées (90% *neutre*). Cette catégorie de désignation à tendance à employer des termes objectifs. Malgré tout, il y a une nette tendance *positive* (8.6%) par rapport au *négatif* (1.4%). Cette distribution est très proche de la distribution globale de la polarité, car c'est le procédé de nommage de loin majoritaire. La distribution de la polarité pour la *référence géographique* est très proche de celle de la *localisation* que nous verrons par la suite.

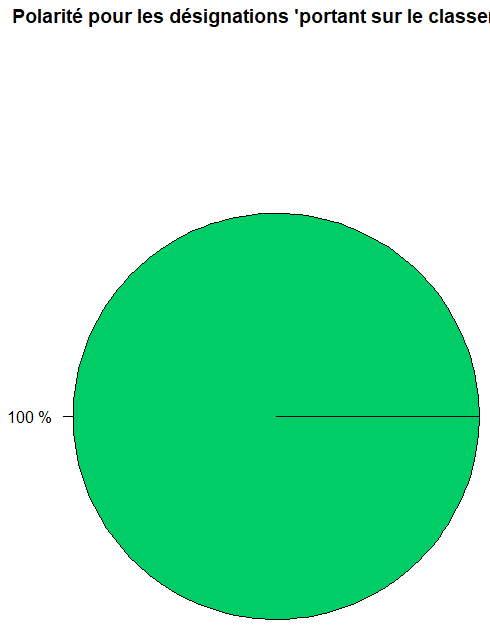

Figure 29 - - Distribution de la Polarité pour le procédé de nommage "classement" dans le Corpus Français

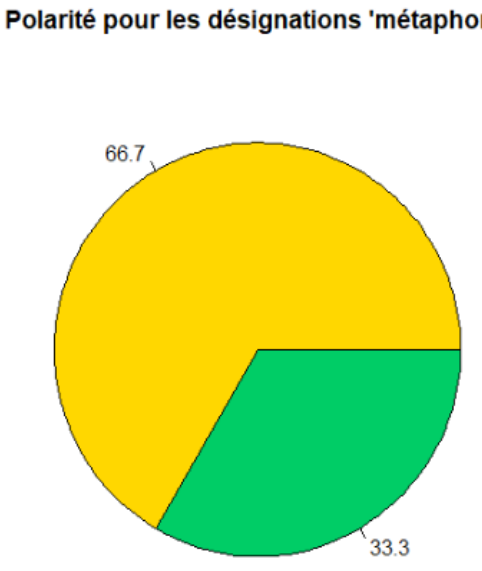

Figure 30 - Distribution de la Polarité pour le procédé de nommage "métaphore" dans le Corpus Français

Les désignations utilisant le procédé de nommage *classement* (figure 30) ont la particularité d'être toutes *positive*. L'absence de polarité *neutre* indique que le classement est toujours utilisé de manière superlative ou bien pour positionner La Défense dans des positions extrêmes (jamais à des classements intermédiaires). L'absence de polarité *négative*, elle, indique que celui-ci ne prend jamais la forme de superlatif négatif (le moins) ni de comparatif d'infériorité ou encore de classement vers un extrême négatif (le dernier)

Les *métaphores* (figure 29) elles, contiennent une majorité de désignations *neutres* mais il faut remarquer qu'elles ne présentent aucune désignation avec une polarité *négative.* **Les *métaphores* sont toujours employés soient à des visées neutres, soit à des visées positives.**

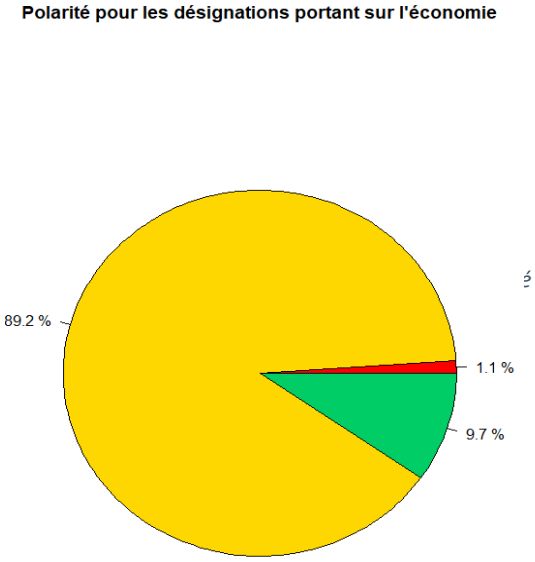

Figure 33 - Distribution de la Polarité des désignations portant sur l'économie dans le Corpus Français

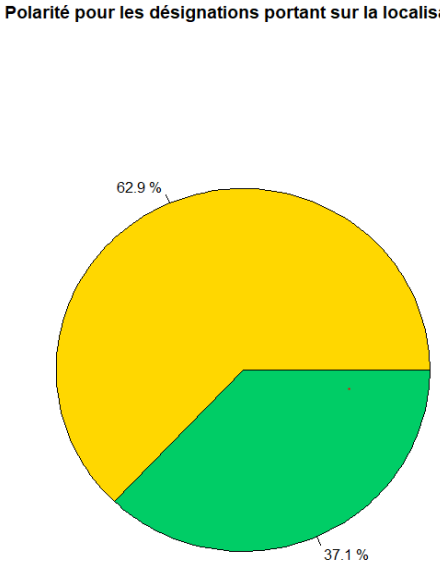

Figure 32 - Distribution de la Polarité des désignations portant sur la localisation dans le Corpus Français

L'*économie* possède une distribution similaire à la distribution globale, elle est peu polarisée, encore largement dû à la désignation « quartier d'affaire ». La *localisation* en revanche possède une part importante de *positif* (37%). Comme attendu, sa distribution est très similaire à celle des *références géographiques*. Cependant, elle n'est pas strictement identique, cela indique qu'il n'y a pas une correspondance parfaite, certaines désignations ciblent la localisation de la Défense sans utiliser une référence géographique.

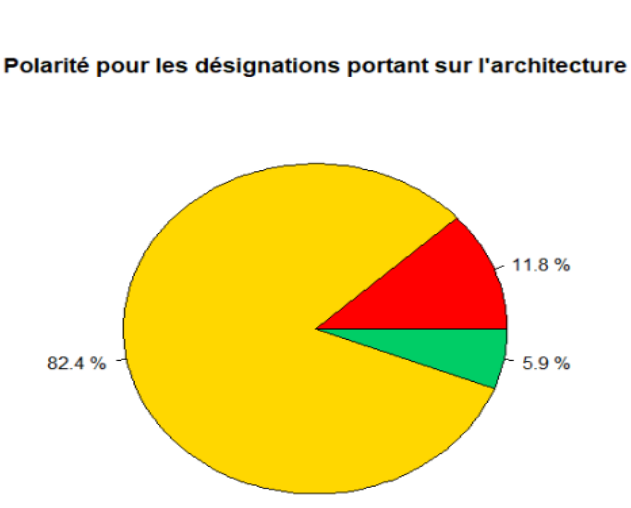

Figure 35 - Distribution de la Polarité des désignations portant sur l'architecture dans le Corpus Français

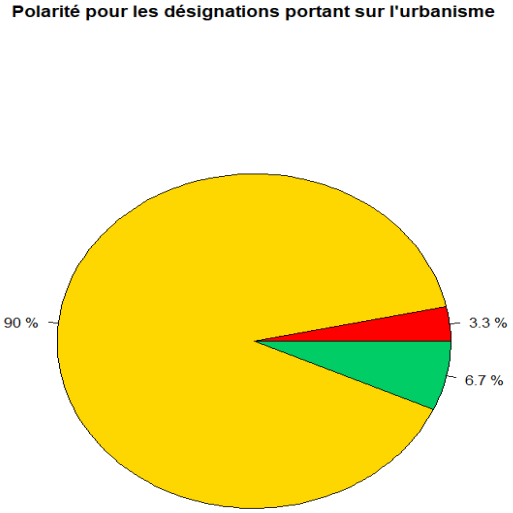

Figure 34 - Distribution de la Polarité des désignations portant sur l'urbanisme dans le Corpus Français

Les cibles *architecture* et *urbanisme* ont une distribution polaire assez proche, l'urbanisme ayant toutefois une plus large proportion de désignations *neutres*. En revanche L'aspect architectural à une plus grande part de désignations *négatives* (11.8%) que de désignations *positives* (5.9° %), cela tend à laisser penser que **les désignations de la Défense en lien avec l'*architecture* donnent une représentation plutôt péjorative des bâtiments du quartier**, ou du moins que lorsque la Défense est nommée en attirant l'attention sur ses buildings et gratte-ciel, l'accent est mis sur le négatif.

10.3.1.2 Analyse du Contexte

Nous avons aussi analysé le discours portant sur la Défense à l'aide de l'annotation du contexte. Nous avons utilisé les mêmes méthodes avec RStudio, en observant d'abord la distribution des catégories globales.

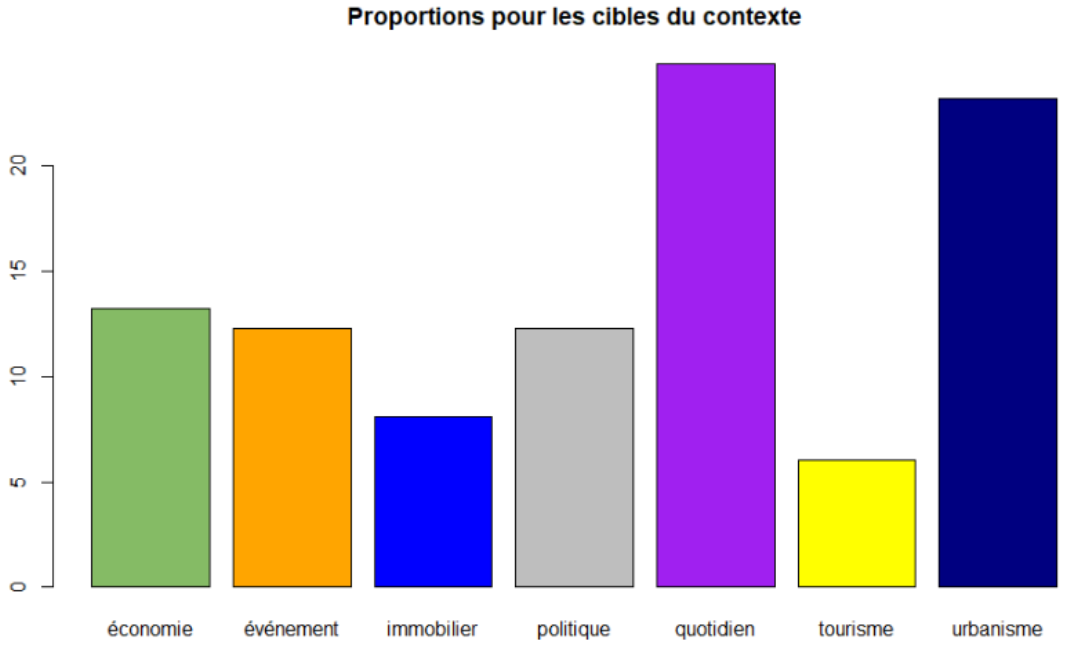

Figure 38 - Distribution des cibles du Contexte de la Défense dans le Corpus Français

La cible du contexte comporte certaines catégories communes avec les cibles de la désignation mais leur distribution est très différente : Si l'*économie* était de très loin majoritaire pour les désignations, ici elle n'est que la troisième cible la plus fréquente. Les sujets en rapport avec la Défense dont en premier lieu la *vie quotidienne* (des habitants et travailleurs) et en second lieu l'*urbanisme*, soit la gestion du quartier. **C'est donc plutôt le côté local et animé qui ressort avant l'aspect économique.** Les évènements qui ont lieu à la Défense sont aussi un sujet récurrent. Le côté touristique de la Défense celui dont on parle le moins dans notre corpus.

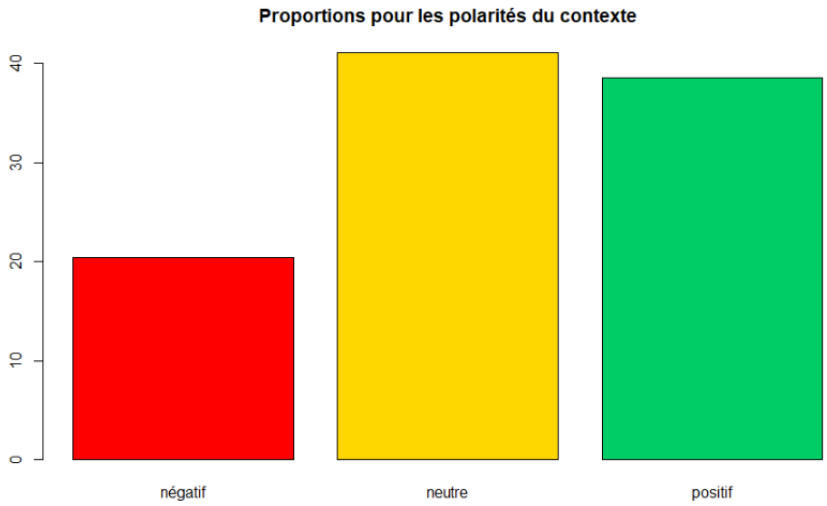

Figure 39 - Distribution de la polarité du Contexte de la Défense dans le Corpus Français

Avec seulement 40% de phrases *neutres* et presque autant de phrases *positives*, et une portion non négligeable de phrases *négatives* (20%) le contexte est bien plus polarisé que les désignations. La Défense est majoritairement évoquée de manière laudative : presque 4 fois sur 10 la Défense ou un de ses aspects est mentionné de manière positive. Nous verrons par la suite comment la polarité se répartit selon les différentes catégories.

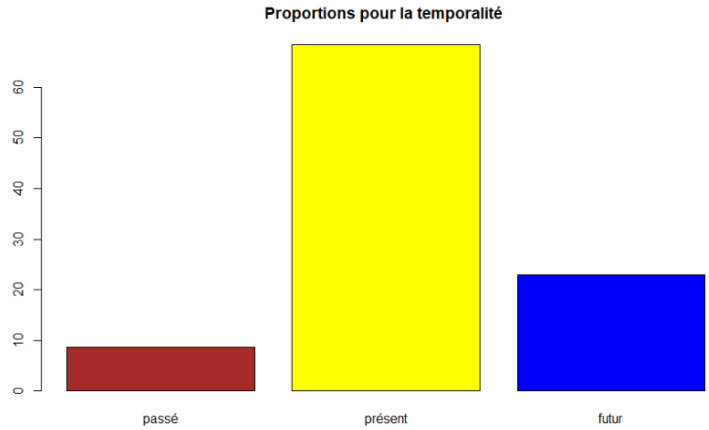

Figure 40 - Distribution de la temporalité associée au contexte de la Défense dans le Corpus Français

De manière attendue pour un corpus de presse, les journalistes parlent surtout du *présent* de la Défense. Cela inclut le présent d'habitude, le présent duratif mais aussi et surtout le présent lié au passé immédiat (lorsque le journaliste rapporte un évènement par exemple). Le *futur* est abordé avec une fréquence conséquente (25%). Le passé est la temporalité la moins employée, avec moins de 10%, **il semble donc y avoir plus projection dans le futur que de rétrospection.**

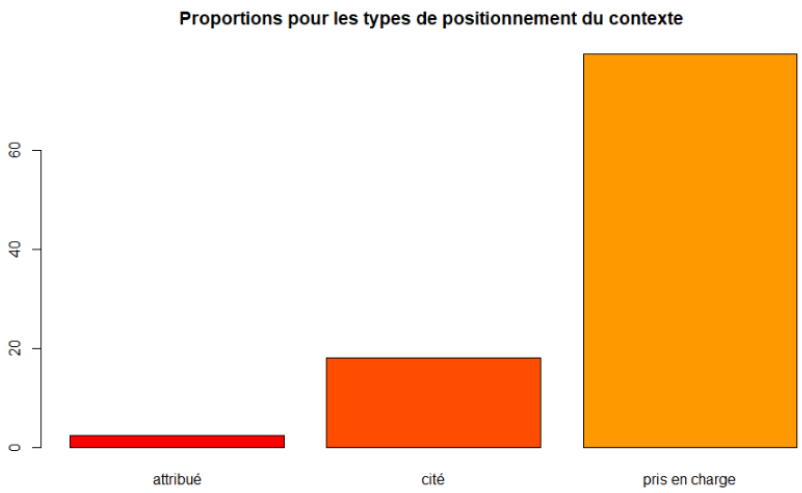

Figure 41 - Distribution du positionnement sur le contexte de la Défense dans le Corpus Français

Comme pour les désignations, le discours est majoritairement *pris en charge*. Toutefois, la part de discours *cité* est bien plus importante (20%) : le discours rapporté portant sur La Défense est non-négligeable. Le discours attribué atteint presque 5%, le journaliste adopte donc ponctuellement une posture assez peu rigoureuse (en attribuant un discours à un locuteur vague) par rapport au quartier d'affaires.

Nous avons suivi la même méthode que précédemment pour le contexte en réalisant une analyse des correspondances multiples.

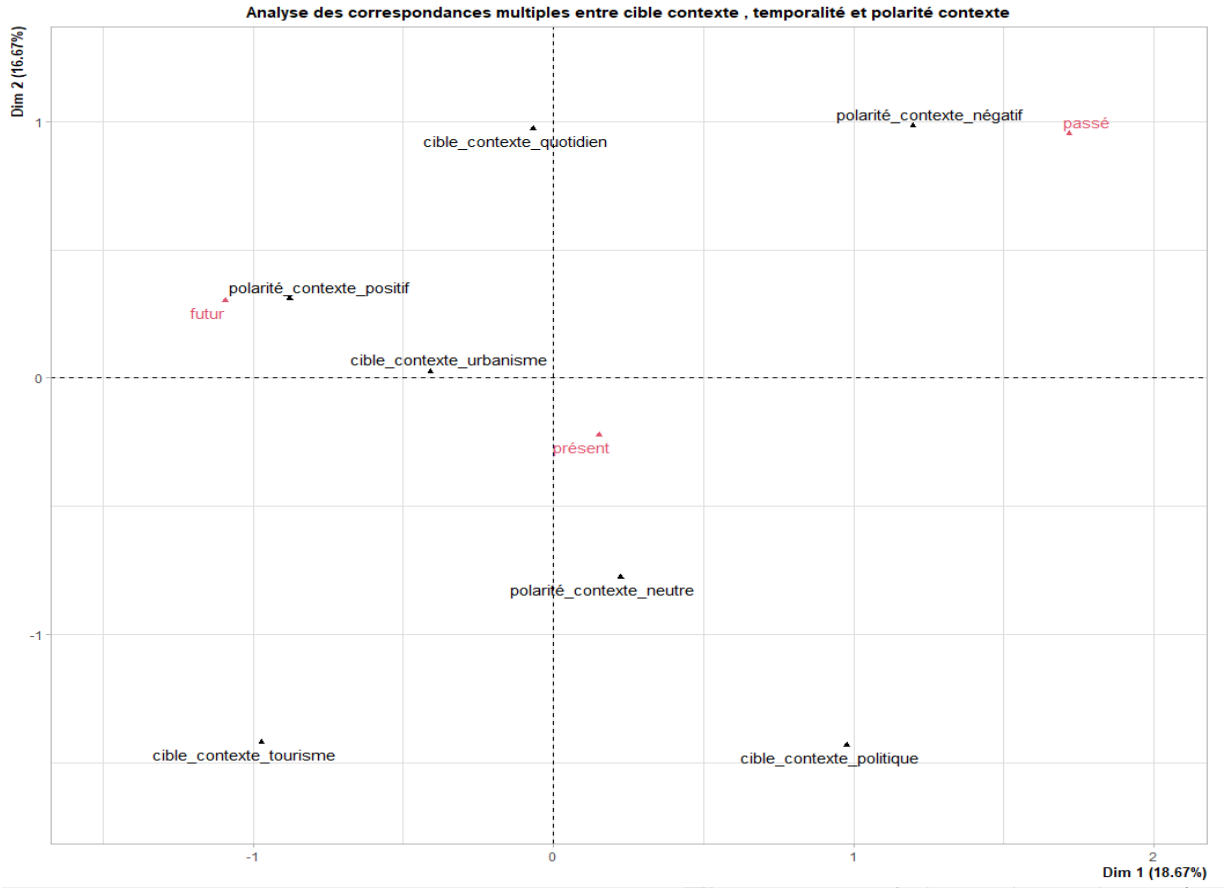

Figure 42 - Deux premières dimensions de l'Analyse des correspondances multiples entre la cible du contexte, la temporalité et la polarité du Contexte dans le corpus Français

Ces deux premières dimensions nous ont incités à explorer la temporalité en fonction de la polarité puisqu'nous pouvons distinctement observer que *passé* est à côté de *négatif*, *présent* proche de *neutre* et *futur* à proximité de *positif*.

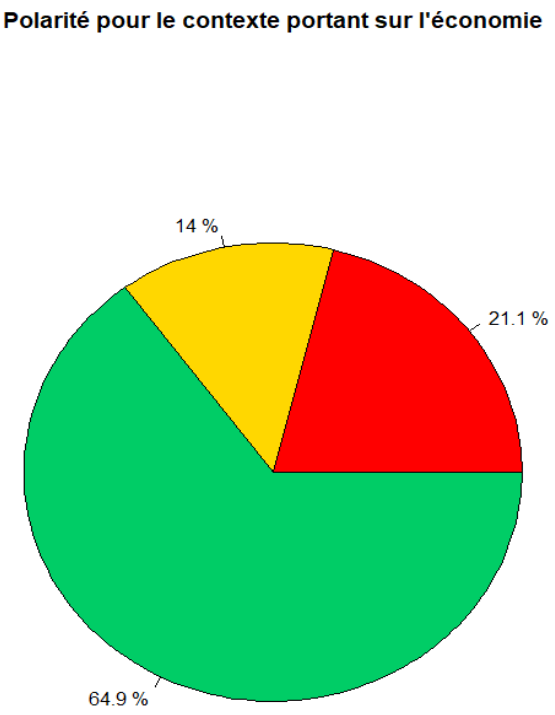

Figure 44 - Distribution de la Polarité pour le Contexte Portant sur l'Economie dans le Corpus Français

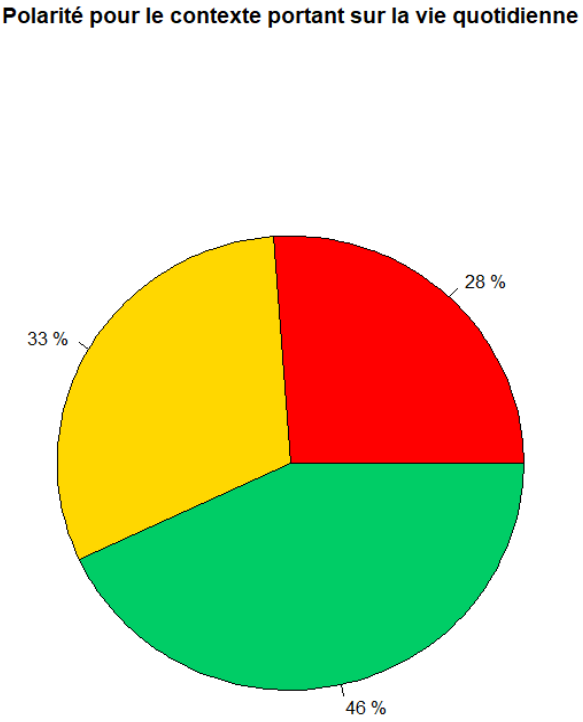

Figure 43 - Distribution de la Polarité pour le Contexte Portant sur la vie quotidienne dans le Corpus Français

L'économie de la Défense est un sujet extrêmement polarisé avec très peu de contexte *neutre*. Il est majoritairement *positif* (65%) mais présente tout de même 21% d'exemples *négatifs*. Pour la *vie quotidienne*, les trois polarités sont bien plus équilibrées avec une tendance *positive*. Ce sujet paraît polémique avec presque 30% de discours *négatif* lié à la *vie quotidienne*. **Il semble donc y avoir beaucoup d'aspects positifs dans la vie à la Défense mais également un nombre important d'aspects négatifs.**

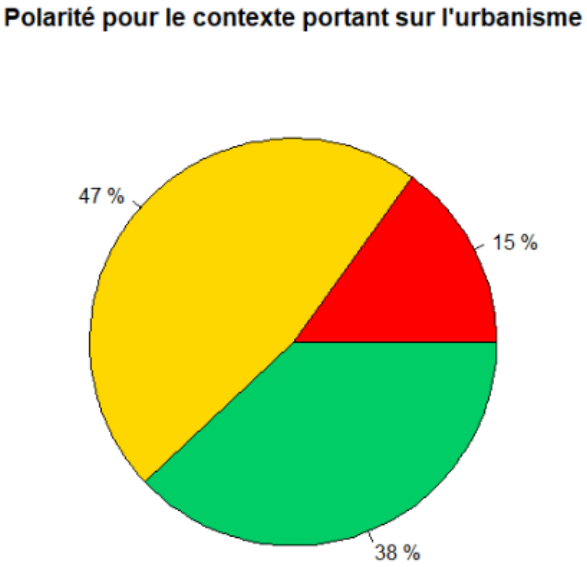

Figure 46 - Distribution de la Polarité pour le Contexte Portant sur l'Urbanisme dans le Corpus Français

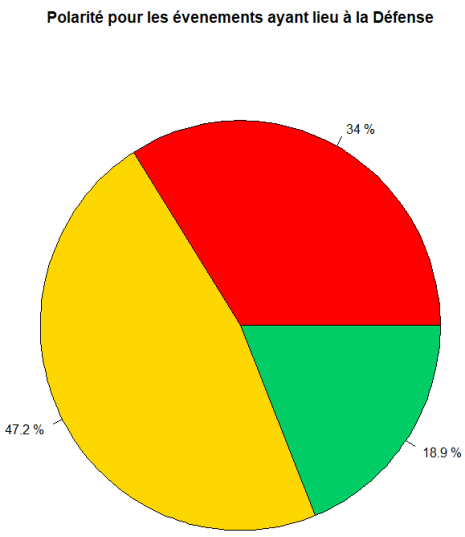

Figure 45 - Distribution de la Polarité pour le Contexte Portant sur les événements ayant lieu à la Défense dans le Corpus Français

L'aspect *urbanistique* de la Défense, qui comprend majoritairement les aménagements dans le quartier semble moins *positif* que l'aspect *vie quotidienne* mais il est aussi moins *négatif*. Les mentions *positives* révèlent que l'aspect positif porte surtout sur la création de parcs et généralement sur le verdissement de la Défense. L'aspect négatif en revanche concerne surtout les problèmes de projets de construction de tours durant plus longtemps qu'anticipé.

Les évènements ayant lieu à la Défense sont majoritairement *neutres* mais ont une part très large de discours *négatif*. Il s'agit pour une grande partie d'événements liés au terrorisme puisque notre corpus couvre la période des attentats de Janvier 2015.

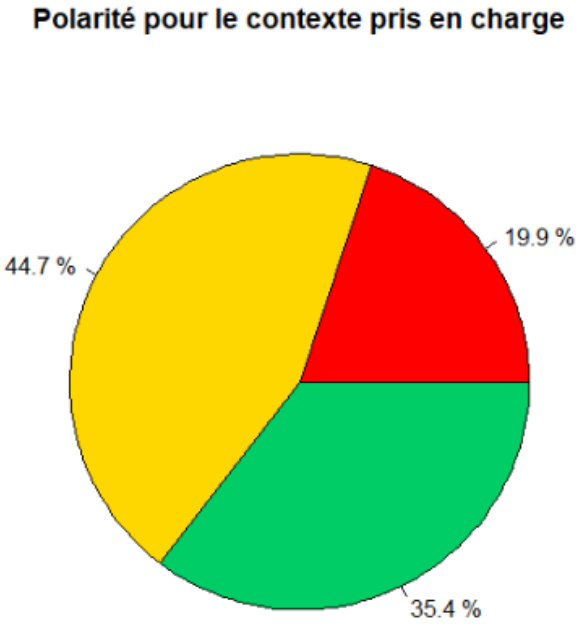

Figure 47 - Distribution de la Polarité pour le discours sur pris en charge sur la Défense dans le Corpus Français

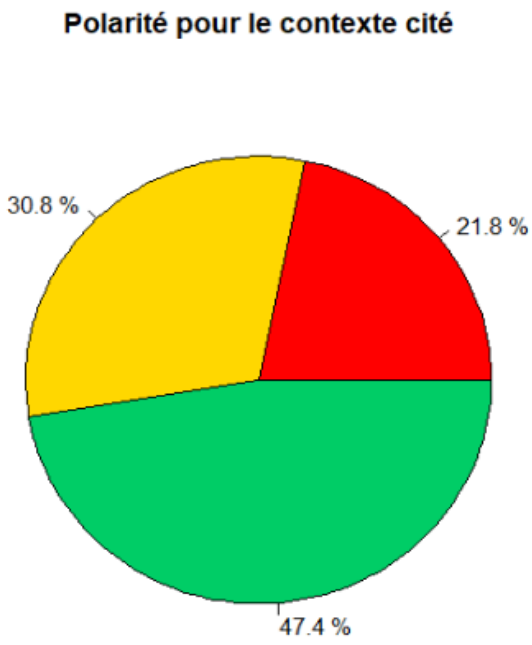

Figure 48 - Distribution de la Polarité pour le discours rapporté sur la Défense dans le Corpus Français

Si la proportion d'exemples *négatifs* est similaire (environ 20%) entre le contexte *cité* et le contexte *pris en charge*, pour le contexte *cité*, la proportion d'exemples *positifs* est quant à elle supérieure pour le contexte *cité*. 35% du discours concernant la Défense produit par le journaliste est positif, mais presque la moitié du discours qu'il cite est *positif*. Cela montre que dans notre corpus, les journalistes ont tendance à citer des discours d'individus plutôt favorables à la Défense **: Le discours que le journaliste rapporte est plus élogieux que celui qu'il tient.**

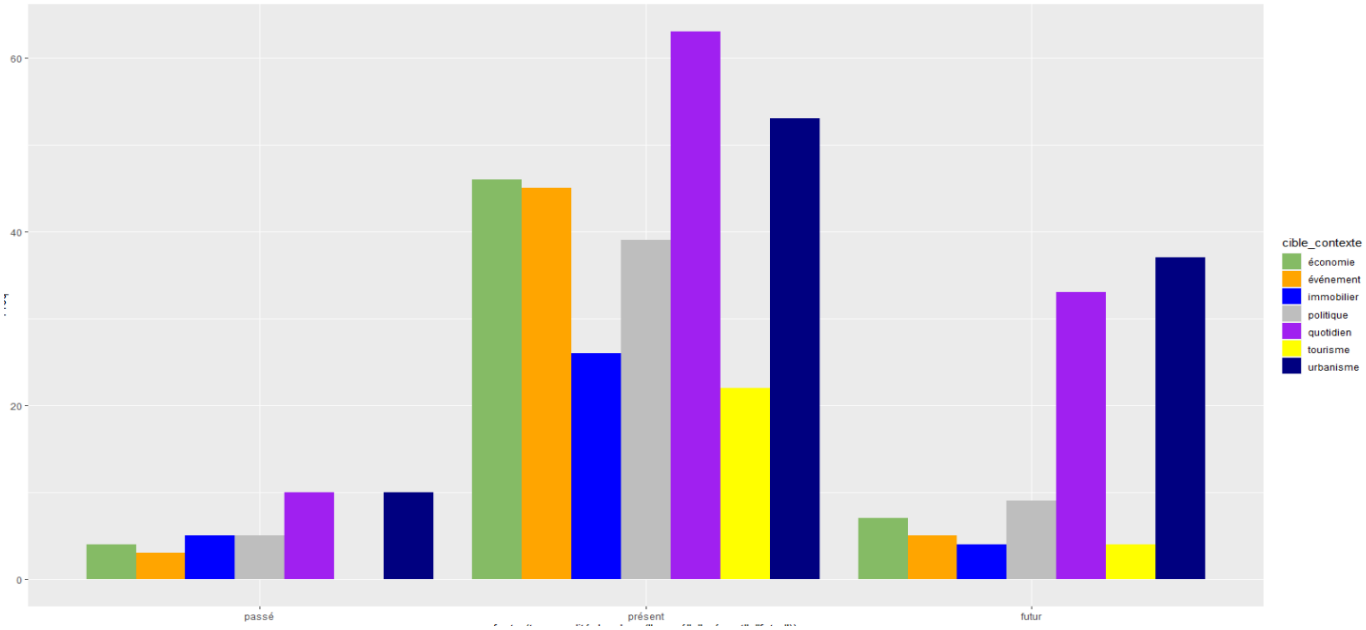

Figure 49 - Répartition des cibles du contexte en fonction de la temporalité dans le Corpus Français

La figure 49 met en lumière la répartition des cibles du contexte par rapport à la temporalité. Pour le *passé*, les thèmes les plus récurrents sont la *vie quotidienne* et l'*urbanisme*. Ces deux thèmes sont les mêmes que pour la répartition générale. En revanche l'*immobilier* et l'*administration politique* sont plus souvent abordés que l'*économie*. Le *passé* du *tourisme* n'est jamais abordé.

La distribution des cibles du contexte avec une temporalité au *présent* est très similaire à la distribution générale, les trois cibles les plus fréquentes étant la *vie quotidienne* puis l'*urbanisme* et l'*économie*. Au *futur* en revanche, les deux premières catégories restent les mêmes, mais elles présentent un écart bien plus grand avec les catégories suivantes. Pour le *futur* on ne parle presque que des aménagements à venir (*urbanisme*) et des projets qui vont impacter la *vie quotidienne* des habitants. L'économie n'est que quatrième de très loin, après l'*administration politique* qui elle aussi est liée à la gestion et les projets du quartier.

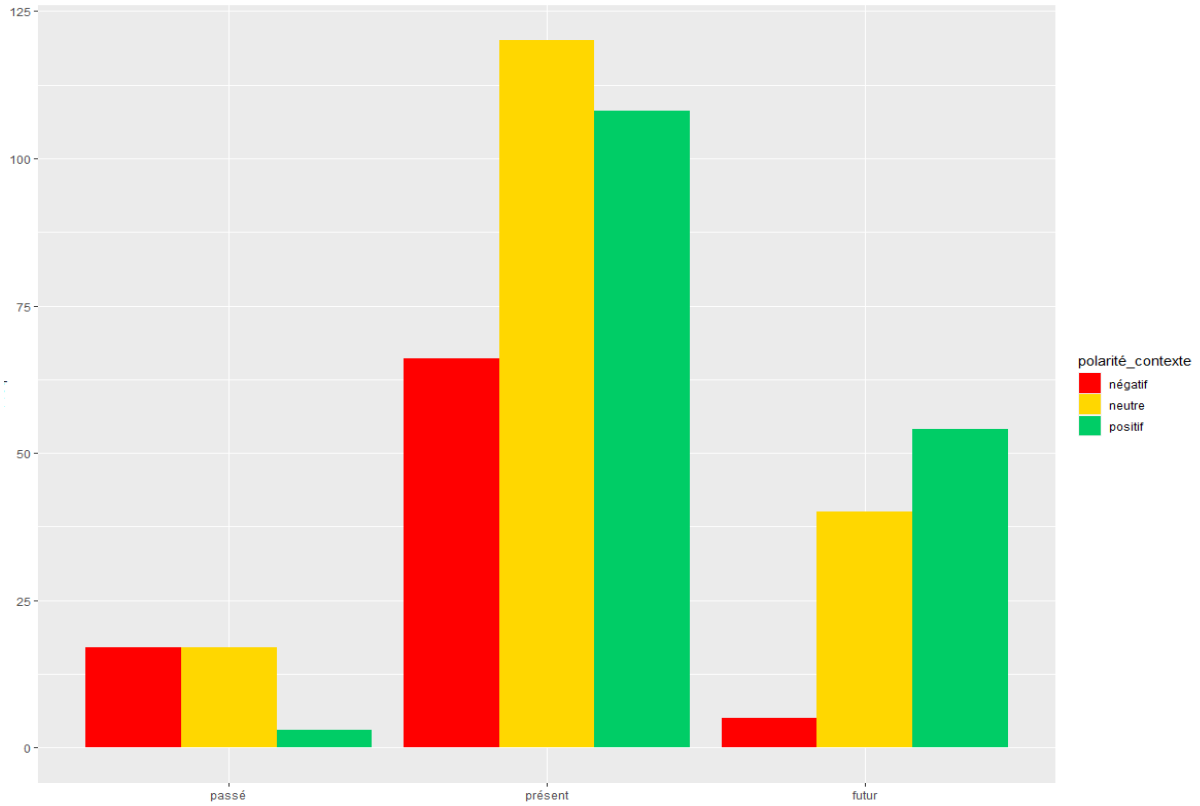

Figure 50 - - Répartition de la polarité du contexte en fonction de la temporalité dans le Corpus Français

La figure 50 représente la polarité du discours en fonction de la temporalité. Le *passé* compte très peu de phrases *positives* et autant de phrases *négatives* que de phrases *neutres*. Le *présent* comporte une majorité de phrases *neutres* et légèrement moins de phrases *positives* avec deux fois moins de phrases *négatives*. Le *futur* quant à lui compte très peu de discours *négatif* et une majorité de discours *positif*. **De ces résultats émerge de manière très nette la notion de progrès** : Le présent a une image plus méliorative que le passé, et le futur à son tour à une image plus positive que le présent.

10.3.1.3 Synthèse des résultats pour la presse française

L'analyse des résultats recueillis après l'annotation du corpus de presse en français constitue une étude à part entière et nous a renseignés sur la représentation du quartier d'affaire de la Défense et sur les signes et stratégies linguistiques employés pour parler du quartier.

Nous avons identifié la présence de deux normes dans les procédés de nommage, d'une part le *type de lieu* et d'autre part *la norme* (« La Défense »). Le procédé de nommage *classement* est toujours employé pour placer la Défense dans une position positive. Il est aussi très souvent associé à une *référence géographique* et attribue ainsi une polarité *positive* à la *localisation*. Les désignations portant sur l'*Architecture* sont quant à elles très souvent

En ce qui concerne le contexte, la cible qui est la plus abordée est la *vie quotidienne*, cela étant possiblement dû au nombre importants d'articles du Parisien. Celle-ci apparaît comme un sujet très polémique avec une part importante d'exemples positifs et négatifs. L'*économie* de la Défense est un sujet avec une polarité majoritairement *positive* même s'il n'est que le troisième sujet le plus évoqué contrairement à ce qu'on aurait pu attendre. Enfin La Défense est vue comme un lieu en progrès, avec un *futur* bien plus désirable que son *passé*.

10.3.2 Les résultats pour la presse anglaise

10.3.2.1 Analyse des Désignations

Nous avons utilisé la même méthode d'analyse avec les annotations obtenues sur le corpus anglais. Le nombre total de désignations est inférieur (465 pour l'anglais contre 1033 pour le français) mais nous estimons que la taille du corpus est suffisante pour que l'analyse comparative des fréquences relatives demeure pertinente.

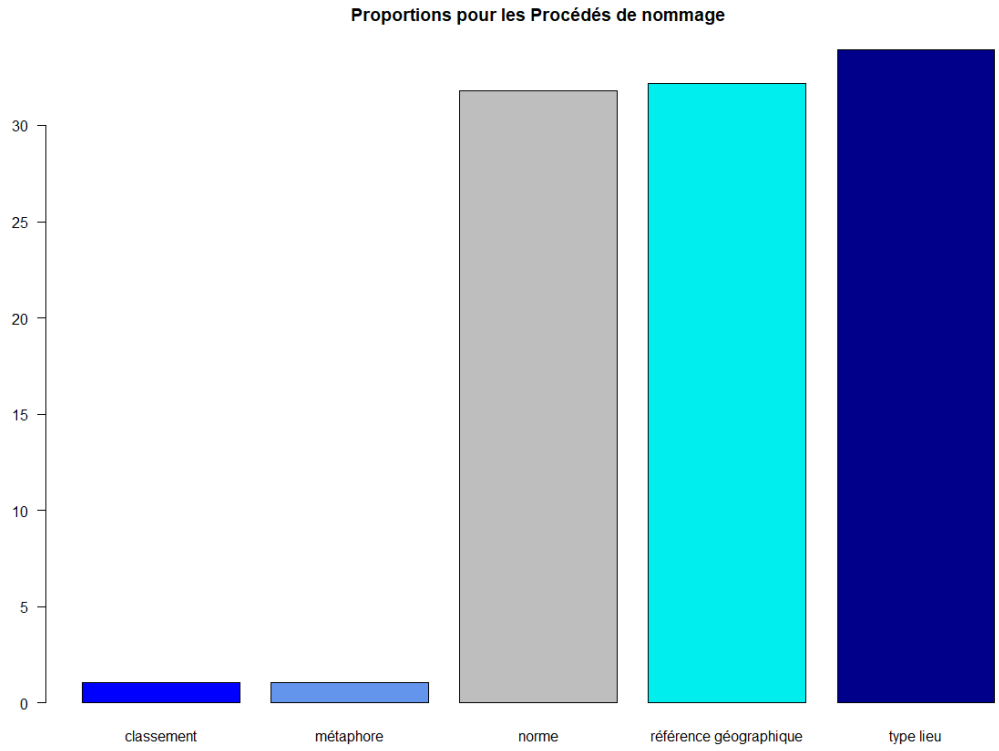

Figure 51 - Distribution des Procédés de Nommage dans le Corpus Anglais

Il existe trois manières de nommer majoritaires, chacune totalisant plus 30% des désignations : norme, référence géographique et type de lieu. Dans le corpus anglais il y a donc **trois manières de nommer différentes qui pourraient être considérées comme la norme**. Cela diffère du corpus français où seuls deux procédés se détachaient des autres.

La norme supplémentaire est *référence géographique*, cela fait sens d'un point de vue pragmatique puisque La Défense n'est pas un lieu familier pour les lecteurs anglophones, puisqu'elle n'est pas géographiquement proche d'eux. C'est pourquoi il y a une nécessité pour le journaliste de préciser la place géographique du quartier : du point de vue britannique, c'est un élément spécifique.

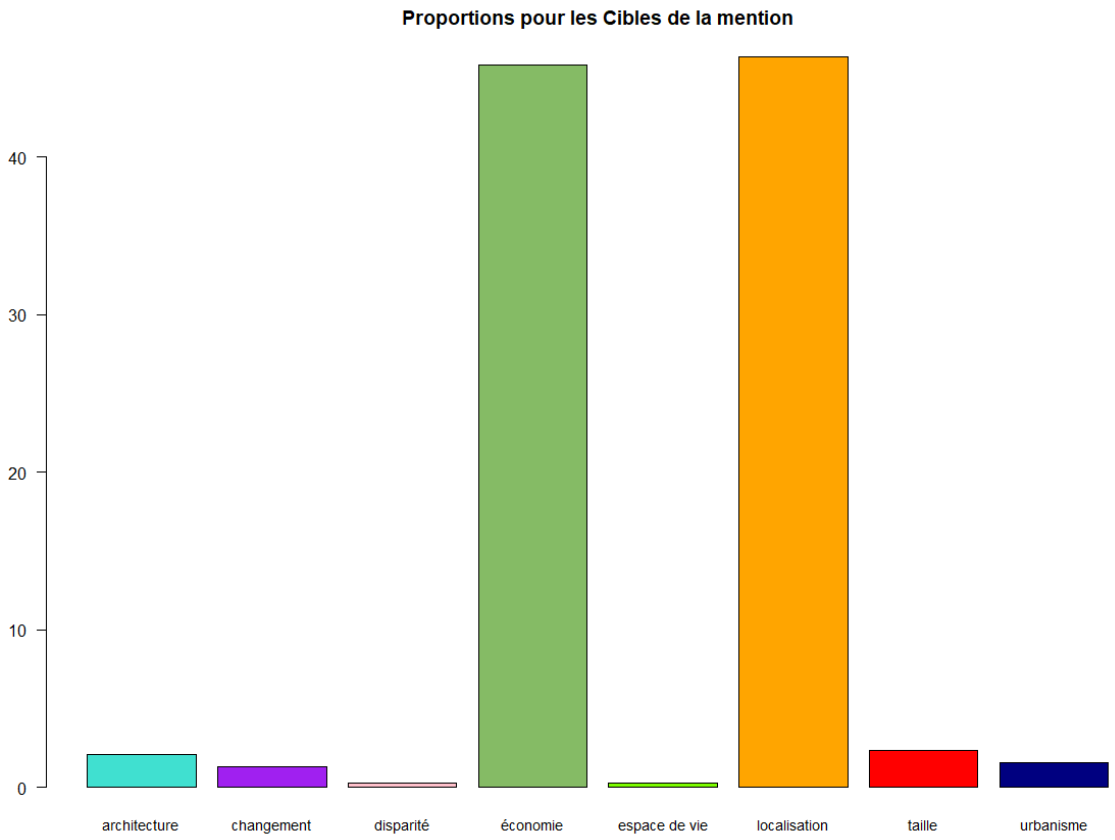

Figure 52 - Distribution des Cibles des Désignations dans le Corpus Anglais

Pour les deux cibles principales des désignations sont la *localisation* et *l'économie*. La *localisation* est omniprésente pour les mêmes raisons que nous venons d'évoquer à propos de la référence *géographique*. La grande symétrie tend à faire penser que les deux aspects vont de pair : Lorsque l'un est évoqué, c'est également le cas de l'autre.

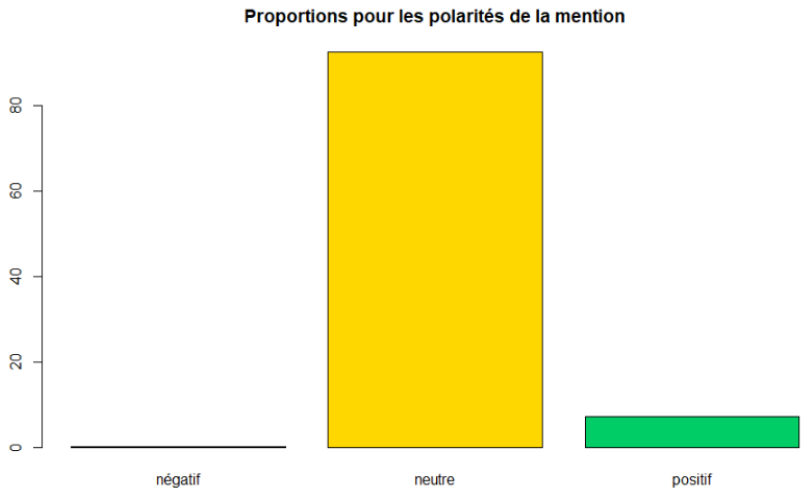

Figure 53 - Distribution de la polarité des désignations dans le Corpus Anglais

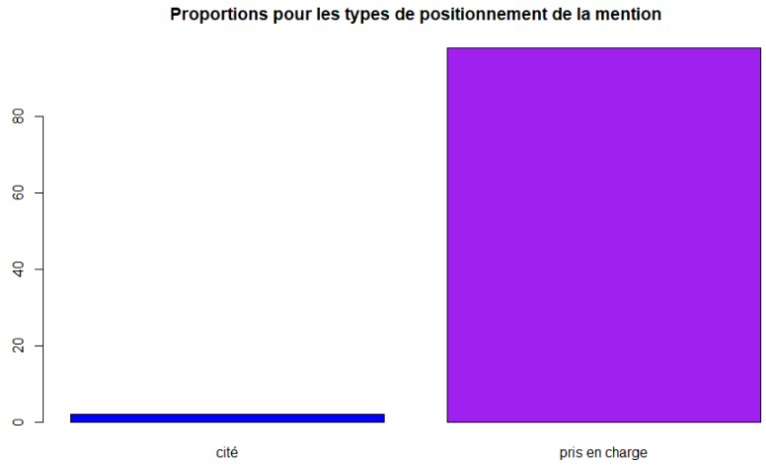

Figure 54 - Distribution du positionnement sur les désignations dans le Corpus Français

Les polarités pour la mention sont assez similaires à celles du corpus français avec une faible polarisation et une majorité de désignations *neutres*. Les auteurs prennent en charge leurs désignations dans la grande majorité des cas. Ces deux variables semblent plus ou moins constantes dans les deux langues, ce qui peut être le signe d'une manière similaire de répartir les signes de subjectivité dans les groupes nominaux.

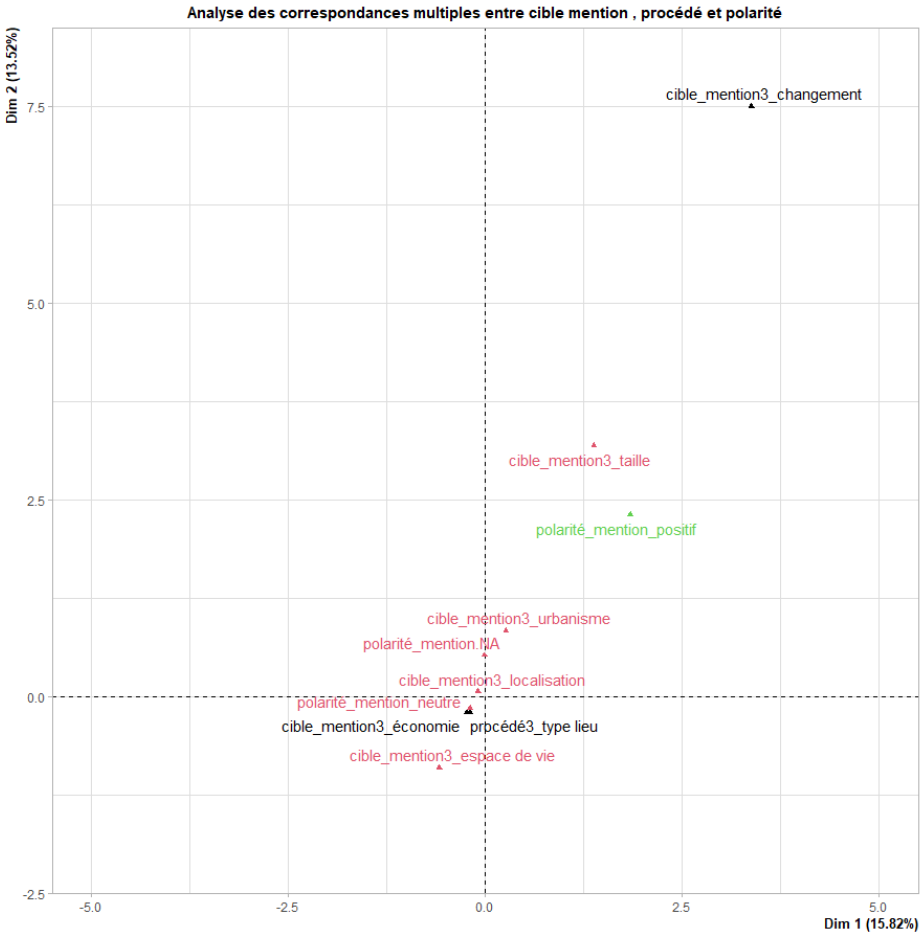

Figure 55 - Deux premières dimensions de l'Analyse des Correspondances Multiples entre les cibles des désignations, procédés de nommage et polarité des désignations dans le Corpus Anglais

L'Analyse des correspondances multiples nous permet d'identifier les points d'intérêts. La figure 55 qui représente 15% (dimension1) et 13% (dimension2) de la variance montre par exemple que taille est souvent associé à une polarité *positive* ou que la polarité *neutre* est très proche de la cible *économie*.

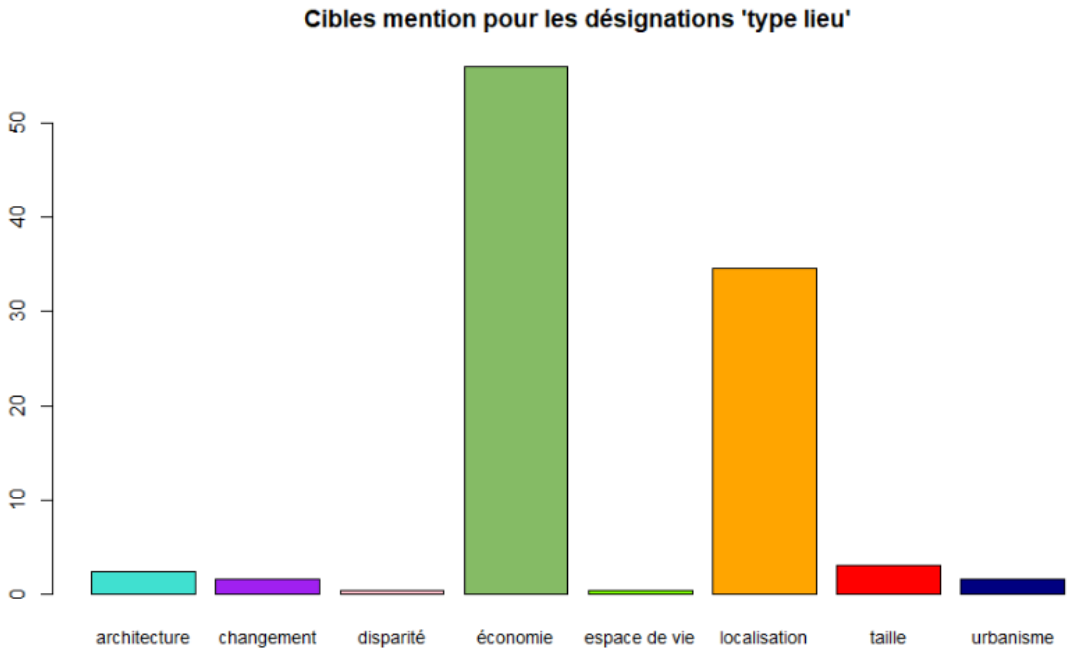

Figure 56 - Distribution des cibles des désignations pour le procédé de nommage "type de lieu" dans le Corpus Anglais

Les cibles principales des procédés *type de lieu* sont l'*économie* (56%) et la *localisation* (37%). Cela tend à nuancer notre hypothèse selon laquelle le *type de lieu* « buisness district » et la *localisation* vont de pair. La *localisation* reste une cible privilégiée pour les désignations en anglais (36%) par rapport aux désignations en français (6%).

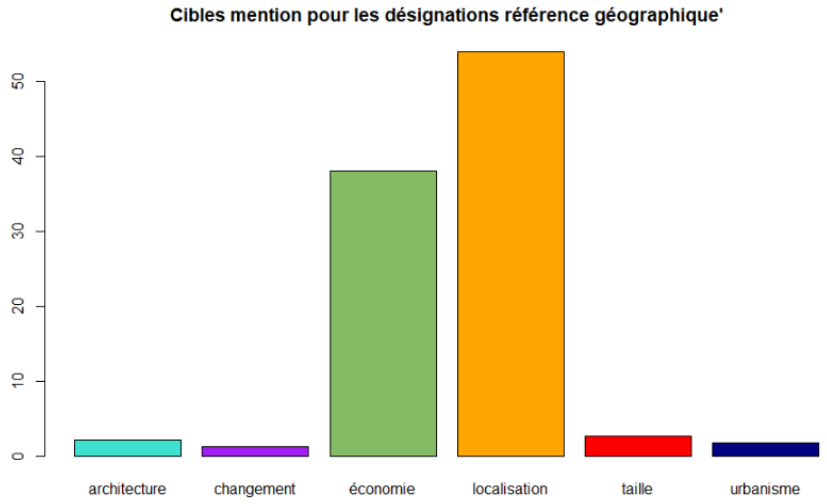

Figure 57 - Distribution des cibles des désignations pour le procédé de nommage "référence géographique" dans le Corpus Anglais

Les cibles des désignations employant des *références géographiques* présentent les pourcentages inverses de celles employant des *types de lieu* : *localisation* est la cible la plus courante (vers 56%) et *économie* la deuxième plus courante (38%). Cela explique les

pourcentages manquants pour arriver à des proportions identiques entre ces deux cibles puisqu'en additionnant les deux (en sachant que la fréquence relative des deux est aussi presque la même) nous obtenons des chiffres très proches.

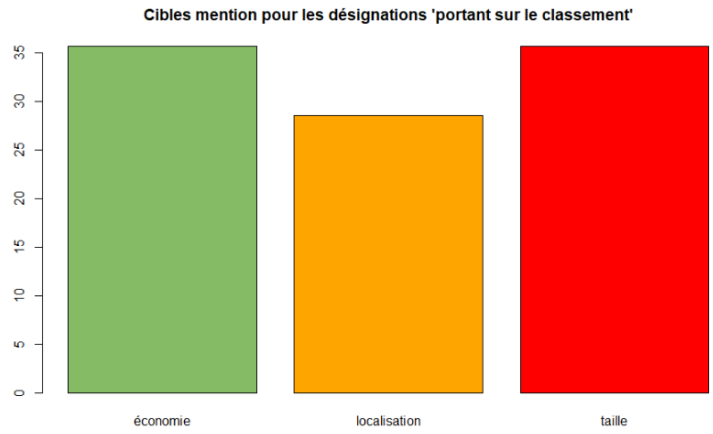

Figure 58 - Distribution des cibles désignations pour le procédé de nommage "classement" dans le Corpus Anglais

Les désignations utilisant le classement ne ciblent que trois aspects de La Défense : D'abord l'*économie* et la *localisation* comme dans le corpus français. En revanche la cible *espace de vie* présente en français est ici totalement absente. La *taille* représente 35% des cibles de désignation utilisant un *classement*, ce qui est bien plus élevé que pour le corpus français (4%). Les journalistes britanniques utilisent bien plus de superlatifs mettant l'accent sur la grandeur de l'espace que couvre la Défense que les journalistes français.

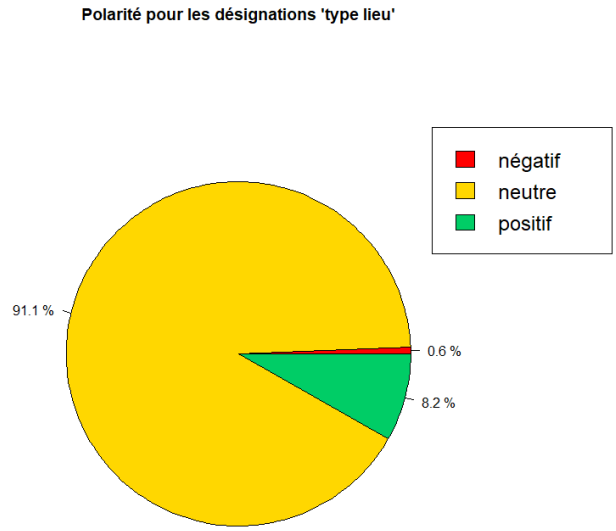

Figure 60 - Distribution de la polarité des désignations pour le procédé de nommage "type lieu" dans le Corpus Anglais

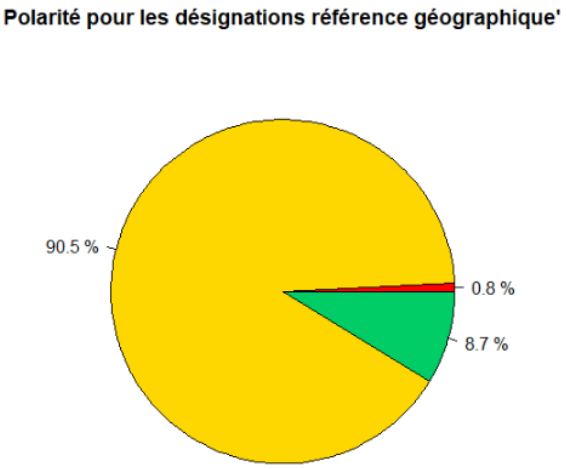

Figure 59 - Distribution de la Polarité pour le procédé de nommage "référence géographique" dans le Corpus Anglais

Les désignations avec un procédé *type de lieu* ont des proportions similaires à celles du corpus français en termes de polarité. **La distribution de la polarité pour *références géographiques* est presque identique à celle de *type de lieu* ce qui recoupe l'idée de symétrie entre ces deux normes** que nous avons évoqué plus haut. Là ou pour le français les normes étaient « La Défense » et « quartier d'affaires », ici les normes sont « La Défense » et le *type de lieu* « business district » associé à une *référence géographique* (typiquement « Paris »).

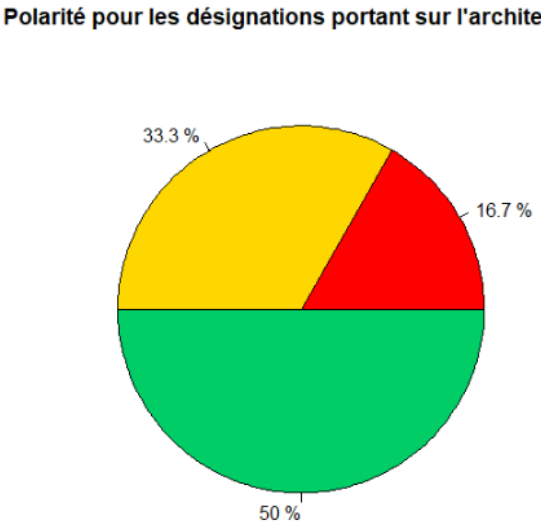

Figure 62 - Distribution de la Polarité portant sur l'architecture dans le Corpus Anglais

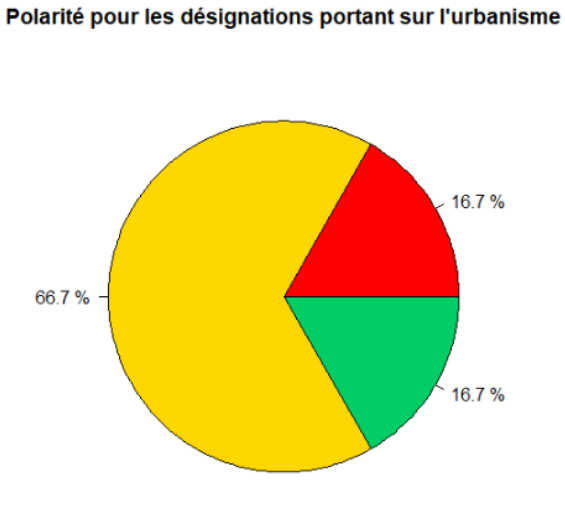

Figure 61 - Distribution de la Polarité portant sur l'urbanisme dans le Corpus Anglais

L'*Architecture* est plus mise en avant dans les désignations en anglais mais elle présente aussi une polarité nettement plus positive. Les références aux gratte-ciels et autres structures marquant l'horizon de la Défense sont pour la moitié d'entre elles mélioratives (seulement 5% de positif en français et plus de négatif). L'aspect *urbain* de la Défense est caractérisé par une vision bien plus nuancée avec autant que *négatif* que de *positif*. Cela contraste avec la vision du corpus français qui avait une image plutôt *négative* de l'*architecture* et ne comportait pas une aussi grande différence entre ces deux aspects de la Défense qui sont connexes à première vue.

10.3.2.2 Analyse du Contexte

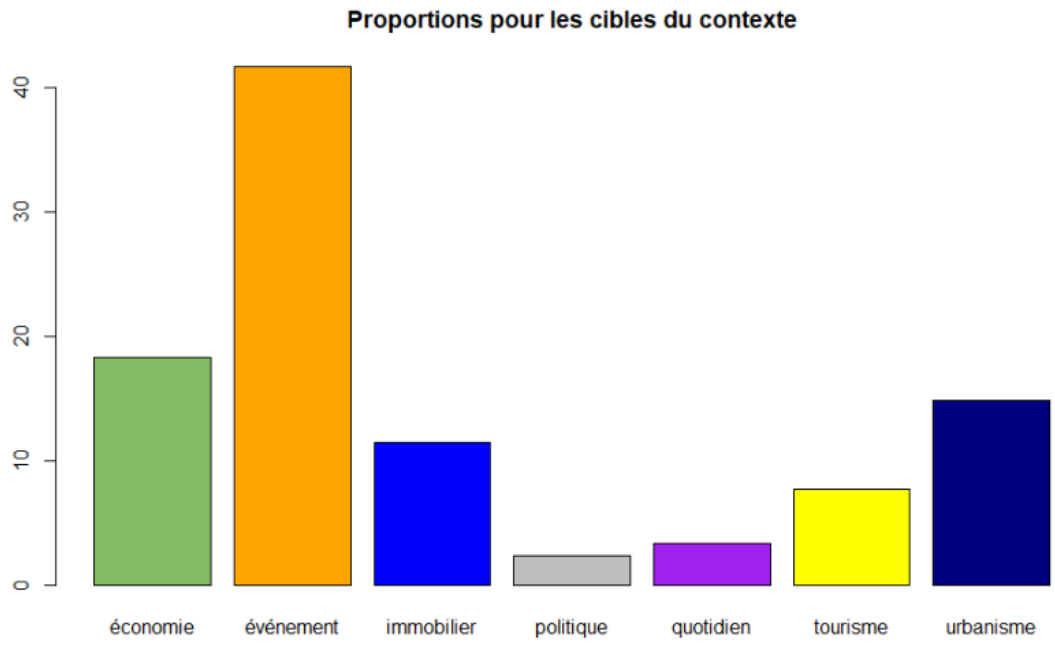

Figure 63 - Distribution des cibles du Contexte de la Défense dans le Corpus Anglais

Contrairement au corpus français qui parlait plus de de la *vie quotidienne* et du caractère *urbain* de la Défense, **40% de contenu mentionnant la Défense porte sur les *événements* y ayant lieu**. C'est très surprenant au vu de la répartition des journaux du corpus : le corpus anglais possède une plus grande partie d'articles de journaux économiques. Nous nous attendions donc à ce que l'*économie* soit le sujet le plus présent. **La Défense semble donc être évoqué majoritairement dans la presse anglaise dans le cadre d'évènements qui s'y déroulent ou s'y sont déroulés.**

Les troisièmes et quatrièmes catégories les plus fréquentes sont l'*urbanisme* (18%) et l'*immobilier* (13%). Le premier était déjà présent pour le français en revanche l'immobilier était une catégorie mineure (7%) en français, cela indique que les lecteurs britanniques sont plus intéressés par les évolutions et opportunités au niveau des locations et prix des immeubles de la Défense que les français. L'aspect *vie quotidienne* est totalement absent alors qu'il était omniprésent dans les articles français : cela est attendu puisqu'une partie considérable des articles en français sont issus de journaux locaux.

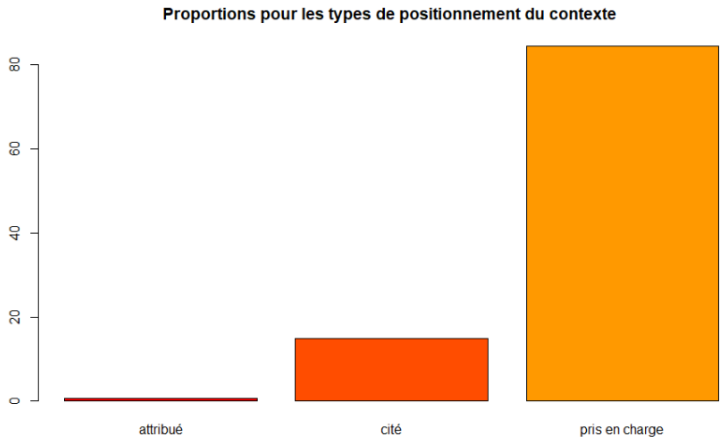

Figure 64 - Distribution du positionnement sur le contexte de la Défense dans le Corpus Anglais

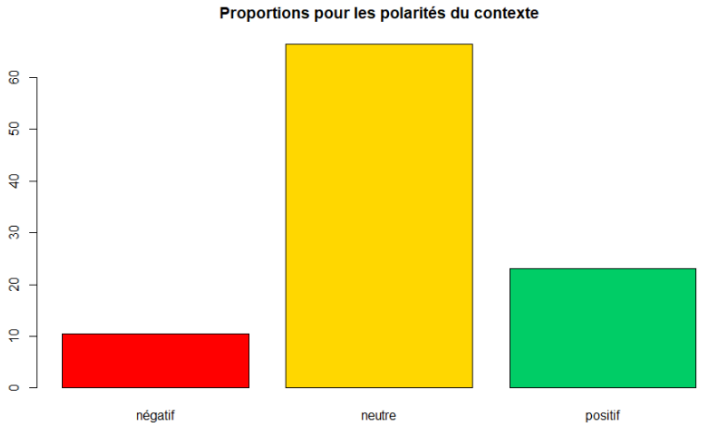

Figure 65 - Distribution de la polarité du Contexte de la Défense dans le Corpus Anglais

La distribution globale similaire pour le positionnement témoigne de la similarité dans l'utilisation du discours rapporté dans les deux langues**. La polarité est bien plus neutre dans le corpus britannique que dans le corpus français** (65% contre seulement 40%). Les proportions relatives de négatif et positif sont similaires.

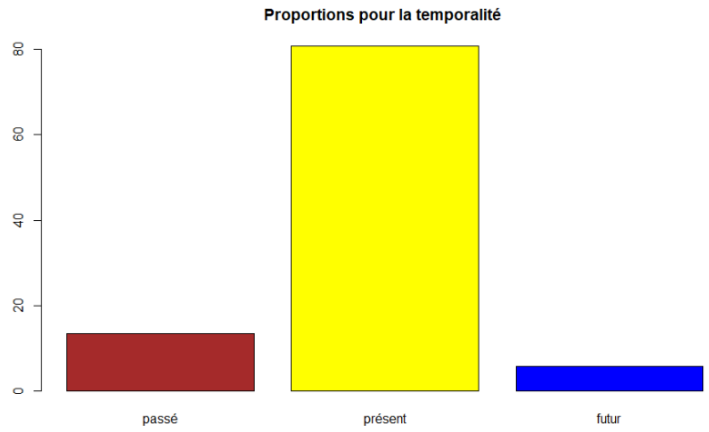

Figure 66 - - Distribution de la temporalité associée au contexte de la Défense dans le Corpus Anglais

Enfin, Si le présent est toujours l'axe temporel dont on parle le plus, nous remarquons que le passé et le futur sont inversés par rapport au corpus français. Dans le corpus anglais, le futur de la Défense est très peu abordé alors que son passé est évoqué presque 15% du temps. Cela indique que **les journaux anglais ont une approche plus rétrospective du quartier et tendent plus à citer les événements de son passé qu'à se projeter dans son avenir**.

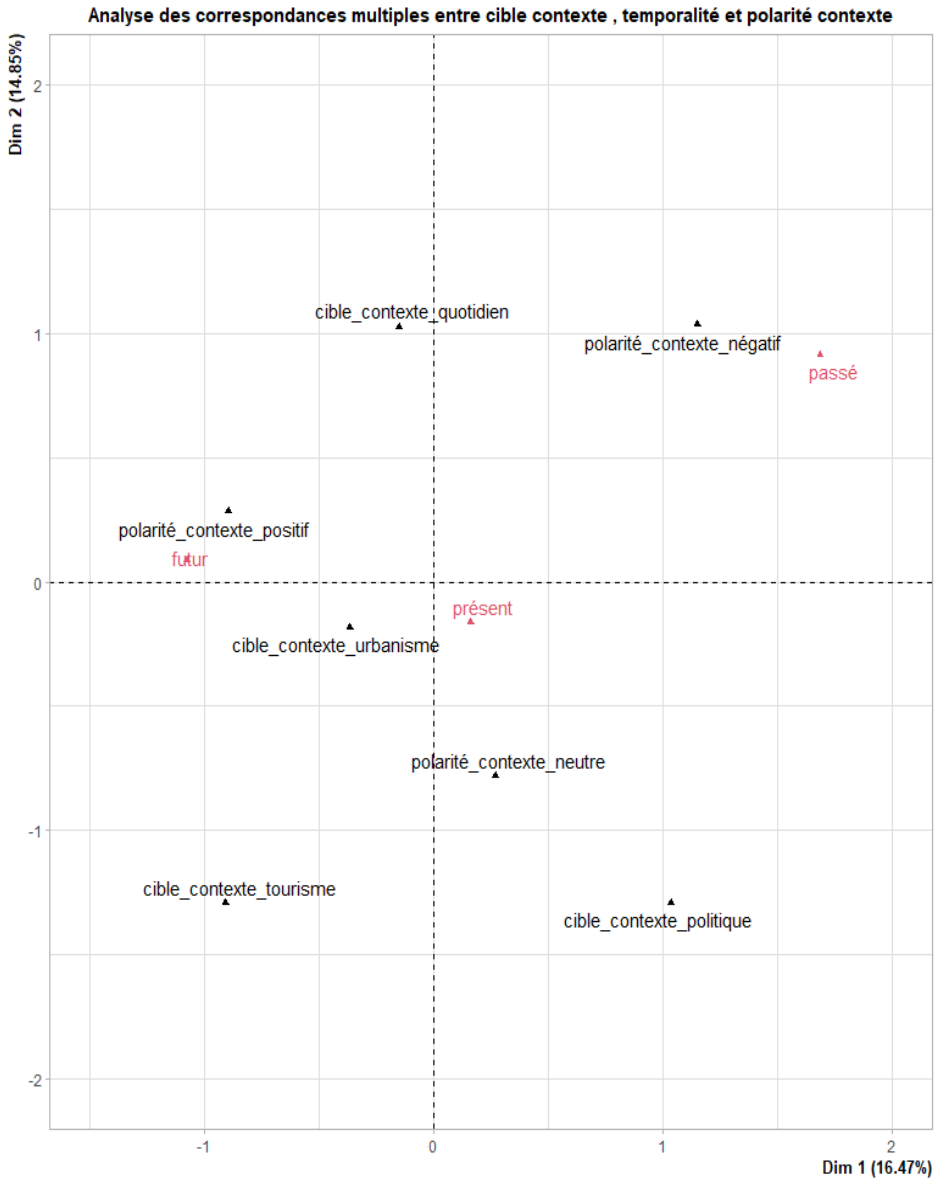

Figure 67 - Deux premières dimensions de l'Analyse des correspondances multiples entre la cible du contexte, la temporalité et la polarité du contexte dans le corpus Anglais

La figure 67 présente 2 des dimensions explorées. Nous constatons par exemple via la dimension 1 que la polarité *positive, le tourisme* et le f*utur* sont sur le même axe. La dimension 2 oppose d'un côté la *vie quotidienne*, la polarité *négative* et le *passé* et de l'autre *l'administration politique* et le *tourisme*.

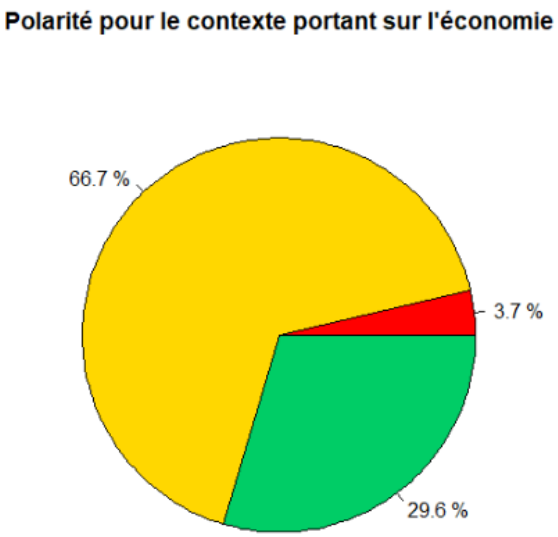

Figure 69 - Distribution de la Polarité pour le Contexte Portant sur l'Economie dans le Corpus Anglais

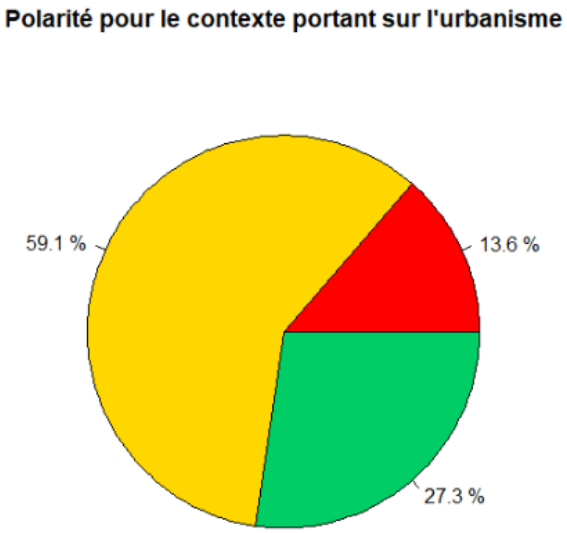

Figure 68 - Distribution de la Polarité pour le Contexte Portant sur l'Urbanisme dans le Corpus Anglais

Avec une polarité *neutre* de 66% *en* anglais contre 21% en français, L'aspect économique dans la presse anglaise est bien moins polarisé que pour la presse française. L'image p*ositive* (29% en anglais contre 65% dans le corpus français) associée à l'économie semble bien plus ténue pour laisser place à un regard plus *neutre* sur l'aspect financier de la Défense. Il faut toutefois noter que l'image négative est aussi plus faible (3.7 contre 21%) ce qui permet de conserver une image plus *positive* qu'elle n'est *négative*. La très forte neutralité à ce sujet est possiblement due à un parti pris des journalistes en rapport avec leurs lecteurs ou leur propre origine. Une autre explication possible est la présence plus importante de journaux de type économiques (*the economist* notamment) ou les articles sont écrit par des personnes qui maîtrisent supposément mieux le sujet est adoptent un regard plus objectif.

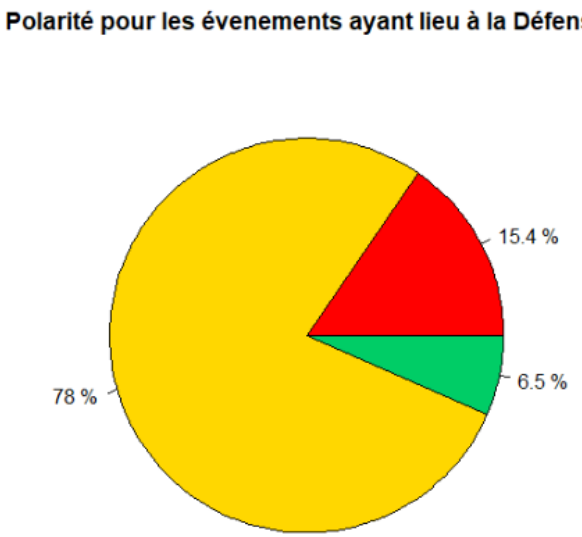

Figure 70 - Distribution de la Polarité pour le Contexte Portant sur les évènements ayant lieu à la Défense dans le Corpus Anglais

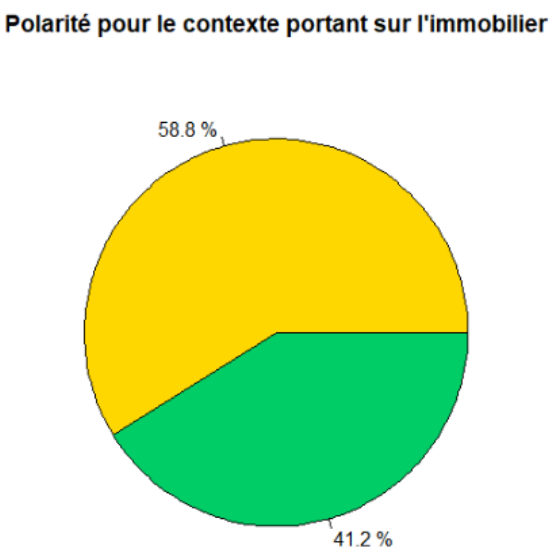

Figure 71 - Distribution de la Polarité pour le Contexte Portant sur l'Immobilier dans le Corpus Anglais

L'urbanisme est lui aussi caractérisé par une plus grande *neutralité* (59% VS 47%) . Le contexte négatif reste similaire mais le contexte positif est notablement moins présent (27% contre 38%), il s'efface donc face au contexte neutre.

Les journalistes font mention d'*évènements* en majorité *neutre* à la Défense. Comme dans le corpus français, la fréquence d'évènements *négatifs* est plus élevée que la fréquence d'évènements *positifs* ; en revanche ces deux fréquences sont plus basses de manière globale, pointant là aussi une plus grande neutralité du côté du corpus britannique.

Les phrases portant sur l'*immobilier* ne présentent aucune polarité *négative* (contre 17% de négatif en français). Elles sont pour la plupart *neutres* mais ont aussi une très forte portion de *positif*. Encore une fois, ce sujet qui semblait presque absent du corpus français est ici non seulement bien présent mais aussi associé à une image méliorative. **L'attractivité des locaux de la Défense semble donc au cœur de la représentation du quartier dans la presse britannique.**

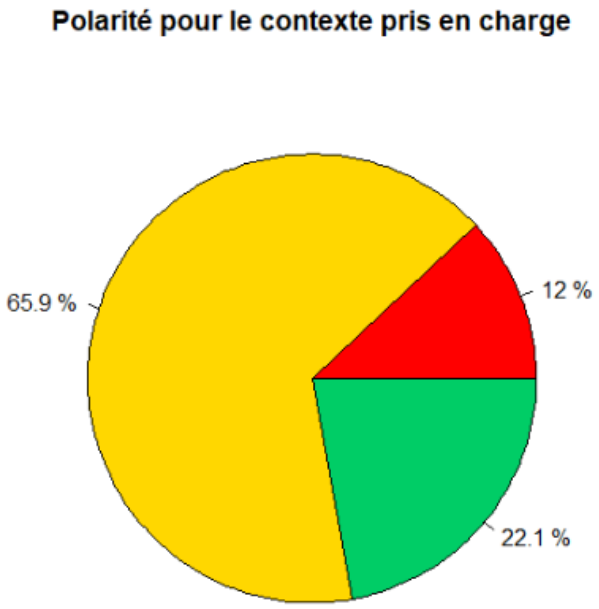

Figure 73 - Distribution de la Polarité pour le discours pris en charge sur la Défense dans le Corpus Anglais

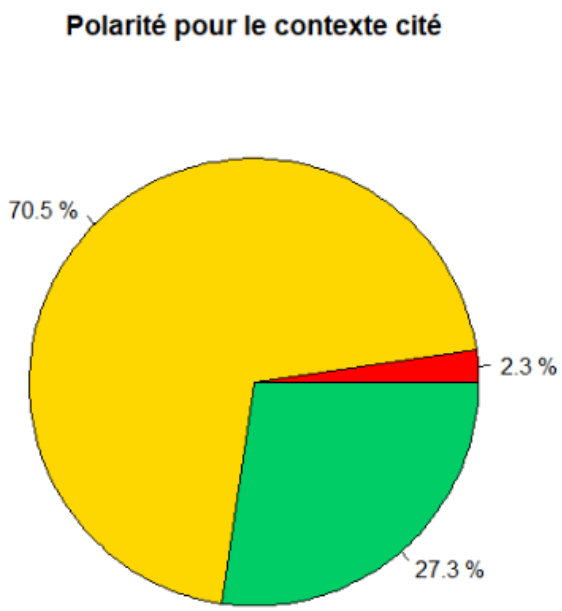

Figure 72 - Distribution de la Polarité pour le discours rapporté sur la Défense dans le Corpus Anglais

Contrairement aux résultats pour le positionnement des désignations, nous constatons une différence entre le corpus français et le corpus anglais. Si la distribution du contexte *pris en charge* n'est est peu éloigné (mais qui penche plus vers le neutre) de celle du français, le discours *cité* compte seulement 2.3% de négatif en anglais contre 21% en français. Le discours *cité positif* présente quant à lui 27% de positif pour l'anglais contre 47% pour le français :

- D'une part les avis négatifs *cités* sont presque absents des articles (possiblement dû à une sélection plus stricte des propos cités).
- D'autre part, il n'y a pas de différence significative entre la proportion de discours positif pris en charge et de discours positif cité.

Cela signifie que **les journaux anglais rapportent à la fois moins de discours négatif sur la Défense que le corpus français** et qu'**ils produisent plus de discours négatif qu'ils n'en rapportent**.

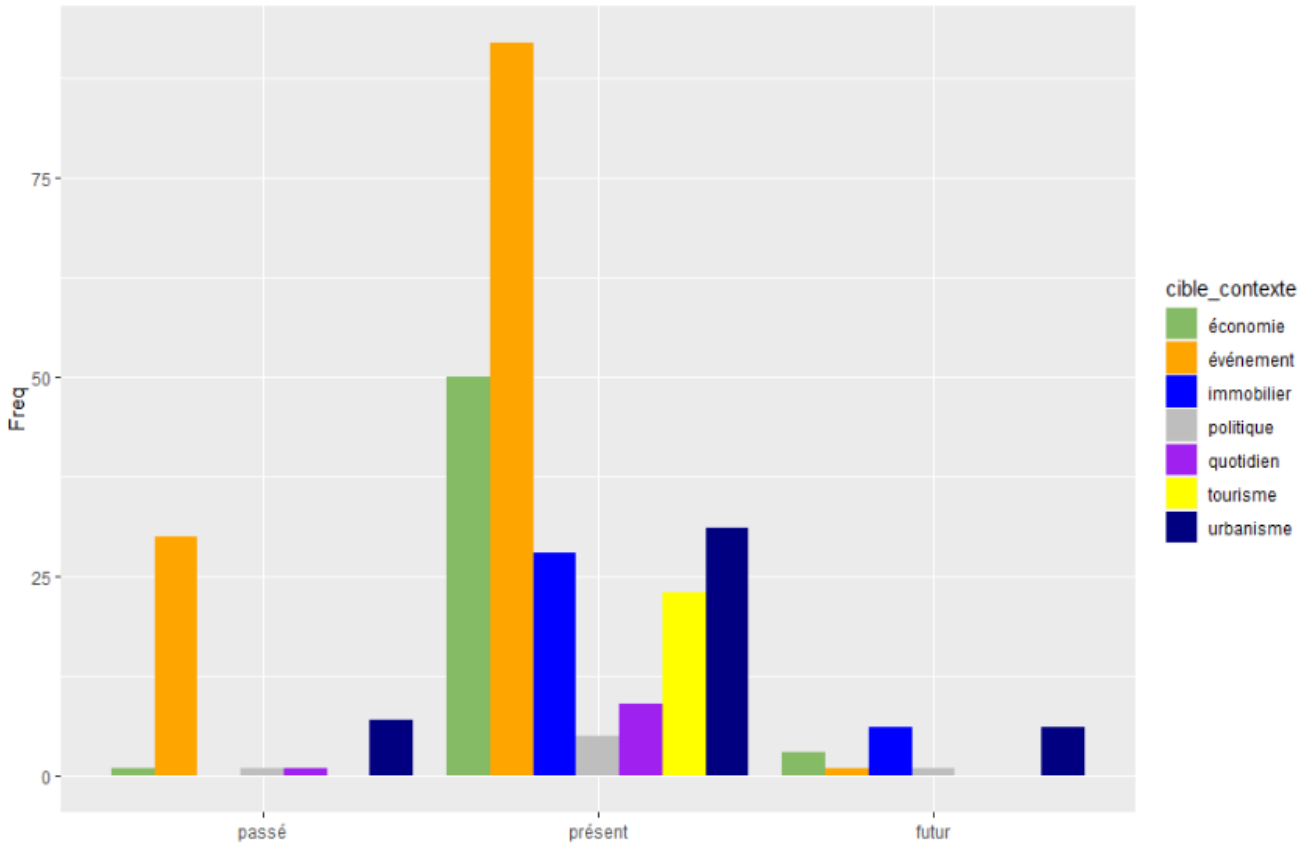

Figure 74 - Répartition des cibles du contexte en fonction de la temporalité dans le Corpus Anglais

L'aspect économique de la Défense n'est mentionné que dans une perspective de *présent*. Il est totalement absent dans la temporalité *passée* et négligeable dans le *futur.* Le tourisme également est un sujet dont on évoque exclusivement le présent ; pourtant nous attendions des projections futures à ce sujet. Le futur est surtout éclairé par l'aspect *immobilier* (positivement comme nous l'avons déjà évoqué) et par l'*urbanisme*, soit des projets d'aménagement du quartier, ceux-ci sont d'ailleurs souvent liés à l'*immobilier*.

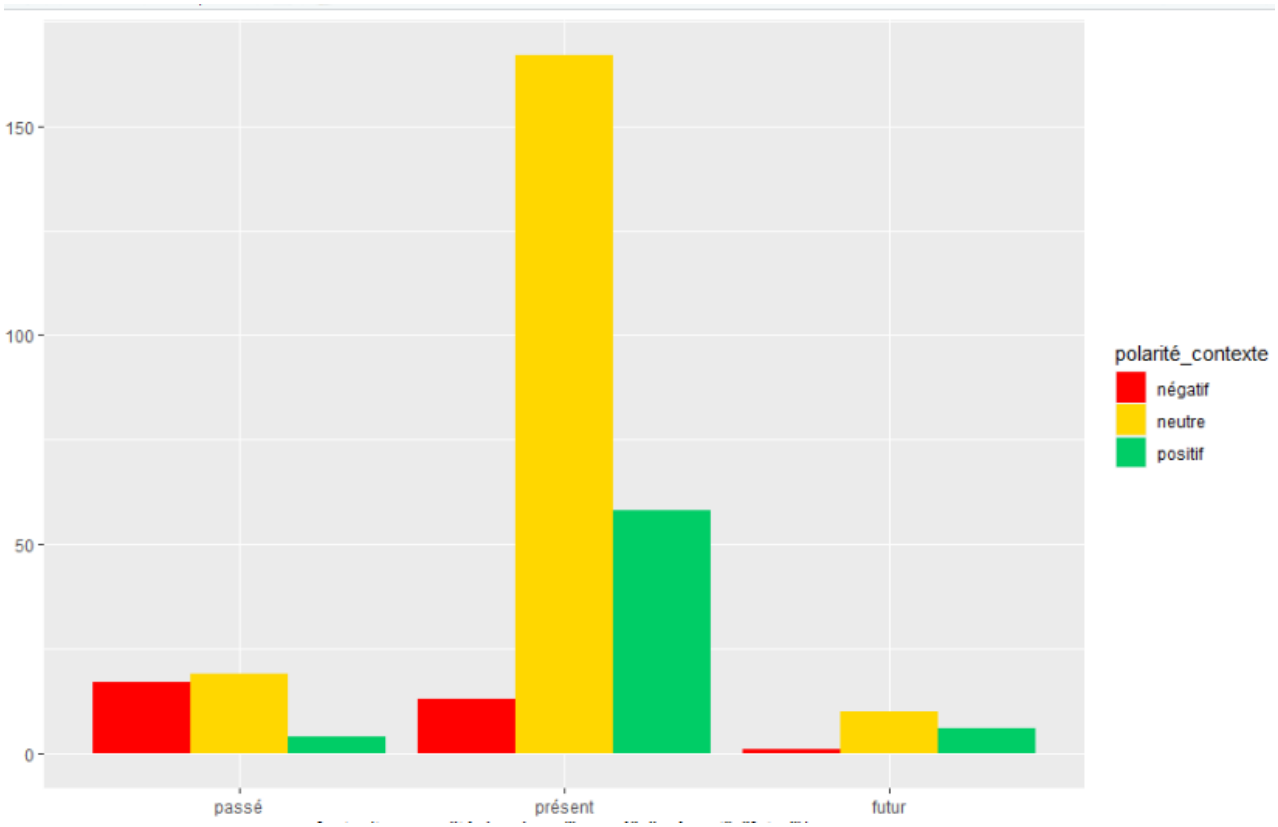

Figure 75 - Répartition de la polarité du contexte en fonction de la temporalité dans le Corpus Anglais

Comme dans le corpus en français la polarité *négative* diminue avec le temps. Cependant les fréquences relatives du neutre et du positif ne varient pas beaucoup dans le temps. Le *neutre* reste majoritaire pour les trois temporalités et la fréquence relative de la polarité *positive* s'accroit un peu pour le présent mais pas de manière significative pour le futur. **Le progrès n'est donc plus très net contrairement au corpus précédent corpus ; il s'agit ici d'un progrès nuancé qui s'arrête au présent et s'élance timidement vers le futur.**

10.3.2.3 Synthèse des résultats pour le corpus anglais

Les désignations de la Défense en anglais se caractérisent par la présence d'une triple norme. En plus de « La Défense », les journalistes britanniques usent la plupart du temps du temps des procédés *type de lieu* et *référence géographique*, le plus souvent en conjugaison. La *référence géographique* dénote une nécessité de rappeler au lecteur la *localisation* du quartier.

Parmi les références à La Défense, Le sujet de prédilection de la presse anglaise semble être les *évènements* qui se déroulent dans le quartier d'affaires parisien. L'*économie*, l'*urbanisme* et l'*immobilier* sont aussi des thèmes qui reviennent très régulièrement à propos du quartier. Ce dernier est d'ailleurs caractérisé par une polarité fondamentalement positive. Cela va de pair avec la polarité très positive désignations indexant l'architecture. L'Architecture et l'immobilier représentent donc un atout majeur pour le quartier d'affaires dans le portrait que s'en font les journalistes britanniques. Ce thème est d'ailleurs également associé au futur de la Défense, ce qui laisse une idée globale de promotion des espaces louables du quartier.

10.3.3 Synthèse Etude comparative : français vs anglais

Etudier un corpus d'articles en français et un corpus d'articles en anglais en utilisant les mêmes méthodes d'analyse nous a permis de mettre en évidence des similarités et des différences qui nous renseignent sur les ressemblances et dissemblances dans la représentation du quartier de la Défense dans la presse Britannique et dans la Presse française. Celles-ci sont valables uniquement pour ce corpus et nous demeurons prudents en ne généralisant pas à outrance et en gardant à l'esprit les possibles biais dû aux choix des journaux, de leur type, ou encore des périodes considérées.

Le positionnement et la polarité présents dans les mentions sont globalement très proches et nous signalent une répartition des marques de subjectivité dans les groupes nominaux référant à La Défense très proche entre les deux langues. Il nous faudrait étudier celles-ci pour des référents plus divers pour voir s'il s'agit là d'une particularité de la manière de nommer la Défense ou bien si ce sont des propriétés relatives aux deux langues. Dans les deux cas la manière de nommer est surtout *neutre* mais exprime plus volontiers des connotations *positives* que des connotations *négatives*.

Alors que le corpus français présente deux manières de nommer dominantes : Le nom « officiel » du quartier (« La Défense ») et le *type de lieu* (quartier d'affaires), le corpus britannique présente quant à lui une troisième manière dominante de nommer : La *Référence géographique*. **Si le corpus français possède une double norme, le corpus anglais en possède une triple.** Une analyse plus poussée nous a montré que dans le corpus britannique la norme « type de lieu » et la norme « référence géographique » se combinent la plupart du temps, de sorte que ces deux normes pourraient presque être considéré comme **une norme en soit : le type de lieu avec référence géographique** (ex « Parisian Buisness district »). Cette combinaison n'a toutefois pas toujours lieu. La prévalence de la *référence géographique* dans le corpus anglais et pas dans le corpus français est signe de la nécessité de pointer la localisation lorsqu'elle n'est pas familière.

Le corpus anglais est en général plus *neutre* dans son discours que le corpus français qui est bien plus polarisé. Cette tendance vers la neutralité est aussi présente à plus petite échelle pour la plupart des catégories, excepté quelques-unes (comme *immobilier*). Il faudrait là aussi regarder d'autres corpora, et en ciblant des éléments différents pour vérifier s'l s'agit là d'une spécificité du discours journalistique sur la Défense, d'une particularité du discours journaliste anglais par rapport au français ou bien encore une caractéristique des deux langues.

Le corpus français parle le plus de la *vie quotidienne* du quartier, ce qui est explicable par la large part d'articles issus du Parisien. **Le corpus anglais parle plus des évènements ayant lieu à la Défense** alors que l'aspect économique de la Défense, que nous nous pensions le plus fréquent n'est que deuxième plus fréquent dans le corpus anglais et troisième plus fréquent dans le corpus français, dans les deux cas assez loin de la cible principale. Le corpus français semble donc avant tout mentionner la Défense dans le cadre de problématiques locales, comprenant la vie quotidienne et l'aménagement du quartier tandis que le corpus anglais mentionne avant tout la Défense dans le cadre des évènements et faits divers qui s'y déroulent.

Les désignations du corpus français liées à l'architecture sont caractérisées par une claire tendance dépréciative. A L'inverse ces désignations sont associées à une connotation favorable dans le corpus anglais. Ce contraste signale une représentation plus positive des bâtiments de la Défense dans la presse anglaise. De plus, le contexte révèle la même dynamique avec la cible *immobilier* : celle-ci est associée à une polarité très positive dans le corpus anglais qui n'est pas présente en français. Cette catégorie, en plus d'avoir une meilleure polarité est relativement bien plus présente dans le corpus anglais. L'*architecture* dans les désignations conjuguée à l'*Immobilier* dans le contexte nous pousse à conclure que **la presse anglaise accorde une place centrale aux thématiques liées à l'attractivité des locaux de la Défense, plus que la presse française**.

Une idée de progrès net se dégage de l'étude de la polarité liée à la temporalité du corpus français. En avançant chronologiquement, le *négatif* diminue et le *positif* augmente. **Lle corpus anglais trahit également une tendance vers le progrès mais c'est une tendance bien plus mitigée** à travers laquelle si le *présent* est bien meilleur que le *passé*, le *futur* n'est pas significativement meilleur que le présent. Cette différence peut s'expliquer par une plus grande considération pour le futur de la Défense par les journalistes français et les personnes qu'ils citent, ce qui a pour effet d'insister plus sur les projets positifs en rapport avec le quartier. A l'inverse les journalistes anglais, moins concernés par le futur du quartier évoqueraient moins le futur. Le corpus français évoque plus le *futur* que le *passé* alors que pour le corpus anglais c'est l'inverse, ce qui semble corroborer notre hypothèse.

10.4 Vers l'automatisation

10.4.1 Vectoriser les documents

Une fois l'analyse des désignations et de leur contexte complété, l'objectif suivant du projet fût d'automatiser l'annotation nécessaire pour l'analyse afin de rendre notre travail reproductible avec un corpus différent. Dans un premier temps, nous avons transformé les documents pour en extraire **une représentation vectorielle**. Nous avons d'abord opté pour deux représentations différentes :

- Une représentation de type *Bag of Words* avec une pondération *Tfidf* (*term frequency / inverse document frequency*) grâce à la bibliothèque *python scikit-learn*.
- Une représentation sémantique basée sur *Word2Vec* grâce à la bibliothèque *Gensim*. Celle-ci est basée sur un modèle pré-entraîné par Jean-Philippe Fauconnier (500 dimensions).

Chacune de ces représentations est appliquée séparément aux mentions et au contexte en vue d'appliquer un traitement différent par la suite à chacune des représentations. Le prétraitement et la vectorisation présentent quelques points communs :

- Un filtrage des mots-vide a été effectué dans les deux cas
- Dans le cas de la représentation du contexte, nous avons retiré les mots composants la mention avant la vectorisation. Cette étape permis de mimer le raisonnement de l'annotateur humain qui, comme le stipule la Convention d'Annotation, ne doit pas prendre en compte la mention lorsqu'il annote les catégories relatives au contexte.

Nous avons donc obtenus quatre représentations vectorielles pour chaque exemple :

- *Bag of Words* pour la désignation
- *Bag of Words* pour le contexte
- *Word2Vec* pour la désignation
- *Word2Vec* pour le contexte

10.4.2 Détecter le procédé de nommage

Suivant la méthodologie d'annotation, nous avons d'abord prédit le procédé de nommage. Il s'agissait d'une tâche de classification multi-étiquettes car chaque exemple pouvait prendre plusieurs étiquettes parmi plusieurs.

10.4.2.1 Détecter la norme

Nous nous sommes rapidement rendu compte après quelques essais sur les différents types à prédire que la catégorie *norme* posait problème. En effet, du fait des choix d'annotation, la *norme* à la particularité d'affecter les autres catégories : A chaque fois que le procédé de nommage est détecté comme « norme », toutes les autres catégories de la mention doivent impérativement être « NA ».

De là est née la **nécessité de détecter la norme séparément des autres procédés de nommage**, de manière à pouvoir extraire les exemples correspondant à la norme pour la suite. Pour cela, nous avons utilisé la méthode symbolique suivante :

- Transformation en minuscule de chaque mot de la mention
- Lemmatisation de chaque mention
- Filtrage des mots-vide
- Retrait des accents
- Test pour vérifier si le seul mot restant est « defense »

Ainsi **nous obtenons un rappel et une précision parfaits pour repérer la norme**.

	Précision	Rappel	F-Mesure	Support
norme	1.00	1.00	1.00	388
pas norme	1.00	1.00	1.00	505
accuracy			1.00	893
Macro Moyenne	1.00	1.00	1.00	893
Moyenne Pondérée	1.00	1.00	1.00	893

Figure 76 - Résultats pour la Détection de la Norme

10.4.2.2 Détecter les autres procédés

En étant assuré que la norme est isolée correctement, il ne restait que quatre procédés de nommage. **La détection des autres procédés était une tâche de classification multi-étiquettes** ou *multi-labels* car à chacun des exemples on associe une ou plusieurs étiquettes parmi les quatre possibles.

Etant donné que peu d'algorithmes supportent la classification multi-labels, nous nous sommes tournés vers le perceptron multicouche qui, lui, supporte nativement ce type de classification. Pour la suite, chacune des évaluations a été effectuée par validation croisée avec 5 plis.

	Précision	Rappel	F-Mesure	Support
classement	0.50	0.10	0.17	20
métaphore	0.00	0.00	0.00	17
Référence géographique	0.40	0.07	0.12	57
Type de lieu	0.94	1.00	0.96	411
Micro Moyenne	0.92	0.82	0.87	505
Macro Moyenne	0.46	0.29	0.29	505
Moyenne Pondérée	0.83	0.82	0.81	505
Moyenne échantillons	0.92	0.88	0.89	505

Figure 77 - Résultats pour la détection des procédés de nommage avec le Perceptron multicouche sur Bag of Words

La figure 77 présente des résultats très médiocres. Seule la classe majoritaire « type lieu » est bien reconnue. Les autres classes sont éclipsées par celles-ci.

	Précision	Rappel	F-Mesure	Support
classement	0.71	0.60	0.65	20
métaphore	0.50	0.29	0.37	17
Référence géographique	0.65	0.65	0.65	57
Type de lieu	0.99	0.99	0.98	411
Micro Moyenne	0.93	0.90	0.92	505
Macro Moyenne	0.71	0.63	0.66	505
Moyenne Pondérée	0.92	0.90	0.91	505
Moyenne échantillons	0.94	0.94	0.93	505

Figure 78 - Résultats pour la détection des procédés de nommage avec le Perceptron multicouche sur Word2Vec

Les résultats avec *Word2Vec* sont bien meilleurs. Nous pensions, contrairement aux résultats, que pour le procédé de nommage qui est une notion purement linguistique, une représentation se basant sur le lexique comme *Bag of Words* soit plus pertinente qu'une représentation plus sémantique comme *Word2Vec*. Il semble toutefois que la représentation sémantique de Word2Vec réussisse malgré tout à capturer les nuances linguistiques comme l'idée de *classement* grâce à son espace sémantique.

Cela s'explique probablement par le fait que de manière sous-jacente aux procédés choisis, il existe des contextes communs : les mots en rapport avec le classement apparaissent dans des contextes similaires, de même pour *référence géographique* ou *type de lieu*… En revanche le procédé *métaphore* (au-delà du fait d'avoir un plus faible nombre d'exemple, puisqu'il en a à peine moins que *classement*) est très mal reconnu. Cela confirme notre hypothèse car le procédé de *métaphore* peut apparaître dans des contextes bien plus variés que les trois autres : Il s'agit en effet d'une analogie sans mot de comparaison, cela laisse la possibilité d'un espace sémantique extrêmement vaste, il faudrait donc trouver une représentation qui donne des indices plus fins pour cette catégorie.

10.4.3 Détecter la cible de la mention

La cible de la mention est intimement lié aux champs sémantiques (puisqu'il s'agit de détecter le thème de la mention), c'est pourquoi cette fois on s'attend à avoir de biens meilleurs résultats avec *Word2Vec*. On utilise le même algorithme de classification multi-labels :

	Précision	Rappel	F-Mesure	Support
architecture	0.00	0.00	0.00	19
changement	0.00	0.00	0.00	6
disparité	0.00	0.00	0.00	4
espace de vie	0.00	0.00	0.00	14
localisation	0.20	0.02	0.03	55
taille	0.00	0.00	0.00	8
urbanisme	0.40	0.07	0.12	59
économie	0.77	0.98	0.86	336
Micro Moyenne	0.76	0.66	0.71	505
Macro Moyenne	0.17	0.13	0.13	505
Moyenne Pondérée	0.59	0.66	0.60	505
Moyenne échantillons	0.75	0.70	0.72	505

Figure 79 - Résultats pour la détection des cibles de la mention avec le Perceptron multicouche sur Bag of Words

	Précision	Rappel	F-Mesure	Support
architecture	1.00	0.58	0.73	19
changement	0.00	0.00	0.00	6
disparité	0.00	0.00	0.00	4
espace de vie	0.17	0.14	0.15	14
localisation	0.69	0.60	0.64	55
taille	0.62	0.62	0.62	8
urbanisme	0.81	0.85	0.83	59
économie	0.98	0.96	0.97	336
Micro Moyenne	0.90	0.85	0.87	505
Macro Moyenne	0.53	0.47	0.49	505
Moyenne Pondérée	0.88	0.85	0.86	505
Moyenne échantillons	0.91	0.89	0.89	505

Figure 80 - Résultats pour la détection des cibles de la mention avec le Perceptron multicouche sur Word2Vec

Cette tâche était plus difficile que la précédente car le nombre d'étiquettes possibles à prédire n'est plus de quatre mais de huit. D'une part, notre hypothèse a été confirmée puisque **la représentation sémantique (figure 80) a donné de bien meilleurs résultats que la représentation lexicale (figure 79),** cette dernière semblant encore une fois et décidément très sensible au déséquilibre du nombre d'exemple par classes.

Le problème de classes minoritaires demeure présent avec la représentation sémantique. Les classes *changement* et *disparité* sont totalement ignorées. Toutefois la classe *taille* qui n'a que quelques exemples de plus est assez bien reconnue. *Espace de vie* en revanche a plus d'exemples mais est très mal reconnu. Cela nous montre que les résultats ne sont pas totalement explicables par le nombre d'exemples.

De manière générale il semble que les étiquettes ayant un référent plus concret, c'est-à-dire qui sont plus facilement associées à des réalités sensibles comme *économie* ou *architecture* sont mieux reconnues que les étiquettes plus abstraite comme *changement* ou *espace de vie*.

10.4.4 Détecter la polarité de la mention

10.4.4.1 Méthode supervisée

La polarité de la mention, contrairement aux catégories précédentes ne prend qu'une seule et unique étiquette pour chaque exemple. Cela simplifie le problème et élargit aussi le champ des algorithmes possibles. Nous avons évalué plusieurs algorithmes dont les forêts aléatoires, les machines à vecteurs de support, le perceptron multicouche et la régression logistique. C'est cette dernière qui donne les meilleurs résultats, ceux-ci figurent sur la figure 81. La représentation utilisée est encore *word2Vec* qui donne de bien meilleures performances pour la polarité.

	Précision	Rappel	F-Mesure	Support
neutre	0.95	0.97	0.96	386
Négatif	0.00	0.00	0.00	7
Positif	0.64	0.63	0.64	46
Accuracy			0.92	439
Macro Moyenne	0.53	0.53	0.53	439
Moyenne Pondérée	0.90	0.92	0.91	439

Figure 81 - Résultats pour la détection de la polarité de la mention avec la régression logistique sur word2vec

Les performances pour chaque classe semble fortement corrélée au nombre d'exemples, ce qui signale encore un problème d'équilibre des classes. La polarité *négative*, avec seulement 7 exemples n'est pas détectée du tout et il est difficile de contrer ce phénomène. La reconnaissance du *positif* est décente mais pourrait être améliorée.

10.4.4.2 Méthode non-supervisée

En vue de remédier au problème que nous venons d'évoquer nous avons envisagé d'utiliser une méthode non-supervisée qui ne serait par conséquent pas influencée par le nombre d'exemples dans le corpus d'entraînement. La méthode suivie est schématisée par la figure 81 :

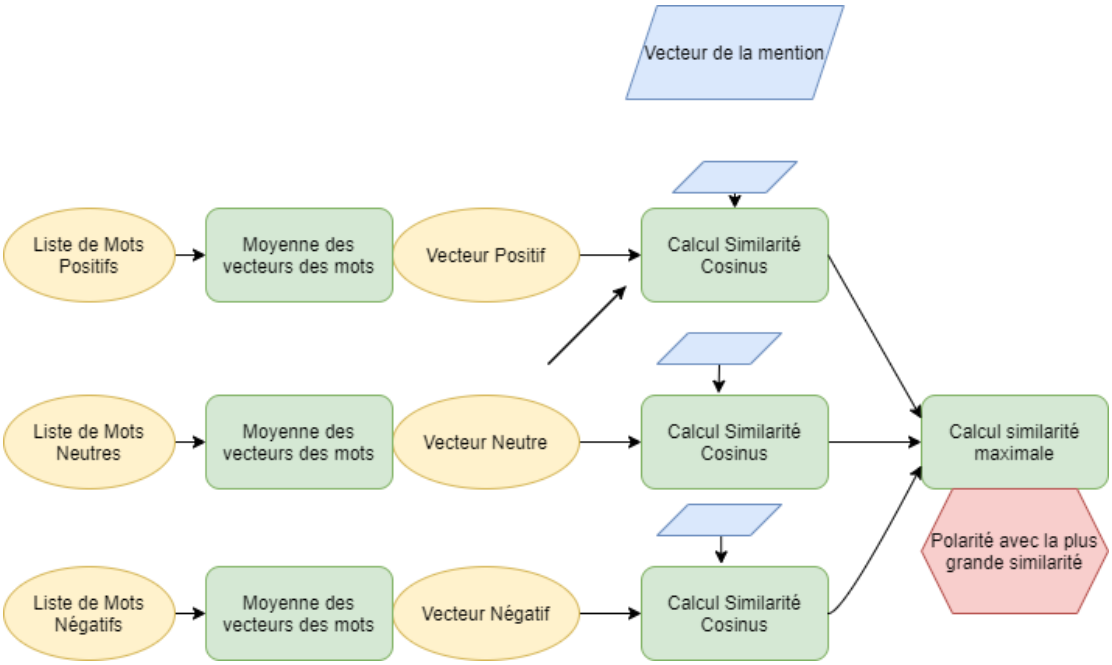

Figure 82 - Détection non-supervisée de la polarité

Nous avons utilisé une liste de mots positifs, une liste de mots neutres et une liste de mots-négatifs (obtenus sur le site polarimots) en français. Le même modele Word2Vec que celui utilisé pour la vectorisation de la mention a été utilisé pour obtenir les vecteurs de chacun des mots de chaque liste, qui sont ensuite agrégés pour en faire un vecteur moyen. Nous avons donc trois vecteurs, un pour chaque polarité. Nous espérions pouvoir tirer partir de l'espace vectoriel sémantique et de la possibilité d'y effectuer des opérations arithmétiques mentionné dans les articles sur word2Vec.

Nous calculons ensuite la Similarité Cosinus entre chacun de ces vecteurs et la mention dont nous cherchons à déterminer la polarité. Puis nous attribuons à la mention la polarité associée au vecteur ayant la plus grande similarité cosinus. Nous obtenons les résultats de la figure 83.

	Précision	Rappel	F-Mesure	Support
neutre	0.77	0.22	0.34	386
Négatif	0.01	0.43	0.02	7
Positif	0.51	0.39	0.44	46
Accuracy			0.24	439
Macro Moyenne	0.43	0.35	0.27	439
Moyenne Pondérée	0.73	0.24	0.35	439

Figure 83 - Résultats de la détection non-supervisée de la polarité de la mention

Les résultats sont bien inférieurs à ceux de la méthode supervisée, avec une exactitude de seulement 0.24. Ces résultats sont étonnants car cette méthode marche très bien pour détecter la polarité à l'échelle du mot. Les faibles performances peuvent s'expliquer par le fait que nous avons comparé des vecteurs issus de différents mots dans chaque mention ; hors, ceux-ci mélangent typiquement une majorité de mots neutres et éventuellement un mot négatif ou positif, ce dernier donnant la polarité à toute la mention. De ce fait, une meilleure méthode serait de considérer le neutre par défaut et de chercher au moins un mot négatif ou positif pour attribuer une polarité différente.

Ces résultats nous ont tout de même permis de sortir de l'écueil de la classe minoritaire ignorée. Un rappel de 0.43 avec une très mauvaise précision indique que beaucoup de mentions ont été considérées comme négatives en dépit du faible nombre d'exemples de cette classe.

11 Limites et Perspectives

La constitution du corpus n'a pas été nécessairement évidente pour des raisons de disponibilité. En effet le quartier d'affaires de la Défense est un sujet assez spécifique, ce qui limite le nombre d'articles à disposition évoquant le sujet. Nous avons voulu dans le même temps avoir un corpus de taille suffisante pour pouvoir faire des analyses significatives. Ces deux contraintes nous ont poussées à choisir un corpus avec des sources non équitablement réparties : Par exemple le Parisien constitue la moitié du corpus en français. Cela pose des problèmes de représentativité du corpus et relativise notre analyse.

Ce souci de représentativité était d'ailleurs au cœur du choix des journaux (trois journaux par type de journaux) pour le corpus français. Pour le corpus en anglais en revanche, le nombre total d'articles mentionnant la Défense étant bien inférieur, nous avons dû faire le compromis de ne pas faire la même sélection de journaux afin d'avoir tout de même un corpus de taille comparable. Ce choix a été fait en considérant qu'une plus grande quantité d'exemples réduit le risque de biais causé par chaque article et ses spécificités (date, journal…). Ce biais demeure tout de même et nous devons le garder à l'esprit. A l'avenir, une étude encore plus précise pourra prendre en compte chacune de ces variables pour obtenir des résultats qui permettront de comparer la variation selon la date ou le journal.

Le Balisage effectué par *CamemBERT* et par *RoBERTa* donne d'excellents résultats. Il est encore possible de les améliorer voire de les parfaire. Pour améliorer les modèles nous pouvons penser à créer une plus grande variation dans le corpus d'entraînement afin de mieux reconnaître les exemples avec des formes plus rares (typiquement les métaphores) qui sont les moins bien reconnues. Des pistes pour arriver à cela peuvent être d'augmenter les données (créer des nouveaux exemples) ou sous-échantillonner les exemples fréquents.

La typologie établie est basée sur une partie des exemples du corpus et en faisant l'hypothèse que le reste du corpus suit également cette typologie. Cependant nous avons vu en annotant que nous avons dû mettre à jour notre typologie (rajouter *immobilier* dans les cibles du contexte) en cours de route, ce qui montre que notre typologie n'est pas absolue et est vulnérable à la non-adéquation face aux exemples ultérieurs. Cela est très important à garder à l'esprit si nous voulons élargir le corpus à l'avenir (et particulièrement si on veut faire un outil de veille où le corpus est mis à jour quotidiennement). Nous avons d'ailleurs vu que certaines étiquettes de la typologie établies pour le français étaient très peu pertinentes pour le corpus anglais, cela témoigne encore une fois du besoin d'avoir une mise à jour ponctuelle par un être humain de la typologie

Nous espérons, lors de la suite du stage, automatiser complétement chacune des étapes effectuée pour arriver aux résultats et à leur analyse afin de concevoir un outil de veille portant sur La Défense dans la Presse. Nous avons déjà commencé à tenter d'automatiser l'annotation de trois des 8 catégories. Pour celles-ci un problème majeur auquel nous avons été confronté est le déséquilibre des classes, nous cherchons donc un moyen de le résoudre : Nous pourrons par exemple donner plus de poids aux exemples des classes minoritaires. Nous songeons aussi à améliorer la reconnaissance des catégories en utilisant d'autres représentations, celles-ci conçus par la création de traits linguistiques pertinents pour chacune des catégories. Par la suite nous tâcherons d'automatiser l'annotation des 5 catégories restantes. Enfin notre outil de veille se devra de posséder une interface visuelle afin de pouvoir être utilisé par les partenaires de la Chaire, qui ont financé ce projet.

12 Conclusion

Nous avons pu répondre à la problématique énoncée au début du stage en établissant une méthode pour étudier la perception du quartier de la Défense et en l'implémentant afin d'obtenir des résultats très satisfaisants.

Ce stage de six mois (Quatre mois et demi au moment du rendu du rapport) m'a permis de mettre en application au cours d'un projet de nombreuses techniques et compétences du traitement automatique des langues et de la linguistique acquises au cours de ma formation. Ce stage nous a permis de mener à bien un projet depuis son initiation et (nous l'espérons) nous permettra dans le futur de le conduire jusqu'à son aboutissement.

Ce projet a brassé de très nombreux aspects de la linguistique et du traitement automatique des langues. De l'analyse linguistique à l'analyse statistique en passant par l'apprentissage Automatique , ce sujet m'a vraiment permis de mettre en œuvre de nombreuses compétences et d'élargir mon horizon dans le domaine.

Les résultats obtenus étant déjà très prometteurs, nous espérons dans l'avenir pouvoir implémenter un outil de veille capable d'analyser automatiquement la perception du quartier de la Défense dans la presse. Notre travail pourra ensuite servir de base pour des travaux ultérieurs portant sur la perception des lieux en général.

13 Références

BORILLO, A. (2001). « Il y a prépositions et prépositions ». *Travaux de linguistique*, 141-155.

BRANDO, C., DOMINGUES, C., & CAPEYRON, M. (2016). Evaluation of NER Systems for the Recognition of Place Mentions in French Thematic Corpora. *Proceedings of the 10th Workshop on Geographic Information Retrieval*, (pp. 7:1-7:10). doi:10.1145/30

CADIOT, P. (2002). Éléments d'une critique de la notion de préposition spatiale. *Syntaxe et Sémantique 2002/1 (N° 3)*, 117-129.

DOMINGUES, C., & ESHKOL-TARAVELLA. (2015). Toponym recognition in custom-made map titles. *International Journal of Cartography, vol. 1, n. 1,*, 109-120. doi:10.1080/23729333.2015.1055935

DOMINGUES, C., JOLIVET, L., BRANDO, C., & CARGILL, M. (2019). Place and Sentiment-based Life story Analysis. *Revue française des sciences de l'information et de la communication*. doi:https://doi.org/10.4000/rfsic.7228

DOMINGUES, C., WEBER, S., BRANDO, C., JOLIVET, L., & VAN DAMME, M.-D. (2017, Novembre). Analyse et cartographie des sentiments dans des récits de vie de migrants. *Spatial Analysis and GEOmatics (SAGEO) 2017*. doi:https://hal.archives-ouverte

ESHKOL-TARAVELLA, I., & FLAMEIN, H. (2017). "Dis-moi Orléans" Repérage et analyse de la perception d'un lieu dans l'oral transcrit. *Echo des études romanes*, pp. 61-72.

FLAMEIN, H. (2019). *« Étude de la perception d'une ville. Repérage automatique, analyse et visualisation ».* Orléans.

FLAMEIN, H., & ESHKOL-TARAVELLA, I. (2021). Exploitation et analyse du corpus ESLO par les outils du TAL et de la géomatique.

GOUVERT, X. (2008). *Problèmes et méthodes en toponymie française. Essais de linguistique historique sur les noms de lieux du Roannais, Thèse de doctorat, Université Paris 4. .*

GROBOL, L. (2021). DeCofre : Detecting Coreference for Oral French. Récupéré sur https://tel.archives-ouvertes.fr/tel-02928209

Huggingface. (s.d.). Neuralcoref. Récupéré sur https://github.com/huggingface/neuralcoref

LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., . . . STOYANOV, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. Récupéré sur https://arxiv.org/abs/1907.11692

MARTIN, L., MULLER, B., JAVIER, P., SUAREZ, O., DUPONT, Y., ROMARY, L., . . . SAGOT, B. (2020). CamemBERT: a Tasty French Language Model. Récupéré sur https://camembert-model.fr/publication/camembert/

WILKENS, R., OBERLE, B., LANDRAGIN, F., & TODIRASCU, A. (s.d.). COFR: COreference resolution tool For FRench. Récupéré sur https://hal.archives-ouvertes.fr/hal-02486764