MASTER EN SCIENCES MATHÉMATIQUES À FINALITÉ APPROFONDIE

Les statistiques textuelles

LIÉGEOIS, Margaux

Award date:
2022

Awarding institution:
Universite de Namur

UNIVERSITE DE NAMUR

Faculté des Sciences

LES STATISTIQUES TEXTUELLES

Promoteur : Germain VAN BEVER

Mémoire présenté pour l'obtention du grade académique

de master en sciences mathématiques, à finalité spécialisée en Project Engineering

Margaux LIEGEOIS

Juin 2022

Remerciements

Je souhaite remercier mon promoteur, Germain Van Bever, pour son accompagnement et sa disponibilité. Ses conseils et son aide ont été précieux tout au long de ce mémoire de même que son écoute et son soutien.

Je remercie mes parents et ma marraine, Nicole Ianiero, pour leurs relectures et conseils ainsi que pour leurs encouragements tout au long de mon parcours universitaire. Je remercie aussi ma soeur, Estelle, pour son soutien sans faille.

Il est important pour moi de remercier aussi mes amis, Marine Delvosal, Sarah Mallia, Célia Lobefaro, Lara Ernst, Bertrand Mouton, Gaëtan Louvet, Laura Jacquemotte ainsi que Noémie Lambert et Émilie Lambert. Merci pour vos encouragements, votre confiance, de même que pour les bons moments passés ensemble.

Enfin, je tiens à remercier le personnel du département de mathématiques de l'Université de Namur. Leur compréhension et leur accessibilité ont joué un rôle dans le cheminement de mes études. Plus particulièrement, je voudrais remercier Anne Lemaitre et André Fuzfa pour leur écoute attentive.

Merci à toutes et à tous.

Résumé

En textométrie, il est coutumier de structurer le texte, c'est-à-dire de le transformer sous forme d'un tableau. Ce mémoire s'intéresse à deux méthodes statistiques qui permettent d'analyser cette table. Tout d'abord, il y a l'analyse factorielle des correspondances qui appartient à la famille de méthodes des axes principaux. Elle produit une cartographie des mots du corpus. Elle a pour but d'identifier les individus semblables ou dissemblables ainsi que les variables les plus ou les moins explicatives de ces ressemblances entre les individus étudiés. Ensuite, il y a la classification ascendante hiérarchique qui est une procédure de clustering. À l'aide de ces techniques d'analyse, nous traitons un corpus constitué d'articles des journaux *Le Soir*, *FranceSoir*, *Le Figaro* et *Libération* qui concernent le réchauffement climatique.

Les conclusions de l'analyse sont les suivantes : le journal *Le Figaro* se démarque grandement des autres. En effet, celui-ci a une manière de communiquer au sujet du réchauffement climatique qui se détache des autres. Suite aux études, nous remarquons qu'il parle de politique bien plus que les autres. Le journal *FranceSoir*, quant à lui, est très neutre et utilise un langage sans originalité ni conviction. Il est à remarqué que les journaux *FranceSoir* et *Le Soir* sont souvent regroupés ensemble car ils sont neutres. Nous remarquons cependant que, grâce à nos analyses, les deux journaux se distinguent. Le journal *Le Soir* est plus orienté sur l'aspect scientifique et environnemental du réchauffement climatique. Le journal *Libération*, lui, semble moins extrême que le journal *Le Figaro* bien qu'ils aient tous les deux une orientation politique prononcée.

Enfin, une analyse des sentiments est faite sur ce corpus afin de comprendre le sentiment général de chaque journal à propos du réchauffement climatique. Nous constatons qu'ils sont globalement neutres lorsqu'ils évoquent le réchauffement climatique mais avec plus de sentiments positifs que négatifs. Les émotions les plus utilisées par les quatre journaux sont la confiance, la peur et l'anticipation. La joie est l'émotion la moins représentée.

Mots-clefs : corpus, partition, lemme, table lexicale, analyse factorielle des correspondances, dendrogramme, réchauffement climatique.

Abstract

In textometry, it is customary to structure the text, i.e. to transform it into a table. This master thesis focuses on two statistical methods for analysing this table. Firstly, there is the correspondence factor analysis, which belongs to the family of principal axis methods. It produces a mapping of the words in the corpus. Its aim is to identify similar or dissimilar individuals as well as the variables that explain these similarities between the individuals studied. Secondly, there is the hierarchical ascending classification which is a clustering procedure. Using these analysis techniques, we process a corpus consisting of articles from the newspapers *Le Soir*, *FranceSoir*, *Le Figaro* and *Libération* concerning global warming.

The conclusions of the analysis reveal that: the newspaper *Le Figaro* is very different from the others. Indeed, it has a way of talking about global warming that stands out. Studies show that *Le Figaro* talks about politics much more than *Le Soir*, *FranceSoir* and *Libération*. The newspaper *FranceSoir* for its part is very neutral and uses language without originality or conviction. While *FranceSoir* and *Le Soir* are often grouped together because they are neutral, the analysis indicates that these two show differences. *Le Soir* is more oriented towards the scientific and environmental aspects of global warming. On the other hand, the newspaper *Libération* seems less extreme than *Le Figaro*, although both have a strong political orientation.

Finally, a sentiment analysis is made on this corpus in order to understand the general sentiment of each newspaper about global warming. Results reveal that they are generally neutral when talking about global warming though with more positive than negative feelings. The emotions most used by the four newspapers are confidence, fear and anticipation. Joy is the least represented emotion.

Keywords : corpus, partition, token, lexical table, factorial correspondence analysis, dendrogram, global warming.

Table des matières

Index

Dans cet index, nous synthétisons les différentes notations utilisées au cours de ce mémoire avec le tableau repris ci-dessous.

Variables	Définitions
Z	Table lexicale d'un corpus
$z_{i,j}$	Effectif du lemme i dans la partition j
z_{max}	Effectif associé au lemme qui a le plus grand effectif
z_{min}	Effectif associé au lemme qui a le plus petit effectif
F	Tableau des fréquences de la table lexicale Z
$f_{i,j}$	Fréquence du lemme i dans la partition j
$f_{i,.}$	Somme des éléments de la ligne i du tableau de fréquence
$f_{.,j}$	Somme des éléments de la colonne j du tableau de fréquence
n	Nombre de lignes de F
p	Nombre de colonnes de F
k_j	Nombre total de lemmes utilisés dans la partition j
k	Nombre total de lemmes considérés
V	Nombre de mots différents dans le corpus
T	Nombre total de mots dans le corpus
e_{max}	Effectif du mot dans le corpus qui a le plus grand effectif
e_{min}	Effectif du mot dans le corpus qui a le plus petit effectif
t	Trace de la matrice $F'F$
$\tau_{i,j}$	Taux d'association entre le lemme i et la partition j
N_n	Nuage des profils lignes
N_p	Nuage des profils colonnes
G_n	Centre de gravité du nuage N_n
G_p	Centre de gravité du nuage N_p

Variables	Définitions
$F_s(i)$	Coordonnée de la ligne i le long de l'axe de rang s
$G_s(j)$	Coordonnée de la colonne j le long de l'axe de rang s
λ_s	L'inertie associée à l'axe s aussi la $s^{ième}$ valeur propre de la matrice $F'F$
u_s	Vecteur propre associé à la $s^{ième}$ valeur propre de la matrice $F'F$
Q	Nombre de clusters
I_q	Nombre d'éléments dans le cluster q
$L(q, q')$	Lien entre les clusters q et q'
$\delta(q, q')$	Diminution de l'inertie en les deux clusters q et q'
b_q	Poids attribué au cluster q
C_q	Centre de gravité du cluster q

Introduction

Les données disponibles aujourd'hui contiennent en grande partie beaucoup de texte. Nous rencontrons tous les jours des données textuelles, quand nous lisons un article, dans notre fil d'actualité Facebook ou encore dans nos messageries instantanées. Nous en trouvons dans les critiques de films ou plus généralement sur des sites internet. Les données de textes ont une particularité : elles ne sont pas structurées.

Cette caractéristique des données textuelles rend difficile l'utilisation des outils habituels de statistiques et d'analyse de données. Il faut donc adapter les méthodes statistiques développées, pour les variables numériques que nous connaissons, aux données sous forme de textes, phrases, paragraphes, etc.. Nous devons aussi gérer la haute dimensionalité des données textuelles. Différentes méthodes statistiques nous permettent d'atteindre cet objectif.

Dans le premier chapitre, nous introduisons la textométrie. Nous expliquons comment organiser les données textuelles pour qu'elles soient structurées afin de pouvoir s'en servir lors de l'analyse statistique. Les principales méthodes multidimensionnelles en matière de statistiques textuelles sont présentées dans le deuxième chapitre. En particulier, la théorie de deux méthodes est développée en profondeur. Nous nous intéressons à l'utilisation et à la compréhension du logiciel TXM. Ce logiciel intègre déjà des bases de données. Celles-ci sont manipulées de manière à se familiariser avec TXM et à illustrer les méthodes statistiques présentées. Le troisième chapitre consiste à créer et à analyser une base de données traitant d'un sujet de société. Le choix de la thématique s'est orienté sur le réchauffement climatique. Grâce au logiciel TXM et les méthodes statistiques intégrées, l'analyse de ces nouvelles données nous permet de répondre à des questions de recherche. Nous souhaitons comprendre si les orientations politiques différentes des journaux ont un impact sur leur manière de parler du réchauffement climatique. Nous désirons savoir si les journaux abordent plus l'aspect politique, scientifique ou environnemental du réchauffement climatique. Le dernier chapitre porte sur l'analyse des sentiments. La théorie et le choix d'une méthode sont dévoilés. Cette technique est utilisée pour analyser la base de données choisie au chapitre 3.

Nous tenterons de montrer comment les outils de calcul et de gestion des données actuellement disponibles peuvent être utilisés pour aider à décrire, assimiler et enfin évaluer les données de nature textuelle.

Chapitre 1

La textométrie

Dans ce chapitre, nous expliquons le concept de statistiques textuelles et dans quels contextes nous l'appliquons. Nous développons la manière dont nous manipulons des textes en tant que données et nous détaillons la marche à suivre. L'objectif premier consiste à structurer les données.

1.1 Qu'est-ce que la textométrie ?

Sur base des références [1] et [2], cette section explique la notion de textométrie. Les données textuelles sont brutes et complexes. L'inconvénient des données non structurées est qu'elles nécessitent une analyse spécifique et des outils spéciaux pour explorer du mieux possible leurs potentiels. En effet, lorsque nous utilisons des méthodes mathématiques, nous désirons analyser des données qui ont été prédéfinies et formatées selon une structure précise. Dans le cadre de la textométrie, nous désirons établir un tableau reprenant les informations des textes considérés. Dés lors, nous devons d'abord traiter les données textuelles afin d'utiliser des propriétés et des méthodes que nous connaissons.

Le concept de la textométrie a été développé principalement en France depuis 1970. Elle associe les méthodes statistiques à l'analyse des textes et aux statistiques multivariées. Grâce à l'analyse de données textuelles, nous serons capables d'extraire les informations les plus pertinentes et essentielles. La textométrie manipule ces dernières à l'aide de plusieurs procédures : analyses factorielles, classifications, etc.. Elle permet de mettre en évidence certaines caractéristiques du texte. Ces techniques produisent des listes ordonnées, des regroupements, des visualisations cartographiques représentant les mots de la même manière que leur composition dans le texte. La textométrie traite principalement des corpus.

Définition 1.1.1 *Un **corpus** est un grand ensemble de textes et peut être vu comme un recueil de documents. Nous utiliserons et analyserons différents corpus tout au long du travail.*

La singularité de la textométrie est qu'elle s'occupe de corpus nécessitant à la fois la gestion et l'analyse de textes. Beaucoup de domaines sont confrontés à une très grande quantité de données textuelles et nous avons besoin de cette aide pour préparer, détailler et comparer les textes.

En particulier, la textométrie se voit être adaptée notamment dans le domaine des sciences humaines et sociales. Par exemple, elle peut être utilisée pour explorer des livres d'un même auteur ou des réponses à une enquête. Elle analyse ces corpus de manière détaillée et globale et nous donne une exploration complète de ces données textuelles. La textométrie prête attention à la signification des mots, ce qui est très important dans ces disciplines.

1.2 Le logiciel TXM

Introduite en 2009 dans le cadre d'un projet ANR de la Fédération des recherches et développements en textométrie [1], la plateforme TXM est très intéressante dans le cadre de ce mémoire. Elle utilise le langage de programmation R. Il est supporté pour Windows 7, 8 et 10. L'architecture du système Windows convient en 32 bits ou en 64 bits. Je travaille avec Windows 10 en 64 bits et la version 0.8.1 de TXM. La dernière version de TXM (0.8.1) est sortie le 29 juin 2020. Pour l'installation de TXM, l'espace disque nécessite 250 Mo sur Windows. Le logo de la plateforme est repris à la Figure 1.1.

Ce logiciel sert à analyser de grands corpus de textes car il bénéficie de techniques robustes et puissantes. TXM possède beaucoup de fonctionnalités. Par exemple,

- il permet de diviser le corpus en plusieurs partitions à partir de différentes propriétés du texte (par exemple : nom du livre, date de publication, locuteur) ;
- les partitions peuvent être analysées en profondeur grâce à des tables lexicales, des histogrammes, des cartographies de mots, etc. ;
- il donne le vocabulaire (liste des mots) d'un corpus ou d'une partition ;
- il retrouve le contexte d'un mot ou d'une requête ;

Figure 1.1 – Logo du logiciel TXM [3].

- il construit les tableaux de contingence dont nous avons besoin pour appliquer les méthodes statistiques ;
- il permet de faire une analyse factorielle des correspondances et une classification ascendante hiérarchique d'un jeu de données.

Ce logiciel a plusieurs propriétés attrayantes comme le fait qu'il soit en accès libre. Il a plusieurs intérêts comme la personnalisation et l'adaptation. Nous pouvons également l'installer sous Windows, Mac OS X, Linux via l'adresse internet [1] ou en ligne. Il traite tous types de corpus : du texte brut ou structuré (nombreux formats d'import), dans plusieurs langues possibles, écrit (texte) ou oral (retranscrit). Toutes les figures de ce chapitre sont créées à l'aide du logiciel TXM. Le manuel [4] détaille en profondeur comment utiliser ce logiciel.

1.3 Le texte sous forme de données

Les méthodes statistiques nécéssitent la définition d'une norme car ces méthodes sont des propriétés générales qui doivent pouvoir s'appliquer sur tous les textes. Nous avons exprimé le fait que ces derniers soient non structurés. Nous devons donc dans un premier temps, standardiser l'utilisation du texte en tant que donnée. En effet, lorsque nous lisons un texte, nous ne pensons pas forcément à des mathématiques. Il y a des structures grammaticales complexes et des interactions riches entre les mots.

Pour commencer, l'article [5] nous explique que nous devons utiliser la segmentation du corpus.

Définition 1.3.1 *La **segmentation d'un corpus** est le principe de découper le corpus en plusieurs documents individuels.*

La répartition se fait souvent de manière naturelle. Par exemple, si nous avons un texte brut ne comportant que des recettes différentes de cuisine, alors nous diviserons ce dernier en p documents contenant chacun les recettes d'un plat unique.

À présent, l'objectif est de diminuer la dimension du corpus car à ce stade, elle est extrêmement grande.

Définition 1.3.2 *La **dimension d'un corpus** est le nombre de mots différents qu'il contient.*

Nous savons qu'un texte se compose de mots, de chiffres, de ponctuations et d'espaces blancs. Pour réduire la dimension de ces données, nous nous intéressons uniquement aux mots qui composent le corpus.

Ensuite, nous réduisons le nombre d'éléments linguistiques que nous considérons, c'est-à-dire que nous ne gardons pas les mots les plus utilisés. Par exemple, les conjonctions de coordination comme "et", "ou", etc. ou encore les articles "la", "une", etc. ne définissent pas le texte que nous analysons. Ils sont importants dans la structure grammaticale mais ce ne sont pas eux qui donnent du sens au texte. Par contre, si nous travaillons sur un texte concernant les magasins en temps de Covid, les mots comme "faillite", "déception", "essentiel", etc. seront plus intéressants. Certains termes, trop peu fréquents, ne sont pas repris non plus. En effet, ils ont généralement du sens dans le texte mais le coût des calculs serait trop conséquent car nous aurons une dimensionalité encore très grande. Nous spécifions qu'à partir d'un certain nombre d'apparitions, nous mettrons de côté les mots qui ont une occurrence plus faible.

Nous devons aussi essayer de limiter l'encodage de tous les mots qui se ressemblent. Pour cela, nous allons utiliser la lemmatisation du vocabulaire dans le texte.

Définition 1.3.3 *Le procédé de* ***lemmatisation*** *s'élabore de la manière suivante :*

- *chaque verbe se trouve sous sa forme infinitive ;*
- *les noms sont mis au singulier ;*
- *la suppression de la voyelle finale dans certains cas.*

Des mots comme "délicat", "délicats", "délicate" et "délicates" seront représentés par un seul mot, notons ici "délicat". La lemmatisation joue un rôle important dans la réduction de la dimension d'un texte. Cependant, nous devons prêter attention au fait que, parfois, certains mots ont un double sens. Par exemple, le mot "physique" peut être interprété par la science en tant que telle mais aussi comme le physique d'une personne. Le livre [6] nous indique qu'il peut exister plusieurs ambiguïtés avec la langue française. Certaines de nature sémantique peuvent être levées par une simple vérification du contexte. D'autres nécéssitent l'examen de plusieurs paragraphes, voire du texte dans son intégralité.

Concernant la lemmatisation, TXM se sert du logiciel TreeTagger pour automatiser la construction des lemmes d'un corpus. TreeTagger doit être installé séparément. Il a été développé par Helmut Schmid dans le cadre du projet TC à l'Institut de linguistique informatique de l'Université de Stuttgart [7]. Il utilise un dictionnaire pour les créer. Il existe des dictionnaires en plusieurs langues et TXM manie par défaut celui en français. Le lexique est composé de deux parties : un lexique complet et un avec les suffixes. Le premier lexique a été créé sur base d'un corpus de deux millions de mots [8]. Le second lexique est organisé sous forme d'arbre de décision. Un exemple d'arbre est donné par la Figure 1.2 insprirée de la source [8]. Lorsqu'un mot doit être lemmatisé, TreeTagger parcourt tout d'abord le dictionnaire complet. S'il le détecte dans ce dictionnaire, il lui associe un lemme. Sinon, le mot est testé avec uniquement des lettres minuscules. Dans le cas où le mot n'est toujours pas trouvé, TreeTagger cherche dans le lexique des suffixes.

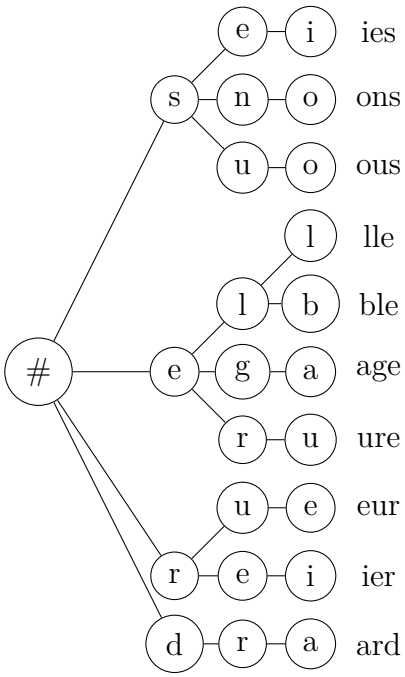

Figure 1.2 – Exemple d'arbre de décision pour le lexique des suffixes de taille trois.

Au fur et à mesure de chaque étape, il y a de moins en moins de mots qui sont pris en considération. Cela implique que la dimension du corpus diminue de plus en plus. Par conséquent, le coût des calculs est fortement réduit et c'est un avantage considérable. Cependant, chacune de ces étapes exige des décisions prudentes sur les éléments susceptibles de porter un sens dans le texte.

Nous garderons une liste de mots définie pour analyser le texte. Nous nommerons "lemmes" les mots que nous avons choisis pour symboliser le texte. Cette procédure nous permet d'accéder à la définition d'une matrice Z. Une ligne z_i de Z est un vecteur numérique dont chaque élément indique le nombre d'apparitions d'un lemme dans le document p. Nous avons maintenant une matrice numérique que nous pouvons manipuler.

1.4 L'analyse quantitative d'un corpus

Nous utilisons dans cette section le jeu de données VOEUX provenant de TXM afin d'illustrer chaque étape du chapitre. Ce corpus se constitue des discours de voeux du Nouvel An entre 1959 et 2012 de sept présidents français : M. Jacques Chirac, M. Charles de Gaulle, M. Valéry Giscard d'Estaing, M. François Hollande, M. François Mitterrand, M. Georges Pompidou et M. Nicolas Sarkozy. Nous divisons donc naturellement ce corpus en sept partitions où chacune comportera les discours de voeux d'un président. Nous observons sur la Figure 1.3, produite à l'aide de TXM, le nombre de mots qu'il y a dans chaque partition.

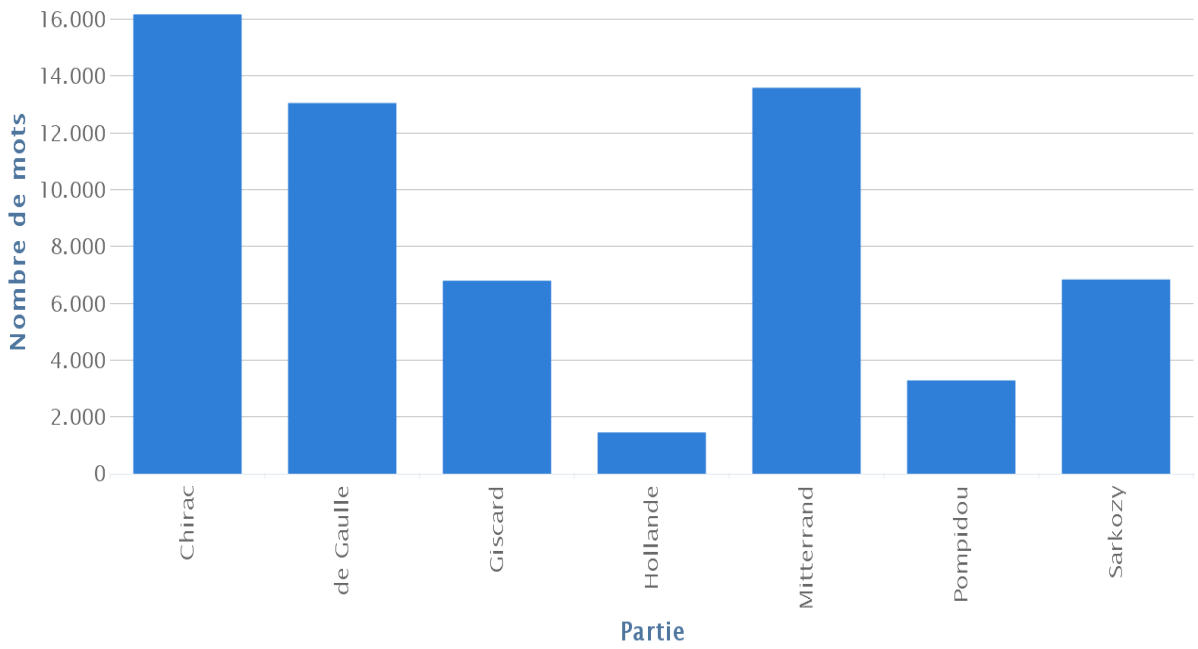

Figure 1.3 – Histogramme du nombre de mots dans chaque partition du corpus VOEUX.

Nous examinons la fréquence des mots d'un corpus pour l'analyser quantitativement. Le mot qui a le plus grand effectif sera représenté par e_{max} ainsi que celui qui a le plus petit effectif sera désigné par e_{min}. Nous avons dans le cas de notre exemple, $e_{max} = 4400$ et $e_{min} = 1$.

Nous désignons par V_1 le nombre de mots de fréquence 1. Dans notre exemple, $V_1 = 3251$. Les nombres correspondant respectivement aux fréquences deux, trois et quatre sont écrits : $V_2 = 1020$; $V_3 = 530$; $V_4 = 306$, etc.. L. Lebart (*et al.*) [6] présente deux formules intéressantes concernant la fréquence des mots. La somme sur toutes les fréquences des nombres V_i est égale au nombre de mots différents contenus dans le corpus, qui est écrit :

$$\sum_{i=1}^{e_{max}} V_i = V.$$

Dans notre exemple, il y a $V = 6406$ mots différents dans le texte. La somme comprise entre les limites 1 et e_{max} des produits de la fréquence par le nombre V_i, est égale à la longueur du corpus :

$$\sum_{i=1}^{e_{max}} V_i \; i = T$$

où T est la longueur du corpus c'est-à-dire le nombre de mots qu'il y a dedans. Pour le corpus VOEUX, $T = 61197$ mots.

Les concordances aident à retrouver les mots ou morceaux de phrases qui entourent un mot. Elles permettent notamment d'examiner plus facilement les relations qui peuvent exister entre les différents contextes d'un mot. Les concordances servent à avoir une vue plus globale des manières dont un mot est utilisé. Sur la Figure 1.4, nous avons un exemple avec une petite partie des concordances du mot "France". Elles sont obtenues via TXM.

L'index établit la liste de fréquences des propriétés d'une requête pour un corpus, sous-corpus ou une partition donnée. Dans l'index, les mots peuvent être classés selon différents critères. Il nous permet de retrouver facilement la fréquence d'un mot ou groupe de mots que nous souhaitons. Pour chaque mot, l'ensemble de ses occurrences dans le corpus peut être localisé grâce à des index. Sur la Figure 1.1, nous avons l'index associé au lemme "souhaiter" produit grâce à TXM. La requête étant : [frlemma = "souhaiter"] pour les partitions du corpus VOEUX. Nous observons les mots repris sous le lemme "souhaiter" ainsi que leur fréquence dans chacune des partitions. Nous remarquons aussi leur fréquence totale dans le corpus. Nous constatons que M. François Hollande n'utilise jamais le lemme "souhaiter". C'est M. Valéry Giscard d'Estaing qui l'emploie le plus souvent.

Un autre aspect que nous observons en statistique textuelle, c'est l'étude de la croissance du vocabulaire. Le nombre de mots contenus dans un corpus n'est pas proportionnel à sa taille. Il y a deux types de variables :

- les variables textométriques de premier type qui augmentent proportionnellement à la longueur d'un texte ;
- les variables textométriques de deuxième type dont le taux de croissance tend à diminuer à mesure que la longueur du texte augmente, c'est-à-dire que le mot n'est pas utilisé tout au long du corpus, mais à certains moments ponctuels.

Sur la Figure 1.5, nous observons la progression des mots "France" et "peuple" construite avec TXM. Ce dernier a une progression de deuxième type tandis que le mot "France" est de premier type.

Table 1.1 – Index pour le lemme "souhaiter" dans le corpus VOEUX.

Mots	Fréquence totale	Chirac	de Gaulle	Giscard	Hollande	Mitterrand	Pompidou	Sarkozy
souhaite	70	15	8	24	0	18	3	2
souhaitez	4	2	0	1	0	0	1	0
souhaitons	4	0	2	1	0	1	0	0
souhaité	2	2	0	0	0	0	0	0
souhaiter	2	1	0	0	0	1	0	0
souhaitant	1	0	1	0	0	0	0	0
souhaitent	1	0	1	0	0	0	0	0
souhaiterons	1	0	0	0	0	1	0	0
Souhaitons	1	0	0	0	0	1	0	0

Requête [word = "France"]

text_id	Contexte gauche	Pivot	Contexte droit
0001	au cours de celle qui finit. En	France	même, nos institutions assurent à l'Etat l'efficacité et l'
0001	la condition humaine et la civilisation. La	France	, dont la population, les ressources, la puissance, sont
0001	. Mais, à toute éventualité, la	France	entend également contribuer à la réforme et, par la suite,
0001	année qui commence, je souhaite à la	France	, à l'Algérie, à la Communauté, l'effort dans
0002	à tous, une bonne année à la	France	. Je le fais en toute confiance. Non point que 1961
0002	Mais si l'univers est troublé, la	France	, elle, ne l'est pas. Solide, laborieuse,
0002	communautés, et qui sera unie à la	France	dans les domaines où celle -ci peut l'aider, nous offrons
0002	besoin de la communauté française, et la	France	, pour son œuvre, a besoin d'elle en Algérie.
0002	, et quoi qu'il arrive, la	France	protégera ses enfants dans leurs personnes et dans leurs biens, quelle
0002	seraient tirées quant à la capacité de la	France	d'assumer la responsabilité des affaires qui la concernent ! Et à
0002	. Voilà l'étranger bien prévenu que la	France	sait ce qu'elle veut. Me voilà moi -même raffermi et
0002	ensemble, nous offrons nos voeux à la	France	. Le 6 janvier nous lui offrirons le oui franc et massif
0002	espoir. Vive la République ! Vive la	France	!
0003	Voici l'année nouvelle ! La	France	en a vécu beaucoup. Cependant elle voit venir celle -ci sans
0003	de l'importance vitale du développement de la	France	et que les travailleurs eux -mêmes et leurs organisations prennent une part
0003	où l'étranger n'ait payé à la	France	plus que celle -ci ne lui a versé. Certes, les
0003	pour le maintenir. C'est pourquoi la	France	aborde l'année nouvelle lucidement et sereinement. Sans se leurrer sur
0003	et dans celui de la défense, la	France	veut, pour sa part, continuer à la développer. Mais
0003	l'organisation commune. En Algérie, la	France	entend que se terminent, d'une manière ou d'une autre
0003	la condition suprême de son avenir, la	France	est résolue à poursuivre l'œuvre immense de la rénovation intérieure qu'
0003	, tous ensemble, nous portons vers la	France	nos souhaits très ardents et très confiants de bonheur et de grandeur
0003	grandeur. Vive la République ! Vive la	France	!
0004	bon sens, marqué le destin de la	France	. Certes ne nous y ont manqué ni les épreuves ni les
0004	raison nous ramène à la puissance, la	France	retrouve son rang, son attrait, ses moyens. Ainsi avons-
0004	nom à tous, je forme pour la	France	le souhait immémorial : « Que l'année lui soit heureuse !
0004	notre ambition nationale. Progrès démographique. La	France	moderne pourrait compter cent millions d'habitants. Combien seront donc bienvenus
0004	mais la structure et la figure de la	France	. Notre plan règle ce développement. Il nous faut l'exécuter
0004	Amérique latine, qui souhaitent celle de la	France	. Ce sont là de bien grands espoirs ? Oui ! Car
0004	nouvelle année sont à la mesure de la	France	nouvelle. Vive la République ! Vive la France !
0004	nouvelle. Vive la République ! Vive la	France	!
0005	Pour la	France	, l'année qui finit a été, en somme, favorable
0005	positif. Pendant ces douze mois, la	France	a continué de monter. En 1963, notre population s'est

Figure 1.4 – Partie des concordances pour le mot “France” dans le corpus VOEUX.

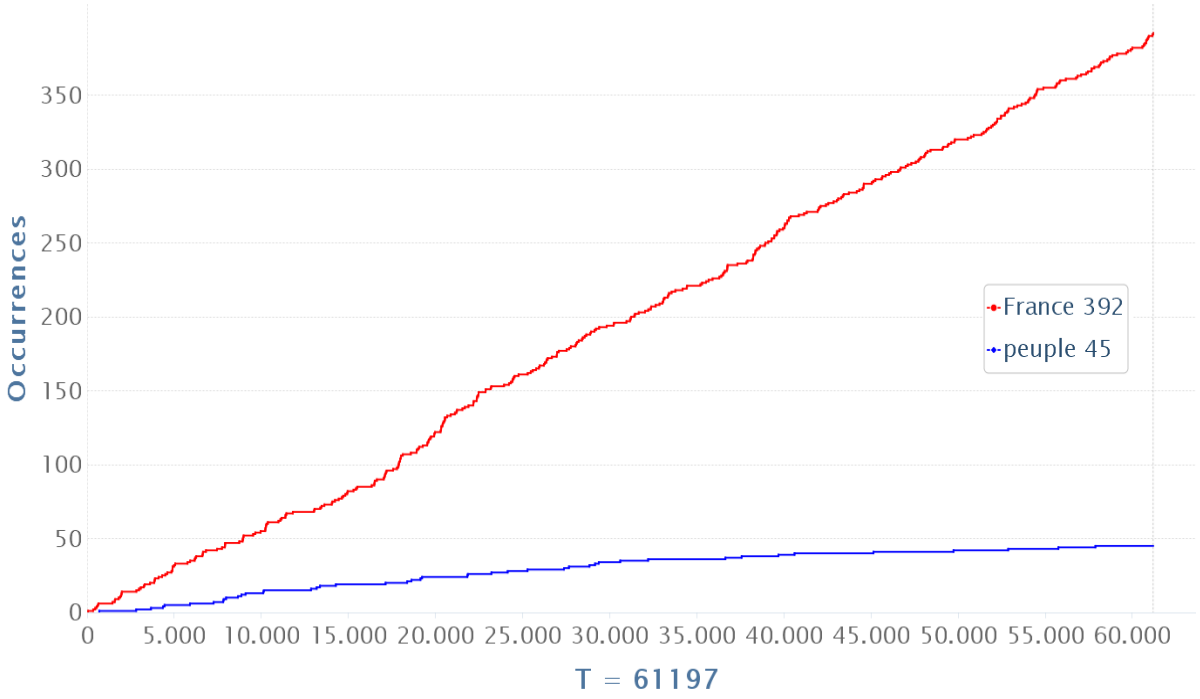

Figure 1.5 – Progression des mots “France” et “peuple” dans le corpus VOEUX.

Enfin, nous générons la table lexicale du corpus et obtenons la Table 1.2. Cette table est un exemple de table de contingence. Nous arrangeons les occurrences de chacun des $n = 31$ lemmes dans chacune des $p = 7$ parties du corpus dans un tableau rectangulaire de n lignes et p colonnes. Le tableau se compose donc de 31 lignes et 7 colonnes. C’est ce dernier qui nous sera fortement utile et que nous analyserons. La dernière colonne donne l’effectif total de chaque lemme et la dernière ligne représente le nombre total de lemmes dans chaque partition. Comme nous l’avons exprimé à la section 1.3, nous mettons de côté les lemmes qui ont trop d’occurrences et ceux qui en ont trop peu. Nous fixons dans le cas de notre exemple, $z_{max} = 490$ et $z_{min} = 50$, qui sont respectivement le seuil d’occurrences maximum/minimum que peut avoir un lemme. Nous prenons en compte ici les lemmes et non les mots. Ce seuil est choisi subjectivement. En effet, TXM nous informe qu’il y a 200 lemmes pour le corpus VOEUX. Cependant, nous devons tenir compte de ce qui est discuté dans la section 1.3 : enlever les déterminants, les conjonctions de coordination et les mots liens ainsi que la ponctuation. Nous ajoutons un seuil avec z_{min} et z_{max} afin de diminuer la dimension du tableau. Il nous reste 31 lemmes. Nous éliminons donc 169 lemmes de l’analyse. Cette dernière partie est faite manuellement.

Table 1.2 – Table lexicale du jeu de données VOEUX divisé en 7 groupes.

Lemmes	Chirac	de Gaulle	Giscard	Hollande	Mitterrand	Pompidou	Sarkozy	Total
France	121	108	87	12	89	25	47	489
année	64	51	50	6	50	25	35	281
pouvoir	37	45	21	0	32	24	31	190
pays	29	34	22	3	37	11	17	153
vouloir	46	15	23	4	21	7	25	141
Europe	55	18	0	1	52	2	13	141
monde	43	23	28	1	20	4	20	139
nouveau	53	22	11	2	18	5	16	127
voeu	13	10	37	1	29	3	7	100
peuple	8	36	8	2	28	6	3	91
vie	18	16	13	2	19	6	15	89
social	34	18	1	3	23	1	9	89
souhaiter	20	12	26	0	22	4	2	86
emploi	46	3	4	3	12	2	13	83
paix	18	18	10	0	25	4	5	80
avenir	41	8	3	3	5	7	12	79
politique	10	20	11	0	25	2	5	73
république	21	17	8	1	20	0	4	71
compatriote	31	0	7	3	5	0	19	65
économique	10	29	3	0	15	2	4	63
liberté	14	4	21	0	15	3	5	62
effort	15	14	6	2	12	3	8	60
confiance	18	6	10	3	7	5	11	60
travail	15	7	10	2	12	1	12	59
gouvernement	22	4	2	3	14	7	7	59
progrès	22	25	7	0	1	3	0	58
ensemble	23	11	6	0	13	2	3	58
crise	9	6	8	2	7	2	24	58
national	14	23	3	0	15	1	0	56
famille	11	5	13	1	11	5	8	54
solidarité	34	0	0	2	8	1	5	50
Total	915	608	459	62	662	173	385	3264

Chapitre 2

Méthodes statistiques pour l'analyse de données textuelles

Dans ce chapitre, nous étudions différentes méthodes statistiques que nous pouvons manipuler pour analyser des textes. Nous commençons par le test χ^2 d'homogénéité. Ensuite, deux outils permettent d'avoir une vision globale des données : les techniques de visualisation (plans d'axes principaux) et les algorithmes de clustering (méthodes regroupant les individus en fonction de plusieurs caractéristiques). Ces deux familles de méthodes peuvent être utilisées sur la même matrice de données. Nous nous intéressons plus précisément à l'analyse factorielle des correspondances et aux dendrogrammes. Nous utilisons le logiciel TXM ainsi que le jeu de données VOEUX pour illustrer ces méthodes.

Rappelons qu'une unité statistique correspond à une case du tableau que nous analysons. La case (i, j) du tableau contient le nombre d'occurrences du mot i dans le groupe de textes j. C'est l'occurrence d'une unité textuelle dans une partition : mot, lemme, etc.. Les lignes du tableau sont les mots dont l'effectif dans les données est supérieure à un seuil donné. Les colonnes sont des groupes de textes : locuteurs, auteurs, documents, etc.. Les totaux des lignes représentent le nombre d'occurrences de chaque lemme tandis que les totaux des colonnes représentent le nombre total de lemmes utilisés par les différentes partitions satisfaisant les contraintes d'effectif. Ce qui correspond bien au tableau lexical de la Table 1.2, nommée Z. Dans ce chapitre, nous travaillons aussi avec la matrice de fréquence relative, nommée F. Elle est calculée à partir de la matrice Z tel que $\frac{z_{i,j}}{k} = f_{i,j}$.

2.1 Le test χ^2 d'homogénéité

Dans cette section, nous expliquons en quoi consiste le test χ^2 d'homogénéité et nous l'appliquons à notre exemple. Les références [6] et [9] ainsi que [10] nous donnent les formules principales dont nous avons besoin.

Le test χ^2 d'homogénéité évalue si les distributions des variables du tableau sont homogènes. Nous allons donc voir si les lemmes sont répartis de manière homogène dans

les différentes partitions du corpus, mais aussi si les partitions utilisent la même répartition des lemmes. Les hypothèses de ce test sont les suivantes :

$$\begin{cases} \mathcal{H}_0 & : \forall i \in 1, ..., n, f_{i,1} = ... = f_{i,p} \\ \mathcal{H}_1 & : \exists i \in 1, ..., n \text{ tels que } f_{i,j} \neq f_{i,j'}. \end{cases}$$

La statistique de test donnée par la source [9] est

$$X^2 = \sum_{i=1}^{n}\sum_{j=1}^{p} \frac{(kf_{i,j} - kf_{i,.}f_{.,j})^2}{kf_{i,.}f_{.,j}} = k\sum_{i=1}^{n}\sum_{j=1}^{p} \frac{(f_{i,j} - f_{i,.}f_{.,j})^2}{f_{i,.}f_{.,j}}, \text{ où}$$

- k est le nombre total de lemmes que nous considérons ;
- $f_{i,j}$ est la fréquence du lemme i dans la partition j, donc l'effectif divisé par k ;
- $f_{i,.}$ désigne la somme des éléments de la ligne i ;
- $f_{.,j}$ est la somme des éléments de la colonne j de ce tableau.

Cette statistique compare les valeurs observées de la table lexicale 1.2 avec les valeurs théoriques calculées. Il calcule la distance entre chaque élément de chacun des deux tableaux. Ensuite, grâce à un seuil, ce test nous indique s'ils sont proches avec une probabilité de confiance de 0.95. Le seuil est calculé à l'aide de la table de distribution de χ^2.

L. Lebart (*et al.*) [6] nous explique que cette statistique peut aussi être réécrite avec la trace, notée t, de la matrice $F'F$:

$$X^2 = kt.$$

Si cette statistique vaut zéro, les fréquences des lemmes dans chaque variable sont toutes identiques. Dans le cas où la statistique est très faible, elles sont proches de l'égalité. Si la statistique est grande, elles sont loin d'être les mêmes. Les deux premières possibilités arrivent très rarement puisque tous les documents auraient presque le même contenu lexical. Le nombre de degrés de liberté est de

$$dl = (n-1)(p-1).$$

La proposition concernant la convergence du χ^2 est tirée de la source [10].

Proposition 2.1.1 *Considérons une situation dans laquelle nous observons p multinomiales mutuellement indépendantes $F_j = (f_{1,j}, ..., f_{n,j}) \sim Mult(k_j, p_{1,j}, ..., p_{n-1,j}), j = 1, ..., p$, où k_j représente le nombre total de lemmes utilisés dans la partition j et $k = \sum_{j=1}^{p} k_j$.*

$$\sum_{i=1}^{n}\sum_{j=1}^{p} \frac{\left(z_{i,j} - \frac{z_{i,.}z_{.,j}}{k}\right)^2}{\frac{z_{i,.}z_{.,j}}{k}} = k\sum_{i=1}^{n}\sum_{j=1}^{p} \frac{(f_{i,j} - f_{i,.}f_{.,j})^2}{f_{i,.}f_{.,j}} \xrightarrow{\mathcal{L}} \chi^2_{(n-1)(p-1)}$$

sous $\mathcal{H}_0$ quand $k \longrightarrow \infty$.

Le test qui en résulte consiste à rejeter l'hypothèse nulle d'homogénéité au niveau asymptotique α si et seulement si

$$k \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(f_{i,j} - f_{i,.} f_{.,j})^2}{f_{i,.} f_{.,j}} > \chi^2_{(n-1)(p-1),1-\alpha}.$$

Nous nous plaçons dans le cadre de l'exemple du corpus VOEUX. Nous connaissons la valeur des paramètres suivants :

- le nombre de lignes est de $n = 31$;
- le nombre de colonnes est de $p = 7$;
- le nombre total de lemmes considérés est de $k = 3264$;
- la trace de la matrice $F'F$ est de $t = 0.2355$.

La valeur du X^2 est égale à

$$kt = 3264 \times 0.2355 = 768.672.$$

Le degré de liberté pour notre exemple est donné par :

$$dl = (n-1)(p-1) = (31-1)(7-1) = 180.$$

La valeur du $\chi^2_{(n-1)(p-1),1-\alpha}$ est égale à 212.3039. La p-valeur est de $1.096842\ 10^{-73}$. Nous devons donc rejeter l'hypothèse nulle du fait que $kt > \chi^2_{(n-1)(p-1),1-\alpha}$ pour le paramètre $\alpha = 0.05$. Par conséquent, les lemmes ne sont pas répartis de la même façon dans les différentes partitions du corpus VOEUX et les partitions.

Dans le contexte des statistiques textuelles, ce test statistique rejette l'hypothèse d'homogénéité la plupart du temps. Par conséquent, les partitions possèdent des mots différents qui sont liés de différentes manières. Comme nous n'avons presque jamais l'homogénéité des variables, nous exécutons l'analyse factorielle des correspondances.

2.2 L'analyse factorielle des correspondances

L'analyse factorielle des correspondances, présentée dans cette section, appartient à la famille de méthodes des axes principaux. Ces dernières permettent d'obtenir un nuage de points d'une forme proche du nuage initial mais dans un repère modifié. Ce dernier est de dimension plus petite car il prend en compte peu d'axes principaux. Ces derniers sont des axes qui expliquent une grande partie de la variabilité des données. L'analyse factorielle des correspondances a été proposée par J.-P. Benzécri [9]. Elle correspond à une technique de description des tableaux de contingence et de certains tableaux binaires, aussi appelés tableaux "présence-absence". Elle se fonde sur le tableau de contingence et cherche à montrer le rapport qu'il y a entre les lignes et les colonnes. L'analyse factorielle

des correspondances d'un tableau permet de trouver la distance entre les lignes du tableau en partitionnant le corpus. Cette description prend essentiellement la forme d'une représentation graphique des associations entre lignes et entre colonnes.

L'objectif principal de l'application de l'analyse factorielle des correspondances (AFC) à une table lexicale est de visualiser la proximité entre les documents, la proximité entre les lemmes ainsi que les associations entre les documents et les lemmes. Si deux documents favorisent ou évitent les mêmes lemmes, alors ils sont proches au sens de la distance du chi-carré. Si deux lemmes sont distribués de la même manière dans tous les documents différents, alors ils sont proches au sens de la distance du chi-carré. Enfin, nous posons qu'une partition et un lemme s'attirent ou se repoussent selon que la partition utilise le lemme avec une fréquence relative supérieure ou inférieure à la moyenne.

Le but de cette méthode est aussi de réduire la dimension de l'espace de départ qui contient toutes les informations tout en en conservant un maximum, ce qui rend les visualisations plus claires et compréhensibles. Afin d'atteindre ces différents objectifs, nous manipulons et analysons la table lexicale pour arriver à cette représentation graphique.

2.2.1 Profils lignes et profils colonnes

La table lexicale Z ne permet pas de mettre en évidence des liens entre les lemmes et les partitions. Pour améliorer cela, nous divisons chaque élément du tableau par la somme de la ligne correspondante ou en divisant les éléments du tableau par la somme de la colonne correspondante. Nous nommerons les matrices obtenues : les profils lignes et les profils colonnes respectivement. L. Lebart (*et al.*) [6] nous indique que nous pouvons exprimer cette construction mathématiquement.

Définition 2.2.1 *Le profil ligne i est l'ensemble des p valeurs :*

$$\left\{ \frac{z_{i,j}}{z_{i,.}} = \frac{f_{i,j}}{f_{i,.}},\ j = 1, ..., p \right\}. \tag{2.1}$$

Définition 2.2.2 *Le profil colonne j est l'ensemble des n valeurs :*

$$\left\{ \frac{z_{i,j}}{z_{.,j}} = \frac{f_{i,j}}{f_{.,j}},\ i = 1, ..., n \right\}. \tag{2.2}$$

La Table 2.1 correspond au tableau des profils lignes de la Table 1.2. En effet, nous pouvons dire que 24.75% de l'utilisation du lemme "France" est faite par la partition "Chirac". Cette dernière utilise moins le lemme "monde" (43 fois) que le lemme "France" (121 fois) à la Table 1.2. Cependant, la partition "Chirac" accorde au lemme "monde" une proportion plus élevée (30.9%) que la "France". Par conséquent, la Table 2.1 est plus

Table 2.1 – Tableau des profils lignes de la table lexicale du jeu de données VOEUX.

Lemmes	Chirac	de Gaulle	Giscard	Hollande	Mitterrand	Pompidou	Sarkozy
France	24.75	22.1	17.8	2.45	18.2	5.1	9.6
année	22.8	18.1	17.8	2.15	17.8	8.9	12.45
pouvoir	19.5	23.7	11.05	0	16.85	12.6	16.3
pays	18.95	22.2	14.4	1.95	24.2	7.2	11.1
vouloir	32.6	10.65	16.3	2.85	14.9	5	17.7
Europe	39	12.75	0	0.71	36.9	1.42	9.22
monde	30.9	16.55	20.15	0.7	14.4	2.9	14.4
nouveau	41.7	17.3	8.65	1.6	14.2	3.95	12.6
voeu	13	10	37	1	29	3	7
peuple	8.8	39.55	8.8	2.2	30.75	6.6	3.3
vie	20.2	18	14.6	2.25	21.35	6.75	16.85
social	38.2	20.22	1.1	3.4	25.85	1.12	10.11
souhaiter	23.25	13.95	30.2	0	25.6	4.65	2.35
emploi	55.42	3.62	4.81	3.62	14.46	2.41	15.66
paix	22.5	22.5	12.5	0	31.25	5	6.25
avenir	51.9	10.1	3.8	3.8	6.3	8.9	15.2
politique	13.7	27.4	15.1	0	34.25	2.75	6.8
république	29.6	23.9	11.3	1.4	28.2	0	5.6
compatriote	47.7	0	10.8	4.6	7.7	0	29.2
économique	15.9	46	4.8	0	23.8	3.2	6.3
liberté	22.6	6.4	33.9	0	24.2	4.85	8.05
effort	25	23.3	10	3.35	20	5	13.35
confiance	30	10	16.65	5	11.65	8.35	18.35
travail	25.4	11.9	16.9	3.4	20.35	1.7	20.35
gouvernement	37.3	6.8	3.4	5.1	23.7	11.85	11.85
progrès	37.9	43.1	12.1	0	1.7	5.2	0
ensemble	39.65	18.95	10.35	0	22.4	3.45	5.2
crise	15.5	10.3	13.8	3.45	12.1	3.45	41.4
national	25	41	5.4	0	26.8	1.8	0
famille	20.4	9.3	24.1	1.8	20.4	9.2	14.8
solidarité	68	0	0	4	16	2	10

représentative de la réalité. En comparant deux lignes de la Table 2.1, nous apprenons comment les lemmes représentés par ces deux lignes sont associés aux partitions. En comparant deux colonnes, nous apprenons les similitudes qui existent entre les différentes partitions des présidents par rapport au vocabulaire utilisé. L'analyse factorielle des correspondances adopte ce point de vue.

Le calcul des profils lignes moyens nous permet de les comparer aux profils lignes des documents, révélant ainsi les partitions utilisant certains lemmes un plus grand ou petit nombre de fois que la moyenne. Nous faisons de même pour les profils colonnes moyens. Les profils moyens nous permettent d'avoir un point de comparaison.

Définition 2.2.3 *Le profil ligne moyen est donné par*

$$\left\{\frac{z_{.,j}}{k} = f_{.,j},\ j = 1, ..., p\right\}.$$

Définition 2.2.4 *Le profil colonne moyen est donné par*

$$\left\{\frac{z_{i,.}}{k} = f_{i,.},\ i = 1, ...n\right\}.$$

Les profils moyens de ligne et de colonne sont également appelés profils marginaux. L'AFC agit comme une synthèse visuelle des similarités ou dissimilarités distributionelles des lignes et des colonnes en comparant tous les différents profils entre eux, mais aussi avec le profil moyen des lignes ou des colonnes.

Par exemple, le lemme "souhaiter" apparait dans la partition "Giscard" ($\frac{26}{86} \times 100 = 30.23\%$ des occurrences) plus que la moyenne ($\frac{459}{3264} \times 100 = 14.06\%$ de toutes les occurrences), et nous disons qu'il l'attire. D'autre part, M. Nicolas Sarkozy utilise le lemme "France" ($\frac{47}{385} \times 100 = 12.2\%$ des occurrences) moins que la moyenne, et nous spécifions qu'il le repousse.

2.2.2 Indépendance des documents

Il existe une autre manière d'étudier les associations entre documents et lemmes, équivalente à la précédente. Elle peut être développée en termes d'écart à l'indépendance [11]. Pour cette méthode, la situation de référence est l'absence de relations entre colonnes et lignes.

Définition 2.2.5 *Il y a indépendance entre les lignes et les colonnes si, pour toute ligne* $1 \leq i \leq n$ *et colonne* $1 \leq j \leq p$*, l'équation est vérifiée :*

$$f_{i,j} = f_{i,.}f_{.,j}.$$

Si nous avons l'indépendance, cela signifie que tous les profils lexicaux sont égaux les uns aux autres.

Nous comparons la table lexicale (observée) vue au chapitre 1 avec la table des effectifs attendus. La case (i, j) de ce tableau est calculée à partir des fréquences relatives de la manière suivante :

$$case(i, j) = kf_{i,.}f_{.,j}$$

ou, à partir des effectifs du tableau Z, de la manière suivante :

$$case(i, j) = \frac{z_{i,.}z_{.,j}}{k}.$$

Ce modèle est appelé modèle d'indépendance. Ce qui est important ici, c'est de voir si nous sommes suffisamment éloignés de l'indépendance pour que des caractéristiques textuelles intéressantes puissent être potentiellement récoltées dans les données.

Les tables des effectifs observés et attendus sont celles que nous comparons lorsque nous manipulons le test χ^2. Cette statistique est une somme des différences relatives entre les éléments de ces deux tables.

Si nous prenons l'exemple du jeu de données VOEUX, nous avons un tableau des effectifs observés qui est la Table lexicale 1.2 et un tableau des effectifs attendus à la Table 2.2. La comparaison des effectifs observés et attendus conduit à la même conclusion que la comparaison du profil ligne (resp. colonne) avec son profil moyen associé. La troisième (resp. la septième) partition utilise le lemme "France" 87 (resp. 47) fois, beaucoup plus (resp. moins) que le compte attendu du modèle d'indépendance de 68.77 (resp. 57.68), en bleu dans le tableau 2.2.

L'AFC étudie toutes les relations entre les documents et les lemmes en termes d'écart par rapport au modèle d'indépendance et fournit une représentation géométrique interprétable de ces relations. Cela conduit ensuite à la construction de ce que nous appelons des axes factoriels.

2.2.3 Taux d'association

Le rapport du nombre observé et du nombre attendu pour une combinaison lemme/partition donnée, mesure l'association entre ce document et ce lemme. Ce rapport $\tau_{i,j}$ est appelé le taux d'association entre le lemme i et le document j et est donné par :

$$\tau_{i,j} = \frac{kz_{i,j}}{z_{i,.}z_{.,j}} = \frac{f_{i,j}}{f_{i,.}f_{.,j}}.$$

Le taux d'association est supérieur à 1 si le document et le lemme s'attirent ou est inférieur à 1 s'ils se repoussent.

2.2.4 Nuage de lignes et de colonnes

Grâce à M. Bécue-Bertaut [9] et STHDA [12], nous définissons la notion de nuage des profils lignes ou colonnes. Nous notons N_n le nuage des profils lignes. Il est composé de l'ensemble des n profils lignes dans l'espace $\mathbb{R}^p$. Nous désignons le centre de gravité de ce nuage de point par G_n qui correspond au profil moyen, c'est-à-dire $f_{.,j}$. Nous avons aussi N_p le nuage des profils colonnes. Il est constitué de l'ensemble des p profils colonnes dans l'espace $\mathbb{R}^n$. Nous désignons le centre de gravité de ce nuage de point par G_p qui correspond au profil moyen, c'est-à-dire $f_{i,.}$. Notre objectif est de comparer la position du profil ligne i ou du profil colonne j avec le centre de gravité correspondant.

Table 2.2 – Tableau des effectifs attendus pour l'indépendance du jeu de données VOEUX divisé en 7 partitions.

Lemmes	Chirac	de Gaulle	Giscard	Hollande	Mitterrand	Pompidou	Sarkozy	Total
France	137.08	91.08	68.77	9.29	99.18	25.92	57.68	489
année	78.78	52.34	39.52	5.34	56.99	14.89	33.14	281
pouvoir	53.26	35.39	26.72	3.62	38.54	10.07	22.4	190
pays	42.89	28.5	21.51	2.91	31.03	8.11	18.05	153
vouloir	39.53	26.26	19.83	2.68	28.6	7.47	16.63	141
Europe	39.53	26.26	19.83	2.68	28.6	7.47	16.63	141
monde	38.97	25.9	19.55	2.64	28.19	7.37	16.4	139
nouveau	35.6	23.66	17.86	2.41	25.76	6.73	14.98	127
voeu	28.03	18.63	14.06	1.9	20.28	5.3	11.8	100
peuple	25.51	16.95	12.8	1.73	18.46	4.82	10.73	91
vie	24.95	16.58	12.51	1.69	18.05	4.72	10.5	89
social	24.95	16.58	12.51	1.69	18.05	4.72	10.5	89
souhaiter	24.11	16.02	12.09	1.63	17.44	4.56	10.14	86
emploi	23.27	15.46	11.67	1.58	16.84	4.4	9.79	83
paix	22.43	14.9	11.25	1.52	16.22	4.24	9.44	80
avenir	22.15	14.72	11.1	1.5	16.02	4.19	9.32	79
politique	20.46	13.6	10.27	1.39	14.81	3.86	8.61	73
république	19.9	13.23	9.98	1.35	14.4	3.77	8.37	71
compatriote	18.22	12.11	9.14	1.23	13.18	3.45	7.67	65
économique	17.65	11.74	8.86	1.2	12.78	3.34	7.43	63
liberté	17.38	11.55	8.72	1.18	12.57	3.29	7.31	62
effort	16.81	11.18	8.44	1.14	12.17	3.18	7.08	60
confiance	16.81	11.18	8.44	1.14	12.17	3.18	7.08	60
travail	16.53	10.99	8.3	1.12	11.97	3.13	6.96	59
gouvernement	16.53	10.99	8.3	1.12	11.97	3.13	6.96	59
progrès	16.27	10.8	8.16	1.1	11.76	3.07	6.84	58
ensemble	16.27	10.8	8.16	1.1	11.76	3.07	6.84	58
crise	16.27	10.8	8.16	1.1	11.76	3.07	6.84	58
national	15.7	10.43	7.87	1.06	11.36	2.97	6.61	56
famille	15.14	10.06	7.59	1.03	10.95	2.86	6.37	54
solidarité	14.02	9.31	7.03	0.95	10.14	2.65	5.9	50
Total	915	608	459	62	662	173	385	3264

Nous définissons maintenant la notion de proximité. Nous expliquons comment sont calculées les distances qu'il y a entre deux lignes ou deux colonnes. Nous utilisons les profils lignes (2.1) ou les profils colonnes (2.2) pour placer les points dans les nuages qui sont des espaces de n ou p dimensions. Cependant, nous voulons réduire la dimension pour obtenir une représentation visuelle en deux dimensions qui déforme le moins possible les distances entre les points. L. Lebart (*et al.*) [6] nous donnent les définitions suivantes.

Définition 2.2.6 *La distance entre deux lignes i et i' est donnée par*

$$d^2(i,i') = \sum_{j=1}^{p} \frac{1}{f_{.,j}} \left(\frac{f_{i,j}}{f_{i,.}} - \frac{f_{i',j}}{f_{i',.}} \right)^2 . \tag{2.3}$$

Définition 2.2.7 *La distance entre deux colonnes j et j' est donnée par*

$$d^2(j,j') = \sum_{i=1}^{n} \frac{1}{f_{i,.}} \left(\frac{f_{i,j}}{f_{.,j}} - \frac{f_{i,j'}}{f_{.,j'}} \right)^2 . \tag{2.4}$$

Cette distance est appelée "distance chi-carré" et elle ressemble à un facteur près à la distance euclidienne qui est la somme des carrés des différences entre les composantes des profils. Pour l'équation (2.3), ce facteur correspond à $\frac{1}{f_{.,j}}$. Concernant la seconde équation (2.4), il est égal à $\frac{1}{f_{i,.}}$. La distance du chi-carré vérifie la propriété d'équivalence distributionnelle [13].

Définition 2.2.8 *Le principe d'équivalence distributionnelle est le suivant :*

- *Dans $\mathbb{R}^n$, s'il existe deux lignes $i' \neq i''$ telles que leurs profils sont identiques (ou encore équivalents distributionnellement) et si ces deux lignes sont remplacées par une seule : i''' avec*

  $$f_{i''',j} = f_{i',j} + f_{i'',j}, \forall j = 1, ..., p$$

  alors les distances définies en (2.3) sont inchangées.

- *Dans $\mathbb{R}^p$, s'il existe deux colonnes $j' \neq j''$ telles que leurs profils sont identiques (c'est-à-dire équivalents distributionnellement) et si ces deux colonnes sont remplacées par une seule : j''' avec*

  $$f_{i,j'''} = f_{i,j'} + f_{i,j''}, \forall i = 1, ..., n$$

  alors les distances définies en (2.4) sont inchangées.

Il y a donc la notion d'invariance.

Proposition 2.2.1 *La distance du chi-carré vérifie la propriété d'équivalence distributionnelle.*

Preuve :
La démonstration suivante est faite pour les lignes. La démarche est identique pour les colonnes.

Soit deux lignes $i' \neq i$ telles que leurs profils sont identiques. Ceci signifie que

$$f_{i',j} = f_{i'',j}, \forall j = 1, ..., p. \tag{2.5}$$

Le principe d'équivalence distributionnelle nous informe que nous considérons une nouvelle ligne i''' telle que

$$f_{i''',j} = f_{i',j} + f_{i'',j}, \forall j = 1, ..., p.$$

Par l'équation (2.5), nous pouvons réécrire cette expression

$$f_{i''',j} = 2f_{i',j}, \forall j = 1, ..., p.$$

L'équation (2.5) implique que le profil moyen de i''' est

$$f_{i''',.} = f_{i',.} + f_{i'',.} = 2f_{i',.}, \forall j = 1, ..., p.$$

Le profil moyen de ligne reste identique. Nous pouvons à présent détailler la distance entre une ligne i et la nouvelle ligne i'''.

$$\begin{aligned} d(i, i''') &= \sum_{j=1}^{p} \frac{1}{f_{.,j}} \left(\frac{f_{i,j}}{f_{i,.}} - \frac{f_{i''',j}}{f_{i''',.}} \right)^2 = \sum_{j=1}^{p} \frac{1}{f_{.,j}} \left(\frac{f_{i,j}}{f_{i,.}} - \frac{2f_{i',j}}{2f_{i',.}} \right)^2 \\ &= \sum_{j=1}^{p} \frac{1}{f_{.,j}} \left(\frac{f_{i,j}}{f_{i,.}} - \frac{\cancel{2}f_{i',j}}{\cancel{2}f_{i',.}} \right)^2 = \sum_{j=1}^{p} \frac{1}{f_{.,j}} \left(\frac{f_{i,j}}{f_{i,.}} - \frac{f_{i',j}}{f_{i',.}} \right)^2 \\ &= d(i, i'). \end{aligned}$$

□

Lorsque deux colonnes (ou lignes) ont des profils identiques, nous pouvons les fusionner en une seule colonne sans bouleverser les positions des autres points du graphique. Les distances calculées avant le changement restent identiques. En effet, la distance entre deux colonnes (ou lignes) qui ont le même profil est nulle. Par conséquent, les deux points n'en forment plus qu'un seul et la distance avec les autres éléments ne change pas.

Dans notre exemple, aucune paire de mots ou de documents n'a un profil identique, donc toutes les distances sont non nulles. Néanmoins, les lemmes "pouvoir" et "France" ont des profils très similaires et sont donc proches. Ils sont utilisés respectivement 489 et 190 fois. Les occurrences de "France" sont réparties entre les sept catégories de documents comme suit : 24.74%, 22.09%, 17.79%, 2.45%, 18.2%, 5.11%, 9.61%. La distribution des occurrences de "pouvoir" parmi les documents est très similaire, égale à 19.47%, 23.68%, 11.05%, 0%, 16.84%, 12.63%, 16.32%. La différence la plus importante, qui dépasse légèrement les 6.74%, est observée dans la catégorie "Giscard". La distance entre ces lemmes est de 0.221. Deux autres lemmes, "Europe" et "social", ont des profils lexicaux qui se ressemblent. Ils sont proches au sens de la distance du chi-carré. La distance entre ces deux lignes est de 0.156. C'est cette paire de lemmes qui a la plus petite distance.

2.2.5 L'inertie d'un nuage

L'inertie est une mesure de la dispersion d'un nuage. Si tous les profils sont égaux les uns par rapport aux autres, alors tous les points fusionnent avec le centre de gravité. Dans ce cas, l'inertie du nuage est égale à zéro. Nous pouvons écrire la distance qu'il y a entre le profil ligne i et le centre de gravité G_n par

$$d^2(i, G_n) = \sum_{j=1}^{p} \frac{1}{f_{.,j}} \left(\frac{f_{i,j}}{f_{i,.}} - f_{.,j} \right)^2.$$

De même, la distance entre le profil colonne j et le centre de gravité G_p est donnée par

$$d^2(j, G_p) = \sum_{i=1}^{n} \frac{1}{f_{i,.}} \left(\frac{f_{i,j}}{f_{.,j}} - f_{i,.} \right)^2.$$

La référence [12] nous montre que l'inertie du nuage des profils lignes N_n par rapport à son centre de gravité, notée $I(N_n)$, est exprimée comme suit :

$$\begin{aligned} I(N_n) &= \sum_{i=1}^{n} I(i) = \sum_{i=1}^{n} f_{i,.} d^2(i, G_n) \\ &= \sum_{i=1}^{n} f_{i,.} \sum_{j=1}^{p} \frac{1}{f_{.,j}} \left(\frac{f_{i,j}}{f_{i,.}} - f_{.,j} \right)^2 \\ &= \sum_{i=1}^{n} \sum_{j=1}^{p} \frac{(f_{i,j} - f_{i,.} f_{.,j})^2}{f_{i,.} f_{.,j}} \\ &= \frac{X^2}{k} = \phi^2. \end{aligned}$$

Lorsque les axes de référence sont orthogonaux, ce qui est le cas pour N_n et N_p, l'inertie d'un nuage est simplement la somme des inerties de tous les axes. Dans notre exemple, $\phi^2 = 0.2355$. Nous pouvons faire le même raisonnement pour le nuage N_p. Nous obtiendrons que

$$I(N_p) = I(N_n). \tag{2.6}$$

Tandis que le χ^2 quantifie l'indépendance entre les documents et les mots, l'inertie ϕ^2 mesure la force de la relation qu'il y a entre eux.

2.2.6 Axes factoriels

Les projections des deux nuages (profils lignes et profils colonnes) sur les axes principaux s'opèrent sur un même graphique. La projection d'un point i sur un axe factoriel noté F_s dans un espace bidimensionnel de variables X_1 et X_2 est représenté sur la Figure 2.1 qui nous permet de mieux visualiser la manière dont l'AFC projette. Grâce aux propriétés géométriques du produit scalaire, nous utilisons ce dernier pour trouver la projection orthogonale du vecteur $\overrightarrow{Gi}$ sur l'axe principal F_s donné par la formule suivante

$$\overrightarrow{Gi'} = \left(\frac{\overrightarrow{Gi} \bullet \overrightarrow{u_s}}{\overrightarrow{u_s} \bullet \overrightarrow{u_s}} \right) \overrightarrow{u_s},$$

où G est le centre de gravité du nuage considéré et i' la projection orthogonale de i [14].

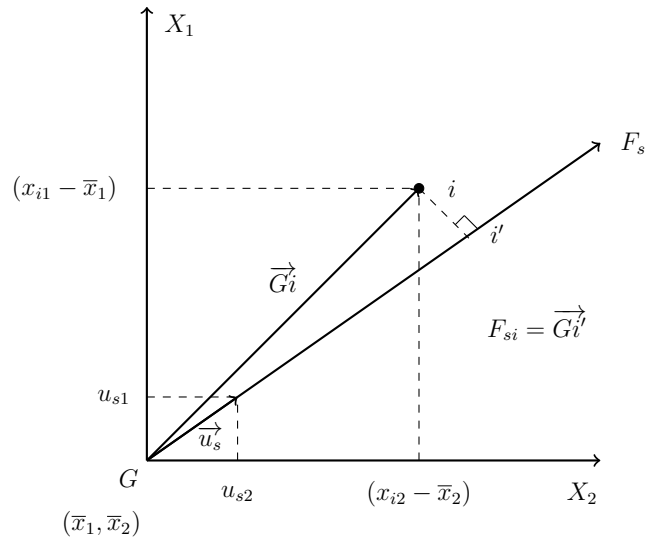

Figure 2.1 – Illustration d'une projection orthogonale, image inspirée de la source [15].

Il est nécessaire de connaitre le vecteur $\overrightarrow{u_s}$ pour chaque axe principal. Dans le cas des lignes, le premier axe u_1 est trouvé en maximisant l'inertie du nuage projeté. Nous sommes dans l'espace $\mathbb{R}^p$ où nous considérons les n lignes de F. Pour réaliser cette maximisation, nous avons besoin de plusieurs éléments. Soit $u \in \mathbb{R}^p$ un vecteur unitaire tel que $u'u = 1$. Soit $y = Fu$. Ses composantes sont les n projections des profils lignes de F sur le sous-espace unidimensionnel engendré par u. Nous souhaitons maximiser $||y||$, c'est-à-dire $y'y = u'F'Fu$, sous la contrainte que $u'u = 1$ pour trouver le vecteur u. La fonction à optimiser est

$$L(u) = u'F'Fu - \lambda(u'u - 1).$$

Nous dérivons cette expression

$$\frac{\partial f}{\partial u} = 2F'Fu - 2\lambda u = 0 \Rightarrow F'Fu = \lambda u \Rightarrow \lambda = u'F'Fu \text{ car } u'u = 1.$$

Nous obtenons

$$y'y = \lambda.$$

Lorsque $y'y$ vaut la plus grande valeur propre λ_1 de $F'F$, il est maximisé. Le vecteur u_1 est le vecteur propre de la matrice symétrique $F'F$ associé à λ_1. Les vecteurs de direction unitaire u_s correspondent en fait aux vecteurs propres de cette matrice. Le premier axe factoriel concorde avec le vecteur propre associé à la plus grande valeur propre de $F'F$. Nous la noterons λ_1. L'inertie pour cet axe vaut λ_1.

Chaque axe suivant, u_s, maximise l'inertie résiduelle. En effet, nous pouvons montrer la maximisation comme telle :

$$\max_{u\in\mathbb{R}^p} \quad u'F'Fu$$
$$\text{s.c.} \qquad \begin{cases} u'u = 1 \\ u'u_1 = 0. \end{cases}$$

La deuxième contrainte vient du fait que tous les axes sont orthogonaux entre eux. Lorsque nous projettons sur un sous-espace, nous avons besoin d'axes orthogonaux. La proposition suivante nous l'indique [10].

Proposition 2.2.2 *Soit $W = (w_1, ..., w_p)$ une matrice $n \times p$ $(p < n)$ telle que $W'W = I_p$ (les colonnes de W sont donc des vecteurs orthonormés). Notons $\pi_{a,W}$ le sous-espace affin*

$$\left\{ a + \sum_{j=1}^{p} \alpha_j w_j | \alpha_1, ..., \alpha_k \in \mathbb{R} \right\}$$

et $P_{a,W}$ l'application linéaire qui est la projection orthogonale sur $\pi_{a,W}$, $P_{a,W} : \mathbb{R}^p \to \pi_{a,W}$.

Nous avons

$$P_{a,W}(x) = a + W[W'W]^{-1}W'(x - a) = a + WW'(x - a).$$

À présent, nous pouvons résoudre le problème d'optimisation :

$$L(u) = u'F'Fu - \lambda(u'u - 1) - \gamma u'u_1.$$

$$\begin{aligned} \frac{\partial f}{\partial u} = 2F'Fu - 2\lambda u - 2\gamma u_1 = 0 &\Leftrightarrow 2u_1'F'Fu - 2\lambda u_1'u - 2\gamma \underbrace{u_1'u_1}_{=1\Rightarrow\gamma=0} = 0 \\ &\Leftrightarrow u_1'F'Fu = \lambda u_1'u \Leftrightarrow F'Fu = \lambda u \\ &\Leftrightarrow \lambda = u'F'Fu. \end{aligned}$$

λ est la deuxième valeur propre de la matrice $F'F$ et $u = u_2$ son vecteur propre associé. Le deuxième axe correspond au vecteur propre associé à la deuxième valeur propre, λ_2, et ainsi de suite.

Nous pouvons généraliser ce qui vient d'être dit. Si nous souhaitons les s meilleurs sous-espace, nous générons les s vecteurs propres associés aux s plus grandes valeurs propres $\lambda_1, ..., \lambda_s$ de la matrice $F'F$.

De plus, nous savons par (2.6) que les profils colonnes et les profils lignes ont la même inertie. Les valeurs propres sont donc identiques. Dans l'un ou l'autre espace, la recherche du premier axe d'inertie maximale suivi du second, nous donne la même solution pour les valeurs propres. C'est pourquoi, de manière analogue au raisonnement qui a été fait précédemment, dans l'espace colonne, le premier axe $v_1 \in \mathbb{R}^n$ maximise l'inertie du nuage

projeté et ${v'}_1 v_1 = 1$. Nous sommes dans l'espace $\mathbb{R}^n$ où nous considérons les p colonnes de F. Le vecteur v_1 sera le vecteur propre associé à la valeur propre λ_1 de la matrice FF'. Nous posons ici $y = F'v$. Ses composantes sont les p projections des profils colonnes de F sur le sous-espace unidimensionnel engendré par v. Nous souhaitons maximiser $||y||$ sous la contrainte que $v'v = 1$ pour trouver le vecteur v. Chaque axe suivant, $v_s \in \mathbb{R}^n$, maximise successivement l'inertie résiduelle.

2.2.7 Représentation simultanée des lignes et des colonnes

En AFC, la représentation simultanée des lignes et des colonnes fonctionne grâce à des relations de transitions aussi appelées "propriété barycentrique". La première concernant les lignes sur l'axe de rang s est

$$F_s(i) = \frac{1}{\sqrt{\lambda_s}} \sum_{j=1}^{p} \frac{f_{i,j}}{f_{i,.}} G_s(j)$$

où

- $F_s(i)$ est la coordonnée de la ligne i le long de l'axe de rang s ;
- λ_s est l'inertie associée à s ;
- $\frac{f_{i,j}}{f_{i,.}}$ est le $j^{ème}$ terme du profil i ;
- $G_s(j)$ est la coordonnée de la colonne j le long de l'axe de rang s.

Nous voyons qu'il est possible de relier les coordonnées des lignes avec celles des colonnes. Plus précisément, afin de calculer la coordonnée de la ligne i le long de l'axe s, nous utilisons la coordonnée de chacune des colonnes j pondérée par les profils. Le barycentre est d'autant plus écarté de l'origine que λ_s est petit : $\frac{1}{\sqrt{\lambda_s}} \geq 1$. En effet, plus le facteur $\frac{1}{\sqrt{\lambda_s}}$ est grand, plus la coordonnée va augmenter et s'éloigner du centre.

Définition 2.2.9 *La propriété barycentrique dit qu'une ligne est du côté des colonnes auxquelles elle s'associe le plus.*

Si $\frac{f_{i,j}}{f_{i,.}}$ est très grand, alors $G_s(j)$ compte beaucoup dans le calcul de la coordonnée de i. En AFC, ligne et colonne jouent des rôles symétriques

$$G_s(j) = \frac{1}{\sqrt{\lambda_s}} \sum_{i=1}^{n} \frac{f_{i,j}}{f_{.,j}} F_s(i).$$

Cette double propriété barycentrique permet une utilisation de la représentation simultanée.

2.2.8 Exemple

Nous appliquons maintenant toute cette théorie à l'exemple. Sur la Figure 2.2, nous observons le graphique fourni par l'analyse factorielle des correspondances pour l'exemple VOEUX. C'est une représentation bidimensionnelle générée par une analyse factorielle des correspondances de la Table 1.2 effectuée à l'aide du logiciel TXM. Il s'agit d'une représentation visuelle qui expose en même temps les similarités entre les profils lignes et les similarités entre les profils colonnes.

Si nous analysons ce qu'il se passe sur le premier axe factoriel, nous pouvons dire qu'il sépare les partitions "Sarkozy", "Hollande" et "Chirac" des autres. Cette séparation montre qu'ils utilisent un vocabulaire plus différent des autres. Nous expliquons ce phénomène par l'époque qui diffère entre les deux groupes. Le groupe "Sarkozy", "Hollande" et "Chirac" sont les discours les plus récents. Par conséquent, ils utilisent des mots différents et ne parlent pas forcément des mêmes évènements. Nous comparons le vocabulaire utilisé et les partitions de la Figure 2.2 grâce à la propriété barycentrique. Cette dernière nous indique qu'un lemme est du côté des partitions auxquelles il s'associe le plus et inversement.

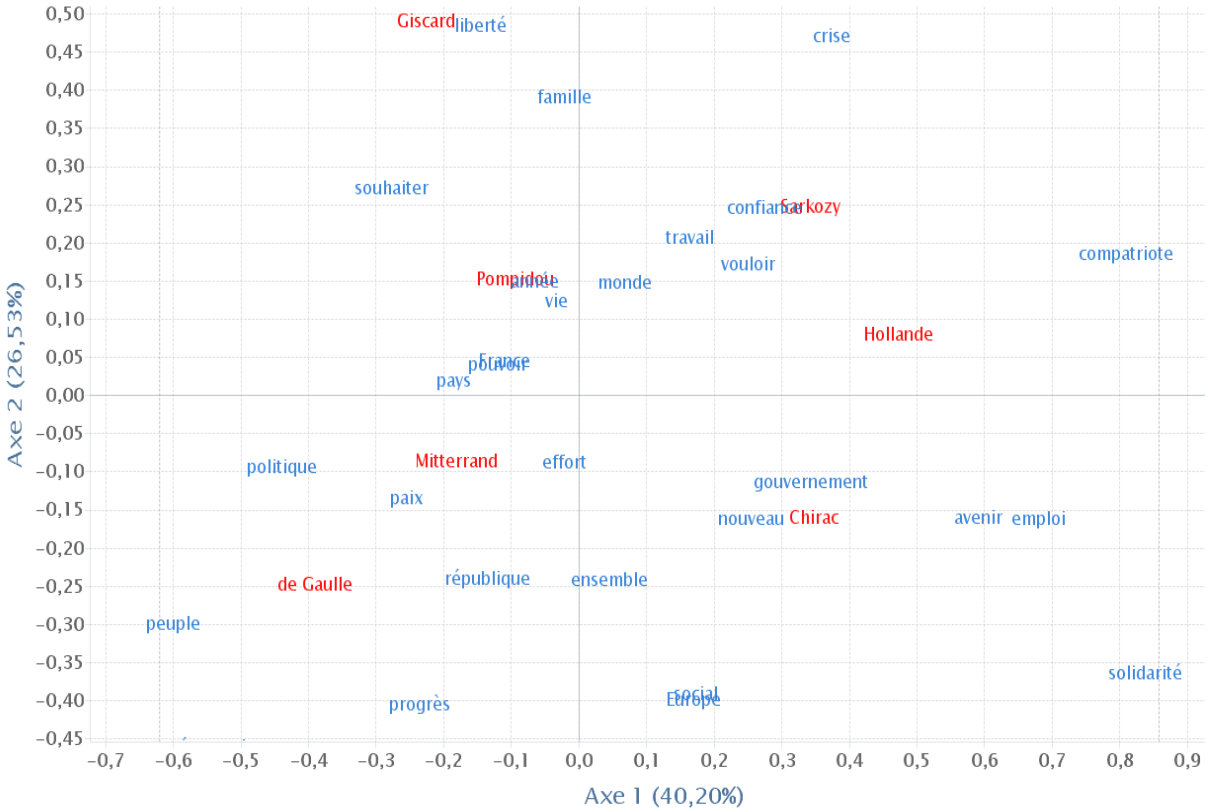

Figure 2.2 – Analyse factorielle des correspondances pour le jeu de données VOEUX réparti en fonction des présidents.

Le deuxième axe factoriel sépare grandement le président M. Valéry Giscard d'Estaing des autres. Nous interprétons ce résultat mathématique par le fait qu'il vient d'un parti très différent des autres (libéral) [16].

Nous observons sur la Figure 2.2 que le lemme "effort" est proche de l'origine, ce qui s'explique par le fait que sa proportion dans chaque catégorie (0.25, 0.23, 0.1, 0.03, 0.2, 0.05, 0.13) est très proche de la proportion globale ($\frac{915}{3264} = 0.28$). De la même manière, nous observons que la partie "Pompidou" a les mêmes propriétés et est donc proche de l'origine.

Si des lemmes ont une répartition presque identique, alors les deux lemmes seront proches. Nous observons sur la Figure 2.2 que les lemmes "France" et "pouvoir" sont très proches. En effet, pour chaque partition le lemme "France" (0.25, 0.22, 0.18, 0.02, 0.18, 0.05, 0.1) a une proportion similaire au lemme "pouvoir" (0.19, 0.24, 0.11, 0, 0.17, 0.13, 0.16). De la même manière, nous remarquons que les parties "Hollande" et "Sarkozy" ont les mêmes propriétés et sont donc proches.

Sur la Figure 2.3, nous trouvons les données associées à la Figure 2.2. La Figure 2.3 indique les valeurs propres de la matrice $F'F$. Ces dernières sont comprises entre 0 et 1 en analyse des correspondances. En effet, les matrices que nous analysons sont des matrices de fréquences. La trace de cette matrice est la somme de ces éléments diagonaux. Par conséquent, la trace sera plus petite que 1. De plus, la trace peut aussi être exprimée comme une somme des valeurs propres de la matrice. Dès lors, elles doivent être comprises entre 0 et 1 pour que leur somme ne dépasse pas 1. Ces valeurs propres mesurent les variances le long de chaque axe principal. Ces valeurs propres sont $\lambda_1 = 0.0947$ pour le premier axe et $\lambda_2 = 0.0625$ pour le deuxième axe dans le cadre de notre exemple. Les pourcentages de variance sont calculés en faisant le rapport entre chaque valeur propre et leur somme globale multipliée par 100. Ils sont égaux à 40.20% et 26.53% pour les deux premiers axes. Les pourcentages de variance mesurent l'importance relative de chaque

#	Valeur propre	%	∑%
1	0,0947	40,20	40,20
2	0,0625	26,53	66,74
3	0,0351	14,93	81,66
4	0,0228	9,67	91,33
5	0,0144	6,12	97,45
6	0,0060	2,55	100,00

Figure 2.3 – Données associées à l'analyse factorielle des correspondances pour le jeu de données VOEUX réparti en fonction des présidents.

valeur propre dans la trace. Dans ce cas, le plan correspondant aux deux premiers axes principaux explique 66.74% de la variance totale.

2.2.9 Comparaison avec l'ACP

Pour terminer cette section, nous pouvons comparer l'analyse factorielle des correspondances (AFC) à l'analyse en composantes principales (ACP). La source [11] nous informe que l'ACP met en évidence certaines relations que nous ne pouvons apercevoir dans les données. Tout comme l'AFC, elle fait en sorte que le jeu de données de taille $n \times p$ soit réduit et qu'il y ait une perte d'informations la plus petite possible. Pour finir, elle permet aussi de détecter les valeurs aberrantes.

L'ACP et l'AFC peuvent se baser sur des tableaux de fréquences ou des tableaux binaires. Lorsque nous utilisons les tableaux binaires, L. Lebart (*et al.*) [6] nous informent que le pourcentage de variance expliquée est très léger pour les deux méthodes. Ce pourcentage donne donc toujours une vision plus pessimiste de l'information extraite.

L. Lebart (*et al.*) [6] nous indiquent que plusieurs expériences montrent qu'il faut privilégier l'utilisation de l'AFC à l'ACP dans le cadre de données binaires ou de tableaux de contingence. Il est important de préciser que ce n'est pas pour autant que l'ACP ne donne pas de bons résultats. Chaque outil fournit un point de vue spécifique sur l'ensemble des données. Dans certaines situations, l'ACP, même si elle n'est pas optimale, pourrait être utilisée et donner des résultats plus intéressants que d'autres techniques.

2.3 La classification ascendante hiérarchique

Les méthodes de classification ascendante hiérarchique sont des méthodes de clustering. Elles créent des regroupements de lignes ou de colonnes d'un tableau Z en clusters. Il s'en suit que soit l'ensemble des colonnes, soit l'ensemble des lignes est analysé. Dans le cadre de notre exemple, nous analyserons dans un premier temps les colonnes (les présidents) et dans un second temps les lignes (les lemmes).

2.3.1 Utilisation des mêmes outils que pour l'AFC

Grâce à [6] et [9] nous savons que la classification ascendante hiérarchique se base sur un ensemble de départ de n ou p éléments en fonction que nous analysons respectivement les lignes ou les colonnes. Tout comme l'analyse factorielle des correspondances, l'analyse en cluster hiérarchique est appliquée à des matrices de lignes et de colonnes telles que les tableaux lexicaux décrits précédemment (Table 1.2). Elle utilise la même métrique du chi-carré que l'analyse factorielle des correspondances, ce qui nous permet d'avoir une compatibilité des résultats. Pour cette méthode, cette métrique nous permettra de faire les regroupements entre les éléments. Ces regroupements se font sur base des distances qui

sont calculées grâce aux équations (2.3) et (2.4) et utilisent donc aussi les profils lignes (2.1) ou colonnes (2.2).

2.3.2 Fonctionnement de l'algorithme

Le fonctionnement de l'algorithme de la classification ascendante hiérarchique est décrit dans cette section et expliqué par [17]. Pour commencer, chaque point représente un seul individu. Nous devons d'abord calculer la distance entre tous les points en créant une matrice de distance [18]. L'objectif est de trouver dans cette matrice la plus petite distance du chi-carré entre deux points. L'arbre hiérarchique commence en regroupant les deux points les plus proches. Ces derniers ne forment plus qu'un groupe. Il faut donc calculer une nouvelle matrice de distance en trouvant la distance entre chaque point et ce nouveau groupe. Pour cela, nous disposons de trois méthodes.

Définition 2.3.1 *La méthode appelée "saut minimum" veut que l'éloignement entre deux groupes q et q' soit égale à la plus petite distance entre un document dans q et un document dans q'. En mathématique, ce lien est donné par*

$$L(q, q') = min(d(l, g)), l \in q, g \in q'.$$

Ce critère peut conduire à des clusters dispersés. En effet, cette méthode vise à regrouper les points de proche en proche et dans certains cas, l'arbre ne mettra pas en évidence des clusters séparés. Il peut en fait y avoir un effet de chaine qui va impacter négativement la classification. Soit, sur la Figure 2.4, la situation de deux classes éloignées (ici bleue et rouge) qui sont reliées par un faible nombre d'observations proches les unes des autres créant une liaison entre les deux groupes. Dès lors, le regroupement basé sur la distance minimale ne considérera pas trois clusters - les deux groupes séparés (bleu et rouge) et le lot d'observations entre eux deux (noir) - mais un unique cluster. Nous aurons donc un cluster très dispersé.

Définition 2.3.2 *La méthode du "lien complet" dit que l'écart entre les clusters q et q' est égale à la plus grande distance entre un document de q et un document de q'. En mathématique, ce lien est donné par*

$$L(q, q') = max(d(l, g)), l \in q, g \in q'.$$

Cela conduit à des clusters concentrés. En effet, nous aurons l'effet inverse de celui de la méthode du "saut minimum". Il y a un risque de fusion entre des clusters proches trop tardive pour être optimale. Il aura tendance à faire beaucoup de petits clusters.

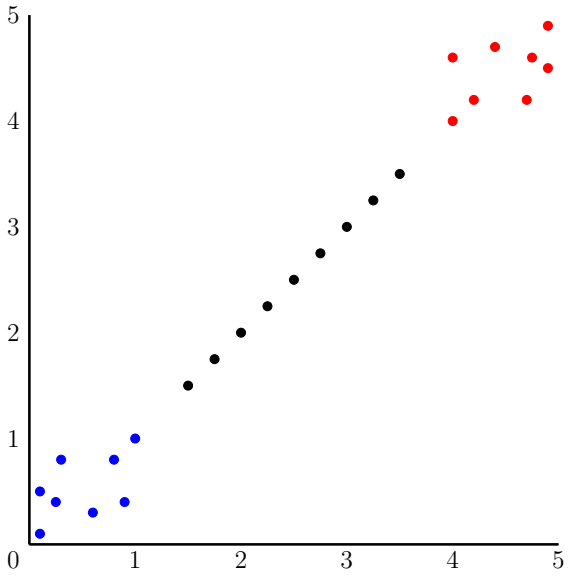

Figure 2.4 – Illustration de l'effet de chaine de la méthode du saut minimum inspirée de la source [12].

Définition 2.3.3 *La méthode de Ward préfère prendre en compte l'inertie. La distance entre deux clusters q et q' est mesurée par la diminution de l'inertie entre les deux clusters. Si les deux clusters sont agrégés, l'inertie entre clusters diminue de*

$$\delta(q, q') = \frac{b_q b_{q'}}{b_q + b_{q'}} d^2(C_q, C_{q'})$$

où C_q est le centre de gravité du cluster q et chaque cluster q se voit attribuer un poids $b_q, q = 1, ..., Q$ égal à la somme des poids des documents qui le composent.

Cette méthode permet de regrouper des clusters ayant des centres de gravité proches et d'éviter les effets de chaînes évoqués précédemment. Nous précisons qu'un individu peut être un cluster déjà créé ou un lemme/partition isolé.

À l'étape suivante, une fois la nouvelle matrice de distance calculée, les deux individus les plus proches sont agglomérés et ainsi de suite jusqu'à ce qu'ils soient tous réunis.

L'analyse en cluster ainsi obtenue peut être représentée de plusieurs manières différentes. La représentation sous forme d'arbre hiérarchique ou de dendrogramme est la plus révélatrice.

Définition 2.3.4 *Un dendrogramme est une représentation en arbre des regroupements hiérarchiques faits lors de l'algorithme de la CAH.*

Les regroupements obtenus à chaque étape de l'algorithme de clustering hiérarchique rassemblent les éléments les plus proches les uns des autres. Ils constituent un noeud de la hiérarchie et leur distance est l'indice attaché au noeud. Lorsque nous évoluons dans la procédure, le nombre d'éléments déjà agglomérés est de plus en plus conséquent. En effet, la plus petite distance entre les clusters qui reste à agglomérer est de plus en plus grande.

Le clustering est très important car il permet de mettre en évidence des groupes de mots qui sont utilisés ensemble ou de la même manière et des groupes de partitions qui utilisent le même vocabulaire ou de la même façon. Un autre point fort du clustering est que la dimension des résultats est réduite. En effet, cela se produit lorsque nous les regroupons en clusters.

2.3.3 Inertie inter-cluster et intra-cluster

Tout comme pour l'analyse des correspondances, nous pouvons définir la notion de nuage et d'inertie pour la classification ascendante hiérarchique. Le nuage de point N_n est partitionné en Q clusters avec $q = 1, ..., Q$. Chacun des clusters comporte I_q éléments. Nous symboliserons le centre de gravité de chaque cluster par C_q. Pour calculer l'inertie totale, M. Bécue-Bertaut [9] nous donne les définitions suivantes.

Définition 2.3.5 *L'inertie inter-cluster est la mesure de dispersion qu'il y a entre deux classes et est calculée de la manière suivante :*

$$I_{inter} = \sum_{q=1}^{Q} \sum_{s=1}^{S} b_q F_s^2(C_q),$$

où la fonction $F_s(x)$ nous donne la coordonnée de x sur l'axe s de l'analyse des correspondances et $b_q = \sum_{i=1}^{I_q} b_i$.

Si les clusters sont fortement dispersés, la coordonnée des centres de gravité de chaque cluster sur les différents axes ($F_s(C_q)$) sera plus grande, c'est-à-dire éloignée du centre. Par conséquent, si nous additionnons ces coordonnées éloignées, nous aurons une inertie inter-cluster importante. Le carré de la somme est crucial car il permet de prendre en compte les coordonnées négatives à leur juste valeur. En effet, sans le carré, deux clusters avec les coordonnées opposées donneraient une somme nulle. Par conséquent, nous dirons que l'inertie inter-cluster est nulle et qu'ils ne sont pas dispersés, ce qui ne serait pas correct. Il est aussi important de préciser que nous considérons chaque axe s. De fait, les clusters peuvent être plus dispersés sur certains axes que pour d'autres. L'inertie inter-cluster tient compte de toutes ces informations.

Définition 2.3.6 *L'inertie intra-cluster est la mesure de dispersion qu'il y a à l'intérieur d'une classe et son équation est la suivante :*

$$I_{intra} = \sum_{q=1}^{Q} \sum_{i=1}^{I_q} \sum_{s=1}^{S} b_i \left(F_s^2(i) - F_s^2(C_q) \right),$$

où la fonction $F_s(x)$ nous donne la coordonnée de x sur l'axe s de l'analyse des correspondances.

L'inertie intra-cluster calcule l'inertie qu'il y a dans chacun des clusters. Elle compare donc la coordonnée d'un élément du cluster avec son centre de gravité. Si l'élément est proche du centre, alors le cluster est peu dispersé. De même que pour l'inertie inter-cluster, nous considérons tous les axes s.

Théorème 2.3.1 *Théorème de Huygens*
L'inertie totale est la somme de l'inertie inter-cluster et de l'inertie intra-cluster.

$$I_{tot} = \sum_{q=1}^{Q} \sum_{i=1}^{I_q} \sum_{s=1}^{S} b_i F_s^2(i).$$

Preuve :

$$\begin{aligned}
I_{tot} &= \sum_{q=1}^{Q} \sum_{i=1}^{I_q} \sum_{s=1}^{S} b_i F_s^2(i) \\
&= \sum_{q=1}^{Q} \sum_{i=1}^{I_q} \sum_{s=1}^{S} b_i \left(F_s^2(i) + F_s^2(C_q) - F_s^2(C_q) \right) \\
&= \sum_{q=1}^{Q} \sum_{i=1}^{I_q} \sum_{s=1}^{S} b_i \left(F_s^2(i) - F_s^2(C_q) \right) + F_s^2(C_q) \\
&= \sum_{q=1}^{Q} \sum_{i=1}^{I_q} \sum_{s=1}^{S} b_i \left(F_s^2(i) - F_s^2(C_q) \right) + \sum_{q=1}^{Q} \sum_{i=1}^{I_q} \sum_{s=1}^{S} b_i F_s^2(C_q) \\
&= \sum_{q=1}^{Q} \sum_{i=1}^{I_q} \sum_{s=1}^{S} b_i \left(F_s^2(i) - F_s^2(C_q) \right) + \sum_{q=1}^{Q} \sum_{s=1}^{S} b_q F_s^2(C_q) \\
&= I_{intra} + I_{inter}.
\end{aligned}$$

□

2.3.4 Partitionnement

Dans cette section, nous découpons l'arbre hiérarchique en plusieurs clusters. En définissant un niveau de coupure sur l'arbre, nous construisons une partition. Il n'y a pas une seule réponse et un seul nombre de clusters correct. L'objectif est de faire le choix le plus judicieux. Une partition est de bonne qualité si elle dispose de deux critères :

- les individus d'un même groupe sont très proches (variabilité intra-cluster petite) ;
- deux groupes différents sont éloignés (variabilité inter-cluster grande).

Donc, chaque partition possède des caractéristiques identiques et différentes des autres clusters. Par conséquent, le critère de décision pour choisir le nombre de clusters devra prendre en compte l'inertie intra-cluster et l'inertie inter-cluster.

Pour un nombre fixe de clusters Q, plus l'inertie intra-cluster est faible, meilleure est la qualité de la partition. En effet, une faible inertie intra-cluster signifie que les points ayant des profils lexicaux similaires sont placés dans les mêmes groupes. Étant donné que l'inertie totale est constante, si l'inertie intra-cluster diminue, il en résulte que l'inertie inter-cluster augmente. Ainsi, pour un Q fixe, minimiser l'inertie intra-cluster est équivalent à maximiser l'inertie inter-cluster.

Définition 2.3.7 *Pour un Q fixe, la mesure de la qualité de la partition est le rapport suivant :*

$$\frac{\textit{inertie inter-cluster}}{\textit{inertie totale}} = \frac{\sum_{q=1}^{Q}\sum_{s=1}^{S} b_q F_s^2(C_q)}{\sum_{q=1}^{Q}\sum_{i=1}^{I_q}\sum_{s=1}^{S} b_i F_s^2(i)}.$$

Cet indicateur prend des valeurs comprises entre 0 (pas d'inertie entre clusters) et 1 (l'inertie entre clusters est égale à l'inertie totale). Ce rapport correspond au pourcentage de la variabilité totale représentée par la partition donnée.

La décision du nombre de clusters Q doit être le juste milieu entre une inertie inter-cluster maximale et ne pas avoir trop peu de classes. En fait, il faut un juste milieu entre prendre un seul cluster avec toutes les partitions et prendre un cluster pour chaque partition. Ces deux extrêmes ne conviennent pas. Le choix de Q se fait à partir de l'inertie inter-cluster que nous souhaitons maximiser. Nous regardons lorsque le saut de l'inertie inter-cluster entre deux choix de nombre de classes est grand. Le nombre de classes associé à un grand saut sera choisi.

2.3.5 Exemple

Dans cette section, nous appliquons la théorie vue précédemment à l'exemple sur le jeu de données VOEUX. Nous représentons la Table 1.2 sous forme d'arbre. En pratique,

l'interprétation d'une analyse en cluster est effectuée en considérant les regroupements qui apparaissent aux deux extrémités du dendrogramme :

- la partie inférieure de la hiérarchie où il y a une distance très petite entre groupes ;
- la partie supérieure de la hiérarchie où il y a les ensembles de lemmes ou de parties qui sont très éloignés.

Sur la Figure 2.5, nous présentons la classification ascendante hiérarchique de la Table 1.2 où nous analysons les colonnes. Nous observons que la quatrième partition "Hollande" et la septième "Sarkozy", étant très proches, elles sont agglomérées à la première itération et forment un nouveau groupe. Nous le noterons le groupe 8. À la deuxième étape, ce huitième groupe est couplé avec la sixième partition "Pompidou" et, à elles deux, formeront le groupe 9. Pour l'étape suivante, ce sont les partitions 2 "de Gaulle" et 5 "Mitterrand" qui sont réunis et ainsi de suite. Nous appelerons la Figure 2.5 le dendrogramme pour les colonnes du jeu de données VOEUX. Il n'est pas forcément unique et peut, dans certains cas, être représenté différemment.

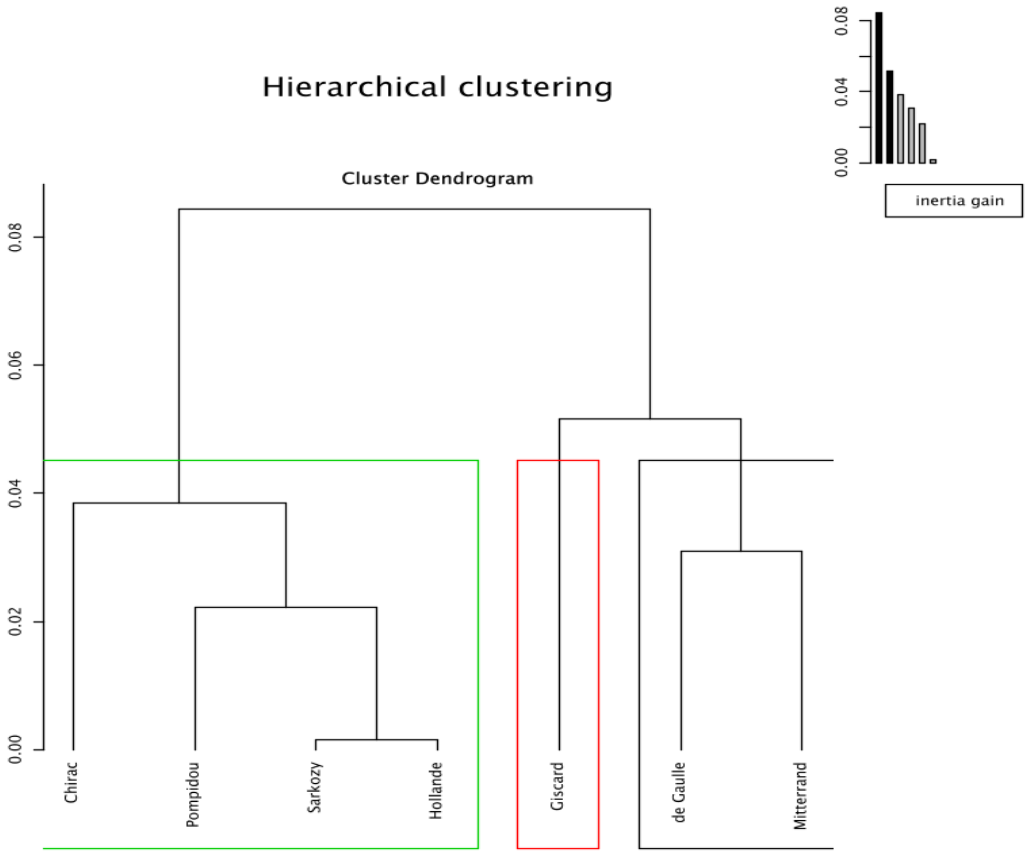

Figure 2.5 – Dendrogramme du jeu de données VOEUX réparti en fonction des présidents (colonnes).

Une fois le dendrogramme réalisé, nous pouvons décider de définir un niveau de découpage de l'arbre. Nous aurons donc un ensemble de plusieurs clusters. Le découpage effectué sur le dendrogramme de la Figure 2.5 se compose de trois clusters : les parties "Chirac", "Pompidou", "Sarkozy" et "Hollande" (en vert), la partie "Giscard" (en rouge) et les parties "de Gaulle" et "Mitterrand" (en noir). Nous observons que la partition "Giscard" forme un cluster à elle seule. Nous pouvons interpréter cela par le fait que M. Valéry Giscard d'Estaing vient d'un parti libéral et pas les autres présidents [16].

Nous pouvons observer la présence d'un histogramme en haut à droite de la Figure 2.5. Il y a six bâtons sur ce graphique et ils correspondent aux indices des noeuds les plus élevés jusqu'au noeud le plus bas du dendrogramme. Ce diagramme montre la progression de la valeur du critère de Ward, ici égale aux gains d'inertie entre clusters. Cela permet de choisir une coupe dans l'arbre correspondant à un grand saut entre les valeurs successives de celui-ci.

Pour plus de précisions sur le choix du nombre de classes, nous représentons à la Figure 2.6 l'évolution de l'inertie au fur et à mesure du processus de regroupement. Lorsque le nombre de classes est de trois, le saut de l'inertie inter-cluster est grand et un nombre de classes pas trop petit. Nous aurions pu choisir deux ou quatre clusters car les conditions énoncées sont remplies. Il s'agit ici d'un choix subjectif entre ces trois possibilités. Ce graphique a été créé grâce à la fonction *hclust* en R qui calcule le tableau de distance

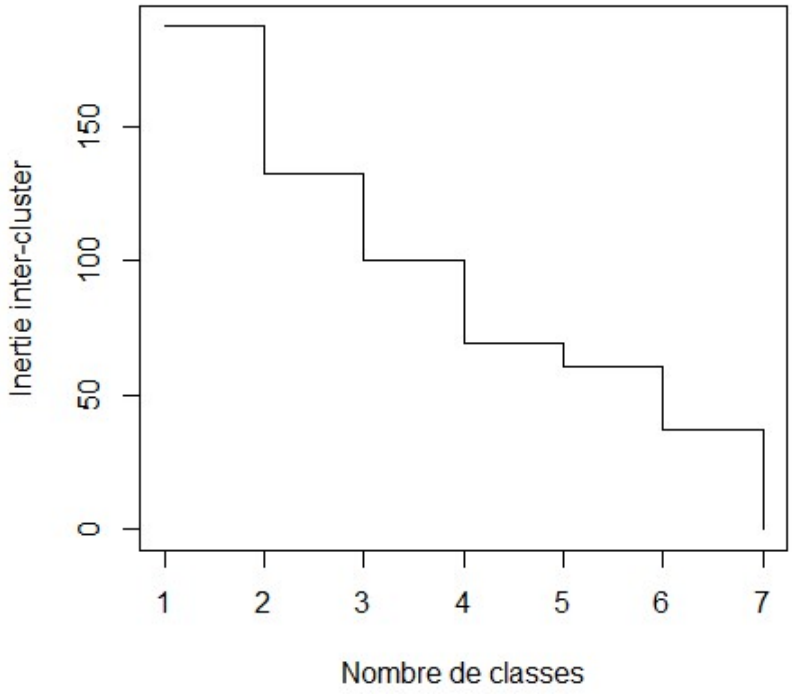

Figure 2.6 – Evolution de l'inertie inter-cluster au cours des regroupements des partitions.

avec la méthode de Ward. Elle renvoie un objet de classe hclust qui décrit l'arbre produit par le processus de clustering. L'objet est une liste comportant plusieurs composants. Celui qui nous intérresse ici est la "hauteur". C'est un ensemble de valeurs réelles qui représente la hauteur de regroupement, c'est-à-dire la valeur du critère de Ward associée au regroupement pour l'agglomération particulière. Cette fonction nécessite le package *fastcluster*.

Jusqu'à présent, nous avons analysé les colonnes de notre tableau lexical, c'est-à-dire les différents présidents. Maintenant, nous pouvons aussi dessiner le dendrogramme pour les lignes et donc les lemmes. Il est représenté par la Figure 2.7 avec trois clusters mis en évidence par les encadrés de couleur vert, noir et rouge. En effet, le choix de trois clusters correspond au saut le plus grand dans l'histogramme.

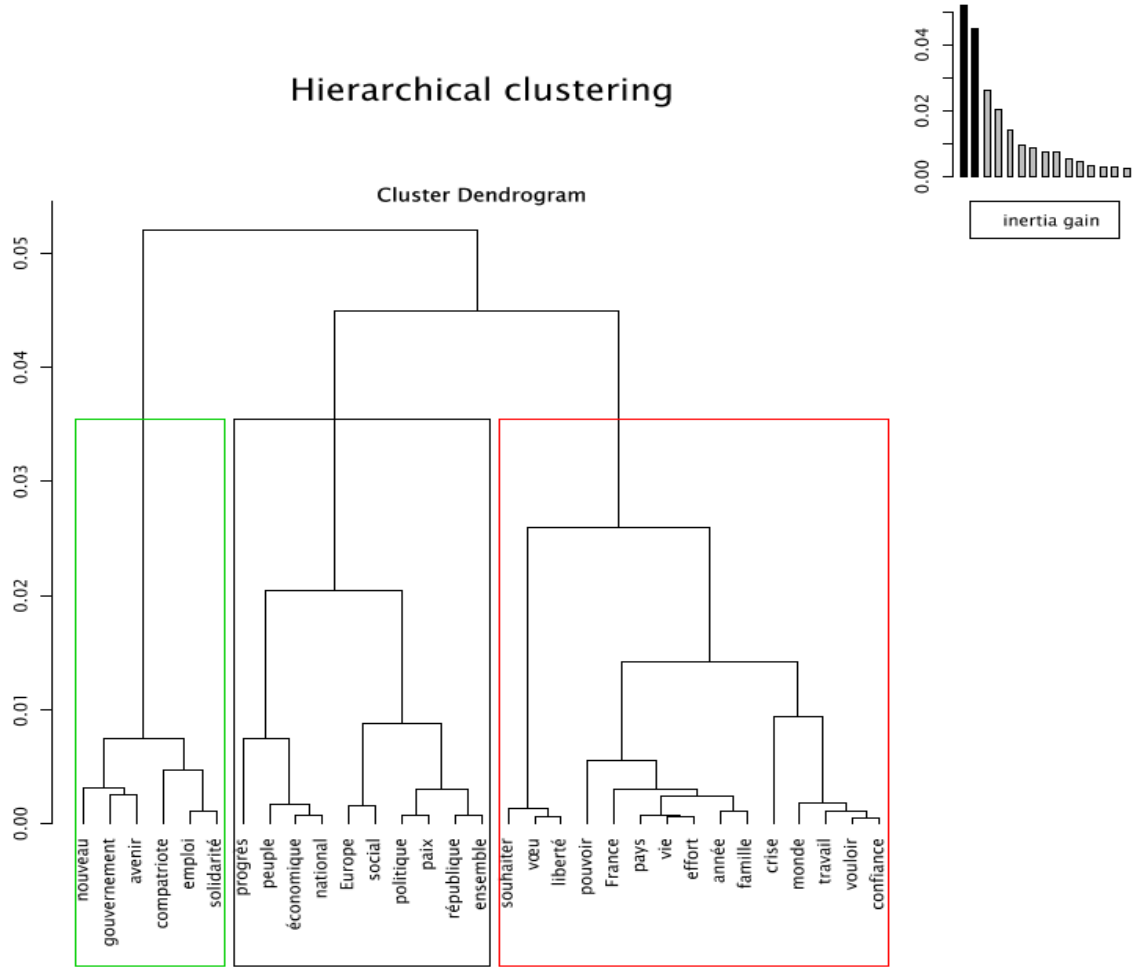

Figure 2.7 – Dendrogramme du jeu de données VOEUX réparti en fonction des lemmes (lignes)

Sur la Figure 2.7, nous lisons que, dans le troisième cluster, les mots “pays”, “vie” et “effort” s’agglomèrent très tôt et sont donc très proches. Nous observons dans le troisième cluster que les mots “vouloir” et “confiance” se regroupent très vite, mais que le cluster {monde, travail, vouloir, confiance} ne s’unit au mot “crise” que beaucoup plus tard. C’est ainsi que les regroupements se forment au fur et à mesure.

2.3.6 Association AFC et clustering

Une comparaison avec la Figure 2.2 de la section précédente est intéressante. Les principaux regroupements observés sur la carte de cette figure sont ceux qui décrivent le processus de clustering hiérarchique mais avec quelques petites différences. Nous pouvons comparer l’analyse des correspondances et l’analyse en cluster car elles décomposent la même quantité de deux manières différentes (chi-carré).

Nous remarquons que les lemmes qui sont contenus dans le premier cluster du dendrogramme de la Figure 2.8 se trouvent bien au même endroit sur la représentation graphique de l’analyse des correspondances (coin inférieur droit). Le deuxième cluster contient lui aussi des mots qui sont regroupés dans la Figure 2.2 (coin inférieur gauche). Nous pouvons conclure la même chose pour le troisième cluster (haut du graphique). Pour résumer ceci, nous observons la Figure 2.8 où une analyse factorielle des correspondances est faite avant la classification ascendante hiérarchique sur les lignes.

Concernant les colonnes, nous observons que les parties “Hollande” et “Sarkozy” sont proches sur la représentation graphique de l’analyse factorielle des correspondances et que ce sont les premières à se regrouper sur le dendrogramme de la Figure 2.9. Le cluster à droite reprenant “de Gaulle” et “Mitterrand” est bien proche sur la Figure 2.2. Pour l’analyse des correspondances, ce dernier est plus proche de la partie “Pompidou” que ne l’est le cluster en rouge des parties “Chirac”, “Hollande” et “Sarkozy”. L’analyse factorielle des correspondances et le clustering se complètent réellement. Nous le constatons plus clairement à l’aide de la Figure 2.9.

La classification ascendante hiérarchique vient compléter les informations obtenues avec l’analyse factorielle des correspondances. Cette dernière met en évidence la proximité entre les lemmes et entre les parties. Lorsqu’il y a un grand nombre de lemmes ou de documents, la représentation graphique peut vite devenir illisible. Il devient difficile de voir facilement les positions des lemmes en observant juste le graphique. De la même manière, lorsque le corpus est très long, le nombre de lemmes augmente très rapidement même si le nombre de parties n’est pas trop grand.

Revenons maintenant aux quelques différences de ces méthodes. Par exemple, les points “France” et “pouvoir” sont proches sur la Figure 2.2. Le dendrogramme montre au contraire que le point “France” est plus proche du cluster {pays, vie, effort, année

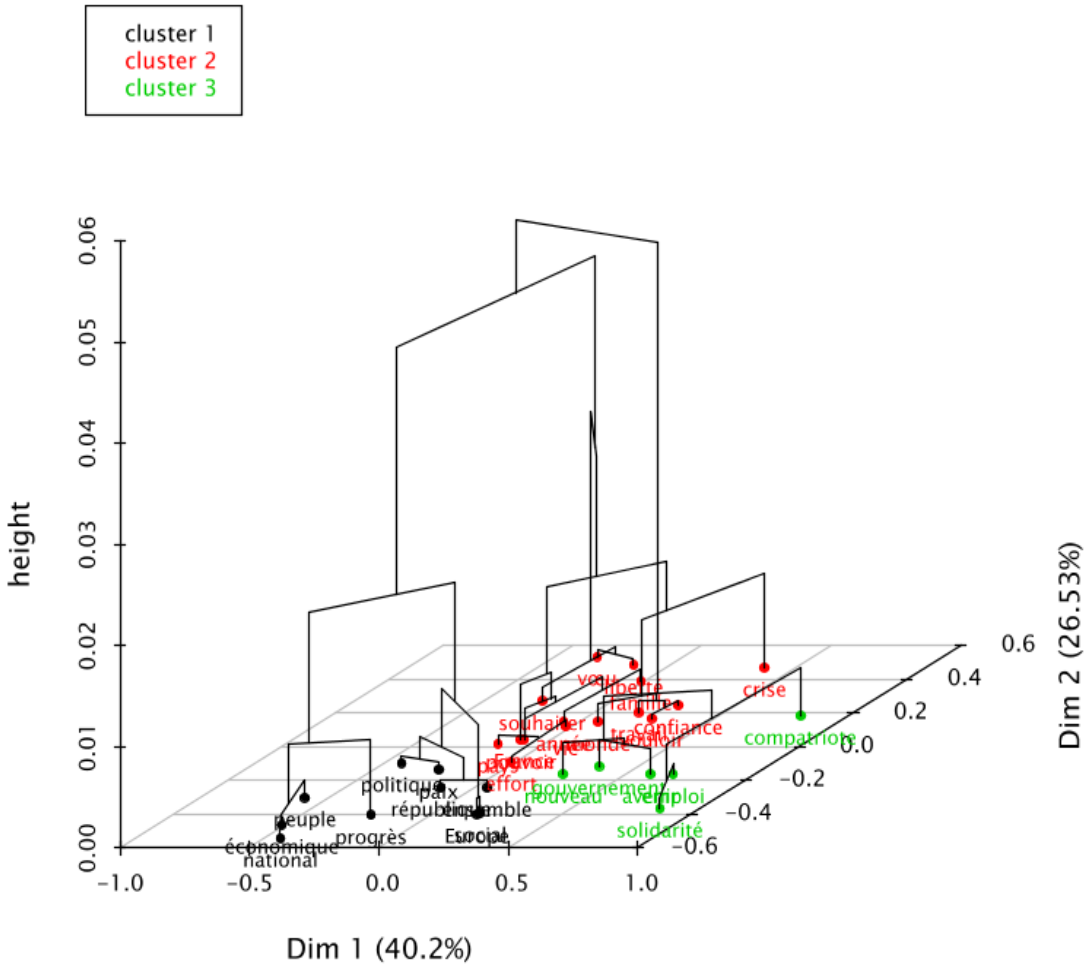

Figure 2.8 – Regroupement hiérarchique sur le plan factoriel pour les lemmes.

et famille} que du point “pouvoir”. Sur la représentation de l’analyse factorielle des correspondances, nous observons aussi que les points “avenir” et “emploi” sont très proches mais ne sont regroupés qu’après plusieurs itérations sur le dendrogramme de la Figure 2.7. Ces différences viennent du fait qu’avec l’AFC, il y a une petite perte d’informations qui explique ces changements. Il ne faut donc pas en conclure que l’une des deux méthodes donne de mauvais résultats. Ce sont deux visualisations complémentaires.

Il est important de noter que nous pourrions observer une différence entre le dendrogramme d’une table lexicale et le dendrogramme fait à partir de l’analyse factorielle des correspondances. En effet, l’AFC perd de l’information et donc lorsque nous créons la classification hiérarchique dessus, nous avons moins d’informations. Cependant, nous estimons que ce dernier est suffisant et représente suffisamment la réalité. Par conséquent, dans la suite de ce manuscrit, nous n’observerons pas systématiquement les deux figures. L’association des deux méthodes semble être un bon résumé.

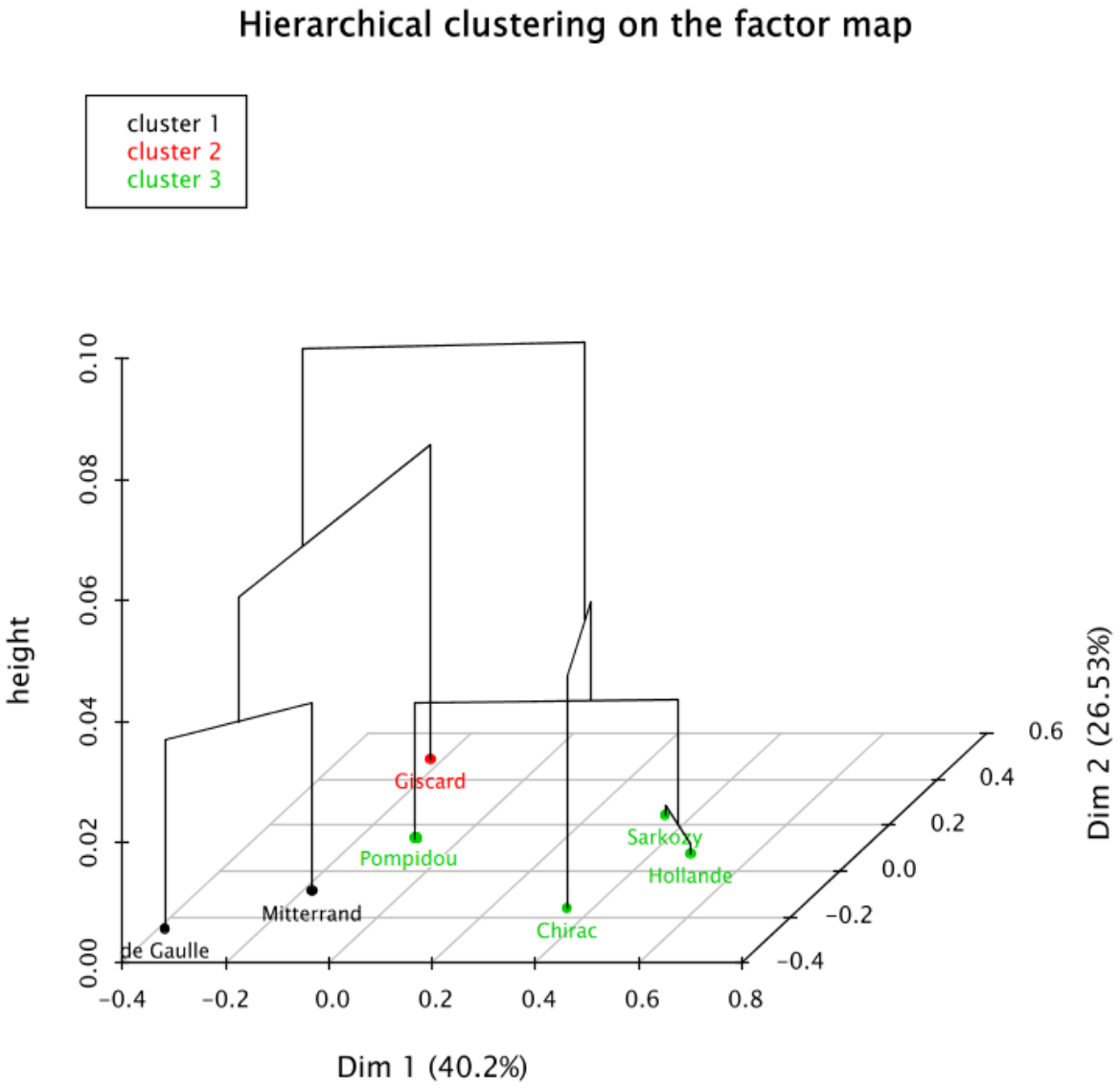

Figure 2.9 – Regroupement hiérarchique sur le plan factoriel pour les présidents.

Les méthodes d'axes principaux et de clustering sont utilisées ensemble. En effet, le clustering produit un regroupement basé sur la distance calculée dans l'ensemble de l'espace tandis que les méthodes d'axes principaux calculent la distance sur les premières composantes principales. La classification ascendante hiérarchique permet de rectifier les petites erreurs qui auraient pu être commises lors de la projection dans un espace de dimension inférieure.

Chapitre 3

Traitement d'un sujet de société

À présent, l'objectif est d'appliquer les méthodes statistiques vues dans les chapitres précédents à un jeu de données plus conséquent. Nous analysons et interprétons les résultats obtenus. Nous pouvons, grâce à ces derniers, répondre à différentes questions d'intérêts. Le sujet choisi concerne la manière dont la presse quotidienne évoque le réchauffement climatique. Nous travaillons dans ce chapitre avec le logiciel TXM et des codes constitués en R.

3.1 Création du Corpus

Pour commencer, nous construisons la base de données associée à la thématique. Pour cela, nous sélectionnons quatre journaux quotidiens. Cette décision s'est réalisée de manière simple. Tout d'abord, le choix s'est orienté vers les journaux français qui ont une orientation politique plus marquée que ceux de la Belgique. Par conséquent, les journaux d'orientations politiques différentes ont été retenus. Ensuite, il est intéressant de comparer ces trois journaux à un journal belge. Les caractéristiques des quatres journaux désignés sont données à la Table 3.1.

Table 3.1 – Table des spécificités de quatre journaux.

Journal	Pays	Orientation politique	Nombre articles sélectionnés	Référence
Le Figaro	France	droite	102	[20]
FranceSoir	France	centre	74	[21]
Libération	France	gauche	103	[22]
Le Soir	Belgique	progressiste[1]	103	[23]

[1]Le progressisme est une philosophie politique fondée sur l'idée que les progrès scientifiques, technologiques, économiques ainsi que les réformes sociales permettront une amélioration des conditions de vie de l'humanité. [19]

Le journal *Libération* est paru pour la première fois le 18 avril 1973. C'est une presse quotidienne française qui produit une version papier et une version en ligne tous les matins. Les fondateurs de ce journal sont Jean-Paul Sartre et Maurice Clavel. Au départ, ce journal était dirigé vers l'extrême gauche et a, par la suite, quoique plus modéré, conservé son orientation de gauche [24].

Le journal *FranceSoir* est mené par Xavier Azalbert depuis 2019. En effet, avant cette date, ce journal était connu sous le nom de "*France-Soir*" et a été créé en 1944. Ce journal français publiait la version papier et c'est seulement en 2013 que la version numérique est arrivée. L'écriture des articles est faite par l'entrepreneur lui-même mais aussi par des bénévoles parfois anonymes. C'est une presse quotidienne généraliste. Ce journal est un média indépendant. Il n'appartient ni à un grand groupe ni à de grands chefs d'entreprise. Ce journal est gratuit dans le but d'être accessible à tous. Cependant, le nombre d'articles disponibles est plus restreint que les autres [21][25].

Le journal *Le Figaro* a débuté la publication de ses articles de presse en 1826 sous le régime de Charles X. Il fait partie des plus anciens journaux français qui publient encore actuellement. Son point de vue s'oriente vers la droite gaulliste, libérale et conservatrice. Ce sont donc des personnes de droite ou de centre droite qui consultent ce journal [26].

Le journal *Le Soir* est fondé par Emile Rossel en 1887. Ce journal belge est généraliste et indépendant. Il se veut progressiste. Son orientation politique est du centre et de tradition libérale. Il fait partie des journaux les plus consultés en Belgique. Il est vendu en version papier mais est aussi disponible en version numérique [27].

Pour les journaux *Le Figaro* et *Libération*, j'ai acheté un abonnement d'un mois dans le but d'avoir accès à tous les articles numériques disponibles sur le site internet. Le journal *FranceSoir* étant gratuit, je n'ai pas eu le besoin de prendre un abonnement. Concernant le journal belge, j'ai séléctionné le journal *Le Soir* car je lis souvent ce dernier dans ma vie quotidienne et il est gratuit pour les étudiants.

Le choix des articles s'est réalisé à l'aide d'une recherche sur les sites internet des journaux respectifs. En écrivant "Réchauffement climatique" dans la barre de recherche, j'ai sélectionné les articles les plus pertinents et les plus récents. Les liens de tous les articles de presse sélectionnés peuvent être obtenus sur simple demande. La liste, peu informative, n'est pas reprise en annexe.

Les adresses urls des articles étant choisies, il ne reste plus qu'à créer quatre fichiers contenant les textes sous format ".txt" afin de les importer dans le logiciel TXM. Pour le journal *FranceSoir* qui ne nécessite pas d'abonnement, le logiciel Gromoteur [28] est utilisé de manière à extraire tout le texte en lien avec les urls données. Le logiciel libre Gromoteur rencontrant des problèmes lorsque le journal nécessite un abonnement (pour

les trois autres journaux), il s'agit de copier-coller les articles dans un fichier format ".txt". Le corpus composé de ces quatres fichiers sera nommé "JOURNAUX".

3.2 Analyse du Corpus

Nous divisons le corpus en quatre partitions, une pour chaque journal. Nous observons sur la Figure 3.1 l'histogramme du nombre de mots dans chaque partition. Nous remarquons que le journal qui contient le plus de mots dans le corpus est *Libération* et celui qui a le moins de mots est *FranceSoir*. En effet, c'est le journal pour lequel il y a moins d'articles considérés pour l'analyse. La longueur totale du corpus est de $T = 283568$ mots.

La Table A.1, en annexe, représente la table lexicale associée au corpus JOURNAUX. Nous organisons dans cette table, de dimension $n \times p$, les $n = 143$ lemmes dans chacune des $p = 4$ parties. La dernière colonne donne le nombre total de chaque lemme et la dernière ligne représente le nombre total de lemmes dans chaque partition. Nous mettons de côté les lemmes qui ont trop d'occurrences et ceux qui en ont trop peu. Par conséquent, nous fixons dans le cas de notre exemple, $z_{max} = 1720$ et $z_{min} = 122$. La table lexicale sera représentée par une matrice Z. La quantité de z_{min} correspond au lemme qui a la plus petite valeur d'effectif donné par TXM. Nous ne perdons pas de lemmes à ce niveau

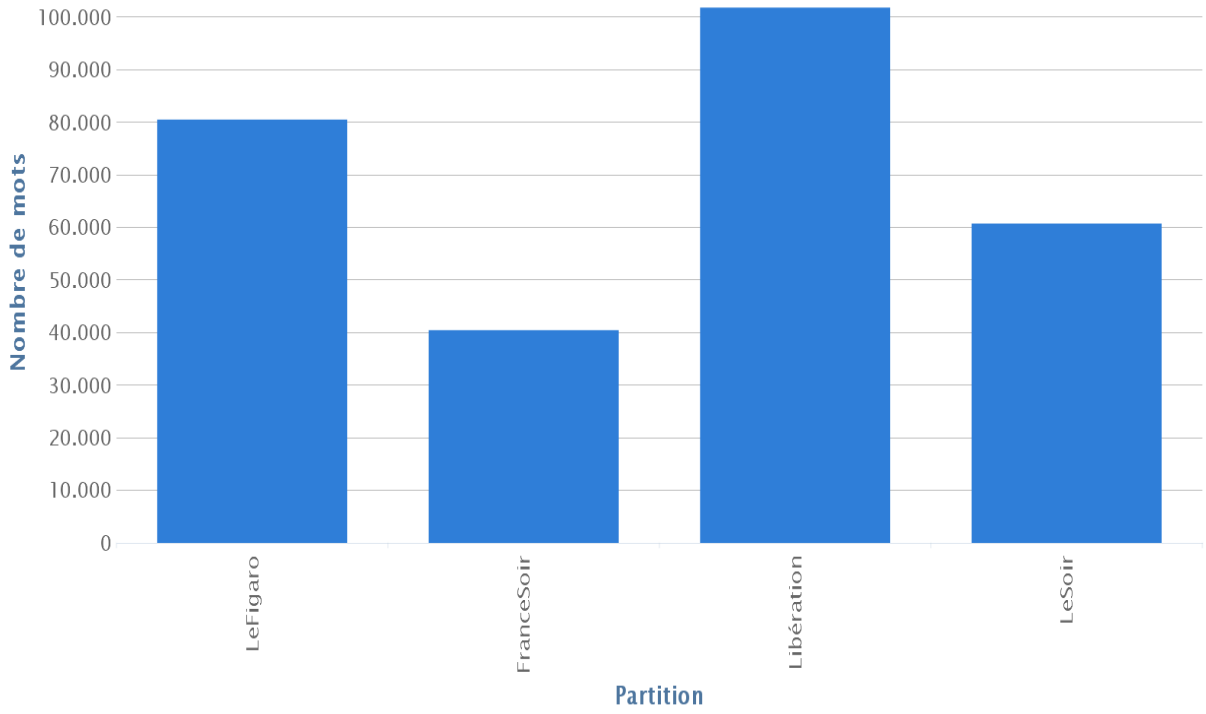

Figure 3.1 – Histogramme du nombre de mots en fonction de la partition.

là. Par contre, les lemmes ayant une fréquence plus grande que le z_{max} choisi sont des déterminants, des ponctuations, des mots qui ne définissent pas vraiment le texte. Nous nous arrêtons au premier mot important, ici "climatique". Ensuite, les mots et ponctuations qui n'influencent pas la compréhension du texte sont supprimés. Donc, sur les 219 lemmes proposés par TXM, nous en gardons 143.

Nous observons les concordances du mot "climatique" dans le corpus à la Figure 3.2. Seule une petite partie des concordances est présentée sur cette figure. Sur la gauche, nous remarquons que les journaux associés aux parties de phrase sont stipulés. Lorsque nous parcourons les concordances de ce mot, nous constatons qu'il est très souvent précédé des mots "réchauffement","changement","dérèglement" ou "urgence". Ce champ lexical signifie une constatation d'une modification et d'une alerte.

Certains lemmes, comme "pouvoir", peuvent porter à confusion. Celui-ci peut représenter le verbe ou alors le nom. Grâce à la table lexicale, nous observons que le lemme "pouvoir" est utilisé 1014 fois dans le corpus. Cependant, le mot "pouvoir" n'apparait que 48 fois dans le texte tel quel sans lemmatisation. Grâce aux concordances de ce mot à la Figure 3.3, nous remarquons qu'il est principalement utilisé comme le verbe et beaucoup moins en tant que nom. De plus, le verbe "pouvoir" peut avoir différentes conjugaisons. Ce qui explique que lorsque nous lemmatisons, nous obtenons beaucoup plus d'effectifs pour le lemme "pouvoir". Grâce aux concordances, nous comprenons que le lemme "pouvoir" représente dans cet exemple principalement le verbe.

3.2.1 Analyse factorielle des correspondances du jeu de données JOURNAUX

Avant de créer l'analyse factorielle des correspondances du jeu de données JOURNAUX, nous procédons tout d'abord au test χ^2 d'homogénéité. Le calcul est donné par

$$X^2 = kt = 39259 \times 0.0533 = 2092.5047$$

et le degré de liberté est

$$dl = (n-1)(p-1) = (140-1)(4-1) = 417.$$

Le seuil de décision est

$$\chi^2_{(n-1)(p-1),1-\alpha} = 475.1215.$$

Nous constatons que nous devons rejeter l'hypothèse nulle du fait que

$$kt > \chi^2_{(n-1)(p-1),1-\alpha}$$

pour le paramètre $\alpha = 0.05$. La p-valeur est de $5.304474\ 10^{-239}$. Elle est obtenue grâce à la fonction *chisq.test* du logiciel R.

Requête [word = "climatique"]

text_id	Contexte gauche	Pivot	Contexte droit
FranceSoir	Le réchauffement	climatique	, " un leurre " ? L'escroquerie climatosceptique de François Gervais
FranceSoir	climatosceptique de François Gervais Non, le réchauffement	climatique	n'est pas lié à l'activité de l'Homme, clame
FranceSoir	Gervais dans son dernier ouvrage " L'Urgence	climatique	est un leurre ". Sauf que le livre pseudo-scientifique de ce
FranceSoir	d'année 2018 un nouvel ouvrage L'Urgence	climatique	est un leurre après celui publié en 2013 L'Innocence du carbone
FranceSoir	CO2 a une influence mineure sur le réchauffement	climatique	passé et à venir auprès d'un public qui n'attend que
FranceSoir	non averti. Son livre, L'Urgence	climatique	est un leurre, est truffé de bêtises voire de mensonges.
FranceSoir	a débat sur l'origine anthropique du changement	climatique	. Nous allons tout de même nous lancer dans cet exercice périlleux
FranceSoir	reprises, l'auteur discute la " sensibilité	climatique	". Par définition, il s'agit de l'augmentation de
FranceSoir	méthodes indépendantes sont utilisées pour estimer la sensibilité	climatique	, sur la base de modèles de climat, sur l'analyse
FranceSoir	air. Pour François Gervais, le réchauffement	climatique	déjà observé n'est pas généré par l'augmentation du CO2 atmosphérique
FranceSoir	a des incertitudes sur l'ampleur du changement	climatique	à venir et ses impacts, mais il est ridicule de vouloir
FranceSoir	Sauver les baleines pour lutter contre le réchauffement	climatique	En absorbant des tonnes de dioxyde de carbone, ces géantes des
FranceSoir	une aide précieuse pour lutter contre le réchauffement	climatique	. Les baleines, plus fortes que les arbres Les baleines bleues
FranceSoir	bleues pourraient -elles sauver la planète du réchauffement	climatique	? Eux -mêmes menacés par le réchauffement et l'acidification des océans
FranceSoir	dans ses poumons le gaz responsable du réchauffement	climatique	. Selon le FMI, chaque baleine vaut d'ailleurs plus de
FranceSoir	pour atténuer et renforcer la résilience au changement	climatique	». Une espèce en voie d'extinction Aussi précieuses soient -elles
FranceSoir	enjeu primordial pour sauver la Terre du réchauffement	climatique	. Elles sont d'autant plus inestimables à la lutte contre les
FranceSoir	Horn alerte sur les conséquences désastreuses du réchauffement	climatique	En 2019, l'aventurier Mike Horn a entrepris de traverser à
FranceSoir	, il alerte sur les dangers du réchauffement	climatique	. Un périple dangereux " C'est la première fois qu'on
FranceSoir	le front ". Les conséquences du réchauffement	climatique	Pour Mike Horn, cela ne fait aucun doute : l'échec
FranceSoir	échec de sa mission est dû au réchauffement	climatique	, dont les conséquences sont nettement visibles en Arctique. " Quand
FranceSoir	souhaite aujourd'hui alerter sur les dangers du changement	climatique	pour notre planète. Il garde cependant espoir : " On peut
FranceSoir	de la Terre se modifie. Le réchauffement	climatique	en cause ? Saviez -vous que l'axe de rotation de la
FranceSoir	. Ils en ont déduit que le réchauffement	climatique	aurait accéléré ce mouvement naturel de manière rapide. Explications. Naturellement
FranceSoir	publier une étude : ils parlent du réchauffement	climatique	, comme une des causes principales de ce changement d'axe.
FranceSoir	des glaciers, une conséquence majeure du changement	climatique	, « a suffisamment redistribué les eaux pour accélérer le déplacement du
FranceSoir	causes naturelles et humaines. Comment le réchauffement	climatique	pourrait déclencher la prochaine crise économique Alors qu'on ne parle que
FranceSoir	ne parle que des conséquences environnementales du changement	climatique	, il devient maintenant évident que l'impact économique du changement climatique
FranceSoir	maintenant évident que l'impact économique du changement	climatique	sera beaucoup plus important qu'il n'y paraît pour l'instant
FranceSoir	. Une série de risques liés au changement	climatique	Plusieurs effets du changement climatique pourraient en effet ébranler le système économique
FranceSoir	liés au changement climatique Plusieurs effets du changement	climatique	pourraient en effet ébranler le système économique. Mark Carney, l'
FranceSoir	contre les entreprises qui ont contribué au changement	climatique	et la baisse de la valeur des actifs liés aux combustibles fossiles

Figure 3.2 – Partie des concordances du mot “climatique” dans le corpus JOURNAUX.

Requête [word = "pouvoir"]

text_id	Contexte gauche	Pivot	Contexte droit
FranceSoir	datée de 2014, François Gervais prétend pourtant	pouvoir	expliquer les variations temporelles de la température moyenne globale avec un cycle
FranceSoir	de descendre à la station de Montenvers pour	pouvoir	profiter de la vue exceptionnelle sur le glacier. Mais depuis 1998
FranceSoir	en métal de plus de 600 marches pour	pouvoir	atteindre le glacier et visiter la grotte Claret, creusée dans la
FranceSoir	Mais cela ne suffit désormais plus. Pour	pouvoir	continuer à y accéder, de gros aménagements du site d'accueil
FranceSoir	choix que de continuer à travailler parfois sans	pouvoir	boire ou s'abriter à l'ombre. " Ceux qui travaillent
FranceSoir	champs risquent de ruiner leur santé juste pour	pouvoir	se nourrir ", a averti Saleemul Huq, chef du Centre
FranceSoir	" Je suis très satisfaite, j'espère	pouvoir	faire entendre ma voix, j'ai de fortes convictions environnementales "
LeFigaro	engagent à suivre des objectifs qu'ils pensent	pouvoir	réaliser mais ces objectifs sont encore trop timides. De plus,
LeFigaro	a besoin de 69 millions de dollars pour	pouvoir	mettre en place l'aide nécessaire sur les six prochains mois.
LeFigaro	d'intéressement dans l'entreprise pour augmenter le	pouvoir	d'achat des salariés, que faire face à l'accroissement effarant
LeFigaro	celle des « bons puits », pour	pouvoir	arroser les cultures. L'ambition du président Rajoelina et son grand
LeFigaro	en Islande. En 2035, il faudra	pouvoir	absorber un quart du CO2 émis, soit 4 gigatonnes, et
LeFigaro	quitter la scène politique après 16 années au	pouvoir	de la première économie européenne. Après s'être tenue en retrait
LeFigaro	faible prix de l'électricité imposé par le	pouvoir	central et les cours des matières premières qui s'envolent. Les
LeFigaro	plus courte que le carbone. Mais son	pouvoir	de réchauffement est très supérieur. « Sa durée de vie dans
LeFigaro	pays pauvres, qui a un temps semblé	pouvoir	faire dérailler les négociations, n'a pas trouvé de résolution.
LeFigaro	plein fouet le décalage entre capacités technologiques et	pouvoir	d'achat. Une étude sur les populations à faible revenu de
LeFigaro	dioxyde de carbone, mais il a un	pouvoir	réchauffant 85 fois supérieur à celui du CO2. Il absorbe mieux
LeFigaro	engrais azotés dans l'agriculture, a un	pouvoir	de réchauffement 300 fois supérieur à celui du CO2. Ce gaz
LeFigaro	le dérèglement climatique. Le méthane a un	pouvoir	réchauffant 20 à 30 fois plus important que le CO2. «
LeFigaro	la finance, qui détient par nature un	pouvoir	de vie ou de mort sur tout projet. Entre 20 %
LeFigaro	« enfermés » dans les énergies fossiles et	pouvoir	avancer avec le reste du monde, a déclaré le président français
LeFigaro	-il ajouté. Mais « nous avons un	pouvoir	immense. Nous pouvons soit sauver notre monde soit condamner l'humanité
LeSoir	de données sur un temps suffisamment long pour	pouvoir	trancher. Or, un ouragan de catégorie 5 reste un événement
LeSoir	vers une société sans carbone, il faut	pouvoir	faire une prédiction du nombre de mégatonnes de CO2 séquestrables par ces
LeSoir	-ci ! " » Ces gens sont au	pouvoir	et « accélèrent le mouvement vers la catastrophe ». Le caricaturiste
LeSoir	la forêt tropicale. À son arrivée au	pouvoir	en 2019, grâce en partie au soutien du puissant lobby de
LeSoir	. En 2019, sa première année au	pouvoir	, une hausse importante des feux en Amazonie a provoqué un émoi
LeSoir	. « Les politiques et les personnes au	pouvoir	s'en tirent depuis bien trop longtemps sans rien faire pour lutter
LeSoir	projet pétrolier ou gazier s'il veut encore	pouvoir	limiter le réchauffement climatique à 1, 5ºC. Grande Barrière de
LeSoir	, mais aussi du maintien du gouvernement au	pouvoir	». Pour la chancelière, il s'agit en effet de
LeSoir	est nécessaire. Parfois nous sommes fiers de	pouvoir	sauver des vies. » Solidarité climatique inédite Désormais tout va plus

Figure 3.3 – Partie des concordances du mot "pouvoir" dans le corpus JOURNAUX.

Par conséquent, les documents possèdent des mots différents. Ces derniers sont liés de diverses manières et les partitions contiennent des mots différents. Comme nous n'avons pas l'homogénéité des variables, nous exécutons l'analyse factorielle des correspondances. C'est ce à quoi nous nous attendons de par ce qui a été discuté au Chapitre 2.

Grâce au logiciel TXM, nous générons l'analyse factorielle des correspondances du jeu de données JOURNAUX réparti en quatre journaux à la Figure 3.4. Nous constatons que l'axe 1 divise les journaux. D'un côté, il y a les partitions "LeSoir", "FranceSoir" et "Libération" qui se situent à droite. D'un autre côté, "LeFigaro" est à gauche sur le graphique. Le journal *Le Figaro*, séparé par l'axe 1, prend la plus grande partie de la variabilité. Le journal *Le Figaro* se démarque des autres de par son orientation politique de droite. L'axe 1 peut donc être interprété comme l'analyse d'un journal qui serait d'orientation politique de droite ou non. Le deuxième axe sépare la partition "Libération" des autres. Ceci peut s'expliquer par le fait que ces deux journaux ont une orientation politique distincte et marquée.

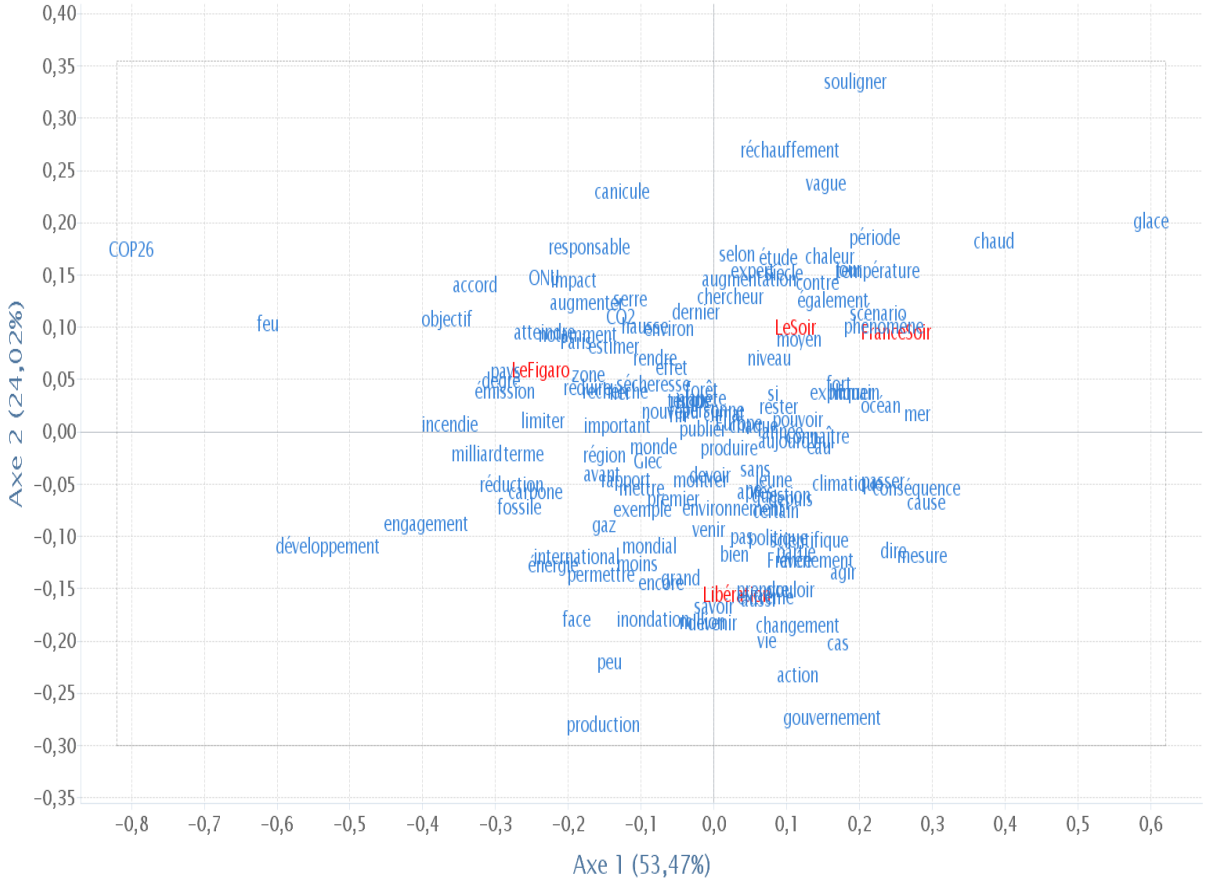

Figure 3.4 – Analyse factorielle des correspondances pour le jeu de données JOURNAUX réparti en fonction des journaux pour les axes 1 et 2.

Les lemmes sont répartis autour des journaux. Cependant, pour les journaux *Le Soir* et *FranceSoir*, nous observons que peu de lemmes sont très proches d'eux (moins que pour les autres journaux). Nous interprétons cela par le fait que ce sont des journaux avec une opinion politique neutre et donc, ils n'utilisent pas des mots propres à eux. Ils emploient des mots ordinaires que tout le monde utilise sans marquer leur propre originalité. Nous pouvons analyser les lemmes et les partitions simultanément grâce à la propriété barycentrique du Chapitre 2 section 2.2.7. Elle relate qu'un lemme est du côté des partitions auxquelles il s'associe le plus et inversement.

Sur la Figure 3.5, nous trouvons les données associées à la Figure 3.4. Elle indique les valeurs propres de la matrice $F'F$. Elles sont comprises entre 0 et 1 en analyse des correspondances. Ces dernières mesurent les variances le long de chaque axe principal. Ces valeurs propres sont $\lambda_1 = 0.0285$ pour le premier axe, $\lambda_2 = 0.0128$ pour le deuxième axe et $\lambda_3 = 0.012$ pour le troisième axe. Les pourcentages de variance sont calculés en faisant le rapport entre chaque valeur propre et leur somme globale multipliée par 100. Ils correspondent à ces valeurs propres et sont égaux à 53.47% et 24.02 % et 22.51% pour les trois premiers axes. Les pourcentages de variance mesurent l'importance relative de chaque valeur propre dans la trace de la matrice $F'F$. Dans ce cas, le plan correspondant aux deux premiers axes principaux explique 77.49% de la variance totale. Nous décidons de regarder le troisème axe pour avoir plus d'information au sujet du jeu de données JOURNAUX.

Sur la Figure 3.6, l'analyse factorielle des correspondances du jeu de données JOURNAUX pour les axes 1 et 3, nous remarquons que l'axe 1 divise toujours les mêmes journaux, ce qui est attendu. Le troisième axe sépare le journal *FranceSoir* du journal *Le Soir*. Une piste d'explication est que ces deux journaux sont neutres mais d'une manière différente. De plus, le pays de parution de ces journaux est différent et peut aussi justifier cette discrimination.

Nous concluons de cette analyse factorielle des correspondances que les orientations politiques différentes sont clairement visibles sur les différents graphiques explorés. De plus, nous pouvons aussi dire que le pays du journal n'a pas un grand impact lors de cette analyse. Nous l'expliquons par le fait que le réchauffement climatique est une thématique qui impacte toute la planète et que tous les journaux s'en soucient. La France et la Belgique ont la même envie de combattre le réchauffement climatique. D'autres éléments dans les

#	Valeur propre	%	∑%
1	0,0285	53,47	53,47
2	0,0128	24,02	77,49
3	0,0120	22,51	100,00

Figure 3.5 – Valeur propre lors de l'analyse factorielle des correspondances pour le jeu de données JOURNAUX réparti en fonction des journaux.

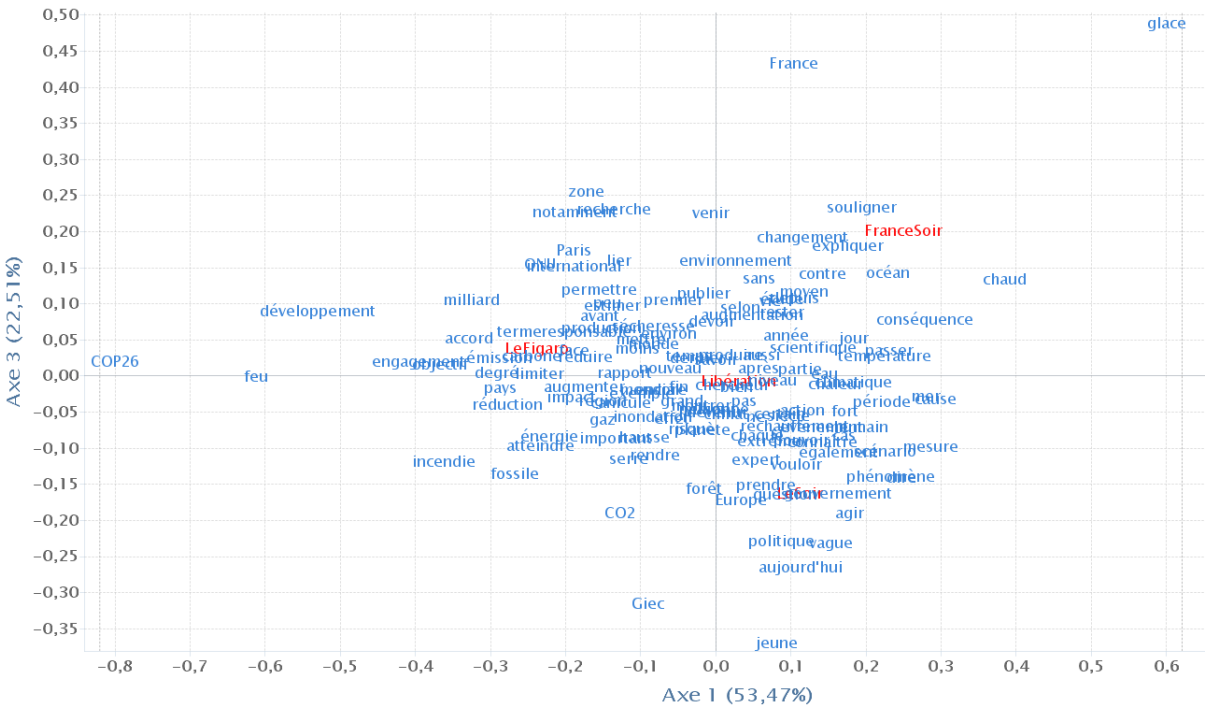

Figure 3.6 – Analyse factorielle des correspondances pour le jeu de données JOURNAUX réparti en fonction des journaux pour les axes 1 et 3.

prochaines sections nous permettrons de mieux comprendre les discriminations faites par les trois axes.

3.2.2 Classification ascendante hiérarchique du jeu de données JOURNAUX

Nous produisons maintenant la classification ascendante hiérarchique à l'aide du logiciel TXM. Sur la Figure 3.7, apparaît le dendrogramme du jeu de données JOURNAUX pour deux classes. Nous observons que le journal *Le Figaro* représente un cluster à lui tout seul alors que les autres journaux (*Libération*, *FranceSoir* et *Le Soir*) sont regroupés ensemble. Nous présentons la classification ascendante hiérarchique uniquement pour les colonnes car les lignes sont trop nombreuses et donc, le dendrogramme devient illisible. Le numéro de cluster associé à chaque lemme est noté dans la dernière colonne de la Table A.1. Le clustering est fait pour six clusters.

Concernant le choix du nombre de classes, nous observons le graphique 3.8. Ce dernier a été produit à l'aide du logiciel R. Nous souhaitons avoir la plus grande inertie tout en regroupant les partitions. En effet, plus nous regroupons les partitions, plus nous diminuons l'inertie. Il nous faut trouver le juste milieu. Lorsque nous choisissons un nombre au-delà de deux classes, l'apport d'inertie supplémentaire est faible. Par conséquent, nous choisirons le nombre de deux classes pour le dendrogramme.

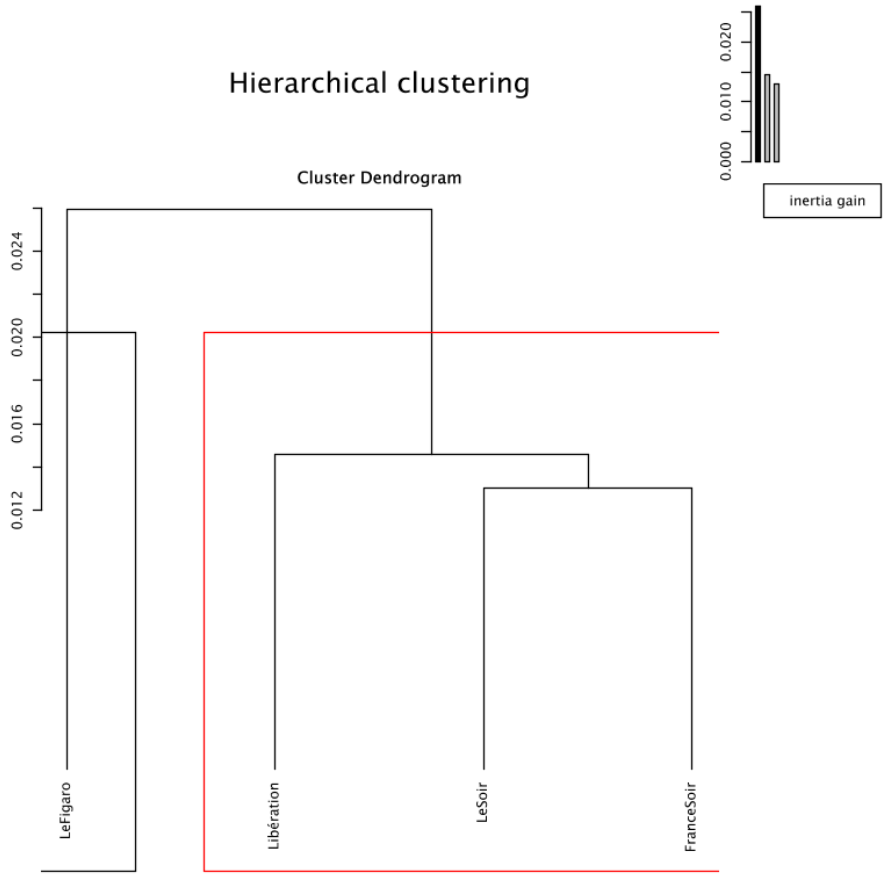

Figure 3.7 – Dendrogramme du jeu de données JOURNAUX réparti en fonction des journaux.

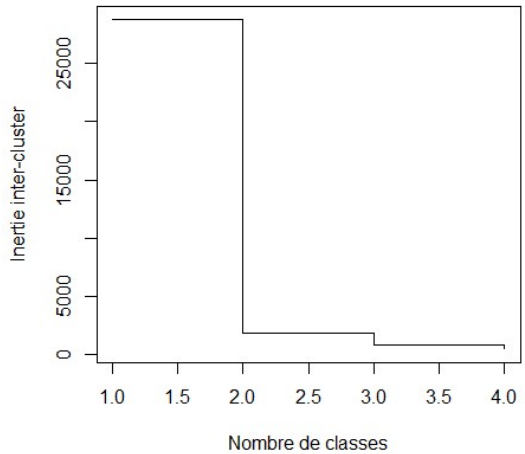

Figure 3.8 – Graphique de l'inertie inter-cluster en fonction du nombre de classe.

Nous pouvons aussi créer un dendrogramme à partir de l'analyse factorielle des correspondances. Nous obtenons la Figure 3.9 qui est par conséquent en trois dimensions. Ce graphique résume et regroupe les informations reçues lors de l'analyse factorielle des correspondances et du clustering. Nous observons que le journal *Le Figaro* est séparé des autres journaux autant sur le premier axe de l'analyse factorielle des correspondances que sur le dendrogramme à deux classes. *Le Figaro* est un journal qui a un vocabulaire qui se démarque des autres journaux. Il montre une manière de parler du réchauffement climatique bien marquée et plus forte que les autres journaux.

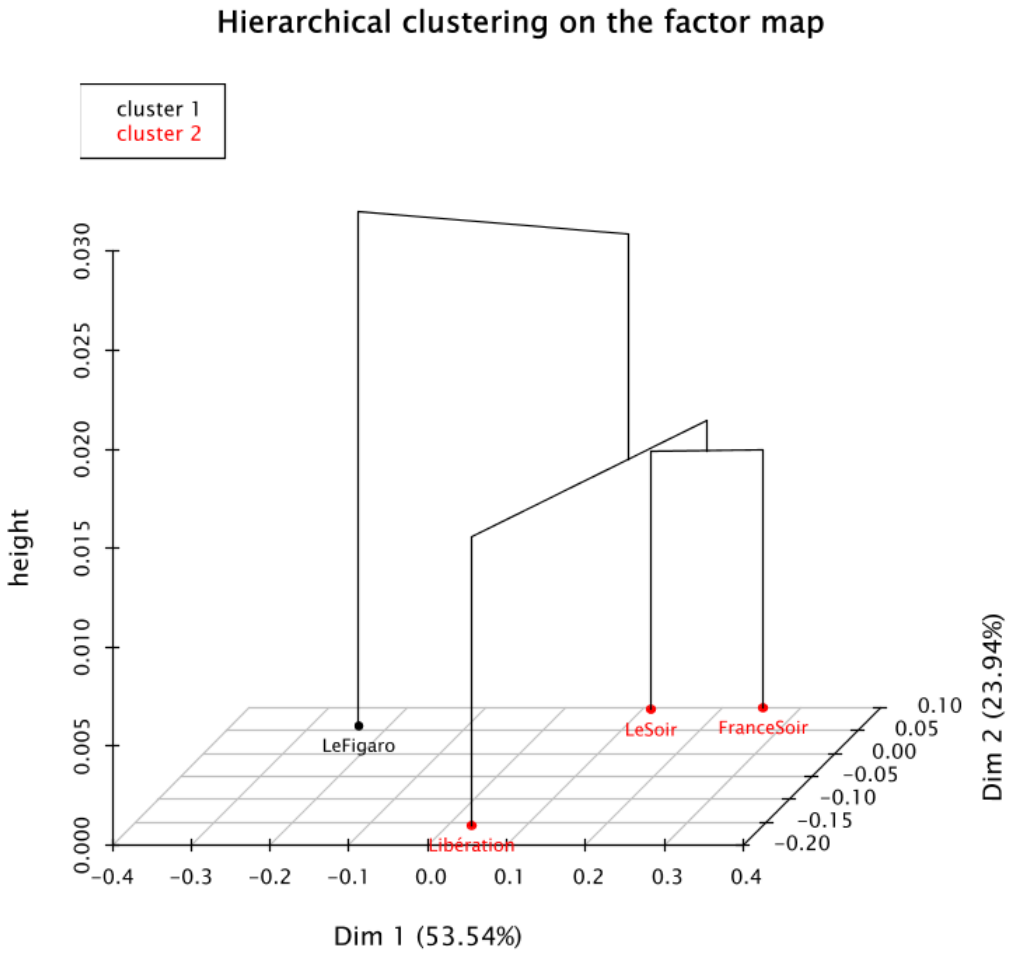

Figure 3.9 – Dendrogramme sur l'analyse factorielle des correspondances pour le jeu de données JOURNAUX réparti en fonction des journaux.

3.3 Quel journal est le plus orienté “politique” ?

Nous allons dorénavant analyser le corpus plus en détails. L'objectif est de diviser la table lexicale A.1 en sous-tables. Elles seront choisies en fonction d'un thème. La Table 3.2 représente les lemmes qui sont plus orientés sur la politique. Elle se compose de 25 lemmes. Le choix de ces derniers est subjectif mais ils font partie, à mon sens, du champ lexical du mot “politique”. Nous nommons cette table “politique”. Cette table nous permettra de découvrir comment les journaux parlent du réchauffement climatique et peut-être de mieux comprendre les sections précédentes de ce mémoire.

Table 3.2 – Table lexicale orientée “politique” du jeu de données JOURNAUX divisé en 4 groupes.

Lemmes	LeFigaro	FranceSoir	Libération	LeSoir	Total
pays	273	53	192	130	648
rapport	212	68	222	124	626
devoir	180	91	216	16	503
France	115	105	164	30	414
monde	135	53	139	80	407
Europe	88	34	114	97	333
objectif	129	22	74	52	277
politique	56	22	102	76	256
Paris	98	37	69	38	242
impact	95	26	63	57	241
accord	105	23	58	43	229
atteindre	88	17	63	54	222
Giec	65	8	78	69	220
limiter	86	20	70	42	218
mesure	29	31	82	55	197
risque	58	22	62	48	190
international	66	22	68	22	178
agir	31	19	72	50	172
engagement	69	8	55	22	154
gouvernement	24	14	69	36	143
COP26	88	0	28	18	134
action	27	16	60	28	131
responsable	52	18	32	28	130
ONU	56	19	33	22	130
dévellopement	64	5	45	12	126
Total	2072	768	2225	1189	6254

3.3.1 Analyse factorielle des correspondances

L'AFC de la table lexicale "politique" est représentée sur la Figure 3.10. Nous observons que le journal *FranceSoir* est éloigné du champ lexical "politique" et des autres journaux. Il parle donc très peu de politique dans ses articles. Nous observons que l'axe 2 de la Figure 3.10 discrimine les journaux *FranceSoir* et *Le Soir*. Nous avons donc la confirmation que l'axe 3 de la Figure 3.6 différencie ces deux journaux sur l'aspect politique.

La partition "LeFigaro" a beaucoup de lemmes de la table lexicale politique qui gravite autour de lui. Nous observons en effet que le premier axe de l'AFC sépare la partition "LeFigaro" des autres partitions. Ce dernier est celui qui est le plus orienté "politique". En effet, l'axe 1 distingue encore "LeFigaro" des autres comme à la section 3.2.1 sur la Figure 3.4. C'est donc en partie bien sur le plan politique que ce journal se détache des autres.

Les informations liées à l'AFC de la table lexicale "politique" sont données par la Figure 3.11. Nous constatons que les deux axes représentent à eux deux 92.4% donc, nous n'avons pas besoin d'observer un troisième axe.

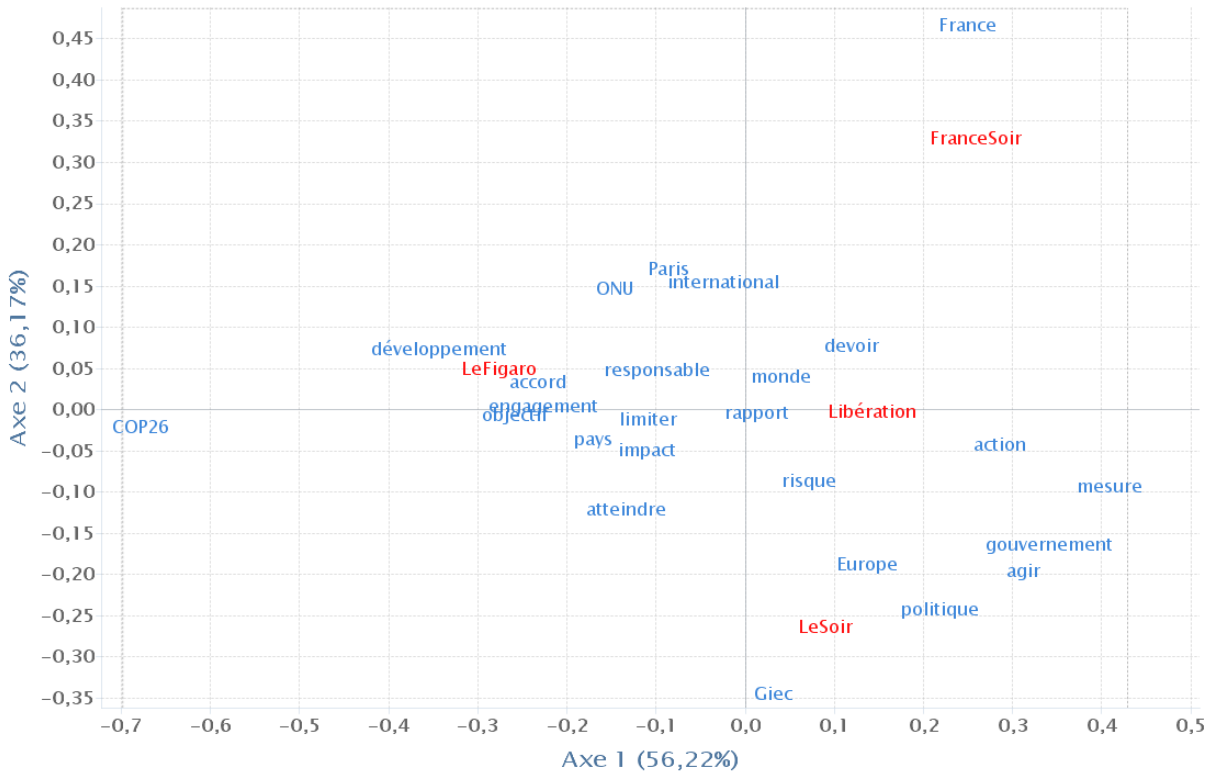

Figure 3.10 – Analyse factorielle des correspondances pour la table lexicale "politique"
3.2.

#	Valeur propre	%	∑%	
1	0,0425	56,22	56,22	
2	0,0274	36,17	92,40	
3	0,0058	7,60	100,00	

Figure 3.11 – Données associées à l'AFC de la Figure 3.10.

Si nous observons en trois dimensions l'AFC et la classification ascendante hiérarchique à la Figure 3.12, nous remarquons que la partition "FranceSoir" ne représente qu'un seul cluster tout comme "LeFigaro". Ils sont chacun dans leur extrême. Les partitions "Libération" et "Le Soir" sont regroupées. Nous interprétons cela par le fait qu'ils parlent de manière assez similaire de l'aspect politique du réchauffement climatique.

La Figure 3.13 est une représentation en trois dimensions de l'AFC et de la classification ascendante hiérarchique des lemmes du corpus JOURNAUX. Le mot "France" se détache fortement. Il n'est pas utilisé de la même manière que les autres lemmes.

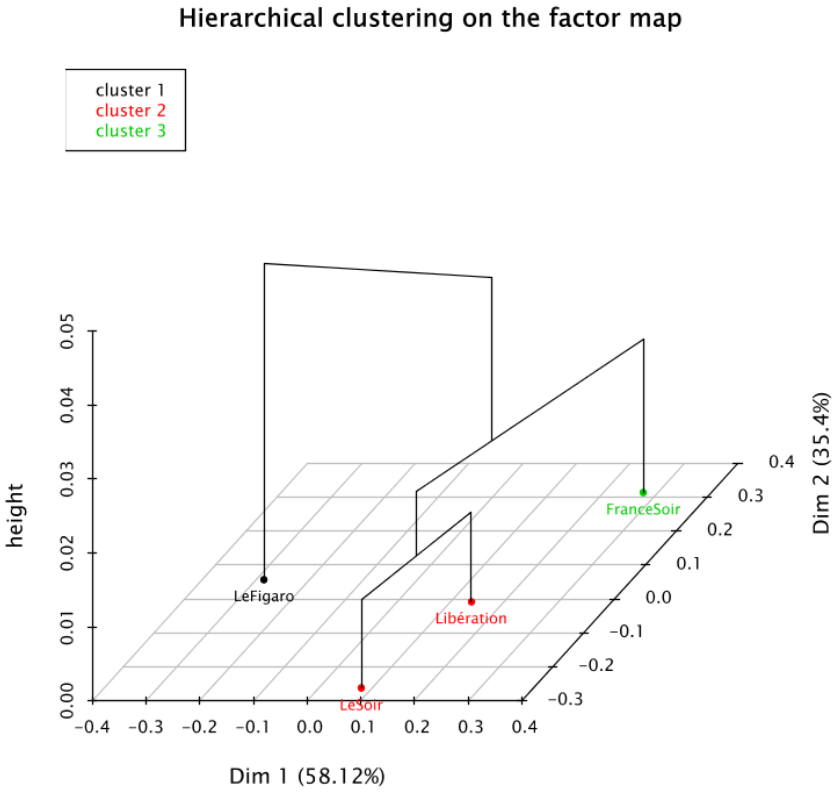

Figure 3.12 – Dendrogramme des partitions sur l'analyse factorielle des correspondances pour la table lexicale "politique" 3.2.

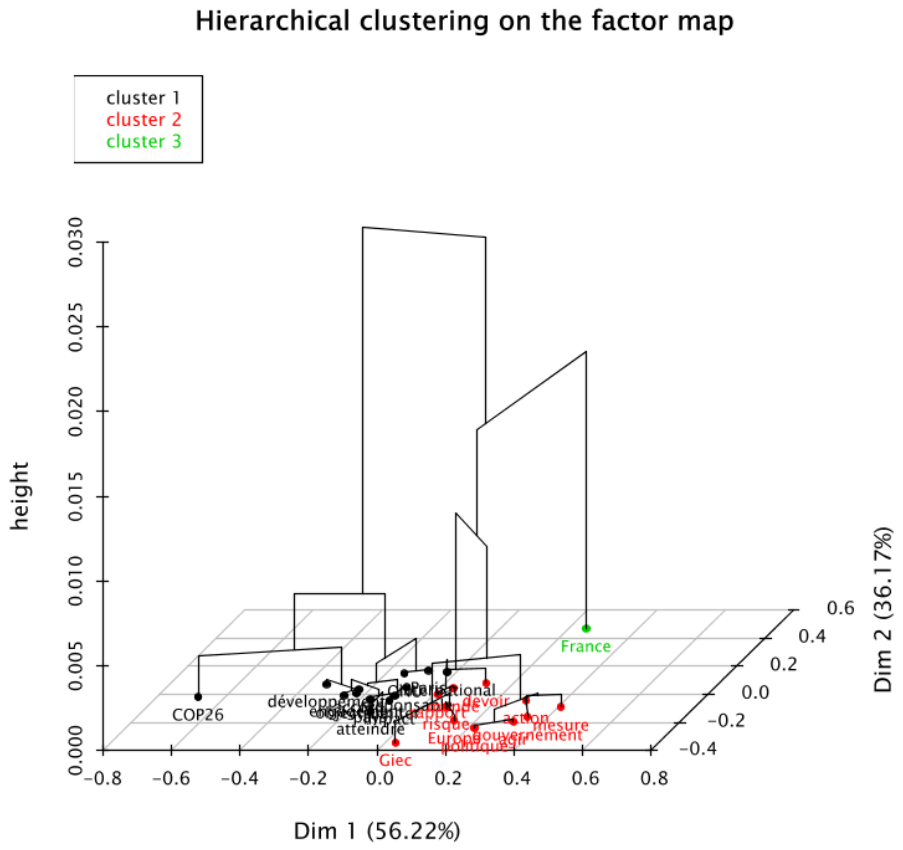

Figure 3.13 – Dendrogramme des lemmes sur l'analyse factorielle des correspondances pour la table lexicale "politique" 3.2.

3.4 Quel journal est le plus orienté "scientifique" ?

Dans cette section, nous analysons la sous-table lexicale orientée "scientifique". La Table 3.3 se compose des lemmes choisis qui concerne le côté scientifique du réchauffement climatique. Elle se compose de 25 lemmes. Nous nommons cette table "scientifique". Cette table nous permettra de découvrir comment les journaux parlent scientifiquement du réchauffement climatique et peut-être de mieux comprendre les sections précédentes de ce mémoire.

3.4.1 Analyse factorielle des correspondances

Sur la Figure 3.14, nous constatons que la partition "Libération" est moins concernée par le côté scientifique du réchauffement climatique. Cela est vraiment observé avec l'axe 2 qui sépare clairement cette partition des autres. De plus, peu de lemmes sont proches de cette partition. Nous observons le même fait pour la partition "FranceSoir". Les partitions "Libération" et "FranceSoir" ne sont pas très proches de lemmes à connotation scientifique.

Table 3.3 – Table lexicale orientée "scientifique" du jeu de données JOURNAUX divisé en 4 groupes.

Lemmes	LeFigaro	FranceSoir	Libération	LeSoir	Total
climatique	359	286	642	423	1710
réchauffement	343	207	267	384	1201
changement	147	116	266	88	617
effet	197	72	189	153	611
émission	245	52	176	106	579
température	117	107	145	145	514
gaz	145	35	161	89	430
degré	160	32	113	72	377
eau	77	55	118	81	331
scientifique	72	51	125	67	315
étude	92	63	83	74	312
serre	109	31	86	85	311
carbone	111	25	100	48	284
objectif	129	22	74	52	277
énergie	95	15	103	52	265
réduire	96	29	79	51	255
CO2	86	19	71	73	249
chercheur	69	35	64	59	227
estimer	81	31	63	41	216
mesure	29	31	82	55	197
expert	56	27	54	60	197
phénomène	32	26	50	53	161
scénario	28	23	41	43	135
développement	64	5	45	12	126
canicule	47	16	28	33	124
Total	2976	1444	3229	2441	10090

Si nous observons l'axe 1, nous constatons que c'est encore une fois la partition "Le Figaro" qui se détache des autres. Ce sont les partitions "Le Figaro" et "Le Soir" qui sont le plus entourés par les lemmes scientifiques. Le journal *Le Figaro* se démarque des autres aussi sur le plan scientifique.

Grâce à la Figure 3.15, nous observons que les deux axes représentent à eux deux 90,78% donc, nous n'avons pas besoin d'observer un troisième axe. Cela ajouterait peu d'informations supplémentaires. Nous resterons donc en deux dimensions pour l'AFC.

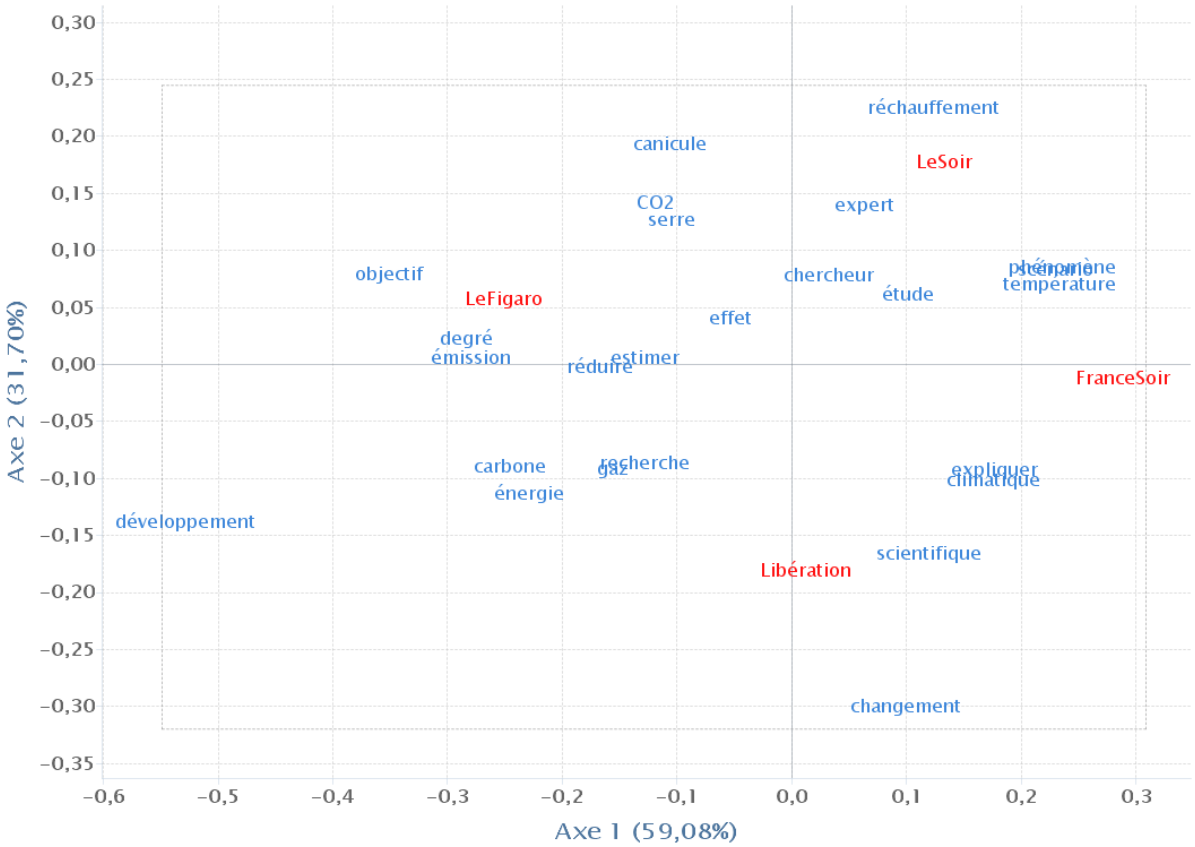

Figure 3.14 – Analyse factorielle des correspondances pour la table lexicale "scientifique" 3.3.

Sur la Figure 3.16, nous observons que la partition "LeFigaro" se détache des autres ainsi que "Libération". Le journal *Libération* étant celui qui est moins proche des lemmes scientifiques que les autres journaux, nous comprenons qu'il se détache des autres.

Sur la Figure 3.17, nous observons l'analyse des lemmes avec l'AFC et le clustering. Les lemmes sont répartis en trois clusters.

#	Valeur propre	%	∑%	
1	0,0351	59,08	59,08	
2	0,0189	31,70	90,78	
3	0,0055	9,22	100,00	

Figure 3.15 – Données associées à l'AFC de la Figure 3.14.

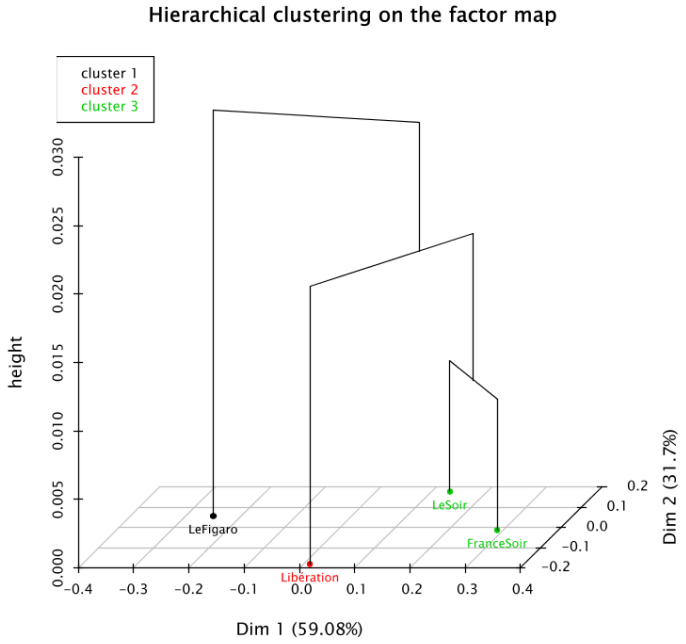

Figure 3.16 – Dendrogramme des partitions sur l'analyse factorielle des correspondances pour la table lexicale "scientifique" 3.3.

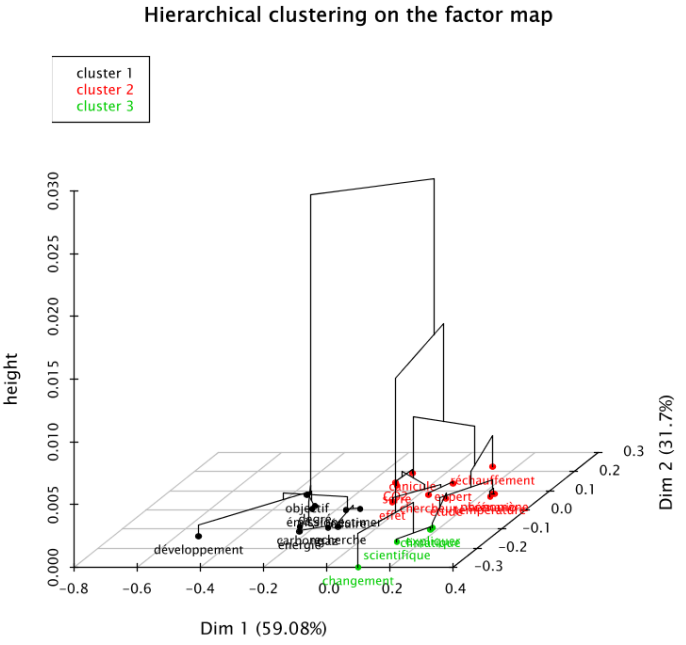

Figure 3.17 – Dendrogramme des lemmes sur l'analyse factorielle des correspondances pour la table lexicale "scientifique" 3.3.

3.5 Quel journal parle le plus de l'environnement ?

Dans cette section, nous analysons la sous-table lexicale orientée "environement". Nous l'observons à la Table 3.4. Elle se compose de 21 lemmes. Nous nommons cette table "environnement". Cette table nous permettra de découvrir comment les journaux évoquent l'aspect environnemental du réchauffement climatique.

Table 3.4 – Table lexicale orientée "environnement" du jeu de données JOURNAUX divisé en 4 groupes.

Lemmes	LeFigaro	FranceSoir	Libération	LeSoir	Total
climat	166	75	194	145	580
température	117	107	145	145	514
eau	77	55	118	81	331
chaleur	80	59	86	91	316
niveau	84	49	96	78	307
planète	68	27	74	58	227
conséquence	37	43	79	48	207
humain	43	33	68	58	202
océan	43	44	67	43	197
sécheresse	67	28	61	39	195
chaud	32	49	50	50	181
mer	31	33	61	49	174
environnement	48	29	62	27	166
inondation	45	13	64	30	152
vague	37	21	37	57	152
forêt	44	15	49	43	151
glace	19	61	41	29	150
fossile	56	6	53	32	147
incendie	63	5	46	31	145
vie	33	21	59	22	135
feu	69	3	31	19	122
Total	1259	776	1541	1175	4751

3.5.1 Analyse factorielle des correspondances

L'AFC de la table lexicale concernant l'environnement est représentée à la Figure 3.18. Elle montre que le journal *FranceSoir* évoque moins la partie environnementale. L'axe 1 met d'un côté "FranceSoir", au milieu "Libération et LeSoir" et d'un autre côté, "LeFigaro". Les deux partitions du milieu semblent plus parler de la partie environnementale du réchauffement climatique. Le deuxième axe discrimine "LeSoir" d'un côté et "Libération" de l'autre. Le journal *Le Soir* semble plus parler de vague de chaleur, de température que *Libération* va être plus global en parlant d'environnement, d'événements et de conséquences.

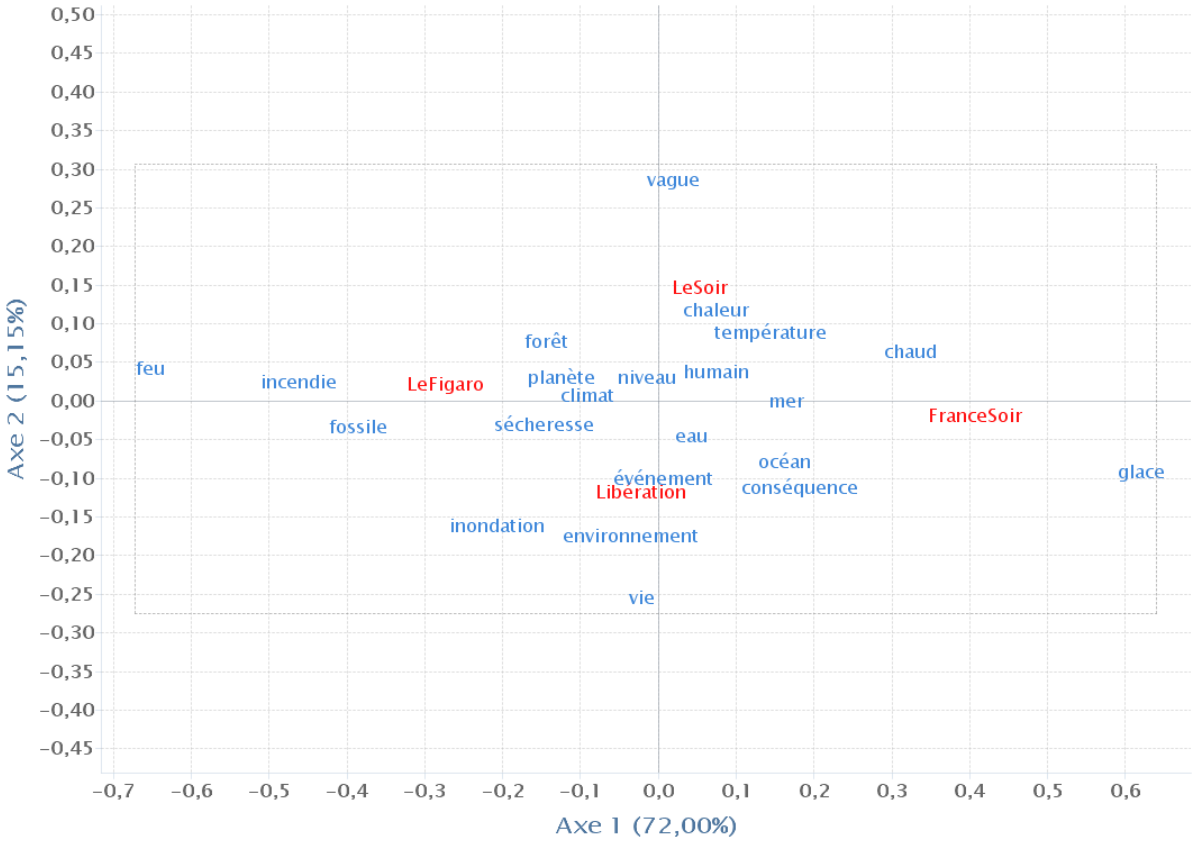

Figure 3.18 – Analyse factorielle des correspondances pour la table lexicale "environnement" 3.4.

La Figure 3.19 nous indique que les deux axes représentent à eux deux 87.16% donc, nous n'avons pas besoin d'observer un troisième axe. Cela ajouterait peu d'informations supplémentaires.

Le dendrogramme 3.20 montre en effet que les partitions "LeSoir" et "Libération" forment un cluster tandis que les autres sont séparés chacun de leur côté. La partition "LeFigaro" est quand même rattachée au cluster "Libération et LeSoir" donc, elle est plus orientée "scientifique" que la partition "FranceSoir".

#	Valeur propre	%	∑%
1	0,0475	72,00	72,00
2	0,0100	15,15	87,16
3	0,0085	12,84	100,00

Figure 3.19 – Données associées à l'AFC de la Figure 3.18.

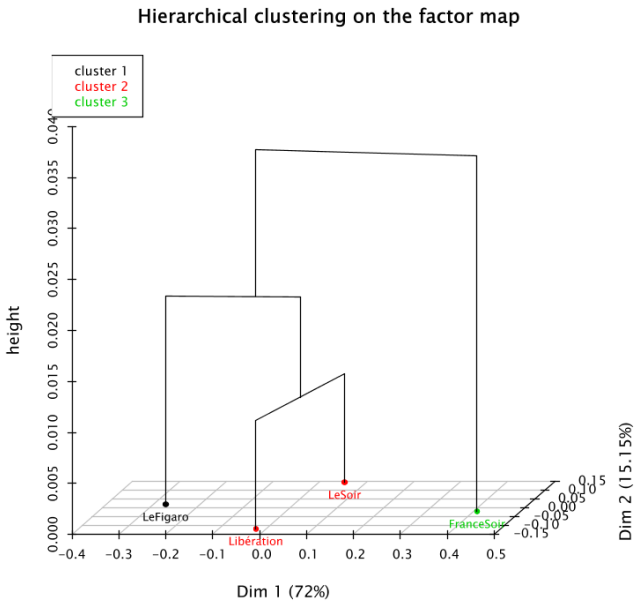

Figure 3.20 – Dendrogramme des partitions sur l'analyse factorielle des correspondances pour la table lexicale "environnement" 3.4.

Le graphique à la Figure 3.21 montre les trois clusters faits sur les lemmes. Celui en noir contient des lemmes autour de l'incendie. Celui en rouge est plus général en évoquant des incidents. Le dernier, en vert, représente les mots autour de l'océan.

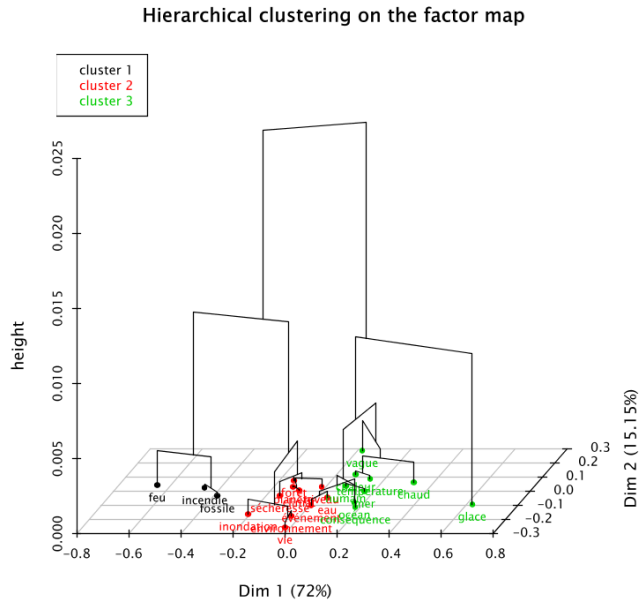

Figure 3.21 – Dendrogramme des lemmes sur l'analyse factorielle des correspondances pour la table lexicale "environnement" 3.4.

3.6 Conclusion

En conclusion, la division de la table lexicale A.1 nous permet de mieux comprendre les comportements observés lors de l'analyse factorielle des correspondances de cette dernière. Tout d'abord, le journal *Le Figaro* est systématiquement discriminé par le premier axe, que ce soit au niveau politique, scientifique ou environnemental. Ce journal est donc différent des autres sur ces trois domaines. Son orientation politique et ses convictions sont fortement marquées (plus que les autres) lorsque nous abordons le sujet du réchauffement climatique. Il évoque plus l'aspect politique et scientifique du réchauffement climatique. Une interprétation possible à ce phénomène est que ce journal est de droite. En effet, il s'adresse a priori à des personnes plus instruites et plus cultivées. Par conséquent, il aborde des aspects plus techniques.

Ensuite, concernant le journal *FranceSoir*, il est souvent peu entouré de lemmes spécifiques. Ce qui convient avec le fait qu'il ait une orientation politique neutre. Il emploie des mots banals que tout le monde utilise. En effet, le journal *FranceSoir* est très neutre et donc se positionne moins sur un aspect en particulier. Ce journal est discriminé par le troisième axe de l'AFC sur la Figure 3.6 ainsi que le journal *Le Soir*. Ils sont tous les deux opposés. Nous observons la même chose uniquement lorsque nous analysons la table lexicale "politique". Par conséquent, nous en déduisons que ces deux journaux diffèrent, en partie, sur leur manière d'aborder la politique liée au réchauffement climatique. Ils sont neutres mais pas de la même manière. En effet, ils n'ont pas excatement la même orientation politique. Le journal *FranceSoir* est neutre alors que le journal *Le Soir* est progressiste. Ce ne sont pas des orientations politiques très marquées mais elles sont différentes. En effet, nous remarquons que c'est le journal *Le Soir* qui aborde le plus la partie environnementale et scientifique. Ce sont les sujets sur lesquels il se démarque.

Enfin, le journal *Libération* se distingue quant à lui uniquement sur le plan environnemental. Le journal *Libération* a une opinion moins marquée lorsqu'il parle de réchauffement climatique que le journal *Le Figaro*. Pourtant ce sont deux journaux qui ont des orientations politiques bien prononcées.

Chapitre 4

Analyse des sentiments

Nous avons étudié les mots et leurs fréquences dans des corpus au cours des chapitres précédents. Maintenant, attardons-nous sur un autre aspect de l'étude de texte : l'analyse des sentiments. Cette technique nous permet de comprendre les émotions d'un écrit. Grâce à cette analyse, nous pouvons savoir si un texte - et donc une combinaison de mots - est plus négatif ou positif. De plus, nous appliquons cette technique d'analyse de texte au jeu de données JOURNAUX. Les graphiques et résultats de ce chapitre sont créés à l'aide du logiciel R.

4.1 La méthode "nrc"

L'analyse des sentiments est un processus qui permet de déterminer la nuance émotionnelle qui se cache derrière une série de mots. Cette analyse est utilisée pour mieux comprendre la perception, les opinions et les émotions exprimées dans un texte. Cette technique utilise l'apprentissage automatique pour identifier et caractériser les états affectifs d'un document. En analyse des sentiments, il existe plusieurs lexiques comme "AFINN", "bing" et "nrc" pour étudier un texte. Nous exploitons dans le cadre de ce mémoire la méthode "nrc".

Définition 4.1.1 *En analyse des sentiments, un* ***lexique*** *est une liste qui recense les mots et leur attribue un sentiment.*

J. Silge et D. Robinson [29] nous indique que ces trois dictionnaires sont biaisés. En effet, certains possèdent plus de mots positifs que de mots négatifs et inversement. Sur la Table 4.1, nous constatons que la méthode "nrc" semble la plus équitable et la plus complète même si il y a plus de mots négatifs que positifs.

Ce dictionnaire est composé de deux colonnes : une pour les mots et l'autre pour l'émotion associée au mot. Il y a 13875 mots qui sont considérés. Les mots non considérés par ce lexique sont jugés comme neutres et sans émotion particulière. Les émotions associées sont au nombre de huit : colère, peur, anticipation, confiance, surprise, tristesse, joie, dégoût. Il attribue aussi deux sentiments : positif et négatif.

Table 4.1 – Table des spécificités de trois dictionnaires.

Méthode / Nombre de mots	AFINN	Bing	nrc
positifs	878	2005	2308
négatifs	1599	4781	3318

En vue d'obtenir ce lexique, nous avons besoin du package *tidytext* afin d'utiliser la fonction *get_sentiments("nrc")* en R.

Certaines limitations au sujet de l'analyse des sentiments sont importantes à mettre en garde et sont expliquées par K. Bannister [30]. En effet, l'automatisation via des programmes informatiques comporte des risques d'erreur. Effectivement, apprendre à une machine comment le contexte peut influencer le sentiment d'un texte est difficile. L'ordinateur n'éprouve pas de sentiments. Ce que nous faisons, c'est lui donner des indications pour qu'il fasse un choix mais cela ne remplacera jamais toutes les émotions complexes que l'être humain peut ressentir. Par exemple, les phrases ironiques ou sarcastiques comme : "La température de la planète a augmenté, je n'ai jamais rien vu d'aussi merveilleux !". La machine ne détectera pas que la personne qui parle a un sentiment négatif au sujet de cette augmentation. Le contexte est très important. De plus, les structures grammaticales complexes peuvent aussi poser ce genre de problème.

Certains mots sont plus positifs ou négatifs que d'autres. La fonction *get_sentiment* en R attribue une valeur à chaque phrase donnée en argument. Cette fonction alloue un poids à chaque phrase. Plus cette valeur est grande, plus la phrase est positive et inversement, plus cette valeur est petite, plus la phrase est négative. Afin d'utiliser le dictionnaire "nrc", nous l'ajoutons aux arguments de la fonction ainsi que la langue du texte, qui est ici le français.

4.2 Résultats

Nous mettons en pratique les notions théoriques vues dans la section précédente au jeu de données JOURNAUX. Grâce à la fonction *get_sentences* en R, nous découpons notre corpus en phrases. Nous analysons le poids associé à chacune d'entre elles par la fonction *get_sentiment*. Plus il est petit, plus le sentiment de la phrase est négatif et inversement. À la Table 4.2, nous montrons les informations statistiques dont nous disposons à propos de ces poids.

Nous remarquons que le journal qui est le plus dans l'extrême est le journal *Le Figaro*. Il a les deux valeurs les plus basses : -12 et -11. Ces deux poids correspondent respectivement aux phrases

Table 4.2 – Table d'informations des sentiments des quatres journaux.

	Le Figaro	*FranceSoir*	*Libération*	*Le Soir*
Minimum	-12	-7	-8	-8
1^{er} quartile	0	0	0	0
Médiane	0	0	0	0
Moyenne	0.2919	0.1953	0.2167	0.1827
$3^{ième}$ quartile	1	1	1	1
Maximum	7	7	8	7
Nombre de négatifs	575	352	928	547
Nombre de neutres	1145	673	1697	1258
Nombre de positifs	1031	480	1385	844

– "On peut alors souffrir d'attaques de panique, d'angoisses, d'insomnies, de pensées obsessionnelles, d'émotions négatives (peur, tristesse, impuissance, désespoir, frustration, colère, paralysie), listent Célie Massini et Antoine Pelissolo."

– "Une proportion importante d'enfants et de jeunes dans le monde témoignent d'une détresse significative et d'une large palette d'émotions douloureuses et complexes (tristesse, peur, colère, impuissance, culpabilité, honte, désespoir, chagrin), écrivent les auteurs de ce travail réalisé en mai et juin et publié hier dans le Lancet Planetary Health."

L'article associé à la première phrase évoque la pollution due au réchauffement climatique et comment cela peut nous rendre fou. Le titre correspondant à la deuxième phrase est : "De nombreux jeunes angoissés par la crise climatique". L'analyse des sentiments a donc bien détecté les différents sentiments utilisés dans ces phrases.

Le journal qui a le plus grand extrême positif est *Libération*. Les phrases correpondant aux poids 8 et 7 sont respectivement :

– "Alors que nous vivons dans un système où chaque classe sociale aspire à passer dans celle du dessus, la transition énergétique, nécessaire au respect des objectifs de réduction de gaz à effet de serre gravés dans l'Accord de Paris, offre une opportunité de changer de paradigme, avec la proposition de la création d'un espace écologique qui définirait un seuil minimum et un plafond maximum de consommation des ressources naturelles, pour que chaque humain ait le même droit d'utiliser les ressources."

– "Cette notion scientifique de neutralité carbone introduite dans l'Accord de Paris de 2015 a réussi à s'imposer comme un horizon politique."

Le premier article porte sur la place de la pollution dans le réchauffement climatique. Le titre du deuxième article est “COP26 : Politiquement, il est impossible de passer outre la question climatique”. Nous remarquons que les deux phrases énoncées sont plutôt positives. Les notions de changement et de réussite sont bien prises en compte par l'analyse.

Il y a plus de phrases positives que négatives lorsque le journal *Le Figaro* parle du réchauffement climatique. Nous observons ce phénomène pour tous les journaux. Celui-ci va plus dans les extrêmes que les autres journaux. Le journal *FranceSoir* est celui qui est le moins excessif au niveau du poids des phrases utilisées. Le journal *Le Soir* est en moyenne le journal le moins positif. Cependant, les quatres journaux ont une moyenne légèrement au-dessus de zéro. Le sentiment de ces derniers n'est pas très tranché.

Sur les histogrammes 4.1, 4.2, 4.3, 4.4, nous remarquons que la part de neutralité est grande mais ceci s'explique par le fait que la méthode “nrc” attribue “positif” ou “négatif” à certains mots comme vu à la Table 4.1. Tous les autres mots sont considérés comme neutres et il y en a donc énormément. De plus, nous construisons le test d'hypothèse pour comparer la distribution des poids des sentiments de chaque journal. Certains éléments du tableau de contingence sont en-dessous de cinq. Par conséquent, nous ne pouvons pas utiliser le test χ^2. Les hypothèses sont données par la source [10].

$$\begin{cases} \mathcal{H}_0 & : \forall i, j, p_{i,j} = p_{i,.}p_{.,j} \\ \mathcal{H}_1 & : \exists i, j \text{ tel que } p_{i,j} \neq p_{i,.}p_{.,j}. \end{cases}$$

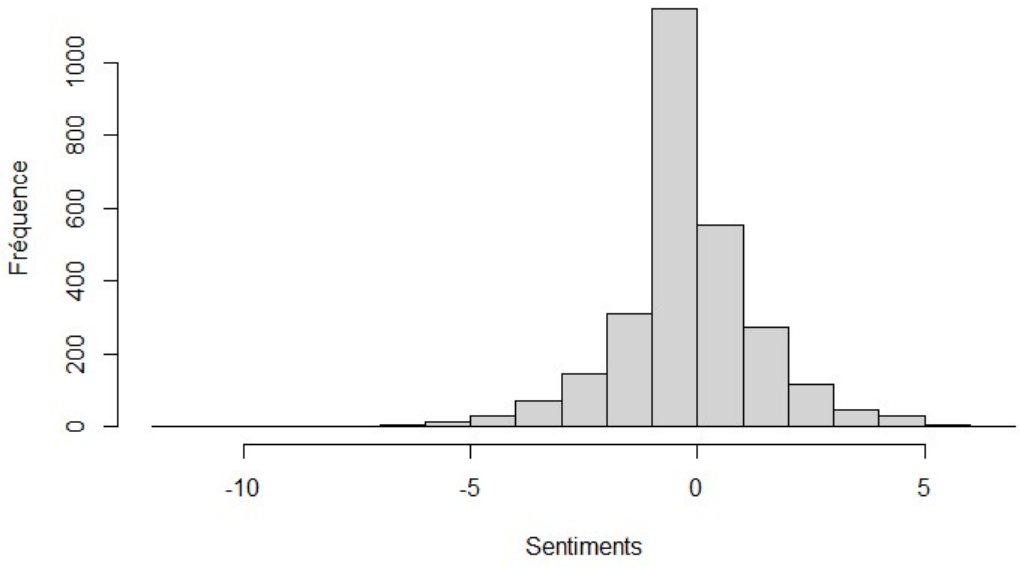

Figure 4.1 – Histogramme de la fréquence de chaque poids de sentiment pour le journal *Le Figaro*.

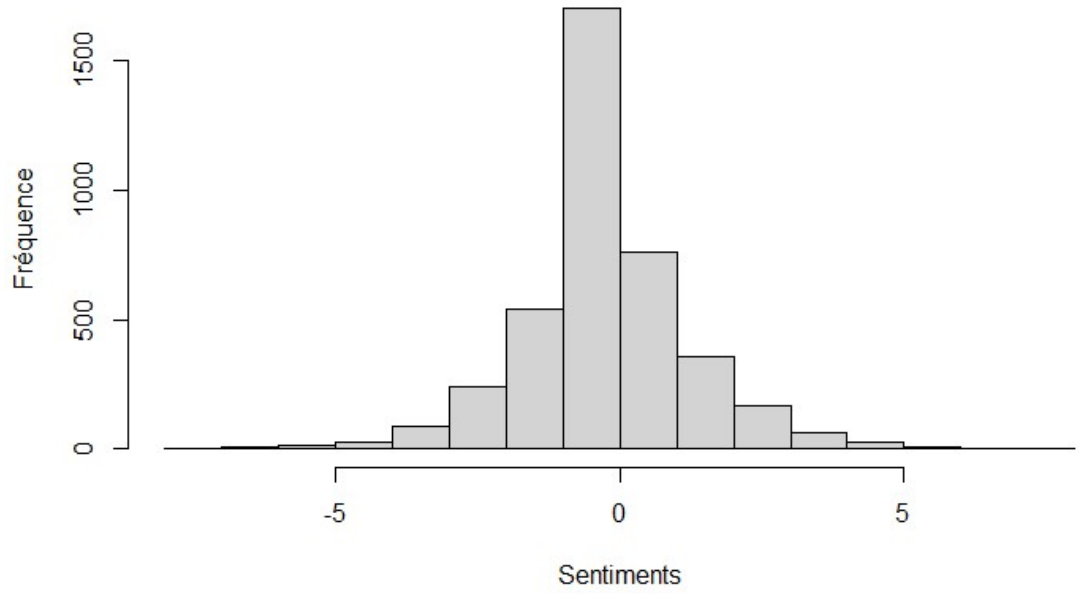

Figure 4.2 – Histogramme de la fréquence de chaque poids de sentiment pour le journal *Libération*.

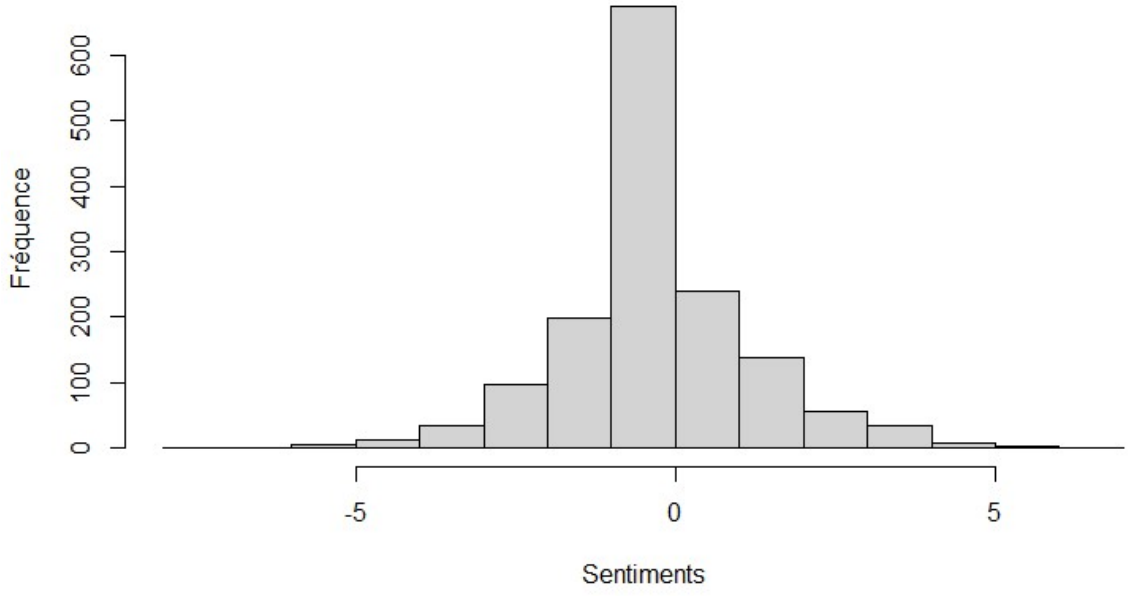

Figure 4.3 – Histogramme de la fréquence de chaque poids de sentiment pour le journal *FranceSoir*.

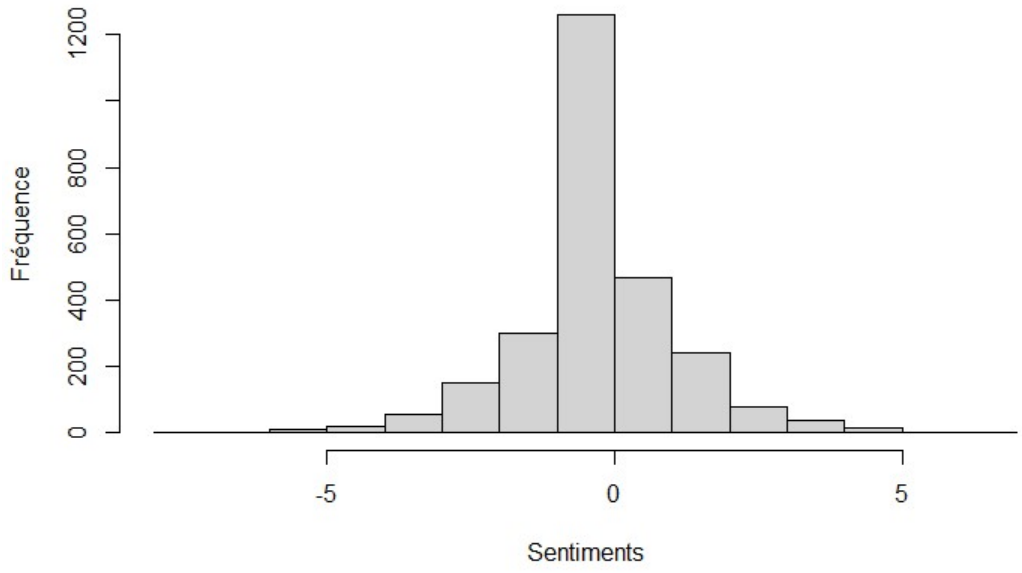

Figure 4.4 – Histogramme de la fréquence de chaque poids de sentiment pour le journal *Le Soir*.

En regroupant les catégories qui possèdent un effectif inférieur à 5, nous sommes capables de faire ce test. Nous obtenons une p-valeur de 0.0005497953 qui est plus faible que le seuil significatif de 0.05. L'hypothèse nulle est rejetée. Nous en concluons qu'il y a un lien entre les journaux et les poids des phrases qui leur sont associés.

Nous procurons une analyse plus poussée concernant les émotions des quatres journaux grâce à la fonction *emotion_by* dans R. Cette fonction nous permet d'obtenir la répartition des différentes émotions du dictionnaire "nrc" d'un journal. Nous effectuons un test χ^2 sur les données liées aux émotions afin de savoir s'il y a un lien avec les journaux. Les informations sont données à la Table 4.3. L'objectif du test est de vérifier si les journaux ont les mêmes émotions. Les hypothèses sont données par la source [10].

$$\begin{cases} \mathcal{H}_0 & : \forall i, j, p_{i,j} = p_{i,.}p_{.,j} \\ \mathcal{H}_1 & : \exists i, j \text{ tel que } p_{i,j} \neq p_{i,.}p_{.,j}, \end{cases}$$

où $p_{i,j}$ est un élément de la Table 4.3.

L'hypothèse nulle peut être réécrite en français de la manière suivante : le fait de connaître le journal ne permet pas d'aider à deviner les émotions d'un individu et inversement. Si l'hypothèse nulle est rejetée, alors il existe une relation journal-émotion. Le nombre de degré de liberté est de 21. La p-valeur est de 0.0003322 qui est plus petite que le seuil de confiance de 0.05. Il y a donc un lien statistique entre le journal et les émotions. L'hypothèse $\mathcal{H}_0$ est rejetée et nous avons démontré qu'il y a une dépendance entre les deux variables.

Table 4.3 – Table des émotions des quatres journaux.

	colère	anticipation	dégoût	peur	joie	tristesse	surprise	confiance
Le Figaro	119	207	109	291	61	129	83	430
FranceSoir	54	93	55	116	19	67	49	148
Libération	169	322	163	369	93	168	168	528
Le Soir	107	136	75	222	43	105	112	243

En effet, nous constatons qu'il y a plusieurs différences entre les journaux. Le journal *Le Soir* se détache légèrement des autres. Les sentiments de colère, surprise et tristesse prennent plus de place que dans les autres journaux. Il utilise plus d'émotions différentes. Le journal *Le Figaro* est celui qui accorde une plus grande partie de ces émotions à la confiance. La notion de tristesse est plus grande pour les journaux *FranceSoir* et *Le Soir*.

De manière plus générale, les émotions principales sont la confiance, la peur et l'anticipation mais il y a aussi de la tristesse et de la colère. Si nous reprenons les sentiments du dictionnaire "nrc", nous constatons que la joie est l'émotion qui a la plus petite proportion pour chaque journal.

Le Tableau 4.4 indique les p-valeurs du test χ^2 de la Table 4.3 uniquement avec les journaux de la ligne et de la colonne associés à la Table 4.4. Le Tableau 4.5 indique les p-valeurs du test χ^2 de la Table 4.3 uniquement avec les émotions de la ligne et de la colonne associées à la Table 4.5. Ces deux tableaux sont arrondis à la quatrième décimale. Les nombres mis en rouge sont les p-valeurs qui rejettent le test χ^2. Il y a donc une dépendance entre les journaux considérés et les émotions considérées. Dans le cas contraire, l'hypothèse nulle n'est pas rejetée mais ce n'est pas pour autant qu'il n'y a pas de lien. Nous ne savons juste pas le démontrer avec ce test.

En conclusion, le sentiment le plus représenté pour tous les jounaux est la positivité face au réchauffement climatique. Il est tout de même à noter que le journal belge l'est moins que les autres. Les journaux ne semblent pas être contre le réchauffement climatique.

Les journaux ont, pour essentiels sentiments, la confiance, la peur et l'anticipation. L'anticipation semble avoir sa place comme sentiment concernant le réchauffement climatique car le combat ne se fait pas de manière immédiate et donc, nécessite une certaine

Table 4.4 – Table des p-valeurs en fonction des associations des journaux.

	Le Figaro	*FranceSoir*	*Libération*	*Le Soir*
Le Figaro		0.0662	0.0296	$2.3428\ 10^{-5}$
FranceSoir			0.3555	0.2468
Libération				0.0068
Le Soir				

Table 4.5 – Table des p-valeurs en fonction des associations des émotions.

	colère	anticipation	dégoût	peur	joie	tristesse	surprise	confiance
colère		0.0872	0.3083	0.7547	0.3043	0.7087	0.1580	0.0226
anticipation			0.8678	0.0481	0.5354	0.0706	0.0009	0.1444
dégoût				0.2493	0.3627	$1.3299\ 10^{-6}$	0.0005	0.7059
peur					0.3211	0.5248	0.0045	0.0580
joie						0.1101	0.0364	0.4412
tristesse							0.0250	0.0197
surprise								$2.8132\ 10^{-6}$
confiance								

anticipation de ce qui pourrait se passer. Le peur est due au dérèglement et aux catastrophes vécues et à vivre à l'avenir. La confiance semble plus surpenante d'autant plus que c'est celle qui prédomine pour les journaux français. La joie est l'émotion qui est la moins utilisée dans tous les journaux. En effet, le sentiment de joie n'est pas le plus adéquat lorsque nous parlons de réchauffement climatique.

Conclusion et perspectives

Un corpus doit tout d'abord être traité afin que nous puissions en tirer facilement des informations. Le corpus est divisé en partitions. Nous procédons à la diminution des mots considérés et à la lemmatisation. Ce processus nous permet d'obtenir la table lexicale du corpus. Le logiciel TXM la crée à partir du corpus. Il produit aussi l'analyse factorielle des correspondances ainsi que la classification ascendante hiérarchique. Ces techniques de visualisation sont illustrées avec un petit jeu de données déjà existant dans TXM. Ces deux méthodes sont complémentaires.

Après avoir parcouru ces concepts, nous examinons un plus grand corpus. La thématique est la manière dont les journaux de la presse quotidienne parle du réchauffement climatique. Plus précisément, le choix des journaux s'est porté sur : *Libération*, *Le Figaro*, *FranceSoir* et *Le Soir*. Nous en concluons que le journal *Le Figaro* est très différent des autres. En effet, il exprime ses opinions plus fortement. De plus, il a un vocabulaire peu commun et orienté sur la politique et le domaine scientifique. Une interprétation possible à ce phénomène est que ce journal est de droite. En effet, il s'adresse a priori à des personnes plus instruites et plus cultivées. Par conséquent, il aborde des aspects plus techniques. Concernant le journal *FranceSoir*, c'est le contraire. Il est très neutre et cela se ressent dans les analyses qui ont été faites. Très peu de mots spécifiques sont utilisés par ce journal. Ensuite, le journal *Libération* est moins démarqué des autres comme nous aurions pu le croire. En effet, son vocabulaire est plus axé sur les évènements qui se produisent dans l'environnement mais pour les autres thèmes, il est éloigné des mots spécifiques. Enfin, le journal *Le Soir* semble neutre mais pas de la même manière que le journal *FranceSoir*. Il emploie des mots scientifiques et relatifs à l'environnement. Ce journal est en effet progressiste.

Quant à l'analyse des sentiments des différents journaux, nous remarquons qu'ils sont globalement neutres lorsqu'ils évoquent le réchauffement climatique mais avec plus de sentiments positifs que négatifs. Les sentiments les plus utilisés par les quatres journaux sont la confiance, la peur et l'anticipation. Le joie est le sentiment le moins représenté.

Les perspectives pour la suite de ce mémoire seraient de prendre en considération plus de journaux belges et couvrir plus d'orientations politiques dans l'analyse. Ce qui nous permettrait de distinguer plus subtilement la différence entre les pays et aussi d'avoir une étude plus complexe au sujet des journaux et leurs regroupements. Plusieurs autres thématiques peuvent être traitées grâce à la textométrie, en français mais aussi dans d'autres langues. Nous pourrions aussi explorer plus en profondeur la théorie de l'analyse des sentiments.

Il est important de noter que nous essayons d'automatiser l'analyse d'un texte. L'expertise du domaine de la linguistique pourrait être intéressante dans ce cadre. De plus, il est à préciser que la corrélation et la causalité sont deux notions différentes. La corrélation établit une relation entre deux variables. Cependant, le fait que ces deux variables ont un lien ne signifie pas forcément qu'une variable est la cause de l'autre. Pour certaines étapes de ce mémoire, le choix de certains paramètres est subjectif. Notamment lors du choix des lemmes pris en considération mais aussi lors de la décision du nombre de clusters. Ces différentes possibilités peuvent être discutées. Le choix de plus de lemmes nous apporte plus d'informations mais ne permet pas de réduire la dimension de la table à analyser. Le nombre de clusters peut aussi parfois être débattu. En effet, malgré la règle mathématique d'un grand saut pour l'inertie inter-cluster, il peut y avoir plusieurs choix en fonction de l'interprétation que nous souhaitons donner. Tout n'est donc pas automatique. Ceci provient du fait que nous analysons des textes. Ils se composent de structures grammaticales complexes et d'interactions riches entre les mots.

Bibliographie

[1] S. Heiden, J-P. Magué et B. Pincemin, *Textométrie*, `http://textometrie.ens-lyon.fr/`, Site du projet Textométrie, consulté le 6 mai 2021.

[2] V. Magri-Mourgues, *Analyse textométrique et interprétation littéraire*, Université de Nice Sophia Antipolis, 77-93, 2011.

[3] Sourceforge, *TXM*, `http://txm.sourceforge.net/doc/ImagesTXM/Logo_TXM.png`, consulté le 23 mai 2022.

[4] S. Heiden, *Manuel de TXM*, ENS de Lyon et Université de Franche-Comté, Février 2018.

[5] M. Gentzkow, B. Kelly et M. Taddy, *Text as Data*, Journal of Economic Literature, 535-574, 2019.

[6] L. Lebart, A. Salem et L. Berry, *Exploring Textual Data*, Kluwer academic publishers, 1998.

[7] 10h11, *TreeTagger : comment lemmatiser une chaîne de caractères*, `https://www.10h11.com/treetagger-comment-lemmatiser-chaine-de-caracteres/`, consulté le 21 mai 2022.

[8] H. Schmid, *Probabilistic Part-of-Speech Tagging Using Decision Trees*, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 1994.

[9] M. Bécue-Bertaut, *Textual Data Science with R*, Chapman and Hall, New York, 2019.

[10] G. Van Bever, *Cours de statistiques*, UNamur, année académique 2019-2020.

[11] G. Van Bever, *Cours d'analyse multivariée et introduction aux logiciels statistiques*, UNamur, année académique 2019-2020.

[12] STHDA, *AFC dans R avec FactoMineR : Scripts Faciles et Cours*, `http://www.sthda.com/french/articles/38-methodes-des-composantes-`

principales-dans-r-guide-pratique/83-afc-dans-r-avec-factominer-scripts-faciles-et-cours, consulté le 14 décembre 2021.

[13] B. Escofier, *Analyse factorielle et distances répondant au principe d'équivalence distributionnelle*, Revue de statistique appliquée, 29-37, 1978.

[14] C. Côté, *Projection orthogonale sur un vecteur*, https://www.youtube.com/watch?v=zhNqS3-KuoI, consulté le 21 mai.

[15] Agence universitaire de la francophonie, *L'Analyse Factorielle des Correspondances* , https://foad-mooc.auf.org/IMG/pdf/M05-3.pdf, consulté le 14 décembre 2021.

[16] P. Bréchon, *Valéry Giscard d'Estaing, le dernier des grands notables de la droite libérale*, https://theconversation.com, consulté le 11 mai 2021.

[17] STHDA, *CAH - Classification Ascendante Hiérarchique dans R avec FactoMineR : Cours*, http://www.sthda.com/french/articles/38-methodes-des-composantes-principales-dans-r-guide-pratique/87-cah-classification-ascendante-hierarchique-dans-r-avec-factominer-cours/, consulté le 14 décembre 2021.

[18] Analyse-R, *Classification ascendante hiérarchique (CAH)*, https://larmarange.github.io/analyse-R/classification-ascendante-hierarchique.html, consulté le 6 décembre 2021.

[19] LAROUSSE, *Définition du progressisme*, https://www.larousse.fr/dictionnaires/francais/progressisme/64216, consulté le 8 avril 2022.

[20] Le Figaro, *Journal Le Figaro*, https://www.lefigaro.fr/, consulté le 8 avril 2022.

[21] France Soir, *Journal FranceSoir*, https://www.francesoir.fr/, consulté le 8 avril 2022.

[22] Libération, *Journal Libération*, https://www.liberation.fr/, consulté le 8 avril 2022.

[23] Le Soir, *Journal Le Soir*, https://www.lesoir.be/, consulté le 8 avril 2022.

[24] Wikipédia, *Libération*, https://fr.wikipedia.org/wiki/Lib%C3%A9ration_(journal), consulté le 11 mai 2022.

[25] Wikipédia, *FranceSoir*, https://fr.wikipedia.org/wiki/FranceSoir, consulté le 11 mai 2022.

[26] Wikipédia, *Le Figaro*, https://fr.wikipedia.org/wiki/Le_Figaro, consulté le 11 mai 2022.

[27] Wikipédia, *Le Soir*, https://fr.wikipedia.org/wiki/Le_Soir, consulté le 11 mai 2022.

[28] K. Gerdes, *Gromoteur*, `http://gromoteur.ilpga.fr`, consulté le 8 avril 2022.

[29] J. Silge et D. Robinson, *Text mining with R, a tidy approach*, O'Reilly Media, 2017.

[30] K. Bannister, *Comprendre l'analyse du sentiment : qu'est-ce que c'est et à quoi ça sert ?*, `https://www.brandwatch.com/fr/blog/comprendre-l-analyse-du-sentiment/`, consulté le 1 mai 2022.

Annexe A

Table lexicale associée au jeu de données JOURNAUX

La table lexicale du corpus JOURNAUX est représentée dans cette annexe. Le corpus est divisé en quatre partitions, une pour chaque journal : “LeFigaro”, “FranceSoir”, “Libération” et “LeSoir”.

Table A.1 – Table lexicale associée au corpus JOURNAUX réparti en 4 groupes.

Lemmes	LeFigaro	FranceSoir	Libération	LeSoir	Total	Numéro de cluster
climatique	359	286	642	423	1710	6
ne	414	205	595	390	1604	6
pas	314	151	472	271	1208	6
réchauffement	343	207	267	384	1201	4
pouvoir	243	142	349	280	1014	6
année	265	173	350	225	1013	5
pays	273	53	192	130	648	1
rapport	212	68	222	124	626	2
changement	147	116	266	88	617	5
effet	197	72	189	153	611	3
devoir	180	91	216	16	503	5
climat	166	75	194	145	580	3
émission	245	52	176	106	579	1
température	117	107	145	145	514	4
si	125	67	160	132	484	6
selon	147	87	121	108	463	4
aussi	107	60	181	84	432	6
gaz	145	35	161	89	430	2
dernier	137	64	120	10	331	3
France	115	105	164	30	414	5

Lemmes	LeFigaro	FranceSoir	Libération	LeSoir	Total	Numéro de cluster
monde	135	53	139	80	407	2
moins	125	43	152	67	387	2
degré	160	32	113	72	377	1
encore	104	35	140	67	346	2
depuis	86	61	128	66	341	5
Europe	88	34	114	97	333	6
eau	77	55	118	81	331	6
mondial	101	33	125	64	323	2
grand	92	33	129	65	319	2
chaleur	80	59	86	91	316	4
scientifique	72	51	125	67	315	6
étude	92	63	83	74	312	4
serre	109	31	86	85	311	3
niveau	84	49	96	78	307	3
contre	81	65	83	68	297	4
sans	81	52	104	53	290	5
carbone	111	25	100	48	284	2
nouveau	92	37	93	62	284	3
premier	90	41	103	48	282	5
objectif	129	22	74	52	277	1
prendre	62	26	114	70	272	6
énergie	95	15	103	52	265	2
certain	62	34	98	62	256	6
politique	56	22	102	76	256	6
réduire	96	29	79	51	255	2
bien	67	32	101	54	254	6
région	88	24	86	53	251	2
CO2	86	19	71	73	249	3
hausse	85	27	72	65	249	3
Paris	98	37	69	38	242	3
impact	95	26	63	57	241	3
mettre	78	29	84	43	234	2
accord	105	23	58	43	229	1
chercheur	69	35	64	59	227	3
planète	68	27	74	58	227	3
siècle	61	36	62	65	224	4
atteindre	88	17	63	54	222	1
Giec	65	8	78	69	220	6
publier	69	36	74	41	220	5
limiter	86	20	70	42	218	2
estimer	81	31	63	41	216	3

Lemmes	LeFigaro	FranceSoir	Libération	LeSoir	Total	Numéro de cluster
extrême	50	24	91	51	216	6
million	58	22	92	44	216	2
expliquer	53	48	71	43	215	4
fin	68	27	71	49	215	3
augmenter	82	24	58	49	213	3
personne	62	27	70	52	211	3
milliard	92	20	69	29	210	1
conséquence	37	43	79	48	207	6
humain	43	33	68	58	202	4
expert	56	27	54	60	197	4
mesure	29	31	82	55	197	6
océan	43	44	67	43	197	4
sécheresse	67	28	61	39	195	3
dire	31	27	80	56	194	6
jour	48	40	54	52	194	4
permettre	67	23	75	26	191	2
important	65	17	63	45	190	2
risque	58	22	62	48	190	3
après	49	27	69	41	186	5
événement	39	25	76	46	186	6
lier	69	29	58	30	186	3
notamment	76	31	52	26	185	3
moyen	50	37	56	41	184	4
augmentation	56	34	50	43	183	4
chaud	32	49	50	50	181	4
international	66	22	68	22	178	2
mer	31	33	61	49	174	6
agir	31	19	72	50	172	6
également	40	27	50	53	170	4
environnement	48	29	62	27	166	5
phénomène	32	26	50	53	161	4
peu	52	17	68	20	157	2
avant	56	19	54	26	155	2
engagement	69	8	55	22	154	1
passer	30	29	58	37	154	6
environ	52	23	44	33	152	3
inondation	45	13	64	30	152	2
vague	37	21	37	57	152	4
aujourd'hui	33	14	54	50	151	6
forêt	44	15	49	43	151	6
fort	34	25	50	42	151	4

Lemmes	LeFigaro	FranceSoir	Libération	LeSoir	Total	Numéro de cluster
glace	19	61	41	29	150	4
question	34	16	56	42	148	6
fossile	56	6	53	32	147	2
partie	34	22	59	32	147	6
exemple	47	15	54	30	146	2
devenir	38	15	62	30	145	2
incendie	63	5	46	31	145	1
jeune	31	8	54	51	144	6
rester	39	26	48	31	144	5
gouvernement	24	14	69	36	143	6
production	46	13	66	18	143	2
temps	45	20	46	31	142	3
chaque	37	18	48	37	140	6
recherche	53	24	43	19	139	3
cause	22	25	54	37	138	6
réduction	55	9	48	26	138	2
venir	42	25	52	18	137	5
scénario	28	23	41	43	135	4
vie	33	21	59	22	135	6
COP26	88	0	28	18	134	1
connaître	30	19	47	37	133	6
action	27	16	60	28	131	6
souligner	38	36	25	32	131	4
ONU	56	19	33	22	130	3
responsable	52	18	32	28	130	3
produire	37	19	45	28	129	5
savoir	35	16	54	24	129	6
rendre	42	13	39	34	128	3
montrer	37	15	46	29	127	3
zone	51	22	38	16	127	3
développement	64	5	45	12	126	1
face	44	11	52	19	126	2
vouloir	27	14	53	32	126	6
cas	23	16	56	30	125	6
canicule	47	16	28	33	124	3
période	28	24	33	38	123	4
terme	50	12	41	20	123	2
feu	69	3	31	19	122	1
Total	11571	5402	13380	8906	39259	