DEUXIÈME PARTIE

Ancrage théorique, description du corpus, méthodologie et outils d’analyse.

PLAN DE LA PARTIE 2

Chapitre 1 : Ancrage théorique, la lexicométrie.

Introduction partielle.

1-Définitions.

2-Aperçu historique : origines et évolution de l’approche lexicométrique.

2-1-Les premiers balbutiements.

2-2-L’Analyse automatique du discours en France.

2-3-Essoufflement de la discipline.

2-4-Les années 1990 et 2000.

2-5-La culturonomique.

3-L’approche lexicométrique, ses principes et fonctionnalités.

3-1-Présentation.

3-2-Débat autour de l’approche lexicométrique.

3-3-Etapes et fonctions de l’analyse lexicométrique.

3-3-1-La numérisation.

3-3-2-La lemmatisation et l’étiquetage.

3-3-3-La distribution du corpus en sous-parties comparables.

3-3-4-Les fonctions documentaires.

3-3-4-1-Les index : la segmentation du corpus en unités graphiques.

3-3-4-2-Les contextes et les concordances.

3-3-5-Les fonctions statistiques : les calculs et les comptages.

3-3-5-1-Le calcul de l’étendue relative des textes.

3-3-5-2-Les graphiques.

3-3-5-3-L’environnement thématique : le calcul des co-occurrences et des réseaux thématiques.

3-3-5-4-L’analyse factorielle des correspondances.

3-3-5-5-L’analyse arborée.

3-3-5-6-Calcul de la distance et la connexion lexicales intertextuelles.

3-3-5-7-Richesse lexicale, hapax et accroissement lexical.

3-3-5-7-1-La richesse lexicale.

3-3-5-7-2-Les hapax.

3-3-5-7-3-L’accroissement lexical.

3-3-5-8-L’analyse des spécificités.

3-4-La lemmatisation, un mal nécessaire !

3-5-Synthèse, vers la logométrie.

Conclusion partielle.

Chapitre 2 : Description et caractéristiques du corpus.

Introduction partielle.

1-Les livres.

1-1-*Guerre et Révolution d’Algérie I : La Nuit Coloniale*.

1-2-*Autopsie d’une guerre : L’Aurore.*

1-3-*L’Indépendance confisquée 1962-1978.*

1-4-*Demain se lèvera le jour.*

2-Les écrits de journaux.

2-1-Articles des années 1920*.*

2-1-1-Le service militaire des Indigènes algériens.

2-1-2-L’exode des ouvriers algériens en France.

2-1-3-L'Intellectuel musulman en Algérie.

2-1-4-La Tragédie d'hier et l'incertitude de demain.

2-1-5-Justice et loyauté d'abord, politique après.

2-2-Articles des années 1930*.*

# 2-2-1-Ecrits parus dans *Ettelmidh*(1931-1933)

# 2-2-2-Ecrits parus dans *L’Entente Franco-Musulmane* (1935-1942)

2-3-Articles des années 1940 et 1950*.*

# 2-3-1-Articles parus dans *Egalité*.

# 2-3-2-Articles parus dans *La République Algérienne*.

# 2-3-3-Ecrits parus dans *El Moudjahid.*

# 3-Les autres publications.

# 3-1-Le Rapport au Maréchal Pétain (ou L’Algérie de demain) (1941)

# 3-2-Le Manifeste du Peuple Algérien (1943)

# 3-3-L’Additif au Manifeste du peuple algérien (1943)

# 3-4-Mon Testament politique (1946)

# 3-5-La Lettre de démission de l'Assemblée nationale constituante(1963)

# 3-6-L’Appel au peuple algérien (1976)

Chapitre 3 : Outils d’analyse et méthodologie.

1-Méthodologie de la recherche.

2-Quelques outils lexicométriques.

2-1-Le logiciel Hyperbase.

2-2-Le logiciel Lexico.

2-3-Le logiciel Tropes.

CHAPITRE 1

Ancrage théorique

La lexicométrie

**Introduction partielle :**

La lexicométrie demeure peu connue et peu exploitée au sein même de sa discipline mère, les Sciences du langage[[1]](#footnote-2). Il n’en demeure pas moins qu’elle constitue depuis son apparition à la fin des années 1960 une approche heuristique particulièrement utile de par la panoplie d’analyses qu’elle permet d’effectuer sur des corpus étendus et le large éventail d’axes de recherche qu’elle peut dégager.

Le peu d’intérêt qu’on accorde à cette approche demeure inexplicable. Il faut noter qu’en France, l’Analyse du discours a été à ses débuts essentiellement automatique et a commencé à se développer, en partie, grâce aux travaux du Centre de Linguistique Quantitative de Paris créé en 1962, ceux effectués en lexicométrie politique par Maurice Tournier à l’ENS de Saint-Cloud, et aux travaux de Michel Pêcheux de l’Université de Paris VII (Analyse Automatique du Discours, l’AAD). On peut donc relever que les premières recherches s’inscrivaient dans une perspective quantitative et statistique du discours puisqu’elles faisaient une large place à l’outil informatique[[2]](#footnote-3).

Par ailleurs, l’originalité de cette approche réside dans le fait qu’elle ouvre la voie à des analyses dont la réalisation était jusque-là impossible. En effet, la consolidation de l’analyse traditionnelle par une analyse automatique quantitative a introduit une rupture dans le champ de l’analyse des données. Plus précisément, avant l’avènement de l’analyse linguistique assistée par ordinateur, il était impossible de pouvoir travailler sur des macro-corpus.

Grâce à la mathématisation de la recherche linguistique, il est devenu possible de travailler sur des corpus plus étendus et d’assurer une certaine objectivité des traitements[[3]](#footnote-4). Avant l’apparition de la statistique lexicale, se posait le problème de la représentativité des échantillons étudiés (Il était alors impossible de travailler sur des macro-corpus). Grâce à la lexicométrie, on peut par exemple appréhender l’œuvre complète d’un auteur sans se soucier du problème de la représentativité des données puisque l’analyse peut prendre en charge l’intégralité de l’œuvre.

Par exemple, comment garantir, lorsqu’il s’agit d’aborder l’œuvre d’un auteur, que les échantillons soumis à l’analyse soient représentatifs de toute sa production littéraire puisqu’on sait que même si elles appartiennent au même auteur, des œuvres peuvent présenter des dissemblances discursives? Avec l’avènement de l’approche lexicométrique, ces obstacles ont été levés.L’analyse automatique du discours est, de l’avis de ses défenseurs, une autre manière, inhabituelle et originale, d’appréhender les discours.

L’outil lexicométrique s’avère, *du point de vue de l’analyse de discours, d’un très grand intérêt, dans trois directions principales* :

*- par les données quantitatives fournies, les comparaisons et les vérifications qu’il permet ;*

*- comme outil de repérage de pistes de recherche, et comme premier bilan d’un corpus ;*

*-comme outil heuristique puissant, entraînant à des allers-retours fructueux entre le texte analysé et les données produites. II incite à une définition plus fine des données et à des comparaisons vers d’autres corpus. Il oblige également à une réflexion sur le statut du « quantitatif » dans le discours à l’écrit et à l’oral ».[[4]](#footnote-5)*

Dans cette partie, nous nous proposons de présenter la lexicométrie, ses méthodes et ses fonctions ainsi que les outils qu’elle met en œuvre, c’est-à-dire les logiciels, tout en brossant un tableau qui retrace succinctement son évolution à travers le temps, depuis ses premiers balbutiements jusqu’à nos jours.[[5]](#footnote-6)

Commençons alors par les définitions.

**1-Définitions :**

Le vocable « lexicométrie » désigne la discipline qui prend en charge l’analyse informatisée du discours et du lexique. Cette jeune discipline est appelée également « analyse du discours assistée par ordinateur » ou encore « traitement automatisé du discours ». Les appellations proposées pour appeler « l’étude scientifique du discours faite avec l’outil informatique » sont aussi nombreuses que diverses et témoignent de l’état de fluctuation dans lequel se trouve la nouvelle discipline qui cherche encore ses contours. En voici quelques-unes : « Textométrie ou Statistique textuelle» (André Salem), « Logométrie » (Damon Mayaffre), « Traitement informatique des données textuelles », « Statistique lexicale (ou lexicostatistique) » (Muller), « ATO (Analyse de Texte Assistée par Ordinateur) », « Traitement électronique de textes », « Analyse du discours assistée par ordinateur » (Pascal Marchand), « Analyse informatisée des textes », « Analyse discursive assistée par ordinateur », « L’analyse textuelle informatisée », « Traitement électronique de textes. », « Traitement automatisé du discours » (Pierre Fiala), « Linguistique quantitative » (Muller), etc.

Il faut noter que des nuances peuvent exister entre ces différentes dénominations qui désignent des approches plus ou moins différentes mais qui se rejoignent toutes dès lors qu’elles ont le même objet d’étude, le discours, et se fixent le même objectif, l’analyse du discours en utilisant l’outil informatique. Ces dénominations ont été proposées par des chercheurs et des spécialistes différents qui ont chacun leur conception de l’approche.

Le vocable « Lexicométrie » est apparu en France dans les années 1970. Il peut être scindé en deux composants : « lexico » et « métrie ». Le premier renvoie, comme on peut le remarquer, au « lexique » tandis que le second renvoie aux « mesures » ou aux « calculs ». On peut dire donc que le mot « lexicométrie » désigne la discipline qui s’intéresse à la mesure du lexique.

Par ailleurs, les définitions existantes sont nombreuses et diverses tout autant que les vocables et les expressions utilisés pour nommer la jeune discipline. Nous nous contentons ici d’en présenter quelques-unes :

-*Le Dictionnaire de la Linguistique* de Georges Mounin définit la lexicostatistique comme étant le :

« *Domaine de la lexicologie dans lequel les procédures de la statistique sont utilisées pour l’étude quantitative du lexique. Ces études vont de simples comptages à des calculs plus élaborés suivant les formules générales de la statistique ou suivant des formules adaptées.* »[[6]](#footnote-7) Cette discipline sert essentiellement à « *la recherche de fréquence pour l’élaboration de vocabulaires fondamentaux, ou en stylistique pour caractériser les styles individuels ou collectifs, apporter des éléments pour apprécier l’évolution d’un auteur, voire contribuer à résoudre des problèmes de datation ou d’attribution d’une œuvre ; en outre, ces études peuvent déboucher sur des théories générales concernant la structure du langage ou la communication.* »[[7]](#footnote-8)

-*Le Grand Dictionnaire de la Linguistique et des Sciences du langage* définit la lexicostatistique (Synonyme de lexicométrie) comme :

« *L’application des méthodes statistiques au vocabulaire d’un texte, d’un ensemble d’énoncés considérés comme représentatifs d’un auteur ou de la langue elle-même. Son domaine d’application sera : la stylistique, si l’on cherche à apprécier la richesse du vocabulaire d’un écrivain, la philologie, si l’on tente de dater des textes les uns par rapport aux autres, la pédagogie, si l’on essaye de définir le vocabulaire fondamental d’une langue, enfin la linguistique générale, si l’on cherche à déterminer la fréquence des mots et leur rang fréquentiel obéissant à des lois générales.*»[[8]](#footnote-9)

-Le *Dictionnaire d’Analyse du Discours* de Patrick Charaudeau et Dominique Maingueneau considère la lexicométrie, non pas comme une théorie, mais comme une méthodologie d’étude du discours, qui se veut exhaustive, systématique et automatisée ; comme une pratique qui consiste à mesurer (métrie) des unités lexicales (lexico).[[9]](#footnote-10)

-Leimdorfer et Salem proposent la définition suivante :

« *On regroupe sous le terme de lexicométrie toute une série de méthodes qui permettent d’opérer, à partir d’une segmentation, des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire.* »[[10]](#footnote-11)

-Enfin, Dominique Maingueneau définit la lexicométrie comme étant une « *Discipline auxiliaire de l’analyse du discours qui vise à caractériser un ensemble discursif (souvent un positionnement) par rapport à d’autres appartenant au même espace grâce à l’élaboration informatique de réseaux quantifiés de relations significatives entre ses unités. Il s’agit par conséquent d’une démarche essentiellement comparative.* »[[11]](#footnote-12)

Cependant, aussi diverses et aussi nombreuses que puissent être les dénominations et les définitions de la jeune discipline, il faut retenir qu’elle s’intéresse au discours en analysant son lexique. Pour ce faire, elle fait appel à des méthodes et des outils mathématiques, statistiques et informatiques. C’est une approche essentiellement quantitative du moment qu’elle privilégie des calculs, des statistiques et des comptages. Mais il ne faut pas perdre de vue qu’elle ouvre la voie à d’autres analyses, notamment qualitatives. Elle n'est donc qu'*une méthode mise au service d'une perspective plus globale qui lui donne sens.[[12]](#footnote-13)*

La lexicométrie se scinde en deux grandes branches. La première s’est constituée autour de Charles Muller de l’Université de Strasbourg et s’intéresse au traitement statistique des textes littéraires, d’où son nom de« lexicométrie littéraire ». La revue *Lexicometrica[[13]](#footnote-14)* est actuellement le principal organe de diffusion de ses travaux[[14]](#footnote-15).La lexicométrie littéraire s’intéresse notamment au style d’un écrivain, à la richesse et à l’évolution de son vocabulaire. La seconde branche, la lexicométrie politique, s’est développée vers la fin des années 1960 grâce aux travaux de Maurice Tournier au laboratoire *Lexicométrie et textes politiques* de l’ENS de Fontenay-Saint-Cloud. La tâche principale d’une telle démarche est d’examiner les rapports entre politique et phénomènes langagiers. Lorsqu’il s’agit d’appréhender un texte sociopolitique, la lexicométrie est chargée « *d’examiner, à partir de corpus de textes soumis à comparaison, comment les termes échangés dans l’espace public autour des enjeux de pouvoir rendent compte des luttes d’appropriation ou de dépossession symboliques qui se jouent dans le lieu même de l’échange*. »[[15]](#footnote-16)

La revue *Les Langages du politique* a constitué pendant longtemps le principal organe de diffusion des travaux effectués dans le cadre de cette branche.

En outre, bien que récente, l’intérêt pour l’approche lexicostatistique commence à se manifester dans plusieurs champs de recherches[[16]](#footnote-17) tels que la médecine, l’histoire, la bibliothéconomie (on parle alors de bibliométrie), les sciences politiques, la sociologie, la psychologie, etc. Toutes ces disciplines et d’autres encore font appel aux méthodes d’analyse statistique des textes pour analyser des corpus volumineux en rapport avec des objets de recherche relevant de leurs domaines respectifs. Ce *boom* est du notamment au perfectionnement des outils lexicométriques et à l’enrichissement de leurs fonctionnalités, rendu possible par la révolution numérique et le développement technologique des dernières décennies.

Cependant, cette discipline peut être qualifiée d’ésotérique parce qu’elle n’est connue et pratiquée que par certains initiés. Sa maitrise nécessite une connaissance assez poussée des techniques sur lesquelles elle s’articule. Par exemple, la manipulation des outils utilisés nécessite une initiation à ces méthodes. Certains logiciels de statistique lexicale ne sont manipulables que par des personnes initiées à l’outil informatique tandis que certains d’autres sont conçus de manière à être faciles à utiliser aussi bien par les informaticiens que par les analystes du discours qui ne n’ont pas effectué une formation spécialisée en informatique.

Pour une bonne compréhension de cette discipline, nous nous proposons d’examiner son évolution chronologique afin de voir comment elle s’est développée à travers le temps.

**2-Aperçu historique : origines et évolution de l’approche lexicométrique :**

**2-1-Les premiers balbutiements :**

Si l’on fait remonter les premiers travaux de l’analyse automatique du discours au milieu du XXe siècle, il n’en demeure pas moins que beaucoup de savants ont effectué, à travers l’Histoire, des mesures sur des textes et calculé leurs mots, surtout dans le but d’étudier le style de leurs auteurs.

Avant le XXe siècle, des tentatives de mesure de vocabulaire sont à relever. Le travail s’effectuait de manière manuelle. Par exemple, des savants ont essayé de calculer le nombre de mots et de syllabes des livres sacrés, tandis que d’autres ont essayé de mesurer les fréquences d’un mot dans un texte ou de calculer le nombre de vers dans des textes de poésie. D’autres s’intéressèrent au style et à la longueur des mots de certains auteurs. Selon Stéphane Sinclair,

«  *On trouve au cours des siècles un grand nombre d’exemples de calculs statistiques sur les divers signes langagiers qui ressemblent beaucoup à certaines branches contemporaines de la stylométrie[[17]](#footnote-18) et de la linguistique computationnelle.*»[[18]](#footnote-19)

Par exemple, *«* *certaines approches remontent au moins jusqu'à un grammairien sanskrit de la période Sutra (500-200 av. J.-C.) qui aurait donné le nombre de vers, de mots et de syllabes du Rig-Veda, livre sacré des Hindous.* »[[19]](#footnote-20)

Sinclair donne d’autres exemples plus ou moins récents de chercheurs ayant tenté d’effectuer des études quantitatives sur des textes appartenant à certains auteurs. On peut citer : Boole (en 1854, il développe des principes pour l'analyse de fréquences)*,* Morgan (en 1859, il propose d'identifier des critères objectifs du style), [Mendenhall](file:///D:\LEXICO\1-G%C3%A9n%C3%A9ralit%C3%A9,%20lexico\2-HISTORIQUE\SINCLAIR\auteurscvbv_fichiers\iaa.htm#_Mendenhall,TC_Characteristics:1887) (en 1887, il examine la longueur des mots de Dickens et de Tackeray) et [Yule](file:///D:\LEXICO\1-G%C3%A9n%C3%A9ralit%C3%A9,%20lexico\2-HISTORIQUE\SINCLAIR\auteurscvbv_fichiers\iaa.htm#_Yule,GU_Sentence-Length:1938) (en 1938, il considère le nombre de mots par phrase comme indicateur pertinent du style d'un auteur.)[[20]](#footnote-21)

Nous pouvons donc dire que cette approche est ancienne et trouve ses racines dans les siècles passés. Mais la discipline telle que nous la connaissons aujourd’hui a traversé plusieurs phases et a dû surmonter des obstacles d’ordre épistémologique et matériel avant d’atteindre son état actuel.

Par ailleurs, le développement de cette branche de recherche a toujours été tributaire de l’évolution de l’informatique, laquelle lui fournit les outils statistiques nécessaires.

Au XXe siècle, les premiers projets d’analyse automatique du lexique sont dus à Roberto Busa (1913-2011), savant et prêtre italien qui voulait :

*« Retracer l'évolution du concept de l'immanence divine chez saint Thomas d'Aquin en étudiant à l'aide de concordances l'usage de la préposition «en». »[[21]](#footnote-22)*

Busa est le pionnier de l’utilisation de l’ordinateur dans l’analyse linguistique et littéraire puisque c’est le premier à avoir entrepris des analyses de textes informatisées. Citons-le :

« *Il faut distinguer la microanalyse herméneutique computérisée que j'ai commencée il y a plus de 50 ans. J'en ai eu l'idée en effet dans les années 1941-1945, pendant la guerre, en faisant ma thèse de doctorat sur "la terminologie thomiste de l'intériorité", thèse que j'ai soutenue en 1946 et publiée en 1949, et cette année-là, j'ai commencé à "bricoler" à New-York avec les machines IBM.* »[[22]](#footnote-23)

Parrainé par l’entreprise américaine IBM, Busa a pu ensuite étendre ses recherches à d’autres écrivains médiévaux qui ont un lien direct avec les textes de Saint Thomas d’Aquin. Aujourd’hui, il est considéré comme le père fondateur de l’analyse automatique des textes.

Mais il faut évoquer également les noms de Georges Kingsley Zipf (1902-1950) et Georges Udny Yule (1871-1951) qui ont tous les deux contribué au développement de la discipline. Le premier, linguiste et philosophe américain, a appliqué la statistique à l’étude des langues. Son nom restera à jamais dans l’analyse automatique du discours car il a inventé la loi qui porte son nom – la Loi de Zipf – Cette loi permet de mesurer la fréquence des mots dans un texte. Elle est largement utilisée dans les recherches de statistique linguistique. Le second, statisticien britannique, a publié en 1944 le premier livre d’analyse automatique des textes. Ce livre s’intitule *The Statistical Study of Literary* *vocabulary.* Son titre peut se traduire en français comme suit : *L’Étude statistique du vocabulaire littéraire.*

Jusqu’aux années 1960, les travaux étaient peu nombreux et restèrent cloitrés entre un nombre très limité de chercheurs épars. C’est à partir de ces années que l’intérêt pour l’analyse informatisée des textes a augmenté parmi les chercheurs. Il était favorisé par la sophistication de quelques outils plus ou moins performants. Néanmoins, ses machines, bien que révolutionnaires aux yeux des chercheurs de l’époque, offraient des fonctionnalités restreintes et le nombre limité de caractères qu’elles possédaient ne permettait pas la transcription des textes dans toutes les langues. L’intérêt des chercheurs se portait essentiellement sur les index et les concordances à cause des limites matérielles des outils disponibles.

Les recherches qui se pratiquaient pendant les années 1960 s’appuyaient sur des applications simples. Par exemple, les machines utilisées ne permettaient pas de désambigüiser les homographes[[23]](#footnote-24) dans un texte. Les premières préoccupations des chercheurs étaient d’objectiver ou de réfuter, par des mesures et des comptages, les appréciations émises auparavant par des spécialistes en littérature, des stylisticiens, par exemple. Le travail consistait surtout à comparer le vocabulaire des différents auteurs ou de mesurer les évolutions ou les permanences du vocabulaire d’un seul auteur à travers la période pendant laquelle son œuvre a été produite.

Durant les années 1970 et 1980, une amélioration nette et importante est à signaler grâce à la prolifération de recherches et à l’augmentation de l’intérêt pour ce type d’analyses. En effet, de plus en plus de chercheurs sont versés dans ce domaine. Les recherches ont fait un bond en avant grâce à l’apparition du micro-ordinateur au début de la décennie 1970. Avant, le matériel était très dispendieux et même impossible à acquérir. Pendant ces années,

« *Le terrain se préparait pour une expansion de l’informatique dans tous les domaines de recherche, car plus les ordinateurs sont bon marché, plus les chercheurs vont pouvoir s’en servir, plus les chercheurs s’en servent, plus les méthodes et les applications d’analyse de texte informatisée auront des chances de se raffiner*. »[[24]](#footnote-25)

**2-2-L’Analyse automatique du discours en France :**

En France, les précurseurs sont notamment Pierre Guiraud (1912-1983) et Charles Muller qui peuvent être considérés comme les fondateurs de la lexicométrie littéraire. Guiraud développa une approche scientifique qui prend en charge l’examen du style d’un auteur en mettant en œuvre une méthode statistique[[25]](#footnote-26). Les travaux de Charles Muller portèrent essentiellement sur le théâtre classique.[[26]](#footnote-27)

En 1962, est créé le Centre de linguistique quantitative de Paris. Il donnera naissance peu après à la revue *Langages* qui est restée, des décennies durant, un lieu de rencontre important pour les linguistes et les analystes du discours. Le dit centre regroupa les noms les plus représentatifs de la linguistique française de l’époque.

Herdan (1964), développe *La linguistique statistique*. Son approche s’intéresse à l’analyse linguistique sur les plans lexical, phonologique et phrastique en utilisant des méthodes statistiques.

Le philosophe Michel Pêcheux (1938-1983) est lui aussi un des promoteurs de la discipline avec sa fameuse Analyse Automatique du Discours (AAD), approche qu’il créa en 1969. Le choix de Pêcheux pour l’automatisation de l’analyse du discours était motivé, si on suit Francine Mazière[[27]](#footnote-28), par sa préoccupation de défendre les Sciences humaines contre ce qu’on appelle les « Sciences dures ». L’automatisation de la recherche était donc incontournableafin de « *garantir une scientificité à l'objet « langue »[[28]](#footnote-29).* » Il développe alors l’Analyse Automatique du Discours (l’AAD qui constitue une entreprise très innovante car elle a eu le mérite de consacrer l’apparition de l’Analyse du discours en France.

Pour élaborer sa propre théorie sur le langage, Pêcheux s’est beaucoup inspiré du philosophe Louis Althusser - dont il est le continuateur – et de sa réflexion sur l’idéologie marxiste et la psychanalyse. Ses premiers travaux s’inscrivaient dans une perspective quantitative.

Il faut rappeler aussi les recherches sur le discours politique initiées par Maurice Tournier et son équipe au Laboratoire « Lexicométrie et textes politiques » de l’ENS de Fontenay Saint-Cloud (depuis 1967). Cette équipe de recherche se donna comme principale tâche de s’intéresser aux rapports entre politique et langue et développa la lexicométrie politique, approche fondée sur le traitement statistique des discours sociopolitiques. Le laboratoire de Saint-Cloud se dota en 1981 d’une revue « *Mots/Ordinateurs/Textes* ». Elle regroupa les principaux collaborateurs du laboratoire et constitua le principal organe de diffusion de leurs travaux. Au fil des années, notamment durant les années 1980, elle est devenue le principal lieu institutionnel de l’approche lexicométrique. Quelques années plus tard, la revue a changé de nom et est devenue « *Mots/Les langages du politique* ». Le laboratoire de Saint-Cloud a disparu en 2000.

Tout comme Pêcheux, Tournier a adopté la mathématisation de la recherche dans un souci d’ordre scientiste. En rattachant l’analyse linguistique à des sciences dures - les mathématiques et l’informatique -, il voulait donner un aspect plus scientifique et plus empirique aux analyses du langage.

Il faut retenir aussi le nom d’Etienne Brunet de l’Université de Nice qui a contribué fortement au développement de la démarche. Ses travaux portèrent sur l’ensemble de la littérature française. Sa thèse intitulée *Le Vocabulaire de Giraudoux. Structure et évolution*, fut soutenue en 1976 et publiée en 1978. Elle constitue une source incontournable pour les chercheurs en lexicométrie littéraire. Il publia en 1981 son ouvrage *Le Vocabulaire français de 1789 à nos jours* (3 tomes)[[29]](#footnote-30). Son œuvre prolifique et ses recherches avant-gardistes lui ont valu une notoriété incontestée dans ce champ de recherche. Ses travaux portent sur le vocabulaire des grands auteurs classiques de la littérature française tels que Proust (1983), Zola (1985), Hugo (1988), etc. On lui doit surtout Hyperbase, le logiciel de lexicométrie le plus perfectionné.

Charles Muller publia en 1973 son *Initiation aux méthodes de la statistique linguistique[[30]](#footnote-31)*. Cet ouvrage présente« *les opérations statistiques appliquées à des faits de langue et de style. Il prend acte du fait que toute manifestation du langage, tout texte, est quantifiable car il se divise en unités (phrases, mots, phénomènes) que l'on peut compter*. »[[31]](#footnote-32)

Il publiera quelques années plus tard un autre ouvrage intitulé *Principes et méthode de la statistique lexicale.[[32]](#footnote-33)*

Ainsi, grâce à, entre autres, Brunet, Guiraud et Muller, la statistique est

« *Entrée* *dans la critique et s'est diversifiée pour s'appliquer à des auteurs particuliers. »[[33]](#footnote-34)* Les possibilités offertes par l'informatique expliquent les progrès de cette approche des textes.[[34]](#footnote-35)

**2-3-Essoufflement de la discipline :**

Les recherches ont connu un déclin pendant les dernières années de la décennie 1980. Les contraintes épistémologiques et matérielles étaient parfois insurmontables. La quête d’outils plus performants commença à se faire sentir parmi les chercheurs qui s’étaient rendu compte de l’insuffisance de l’Analyse Automatique du Discours de Pêcheux. Cette dernière ne permettait pas de rendre compte des relations syntaxiques dans un texte et son analyse était surtout verticale (index des occurrences) alors que le souci des chercheurs était de rendre compte aussi des relations horizontales pour appréhender les séquences discursives et par là même le sens du texte. Il fallait donc inventer des analyseurs syntaxiques performants, vœu que les spécialistes n’ont pas cessé de formuler.

Il est vrai que beaucoup de chercheurs ont contribué au développement de quelques outils en entreprenant des projets assez originaux. Mais ils se heurtaient toujours à une difficulté de taille : la préparation des textes (donc leur numérisation). L’enregistrement des textes se faisait manuellement et prenait une partie importante du temps du chercheur. C’est dire que l’équipement utilisé était primitif et n’avait rien à voir avec les outils très sophistiqués de nos jours. C’est pourquoi nous avons indiqué plus haut que le développement de la jeune discipline a toujours été dépendant du développement de l’outil informatique. Nous pouvons même affirmer que, sans les progrès de l’informatique, l’analyse automatique du discours n’aurait pas vu le jour.

Néanmoins, si l’analyse informatisée de textes doit beaucoup à l’informatique, il n’en demeure pas moins que cette dernière en doit autant à la première, car les chercheurs, en revendiquant sans cesse des machines plus puissantes, ont contribué inversement à l’évolution de l’informatique.

**2-4-Les années 1990 et 2000 :**

De nos jours, les obstacles auxquels la jeune discipline a fait face dans les premières décennies de sa courte vie, ont été en partie levés. La disponibilité de bases de données contenant des textes numériques en grandes quantités, accessibles sur le web, ainsi que la disponibilité de logiciels assez sophistiqués ont rendu possible le redémarrage de la discipline.

La décennie 1990 est marquée par deux moments saillants dans l’évolution de la lexicostatistique : la publication en 1994, par Ludovic Lebart et André Salem, d’un ouvrage intitulé *Statistique textuelle[[35]](#footnote-36) ;* etla création de la revue *Lexicométrica[[36]](#footnote-37)*, principal organe de diffusion des travaux dédiés à l’analyse lexicométrique des textes. Le livre de Salem et Lebart présente les concepts de base et les fondements de la statistique textuelle. Il est illustré d’exemples d’application et est considéré aujourd’hui comme une référence incontournable dans ce domaine de recherche.

Au début du XXIe siècle, le nom à retenir est celui de Damon Mayaffre de l’Université de Nice, auteur d’une thèse en lexicométrie politique publiée en 2000. Elle traite du discours politique français de gauche et de droite dans l'entre-deux-guerres. Il est également l’auteur de nombreux ouvrages comme *Paroles de président. Jacques Chirac (1995-2003)* et *Le discours présidentiel sous la Ve République[[37]](#footnote-38)*. Le plus récent s’intitule *Mesure et démesure du discours*.*[[38]](#footnote-39)* Il analyse le discours présidentiel de Nicolas Sarkozy de 2007 à 2012. Il est aussi l’auteur d’autres publications comme l’article « De la lexicométrie à la logométrie » dans lequel il préconise la refonte de l’approche lexicométrique par le glissement de la lexicométrie «classique» vers la logométrie. Cette dernière, contrairement à la lexicométrie, étend son analyse à d’autres unités du discours en plus des formes graphiques ou mots. Elle prend en compte d’autres unités linguistiques du corpus analysé : les lemmes, les structures grammaticales, sémantiques et rhétoriques, etc. Elle privilégie la prise en compte du contexte linguistique (ou cotexte) dans le travail d’analyse. Son émergence a été rendue possible par le développement technologique qui a favorisé l’optimisation des capacités des outils d’analyse dont l’ergonomie est de plus en plus adaptée aux besoins de l’analyse.

**2-5-La culturonomique :**

Les premières années du troisième millénaire consacrent la naissance d’une approche qui n’arrive toujours pas à recevoir le crédit des spécialistes de l’analyse informatisée du discours. Cette nouvelle approche s’appelle la culturonomique. Elle est née suite au lancement d’un projet gigantesque par la société américaine de l’Internet Google. Cette société lance en 2003 un projet de numérisation de tous les livres publiés jusqu’à nos jours dans le monde. Les livres numérisés constituent une base numérique lexicale de l’ordre de quelques milliards de mots. Google Labs, filiale de Google, développe et met en ligne un outil de traitement statistique des textes appelé Google Ngram Viewer. La nouveauté tient dans le fait que le nouvel outil permet d’effectuer des analyses lexicométriques sur une base numérique lexicale impressionnante constituée de plusieurs millions de livres, appelée GOOFRE et couvrantdes livrespubliés dans le monde depuis l’origine de l’imprimerie.

Google Ngram Viewer est un outil qui permet de « *Retracer sur une très longue durée l’évolution des fréquences d’emploi de termes ou d’expressions, ce qui permet aux porteurs de ce projet de revendiquer un instrument de portée très générale, pouvant prendre en charge de nombreuses problématiques.*»[[39]](#footnote-40)

Ce nouveau protocole ne fait pas l’unanimité des linguistes et des analystes de discours spécialisés en lexicométrie car il présente plusieurs points faibles comme des erreurs de lecture.[[40]](#footnote-41) La culturonomique assume :

« *Une forme de réductionnisme, à une époque où l’on ne cesse de raisonner à partir de cartographies complexes et où les choses ne valent qu’en relation les unes avec les autres*. »[[41]](#footnote-42)

Disons *in fine* que malgré les contraintes matérielles et épistémologiques auxquelles elle se heurte depuis les premières années de son existence, la lexicométrie est en évolution continuelle. Cette évolution dépend notamment de l’évolution technologique et informatique. Une question se pose et s’impose : compte tenu de cette évolution continue, que peut apporter l’avenir à la recherche automatique du discours ? Vers quels horizons se dirige-t-elle ? Dans l’état actuel des choses, il semble difficile de répondre à cette question. Les plus pessimistes des spécialistes de la discipline lui prévoient un avenir certain.

Ce bref aperçu historique n’est pas exhaustif car des noms de grands chercheurs du domaine restent à découvrir. Nous nous sommes contenté de passer en revue les évènements les plus saillants qui ont marqué la vie de l’approche lexicométrique en citant ses plus grands noms, précurseurs et continuateurs.

**3-L’approche lexicométrique, ses principes et fonctionnalités :**

Il faut convenir que les indications données précédemment ne sont pas suffisantes pour comprendre la lexicométrie. Nous allons nous pencher, à travers les lignes qui suivent, sur les différentes méthodes et fonctionnalités de cette dernière.

**3-1-Présentation:**

Il s’agit, dans un premier temps, de voir en quoi consiste un traitement automatique de discours, quelles sont ses différentes phases et caractéristiques, ainsi que ses objectifs.

De manière générale, le traitement lexicométrique se fait sur un macro-corpus constitué de textes numérisés (des dizaines, des centaines, voire des milliers de textes), enregistrés en format texte. Il vise à dégager les caractéristiques linguistiques et lexicales de ce corpus. C’est un traitement informatisé car le chercheur utilise un logiciel d’analyse statistique conçu à cet effet. Ce traitement se fait en fonction des hypothèses formulées par le chercheur et ce, préalablement à l’analyse lexicométrique. Il nécessite de la part de ce dernier des connaissances extralinguistiques relatives au corpus analysé. Par exemple, pour analyser le discours d’un homme politique, il faut bien connaitre ce personnage, ses prises de position, son idéologie, l’histoire politique de son pays, les caractéristiques du discours politique de ce pays, le contexte sociopolitique dans lequel ce discours a été produit, etc. Le programme lexicométrique ne peut pas remplacer l’être humain dans l’appréhension de certaines configurations sociopolitiques et sémantiques liées au texte analysé. Certes, le logiciel fait des comptages quantitatifs et fait ressortir des hypothèses de recherche mais il revient à l’analyste de faire l’effort d’interprétation. C’est ce qu’affirme Pascal Marchand :

« *Il appartient au chercheur d’acquérir aussi une compétence minimale quant aux procédures de traitement, de constituer un corpus susceptible de permettre une interprétation fiable et d’éviter l’écueil qui consisterait à prêter aux résultats issus de l’ordinateur un caractère de vérité absolue. Car, au final, c’est à l’analyste qu’il revient de mener l’interprétation et de valider ou invalider ses hypothèses.* »[[42]](#footnote-43)

On peut donc penser que le travail quantitatif du logiciel et le travail interprétatif du lexicométricien se complètent du moment que les résultats fournis par le premier ne peuvent être considérés comme fiables que si le deuxième vient les valider par une analyse qualitative qui affirme ou infirme les hypothèses émises préalablement. Plus précisément,

« *L’idée que les mots puissent indiquer une tendance macro-sociétale sans que les contextes et configurations sociohistoriques soient accessibles suppose que l’utilisateur soit capable de contextualiser et d’opérer les bons rapprochements et donc qu’il soit en mesure d’utiliser d’autres sources et d’autres outils.* »[[43]](#footnote-44)

L’analyste est amené à suivre plusieurs étapes. Il commence par découvrir et révéler les caractéristiques de son corpus. Il passe ensuite à la formulation des hypothèses en se basant sur les connaissances extralinguistiques qu’il a de la matière qu’il envisage d’analyser. Il termine son travail par la confirmation ou l’infirmation de ses hypothèses.

Le corpus soumis à l’analyse doit présenter certaines caractéristiques : il doit être numérisé (en format texte), délimité et circonscrit (constitué d’un nombre précis de textes), homogène (sa constitution ne doit pas être très hétéroclite, ses textes doivent être comparables.). On insiste surtout sur l’homogénéité générique. La question des genres est primordiale dans le traitement statistique du discours car ils sont prégnants sur les discours. Le corpus doit être découpé en plusieurs sous-parties afin de permettre de rendre compte des convergences ou des divergences lexicales qui peuvent exister entre ces sous-parties. Ce découpage appelé également « balisage » est indispensable à l’analyse par le logiciel. Ainsi distribué, un corpus est contrastif. Le chercheur va procéder à une analyse comparative pour évaluer et établir les clivages ou les cohérences entre les différentes sous-parties les unes par rapport aux autres et par rapport au corpus global qui constitue la référence. Ce corpus global qu’on peut désigner par la lettre T (Texte global) constitue la norme statistique « endogène » qui exerce une loi sur les textes « t » qui le composent selon la formule suivante : T= t1+t2+t3+ etc.

La comparaison peut se faire entre deux auteurs ou deux œuvres. Elle peut aussi être chronologique. Autrement dit, le balisage du corpus se fait en fonction des choix du lexicométricien. Le corpus peut-être fractionné par dates, par auteurs, par œuvres ou encore par genres. Pour exemple, la comparaison peut se faire sur l’œuvre totale d’un auteur pour caractériser chaque partie (ou œuvre) de cette œuvre globale. Mais cette œuvre-là doit être partitionnée en plusieurs sous-parties selon le critère chronologique afin de pouvoir mesurer les particularités de chaque partie (donc de chaque époque) par rapport à l’ensemble du corpus (donc aux autres époques). Cette démarche fait contraster les différentes époques et permet de caractériser chacune d’elle. La comparaison peut fournir des informations importantes sur la matière linguistique employée.

Avec ses forces et ses faiblesses, cette nouvelle manière d’aborder les textes suscite des réactions divergentes. Le débat autour de l’approche automatique du discours est loin d’être terminé et la prise en compte de la matière linguistique dans sa matérialité provoque des réticences de la part de certains spécialistes de l’Analyse du discours. Mais cette branche de recherche bénéficie du soutien de plusieurs spécialistes du domaine. Nous avons donc jugé utile d’exposer ici les avis des uns et des autres.

**3-2-Débat autour de l’approche lexicométrique :**

On a souvent reproché à la lexicométrie son caractère matériel. Selon ses détracteurs[[44]](#footnote-45), l’approche lexicométrique serait sans pertinence scientifique étant donné que le traitement statistique se borne à une description de la matérialité graphique des textes sans essayer de rendre compte du sens de ce matériel.[[45]](#footnote-46) Il serait donc sans utilité scientifique d’appréhender des textes dans un sens graphique, visuel, matériel et informatique[[46]](#footnote-47). Pour ses défenseurs, la lexicométrie est un domaine qui mérite d’être exploité. En effet, grâce aux développements récents de certains outils, la question de la pertinence des analyses et de la fiabilité des résultats ne se pose presque plus.

Par ailleurs, parler du débat entre défenseurs et adversaires de l’approche automatique du discours revient à parler d’une discussion aussi ancienne que l’approche elle-même. En effet, depuis les premiers travaux effectués au milieu des années 1940, confier à un outil informatique d’examiner les phénomènes langagiers a toujours suscité des réserves parmi les sceptiques. En revanche, les spécialistes de la discipline, bien qu’ils reconnaissent ses faiblesses, mettent en relief ses avantages.

Pour les réticents, la première chose qu’on peut reprocher aux traitements automatiques du discours est la technicité des outils mis au service des chercheurs. Le sens serait donc négligé. Sur cette question, Serge Lusignan souligne qu’« *en termes informatiques, l’information est un signal ou une combinaison de signaux. Vu sous cet angle, le texte constitue une longue séquence de bits et la liste de vocabulaire un décompte de combinaisons de bits…Dans les études littéraires, le sens nait de la capacité du spécialiste de lire un texte et de l’interpréter relativement à d’autres textes suivant des modèles théoriques…Replacé dans le processus interprétatif, l’ordinateur apparait comme une machine qui, à partir du texte électronique, produit des informations. L’ordinateur n’accède jamais au niveau du sens. »[[47]](#footnote-48)*

L’automatisation de la recherche va donc à l’encontre de la scientificité des traitements. Pour exemple, lorsqu’il s’agit d’un texte littéraire,

« *La 'substantifique moelle' d'une œuvre échappe par définition à toute tentative de formalisation et, a fortiori, à tout traitement automatique*»[[48]](#footnote-49) Ceci est dû au fait que « *Les données informatiques en critique littéraire informatisée correspondent de la même façon à des faits textuels, mais il ne peut y avoir de corrélation objective entre ces faits et le sens du texte*. »[[49]](#footnote-50)

Il est apparent donc que la question du sens est la plus grande faiblesse de la nouvelle approche car « *Le gouffre entre informations et sens demeure inconciliable sans intervention humaine. L'écart est d'autant plus grave que l'œuvre originale et les données informatiques qui en sont issues restent, pour la plupart, physiquement et conceptuellement séparées: c'est la nature de la représentation linéaire (souvent imprimée) qui le veut.* »[[50]](#footnote-51)

Pour les défenseurs de la lexicométrie**,** l’informatique n’est qu’un outil et non une finalité*.* Elle permet d’*« organiser et de contrôler les parcours de lecture. Elle permet de rechercher et d’extraire de l’information du corpus. Mais l’acte final d’interprétation du chercheur ne peut se faire sans embrasser le texte naturel dans lequel nous sommes toujours replongés.* »[[51]](#footnote-52).

On peut donc comprendre que l’intervention du chercheur vient combler l’écart entre les informations informatiques fournies par l’ordinateur et le sens du matériel linguistique utilisé. Autrement dit, dans l’approche lexicométrique, les nombres ne comptent que dans la mesure où ils ouvrent la voie à des interprétations.

À ceux qui insistent sur *l’opposition insurmontable* entre quantitatif et qualitatif en lexicométrie, il est répondu que cette opposition n’est pas pertinente « *Tant il est vrai que ces deux aspects ne constituent, tout au plus, que deux moments de l’analyse.* »[[52]](#footnote-53)

Pour Maurice Tournier et Simone Bounnafous, loin d’être une faiblesse, le quantitatif est au contraire, une force de l’approche lexicométrique. Ils écrivent :

« *Un texte étant un tout solidaire en système d'oppositions avec d'autres, l'emploi d'un terme ne fonctionne pas isolément, mais à l'intérieur d'un vocabulaire, dans un univers de rapports aux autres emplois et aux autres termes. Dans la mesure de ces rapports, le dénombré devient significatif.* *De quoi ? Retour au sens*. »[[53]](#footnote-54)

Sur la question fondamentale du sens, ils notent qu’« *on ne peut, au départ d'une étude lexicométrique, que cumuler, confronter, réunir ou opposer des occurrences de formes textuelles. Les questions de sens sont à poser ultérieurement, par un travail sur le dépouillement ou les résultats, qui sera spécifique mais dont la statistique ne garantira plus forcément la valeur. »[[54]](#footnote-55)*

Evoquant l’exemple du discours politique, Tournier indique que les programmes lexicométriques,

« *L’inventaire systématique en machine et le calcul lexicométrique permettent seuls de « refroidir » ce domaine brûlant du discours de pouvoir où l’on s’interroge sur les mots pour le dire ou le contredire. Relevons, comptons, comparons ; nous interpréterons ultérieurement.* »[[55]](#footnote-56)

On peut comprendre alors qu’on ne travaille pas sur le contenu (le sens) des unités, les recherches étant systématiques et le sens en contexte étant réfractaire à toute systématicité. La graphie prime sur le sens dans la mesure où l’on travaille sur la surface textuelle dans sa matérialité et non pas sur les référents ou le contenu des unités linguistiques. Le travail sur le contenu sémantique vient après.

Les traitements quantitatifs sont bien entendu effectués par le programme lexicométrique. Mais l’intervention du lexicométricien est indispensable de plusieurs points de vue. C’est à lui de préparer la matière linguistique qu’il envisage d’analyser (le corpus) : en numérisant les textes de sa base de données (au clavier ou avec le scanner) ; en déterminant les balises informatiques en fonctions desquelles il découpe son corpus en plusieurs sous-parties, et en formulant les hypothèses d’analyse.

Pour effectuer une analyse lexicométrie, il est donc nécessaire de suivre plusieurs étapes. C’est ce que nous allons voir à travers les paragraphes qui suivent.

**3-3-Etapes et fonctions de l’analyse lexicométrique :**

Les programmes lexicométriques disposent de deux types de fonctions : les fonctions documentaires (hypertextuelles) et les fonctions statistiques.

Les fonctions documentaires servent à faire le décompte des mots employés dans le corpus. Ces fonctions sont les index (alphabétique et hiérarchique), les concordances et les contextes. Elles permettent à l’analyste de formuler ses hypothèses en lui fournissant des informations sur la matière linguistique utilisée pour mieux mener sa recherche.

Les fonctions statistiques sont l’analyse des spécificités, l’analyse factorielle des correspondances, le calcul de la distance et de la connexion lexicales, l’évolution du lexique et l’analyse arborée (celle-ci est propre à Hyperbase).

Les démarches que nous suivons dans notre analyse sont déterminées par les fonctions documentaires et statistiques du logiciel que nous allons utiliser dans notre analyse, à savoir Hyperbase.

Avant de passer à l’analyse proprement dite, on doit effectuer trois opérations dont l’objectif est la préparation du corpus à l’analyse. Ces opérations sont : la numérisation des données en format texte (le logiciel ne prend en charge que les textes numérisés); la lemmatisation et l’étiquetage du corpus pour transformer le texte brut en texte lemmatisé ; et enfin le découpage (ou la distribution) du corpus en sous-parties comparables.

**3-3-1-La numérisation :**

De nos jours, il existe de plus en plus de bases de données qui mettent à la disposition des lexicométriciens des textes littéraires ou politiques prêts à être analysés, disponibles dans des bases de données gigantesques. Si le corpus n’est pas disponible, le chercheur doit le numériser. Dans certains cas, le corpus existe en format papier – des livres, des archives, etc. – Il suffit alors de le numériser avec un scanner et un logiciel de reconnaissance de texte. Mais parfois, le chercheur doit, en plus de la constitution de son corpus, en faire la saisie laborieuse. Cela nécessite un temps considérable. Il est vrai que cette tâche était plus difficile par le passé, mais il faut reconnaitre qu’elle peut coûter au chercheur des mois de travail. Par exemple, nous avons consacré plusieurs mois à la saisie de presque la moitié de notre corpus[[56]](#footnote-57). Nous ne pouvions pas faire autrement puisque beaucoup de textes de notre corpus (les articles de journaux) ne sont disponibles qu’aux archives, donc impossibles à scanner puisqu’il est interdit de les emprunter. Nous les avons donc photographiés avec un appareil à photos. Les images que nous avons prises, étant l’état fragile, parfois lamentable, des documents qui remontent aux années 1930, 1940 et 1950, ne sont pas assez bonnes pour faciliter une numérisation avec le logiciel de reconnaissance textuelle. Alors, nous avons fait le travail de saisie au clavier.

Une fois numérisé, le corpus est prêt à être découpé en parties selon une variable déterminée par le chercheur. Celui-ci doit également lemmatiser et étiqueter son corpus afin d’avoir à sa disposition deux types de données textuelles: lemmatisées et non-lemmatisées.

**3-3-3-La lemmatisation et l’étiquetage :**

Les outils récents sont assez perfectionnés pour permettre également de fournir, en plus du corpus original, un corpus lemmatisé et étiqueté.

On distingue, en lexicométrie, deux types de corpus: le corpus brut et le corpus lemmatisé et étiqueté. Il est exclu pour nous de nous étaler ici sur ces deux opérations – la lemmatisation et l’étiquetage – que nous allons voir avec plus de détails un peu plus loin. De manière générale, lemmatiser un texte est une opération qui consiste à associer à chaque forme graphique du corpus – ses termes, au sens le plus trivial du mot – son lemme. Le lemme est l’entrée canonique de chaque mot dans le dictionnaire. Ainsi, lemmatiser un texte écrit en français revient à ramener les verbes conjugués à l’infinitif, les substantifs au singulier, les adjectifs au masculin singulier et les formes élidées à la forme sans élision.[[57]](#footnote-58)

L’étiquetage est une opération qui consiste à associer à chaque forme relevée dans le corpus, son lemme, son code grammatical (nom verbe, pronom, adjectif, etc.), et ses informations de flexion (son genre pour les noms, son nombre, son temps et sa personne pour les verbes, etc.) Hyperbase permet deux types d’analyses lexicométriques : analyse sur texte brut et analyse sur texte lemmatisé et étiqueté.

Contrairement à la segmentation classique d’un texte en formes, qui se fait selon des critères graphiques (par repérage des caractères délimiteurs), la lemmatisation et le découpage répondent à des critères linguistiques (Flexions, temps, modes, catégories grammaticales, etc.)

Grâce à l’étiquetage,

« *L’on peut faire des calculs lexicométriques sur le corpus écrit comme une succession de lemmes, ou encore écrit comme une succession de codes grammaticaux, ou tout simplement écrit comme la succession de ses mots. »[[58]](#footnote-59)*

On peut donc comprendreque l’étiquetage permet de multiplier les dimensions d’analyse : forme, lemme, code grammatical; et d’aborder les textes de plusieurs points de vue.

**3-3-3-La distribution du corpus en sous-parties comparables:**

Elle se fait en fonction de « balises » ou « clés » déterminées par le chercheur. Lesquelles correspondent à des hypothèses émises par celui-ci.Les hypothèses sont formulées en fonction d’une variable de l’énonciation déterminée préalablement par le lexicométricien : auteur, époque, lieu, genre, etc. L’intérêt de la partition réside dans le fait que,

« *Dans un corpus de textes comparables* […] *la lexicométrie confronte les emplois des unités du discours.*»[[59]](#footnote-60)

La partition du corpus vise à examiner les ressemblances ou les oppositions lexicales entre les sous-corpus. Le chercheur peut vérifier la pertinence de son découpage et partant valider ou invalider ses hypothèses. Par exemple, si l’on divise un corpus constitué de l’œuvre complète d’un auteur en fonction de la variable « temps »[[60]](#footnote-61), c’est-à-dire en fonction des époques pendant lesquelles les textes de l’écrivain ont été produits, on peut distinguer plusieurs époques, donc plusieurs sous-parties. L’analyse lexicométrique fait ressortir les ressemblances ou les clivages lexicaux entre les différentes sous-parties (donc les différentes époques).

**3-3-4-Les fonctions documentaires :**

Celles-ci sont les index, la mise en concordance et la mise en contexte.

**3-3-4-1-Les index : la segmentation du corpus en unités graphiques:**

Le logiciel réalise une segmentation automatique du corpus en unités graphiques. Il nous fournit ainsi la liste de toutes les formes graphiques du corpus accompagnées de leurs fréquences respectives d’apparition dans le texte. Cette liste est appelée dictionnaire ou index. Un index est ou bien alphabétique ou bien hiérarchique. Il est alphabétique quand il présente la liste des formes graphiques employées classées par ordre alphabétique. Il est hiérarchique quand les occurrences sont classées par ordre décroissant de fréquence. L’avantage de l’index est qu’il permet de mieux appréhender le vocabulaire du texte et qu’il nous donne une idée assez claire sur les formes les plus fréquentes du corpus. Il nous renseigne aussi sur le nombre de mots, leurs fréquences respectives et les distributions de la fréquence de chaque forme dans les sous-parties du corpus. Notons également que les index peuvent engendrer des graphiques et des calculs statistiques.

Afin de segmenter son corpus, le chercheur définit deux caractères (souvent deux blancs ou les ponctuations, les apostrophes et les tirets) pour délimiter les formes graphiques employées. Ces caractères sont appelés « les caractères délimiteurs des formes graphiques » Les autres caractères du texte sont donc « des caractères non-délimiteurs » Les formes graphiques sont ainsi appelées « occurrences ». Une occurrence peut être définie comme étant une *suite de caractères non-délimiteurs bornée à ses deux extrémités par des caractères délimiteurs.* [[61]](#footnote-62) Le texte segmenté en formes graphiques est appelé hypertexte. Les occurrences d’un texte constituent donc son vocabulaire.

Après l’indexation, le programme lexicométrique effectue une segmentation en segments répétés. Un segment répété peut être défini comme étant une suite de formes qui apparaissent l’une après l’autre plusieurs fois de suite.

**3-3-4-2-Les contextes et les concordances :**

Cette fonction permet la mise en « concordance » et la mise en contexte. La mise en concordance se fait en déterminant le contexte linguistique immédiat et étroit de chaque forme. Le contexte immédiat est constitué de la ligne dans laquelle une forme est employée. La mise en contexte consiste en la détermination d’un contexte plus large, constitué d’un ou de plusieurs paragraphes – cinq au maximum -

Dit autrement, l’avantage de la segmentation est qu’elle permet la contextualisation linguistique de chaque item lexical. Par un simple clic, le logiciel nous donne le contexte linguistique (ou cotexte) de chaque occurrence (phrase, passage, paragraphe, texte, etc.) pour en déterminer le sens.

L’objet d’une recherche de contexte ou de concordance peut être une forme graphique simple, une expression (un groupe de mots), un lemme, un code grammatical, une structure syntaxique, le début ou la fin d’un mot, les co-occurrences (présence de deux formes dans le même contexte), etc.

La mise en contexte est importante dans la mesure où elle permet un retour au texte pour déterminer le sens d’une forme ou d’un segment. Ceci facilite à l’analyste l’acte d’interprétation comme le souligne Damon Mayaffre :

« *Dans le domaine sémantique, la* co(n)textualisation *des formes nous semble la philosophie même de nos pratiques, étant entendu que le sens des choses n’émerge qu’en co(n)texte, qu’en corpus.* »[[62]](#footnote-63)

Le contexte permet donc de surpasser l’écueil tant commenté du sens car « *la définition matérielle du contexte – le contexte c’est avant tout un co-texte entendu comme un environnement textuel immédiat et contigu – se trouve équilibrée par une dimension herméneutique – le contexte*, c’est ce qui fait sens; *ce qui sémantise un terme en autorisant l’interprétation.* »[[63]](#footnote-64)

**3-3-5-Les fonctions statistiques : les calculs et les comptages:**

Pour mieux appréhender le lexique de son corpus,  le lexicométricien fait appel, en plus des fonctions documentaires hypertextuelles vues plus haut, aux calculs et aux comptages lexicométriques nécessaires. Ces comptages sont :

**3-3-5-1-Le calcul de l’étendue relative des textes :**

Le logiciel donne les résultats sous forme de listes ou sous forme de graphiques symbolisant l’étendue relative des textes de la base. Il calcule donc l’étendue de chaque texte du corpus (ou de la base) et la confronte à celle des autres textes formant le même corpus. L’étendue est mesurée de deux manières :

1-Par le nombre d’occurrences constituant le vocabulaire du texte en question. La même forme peut être présente des dizaines, voire des centaines de fois. Toutes ses occurrences sont comptabilisées.

2-Par le nombre de formes différentes présentes dans le texte sans tenir compte de leurs fréquences. Ainsi, une forme présente 150 fois dans le texte est prise pour une seule forme et n’est comptée qu’une seule fois. De la même manière, est mesurée l’étendue totale du corpus. Le nombre de formes constitutives d’un corpus doit être inférieur à la somme des formes constitutives des textes dont il est constitué, car il existe inéluctablement des formes communes aux textes du corpus. Lorsqu’il s’agit de mesurer l’étendue totale du corpus, chacune de ces formes communes est comptabilisée une seule fois.

**3-3-5-2-Les graphiques :**

Cette fonction permet de calculer ce qu’on appelle « l’écart réduit ». Le logiciel réalise ce calcul avant de le représenter graphiquement, à travers des histogrammes. Un écart réduit est calculé de la manière suivante :

Premièrement, il faut distinguer dans un corpus deux types de fréquences, une fréquence observée (réelle) et une fréquence théorique (mathématiquement attendue) La fréquence observée est la fréquence réelle d’un mot dans le corpus ou le texte considéré.

Quant à la fréquence théorique, c’est celle à laquelle on s’attend *mathématiquement*, vu le rapport taille du texte/taille du corpus. Ce rapport est mesuré mathématiquement selon la règle de trois suivante : fréquence théorique d'un mot dans un texte = fréquence du mot dans le corpus pondérée par la probabilité *p* ou part du texte dans le corpus.

Etant donné l’inégalité des tailles des textes dans certains cas, ce qui engendre un écart entre les fréquences (théorique et observée), on pondère cet écart absolu selon la formule de l’écart réduit. Cette opération purement mathématique est effectuée par le logiciel et permet de repérer les mots en déficit ou en excédent significatif. Les chercheurs non-initiés aux mathématiques n’ont pas à craindre de faire appel à des opérations qu’ils ne connaissent pas. Le logiciel effectue ces calculs par un simple clic.

**3-3-5-3-L’environnement thématique : le calcul des co-occurrences et des réseaux thématiques :**

La co-occurrence peut être définie comme étant :

« *La co-présence ou présence simultanée de deux unités linguistiques (deux mots par exemple ou deux codes grammaticaux) au sein d’un même contexte linguistique (le paragraphe ou la phrase par exemple, ou encore une fenêtre arbitraire). […] La lexicométrie la constate et l’exprime grâce à des coefficients statistiques à même de mesurer le degré de significativité des co-présences ou attractions trouvées.* »[[64]](#footnote-65)

Autrement dit, cet outil permet de repérer et de présenter les mots qui apparaissent ensemble dans les mêmes phrases ou les mêmes paragraphes, ces mots sont appelés « co-occurents ». Il permet donc de déterminer les thèmes abordés dans le texte et par là même d’en délimiter l’univers thématique.

Le mot dont on cherche le contexte est appelé « mot-pôle ». Le logiciel procède à une comparaison entre « un sous-corpus temporaire » constitué par tous les contextes recensés du mot-pôle, et le corpus global. L’opération statistique consiste à dénombrer les occurrences de tous les mots recensés dans « le sous-corpus temporaire ». Le logiciel effectue alors un calcul de probabilité en confrontant les fréquences réelles (ou observées) des formes recensés dans le « sous-corpus temporaire » à la fréquence mathématiquement attendue (ou théorique) si ce « sous-corpus temporaire » était une sous-partie aléatoire du corpus global de la base. On peut donc, grâce à la méthode des écarts réduits, repérer les mots qui sont en excédent et ceux qui sont en déficit significatif dans les contextes du mot-pôle.

**3-3-5-4-L’analyse factorielle des correspondances[[65]](#footnote-66)** :

En lexicométrie, l’utilisation de l’AFC est d’une grande importance. Après une distribution du corpus en plusieurs sous-corpus ou sous-parties – scission qui se fait en fonction d’une variable déterminée : auteur, époque, évènement, etc., le lexicométricien procède à une analyse contrastive entre ces sous-parties pour rendre compte des antagonismes ou ressemblances entre les vocabulaires des textes constitutifs des différentes sous-parties. Cela permet au chercheur de valider ou d’invalider des hypothèses qu’il a formulées au départ de son analyse, hypothèses formulées en fonction de la variable sur la base de laquelle la distribution du corpus a été faite.

Le but de l’AFC est de « *déterminer, dans les correspondances quantitatives entre un ensemble d’émetteurs (les différents états de la variable qui partitionne le corpus en « textes ») et un ensemble de termes (le vocabulaire de ces textes), quels sont les clivages les plus pertinents.* »[[66]](#footnote-67)

Certes, l’AFC fait ressortir des clivages, des ressemblances, des regroupements et des oppositions lexicaux mais elle est insuffisante. C’est pourquoi, pour la renforcer et l’affiner, on fait appel à l’analyse des spécificités.

Les résultats de l’analyse factorielle sont représentés par un graphique qui se présente comme une cartographie des textes. Il est formé de deux axes orthogonaux : l’axe des abscisses (l’horizontal) et l’axe des ordonnées (le vertical). Les textes de la base sont figurés par des points sur la cartographie. L’information qui se dégage de la représentation est contenue dans la distance entre les points (donc, entre les textes de la base) et leur distribution par rapport aux deux axes du graphe. Il faut avouer que cette opération nécessite d’être expliquée avec des exemples d’application.

**3-3-5-5-L’analyse arborée :**

Il s’agit également des distances entre les textes. L’analyse arborée, propre au logiciel Hyperbase, permet de faire une classification des éléments étudiés selon un ordre ascendant et ce, en regroupant les deux éléments les plus proches, puis les deux suivants, etc. Globalement, cette opération permet d’examiner, de calculer et de représenter graphiquement les distances entre les textes du corpus.

Le graphique engendré par cette analyse se présente sous forme d’arbre à plusieurs troncs, d’où son nom « L’analyse arborée ». Il s’agit d’une opération aussi compliquée que l’analyse factorielle qui nécessite, pour être comprise, d’être illustrée par une application.

**3-3-5-6-Calcul de la distance et de la connexion lexicales intertextuelles :**

Cette opération permet de mesurer la proximité ou l’éloignement des sous-parties du corpus sur le plan lexical les uns par rapport aux autres. La connexion ou la distance lexicales peuvent être mesurées en fonction du vocabulaire commun entre les textes étudiés et du vocabulaire spécifique à chaque texte. Un lexique commun à deux textes les rapproche tandis qu’un lexique spécifique à l’un d’eux, donc non-spécifique à l’autre, les éloigne l’un de l’autre.

**3-3-5-7-Richesse lexicale, hapax et accroissement lexical :**

**3-3-5-7-1-La richesse lexicale :**

La richesse lexicale d’un texte est mesurée et représentée graphiquement dela manière suivante :

Le logiciel dénombre les formes différentes présentes dans chaque texte. Il mesure le vocabulaire théorique (mathématiquement attendu) et le confronte à celui présent réellement (observé). Ce rapport est mesuré en tenant compte de l’étendue relative de chaque texte et des fréquences des formes relevées. L’écart entre la part du vocabulaire réellement observé et celle du vocabulaire théoriquement attendu est mesuré avec la méthode de l’écart réduit. Les résultats apparaissent graphiquement sous forme d’histogrammes.

**3-3-5-7-2-Les hapax :**

Ce sont des formes ou des lemmes qui apparaissent une seule fois dans un texte. Le calcul des hapax prend en considération les écarts et les disparités entre les textes en confrontant, selon la loi de l’écart réduit, les textes les uns aux autres. Ainsi, la machine mesure l’écart entre le théoriquement attendu et le réellement observé. Il est tenu compte, dans ce calcul, de manière mathématique, de l’étendue relative des textes observés et des deux fréquences réelle et théorique.

**3-3-5-7-3- L’accroissement lexical :**

Cette fonction permet d’examiner et de représenter graphiquement la dynamique de la structure du vocabulaire d’un texte ou de celui de la base globale à travers le temps. Le paramètre chronologique est donc pris en compte. De ce fait, il est possible d’observer des cumuls ou des renouvellements du vocabulaire à travers le temps. Le logiciel donne accès à des listes de formes ou de lemmes qui sont en augmentation (en accroissement), ou en diminution (en décroissance) significative. Il est utile de savoir que l’évolution (qu’il s’agisse de croissement ou de décroissement) n’est pas mesurée en fonction des simples fréquences des formes ou des lemmes mais selon un coefficient de corrélation qui varie entre -1 et +1.

Il est possible, grâce au logiciel Hyperbase, d’afficher les résultats (les listes de formes ou de lemmes) par ordre alphabétique ou par ordre hiérarchique décroissant. Rappelons encore une fois que cet ordre hiérarchique tient compte du coefficient de corrélation mesuré par le logiciel et non pas des fréquences. En plus, il est possible de mesurer l’accroissement lexical à l’intérieur de chaque sous-partie du corpus.

**3-3-5-8-L’analyse des spécificités** :

Cette techniquea été mise au point au laboratoire de Saint-Cloud par Pierre Lafont et André Salem. Elle permet de déterminer les lexies qui caractérisent le plus le lexique et celles qui le caractérisent le moins. Une spécificité positive signifie que la lexie en question est sur-employée par rapport à l’ensemble du lexique du texte étudié. Par contre, une spécificité négative ne signifie pas que la lexie n’est pas employée du tout, mais qu’elle est sous-employée en comparaison avec les autres items lexicaux du corpus étudié. La spécificité d’une unité linguistique est donc calculée par rapport aux autres unités du corpus.

Certes, les études contrastives de vocabulaire sont d’une importance capitale dans l’approche automatique du discours mais, en calculant et comparant les fréquences des formes recensées dans le corpus global, on s’expose à un risque susceptible de mettre en cause la fiabilité des résultats. En effet, il se trouve que les textes constitutifs d’un corpus n’ont pas forcément la même longueur ou la même étendue lexicale. C’est pourquoi il convient de calculer la probabilité de fréquence de chaque unité par rapport à l’ensemble des unités qui composent la sous-partie du corpus dans laquelle cette forme a été relevée (donc la sous-fréquence)

La probabilisation est la substitution de la fréquence d’une unité par son jugement de suremploi ou de sous-emploi (probabilité d’apparition de l’occurrence en question). Cette opération s’appelle en lexicométrie la probabilisation par calcul hypergéométrique. Elle consiste à mesurer l’écart entre la fréquence réelle d’une forme et sa fréquence théorique, mathématiquement attendue.

Il s’agit de *« substituer à chaque fréquence locale un jugement[[67]](#footnote-68) sur elle qui tienne compte des rapports quantitatifs dans lesquels elle a été constatée.*»[[68]](#footnote-69).

Grâce à la probabilisation, on peut minimiser les écarts de longueur entre les textes étudiés. C’est là le rôle des mathématiques.[[69]](#footnote-70) Les indices de spécificité sont indiqués par un S+ pour une spécificité positive et par un S- pour une spécificité négative. Un S+ renvoie à une lexie sur-employée. Un S- indique une lexie sous-employée. Certains termes, bien que suffisamment employés, ne sont pas spécifiques, ni positivement ni négativement. Ce sont les mots-outils (prépositions, déterminants, etc.) On dit qu’ils ont un emploi « banal » qui ne les caractérise pas par rapport aux autres lexies. On leur attribue un indice « b ».

**3-4-La lemmatisation, un mal nécessaire ?**

Il ne s’agit pas, dans ce bref passage, de reprendre tout ce qui a été écrit à propos de la lemmatisation. Nous nous limitons donc à la définir en essayant de présenter ses avantages et ses inconvénients ainsi que les arguments avancés par ses défenseurs et ses détracteurs.

En termes simples, la lemmatisation est une opération qui consiste à ramener les formes linguistiques d’un texte à leurs formes du dictionnaire. Lemmatiser un texte, c'est

« *Ramener son vocabulaire à un lexique, au sens où J. Picoche entend ces deux mots, en se proposant «d'appeler* lexique *l'ensemble des mots qu'une langue met à la disposition des locuteurs,* *et* vocabulaire *l'ensemble des mots utilisés par un locuteur donné dans des circonstances données.» (Picoche, 1977: 45).* »[[70]](#footnote-71)

En d’autres mots, la lemmatisation est :

« *L’opération de regroupement qui rassemble les formes différentes appartenant au même vocable - ce que font précisément les auteurs de dictionnaires quand ils établissent leur nomenclature.*»[[71]](#footnote-72)

Le lemme peut être défini donc comme étant :

« *L’étiquette associée à toute forme textuelle, identifiant le lexème auquel elle appartient et correspondant à la forme qui le représente dans un dictionnaire de référence.* »[[72]](#footnote-73)

Par ailleurs, il y a, parmi les défenseurs de la lexicométrie, un autre conflit, celui qui oppose les lemmatiseurs aux non-lemmatiseurs (appelés aussi formalistes)[[73]](#footnote-74). Cette querelle a perdu de son alacrité ces dernières années[[74]](#footnote-75).

Tandis que les premiers rappellent les *avantages multiples* *d’une opération permettant d’ouvrir de nouvelles perspectives d’analyse*, les seconds insistent sur *les limites d’une opération qui constitue un travestissement du texte original et qui risque de fausser les résultats d’analyse.*

Les opposants de la lemmatisation privilégient les traitements sur textes bruts et rejettent toute intervention sur le texte original. Ils considèrent que lemmatiser un texte suppose une intervention du chercheur sur la matière linguistique utilisée. C’est pourquoi ils prônent l’analyse du texte original sans traitement préalable. La lemmatisation serait donc un prétraitement, une intervention du chercheur sur le lexique original du texte qui en perd sa forme initiale et originale. Dès lors, le traitement se fait sur un corpus qui a subi des modifications. En d’autres mots, elle serait l’aboutissement d’une autre analyse.

Mayaffre note que :

« *Faire précéder l’analyse que l’on se propose de faire, d’une « normalisation » sémantique déjà toute faite revient à donner les conclusions du travail avant même l’analyse ! Il y a là une entorse épistémologique dans le procès de la démarche, évidente et insurmontable.*»[[75]](#footnote-76)

Parmi les inconvénients supposés de la lemmatisation, on peut noter :

« *La perte d’informations résultant du remplacement d’un mot par son lemme. Cette perte d’informations est préjudiciable aux algorithmes qui utilisent le contexte des mots (ou, linguistiquement parlant, le cotexte) puisque ce dernier n’est pas nécessairement le même selon la forme des mots.* »[[76]](#footnote-77)

Le lemmatiseur ramène par exemple, les pluriels des substantifs à leur lemme (le singulier). Or, les contextes des deux formes (pluriel et singulier) sont différents. C’est-à-dire que les deux formes différentes (pluriel et singulier) ne sont pas entourées des mêmes formes (qui constituent leur contexte linguistique, ou cotexte), ce qui influe sur l’extraction de l’information sémantique. En d’autres termes, en ramenant les deux formes différentes, pluriel et singulier, au même lemme, on considère qu’elles contiennent la même information sémantique, ce qui affecte grandement la fiabilité de l’analyse. Car le contenu sémantique de la forme dépend de son cotexte. Par conséquent, assigner la même information sémantique à deux formes différentes risque de fausser le calcul.

En revanche, pour ses défenseurs, la lemmatisation constitue un outil d’un intéressant puisqu’elle permet de surmonter certains obstacles et vaut « *une valeur ajoutée à l'exploitation statistique*. » [[77]](#footnote-78)

En effet, tout texte contient des homographes[[78]](#footnote-79), des formes élidées, des formes contractées ou distordues. Cela dit, le logiciel, puisqu’il s’agit d’une machine, ne peut pas distinguer les homographes et les prend pour une seule forme graphique. Par exemple, le programme lexicométrique prend pour une seule forme graphique le verbe « porte » (porter au présent de l’indicatif avec la première personne du singulier) et le substantif « porte » (nom masculin singulier). Les exemples sont légion pour montrer les confusions que peut entrainer une analyse sur textes non-lemmatisés. Les cas d’élision, de contraction et de distorsion sont également nombreux.

L’avantage dela lemmatisation et de l’étiquetageréside dans le fait que :

*« La forme graphique très souvent ambigüe, donc dépourvue de sens, se trouve renseignée sémantiquement, le traitement du texte est étendu au-delà du simple lexique vers la tonalité rhétorique des discours faite notamment de notes grammaticales et de tournures syntaxiques particulières*. »[[79]](#footnote-80)

Dominique Labbé considère, pour sa part, que la lemmatisation est une opération indispensable parce qu’elle ouvre des perspectives nouvelles à la lexicologie et à la stylistique[[80]](#footnote-81). Effectivement, cette opération permet d’étendre l’analyse au-delà du niveau lexical (les lexies au sens matériel et graphique du terme) pour atteindre les niveaux morphosyntaxique et sémantique. Elle permet, de ce fait, de multiplier les niveaux d’analyse : forme, lemme, lexème, code grammatical, etc. Les données étiquetées et lemmatisées permettent d’accéder à la syntaxe, à la sémantique et aux faits de style.

Le logiciel Hyperbas, qui s’associe au lemmatiseur Cordial[[81]](#footnote-82), distribue les données dans des champs appropriés : graphie, lemme, code grammatical et structure syntaxique.

*Les codes grammaticaux :*

La machine détermine et associe à chaque forme graphique le code grammatical correspondant : verbe, substantif, adjectif, déterminant, pronom, adverbe, etc.

*Les structures syntaxiques :*

La machine détermine pour les verbes le statut (auxiliaire ou non), le mode, le temps, la personne et le nombre.

Parmi les avantages de la lemmatisation, on peut citer « *La réduction du nombre de formes à considérer et l’augmentation des occurrences de chaque forme dans le corpus.* »[[82]](#footnote-83)

Pour souligner l’intérêt de la lemmatisation, Mayaffre dit qu’elle permet «*d’une part de décompter des unités à la pertinence linguistique plus avérée (les lemmes qui renvoient à des vocables) et d’autre part de compléter le traitement purement lexical traditionnel par un traitement statistique d’autres régularités linguistiques tels les codes grammaticaux, les modes, les temps, le genre, le nombre, etc.*»[[83]](#footnote-84)

Sylvie Mellet écrit :

« *La lemmatisation permet précisément de dégager les analyses de corpus des aléas de la variation tout en conservant la forme originelle et authentique des textes : elle satisfait donc à la fois aux exigences de la statistique linguistique (en particulier dans le cadre de la comparaison de textes) et à celles de la philologie.* La lemmatisation  *est une nécessité. Elle ouvre en effet des champs d’exploration qui, sinon, restent inaccessibles, elle affine et stabilise les traitements quantitatifs.* »[[84]](#footnote-85)

Pour Etienne Brunet*,*

*« L’accès au lemme et aux codes grammaticaux ouvre des champs plus riches à la recherche. Qu’il s’agisse des fonctions dans la phrase, des parties du discours, ou des temps, des personnes ou des modes verbaux, on explore ici ces perspectives, plus syntaxiques que lexicales*. »[[85]](#footnote-86)

**3-5-Synthèse, vers la logométrie :**

En examinant les arguments des uns et des autres, il devient évident que ni les défenseurs ni les détracteurs de la lemmatisation n’ont entièrement raison. On ne peut se passer des avantages de la lemmatisation, comme on ne peut s’y fier entièrement car elle n’est pas fiable à cent pour cent. Formalistes et lemmatiseurs le reconnaissent.

Maurice Tournier, en tant que formaliste écrit :

«*Oui je l'avoue, les arguments se contrebalancent. L'un* (le lemmatiseur) *crie à la trahison de la langue, l'autre* (le formaliste) *à la trahison du texte [...] Avouons-le ensemble une bonne fois, il y a effectivement trahison des deux côtés*.»[[86]](#footnote-87)

Muller, en tant que lemmatiseur, reconnait *les péchés* de la lemmatisation, sans pour autant céder aux arguments de son adversaire :

«*J'ai dit les objections purement linguistiques qui me font préférer les péchés de la lemmatisation à la pureté du formalisme.»[[87]](#footnote-88)*

La lemmatisation serait donc, pour le lexicométricien, un mal nécessaire ! Sur ce point, la réflexion de Damon Mayaffre est particulièrement intéressante au sens où elle prône des traitements doubles : sur textes bruts et sur textes lemmatisés (donc sur formes et sur vocables) d’autant plus que certains outils lexicométriques, notamment Hyperbase, permettent d’effectuer ce double traitement.

Mayaffre pense qu’« *il est tout aussi impossible de renoncer aux formes graphiques, que de s'en tenir à elles; impossible de se priver des richesses de la lemmatisation - notamment en ce qu'elle permet l'étude des caractères grammaticaux des textes - que d'oublier le texte natif.* » [[88]](#footnote-89)

Dans cette optique, l’auteur recommande le passage d’une lexicométrie originelle classique à une logométrie qui conjugue plusieurs niveaux d’analyse. Il s’agit de compléter le traitement sur les formes graphiques (les lexies, au niveau lexical) par un traitement sur les lemmes et les codes grammaticaux (niveau syntaxique et sémantique)

La logométrieest une approche qui**«***dépasse le traitement des formes graphiques sans les exclure ou les oublier; qui analyse les lemmes ou les structures grammaticales sans délaisser le texte natif auquel nous sommes toujours renvoyés.* *C'est un traitement automatique global du texte dans toutes ses dimensions: graphiques, lemmatisées, grammaticalisées.*»[[89]](#footnote-90)

Ellepeut être définie simplement comme étant la lexicométrie sur textes lemmatisés et étiquetés.

Cette approche est d’autant plus intéressante que les programmes de lemmatisation qui faisaient défaut par le passé sont disponibles de nos jours et permettent de désambigüiser les textes (donc, ils dégroupent les lexies homographes) et de dégrouper les formes flexionnelles verbales, nominales et adjectivales.

En outre, la logométrie, exige que la lemmatisation s’accompagne de deux autres opérations : l’étiquetage morphosyntaxique et le codage grammatical. Ces deux opérations consistent à associer à chaque forme son code et sa catégorie grammaticaux (nom, verbe, adjectif, etc.), son genre et son nombre pour les noms et les adjectifs.

**Conclusion partielle :**

En guise de conclusion, nous disons que la sagacité des programmes lexicométriques permet de déceler, en s’appuyant sur des procédures mathématiques, des stratégies discursives mises en œuvre dans un texte en étudiant les mots employés, lesquels renvoient à des configurations sociopolitiques.

Mais comment réconcilier les lettres avec les chiffres ? Comment aborder les textes avec le prisme des mathématiques? Ne serait-il pas *absurde* de décrire le vocabulaire d’un texte littéraire ou politique à l’aide de formules algébriques ?

À l’inverse, l’approche lexicométrique n’est-elle pas une nouvelle manière, originale de décrire le vocabulaire? Ne permet-elle pas d’apporter de nouveaux éclairages sur les textes et ce, en les abordant autrement et en découvrant certains de leurs aspects difficiles à découvrir avec le cerveau humain?

Les arguments des défenseurs de l’approche lexicométrique semblent surpasser en pertinence ceux de ses adversaires.

Pour une bonne maitrise de l’approche lexicométrique, il importe non seulement de bien connaitre ses fondements et ses principes, mais de la manipuler avec prudence, car il s’agit d’un outil mis au service de l’analyse du discours. Dit autrement, un traitement lexicométrique n’est qu’une étape dans l’analyse des discours. Il ne faut donc pas se fier entièrement à des outils informatiques qui ne peuvent, en aucune manière, se substituer aux subtilités et à la finesse du cerveau humain.

La présentation de la lexicométrie, dont les contours sont si difficiles à cerner, est à vrai dire difficile vu l’évolution continuelle et rapide de ses outils et de ses fonctions, malgré la constance de ses principes et ses fondements. Par conséquent, l’exposé ci-dessus ne prétend pas être exhaustif car il reste beaucoup de choses à découvrir sur la lexicométrie et particulièrement sur ses sciences connexes. Certaines tendances très récentes dans le domaine, notamment les approches développées dernièrement pour l’extraction sémantique n’ont pas été abordées dans cet exposé. En outre, pour être bien compris, certains principes et certaines fonctionnalités vus plus haut nécessitent d’être illustrées par des applications réelles.

Notre espérance est que la partie pratique de notre thèse aura aidé à rendre plus clair ce qui l’est déjà et à éclaircir ce qui ne l’est pas encore.

CHAPITRE 2

Description et caractéristiques du corpus

**Introduction partielle :**

Les données sur lesquelles porte notre analyse constituent un macro-corpus qui comprend des livres, des écrits journalistiques, des lettres, des discours, des rapports, des manifestes et d’autres écrits. Sa composition est donc très hétérogène.

Nous avons décidé de travailler sur l’œuvre complète d’Abbas pour deux raisons principales :

Premièrement, l’approche lexicométrique nécessite que le corpus d’analyse soit assez volumineux pour assurer une bonne fiabilité des résultats.

Deuxièmement, pour mieux cerner l’œuvre d’Abbas, nous jugeons qu’il faut la prendre *in extenso*. La raison en est que les différentes périodes qui constituent l’itinéraire politique de l’auteur algérien sont si étroitement liées les unes aux autres que chacune d’elles constitue le prolongement de celle qui la précède. L’œuvre de l’auteur est donc un tout. Par conséquent, toute analyse qui ne prend en compte qu’une partie de ce « tout » serait à notre avis partielle et incomplète. Faisons nôtre la réflexion de Ortega Y Gasset :

« *Tout texte nous apparait comme un simple fragment d’un tout X qu’il nous appartient de reconstruire. Lire vraiment, de façon authentique, consiste, donc à replacer les mots donnés dans le cadre de ce tout talent au sein duquel ils se précisent et par là même s’expliquent.* »[[90]](#footnote-91)

Il faut noter également que, du point de vue temporel, nos données constituent un corpus diachronique. Le plus ancien texte de ce corpus remonte à l’année 1922. Le plus récent a été écrit en 1985. Tous ces textes ont été écrits en français. Notre corpus est donc, du point de vue lexicométrique, un corpus unilingue francophone.

En outre, l’intitulé de notre thèse pourrait laisser croire que ce corpus est constitué de toute l’œuvre d’Abbas. Nous aurions souhaité inclure dans notre travail tous ses textes mais ils sont si nombreux qu’il serait impossible de les rassembler complètement, d’autant plus que certains d’entre eux ont disparu - des archives surtout – Nous les avons cherchés à la Bibliothèque Nationale d’Algérie, aux Archives Nationales d’Algérie, aux Archives de la Wilaya de Constantine, aux Archives de la Wilaya de Sétif, aux Archives de la Wilaya d’Alger. Toutes nos recherches se sont révélées infructueuses. Nous avons consulté en ligne les répertoires de la Bibliothèque Nationale de France. Ces textes n’y sont pas non plus. Néanmoins, nous nous sommes employé à rassembler le maximum possible de ces écrits (tous les textes disponibles). Nous pouvons donc affirmer que nos données constituent un corpus assez représentatif de l’œuvre de l’homme politique algérien.

Afin de comprendre avec précision nos données, nous avons jugé nécessaire de faire une description détaillée de ce corpus. C’est à quoi nous prêtons attention dans le présent chapitre. Pour ce faire, nous procédons à une répartition générique du corpus.

Trois grandes parties se dessinent :

1-*Les ouvrages ;*

2-*Les écrits de presse ;*

3-*Les* *autres publications*.

**1-Les livres :**

Selon les historiens, on doit à Abbas cinq (05) livres. Néanmoins, nous considérons qu’il n’en a écrit que quatre (04) car le premier, *Le Jeune Algérien[[91]](#footnote-92),* est un recueil d’articles de presse parus durant la décennie 1920.

Voici la liste des quatre livres classés par ordre chronologique de parution :

**-***Guerre et Révolution d’Algérie I :* *La Nuit Coloniale,* Julliard, Paris, 1962, 219 pages. Réédition chez Alger-Livres-Editions, Alger, 2011, 219 pages.

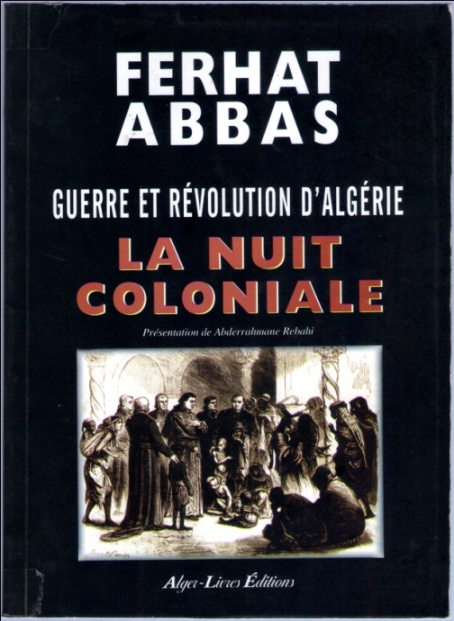
**-***Autopsie d’une guerre : L’Aurore*, Garnier, Paris, 1980, 230 pages. Réédition chez Alger-Livres-Editions, Alger, 2011, 332 pages.

**-***L’Indépendance confisquée 1962-1978*, Flammarion, Paris, 1984, 236 pages. Réédition chez Alger-Livres-Editions, Alger, 2011, 237 pages

**-***Demain se lèvera le jour,* Alger-Livres-Editions, Alger, 2010, 169 pages.

**1-1-*Guerre et Révolution d’Algérie I : La Nuit Coloniale* :**

Connu sous le titre *La Nuit Coloniale*, ce livre a été édité la première fois en 1962 après le cessez-le-feu du 19 mars. Il constitue le premier volet des Mémoires de l’auteur. L’auteur complètera ses Mémoires après l’indépendance en publiant un deuxième volume intitulé *Autopsie d’une guerre : L’Aurore* (1980).



Bien qu’achevé en septembre 1960, ce livre a dû attendre environ deux ans dans les tiroirs de l’auteur avant d’être publié.

Si on avait à résumer le contenu du livre en se fiant aux propos de l’auteur lui-même, cet ouvrage serait :

*« Un essai sur la colonisation française en Algérie, vue par un homme colonisé, resté attaché à ses origines, mais qui ne nie pas les grands apports de l’Europe et de la France à la civilisation humaine. »[[92]](#footnote-93)*

Le contenu du livre est divisé en quatre (04) grands chapitres:

1-Au-delà de l’action psychologique.

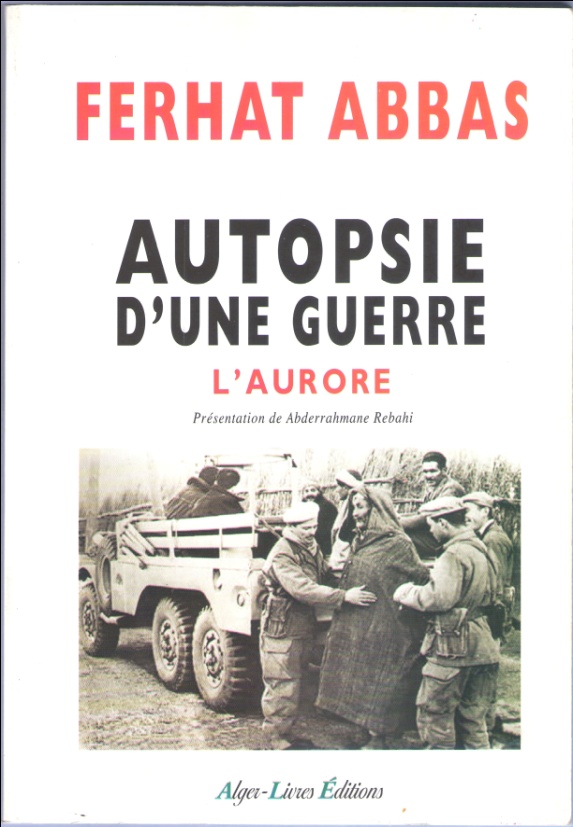
2-Cent trente ans de massacres et de lois raciales en Algérie.

3-L’impossible « Révolution » par la loi ou l’expérience de ma génération.

4-De la clandestinité au Front de Libération Nationale et à la guerre d’indépendance.

**1-2-*Autopsie d’une guerre : L’Aurore :***

Ce deuxième livre constitue le second volume des Mémoires de Ferhat Abbas. On peut être surpris par l’écart temporel qui sépare la publication des deux tomes. Le premier parait en 1962 (avant l’indépendance) et le second en 1980, soit dix-huit (18) ans après. Il semblerait que la résidence surveillée qu’on avait imposée à l’auteur l’eût empêché de terminer le second volume un peu plus tôt.



Comme le premier volume, le second est un réquisitoire virulent contre le système colonial français. Cependant, il n’épargne pas non plus certains dirigeants civils ou militaires de la Révolution tels que Boumediene, Krim, Amirouche, etc. Selon son auteur, *Autopsie d’une guerre* est une :

*« Analyse des évènements qui, à partir du 1er Novembre 1954, ébranlèrent l’Afrique du Nord dite « française » et se répercutèrent jusqu’au cœur de l’Afrique. »[[93]](#footnote-94)*Il ajoute un peu plus loin : *« Il retrace les principales étapes de l’insurrection générale de notre peuple. En même temps, il tente d’insérer ce dernier soulèvement dans l’histoire globale de la Berbérie, pour en saisir la portée sur le présent et concevoir un meilleur avenir».[[94]](#footnote-95)*

Ce livre se structure en douze (12) grands chapitres :

1-Le long cheminement de l’histoire.*Le poids du passé.*

2-L’étincelle et l’incendie. *Effacer le temps du mépris.*

3-Le gouverneur Soustelle face à l’immobilisme colonial.*Aucun choc psychologique ne vient ébranler l’ordre établi.*

4-L’appel de la montagne et l’écho de la plaine.

5-Obstination et échec de Soustelle. Son intégration ressemble comme une sœur au statut colonial.

6-L’Algérie s’installe dans la guerre. Les colons restent hostiles à tout changement- Le congrès de la Soummam.

7-L’internationalisation du conflit. La mission de la Délégation Extérieure- Conséquences de la torture et de la Bataille d’Alger.

8-Le maquisard se bat et meurt debout.*L’assassinat de Abane Ramdane – A Alger, l’orage approche.*

9-Le 13 mai 1958. La main de la Providence et l’épée du Général De Gaulle.*Proclamation du Gouvernement Provisoire de la République Algérienne. (GPRA)*

10-Le GPRA et le droit à l’autodétermination. *Régression du concept colonial dans le monde.*

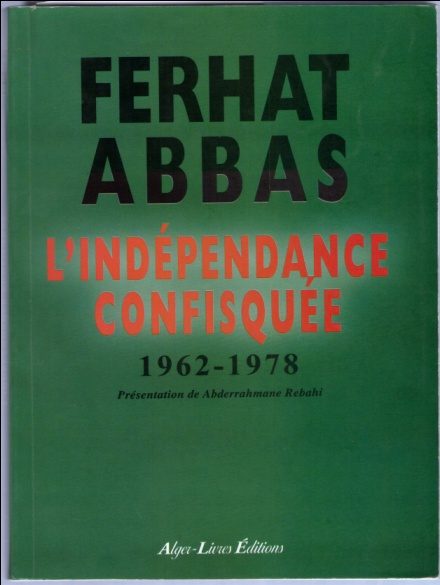
11-Pays mal aimé, l’Algérie sombre dans la guerre civile. *Elle est mutilée par l’OAS, seule responsable de l’exode des Français.*

12-Au-delà de notre temps - La science et l’homme. *L’humanité sera-t-elle une ?*

**1-3-*L’Indépendance confisquée 1962-1978***:

Ce livre est, sans conteste, le plus connu de l’auteur. Il a été édité la première fois en 1984.

Le titre choisi initialement par l’auteur a été attribué au deuxième chapitre de ce livre, c’est-à-dire *Le Sang des chouhadas trahi.*



Abbas y raconte avec beaucoup de détails son arrestation, l’arrestation de son fils, son emprisonnement au Sud algérien et sa libération en 1965. Ce livre est *«un simple témoignage sans aucune ambition personnelle. » [[95]](#footnote-96),* écrit-il.Ferhat Abbas veut apporter son témoignage sur la période de la postindépendance en fustigeant le système du Parti unique et le régime personnel de Boumediene.

*L’Indépendance Confisquée* est divisé en cinq (05) grands chapitres dont les titres sont plus que révélateurs sur l’esprit de ce livre qui constitue une critique « sans merci » du pouvoir de Ben Bella et Boumediene :

1-Grandeur et misère du F.L.N. Les causes de la subversion ou la deuxième mort des chouhadas.

2-Le sang des chouhadas trahi. L’Algérie à l’heure du stalinisme.

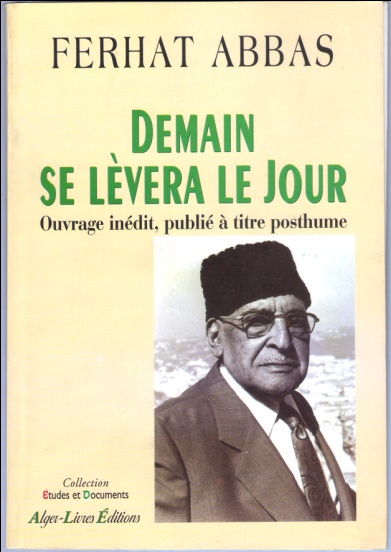
3-Deux dictateurs en compétition.

4-L’arbre et le fruit. Comment le pouvoir personnel camoufle ses échecs.

5-Démence et infaillibilité du pouvoir personnel. Quand Ben Bella explique ses erreurs.

**1-4-*Demain se lèvera le jour*** :

Ce dernier livre a été publié, à titre posthume en 2010, c’est-à-dire vingt-cinq (25) ans après la mort de son auteur qui le laissa sous forme de manuscrit. Selon la postface, il l’aurait écrit durant sa résidence surveillée sous le régime de Boumediene et l’aurait peaufiné dans les dernières années de sa vie.



L’auteur écrit dans l’avant-propos du livre :

« *Je suis au soir de ma vie. Ce livre est le dernier acte de ma vie politique. C'est un adieu à l'Algérie, à mes amis du Maghreb et à tous ceux que j'ai aimés et servis durant ma longue carrière. Et aussi un adieu à mes amis français de France et d'Algérie, et particulièrement à ceux qui ont vécu à nos côtés durant notre terrible guerre de libération, souvent au péril de leur vie.*»[[96]](#footnote-97)

Signalons que ce livre se veut aussi être un réquisitoire contre le régime socialiste de Boumediene :

« *Depuis quinze ans, l'Algérie de Boumediene, comme celle de Ben Bella, fait du «gauchisme» et de «l'aventurisme».*»[[97]](#footnote-98)

Le contenu de *Demain se lèvera le jour* est réparti en cinq (05) grands chapitres :

1-Faire l’État démocratique. *Réhabiliter la Cité musulmane.*

2-Pour une Algérie républicaine.

3-Relations extérieures de l’Algérie.*Diplomatie et liberté.*

4-Vérités permanentes et évolution. *L’éducation passe avant la contrainte.*

5-Foi dans le passé. Espérance dans l’avenir.

**2-Les écrits de journaux :**

Nous employons les expressions « écrits de journaux », « écrits journalistiques » ou encore « écrits de presse » pour désigner tous les textes publiés par Ferhat Abbas dans des organes de presse. En effet, l’auteur n’a pas publié que des articles dans les journaux qu’il a créés et animés ou dans lesquels il a collaboré. Certains écrits ne présentent pas les caractéristiques de l’article ou de l’éditorial. Parmi ces textes, on peut distinguer des messages diplomatiques, des mémorandums, des télégrammes, des appels, des lettres diplomatiques, des discours, etc.

Nous considérons qu’une distinction nette doit être faite entre les éditoriaux et les articles de journaux proprement dits. S’appuyant sur l’exemple du journal français *Le Monde,* Roselyne Koren[[98]](#footnote-99), affirme que :

« *Les prises de position ne pourraient être énoncées que dans le cadre des éditoriaux ou dans celui des analyses et des commentaires.*»

Autrement dit, les éditoriaux expriment un engagement et une prise de position de leur auteur et de l’organe de presse. Ils peuvent donc être classés dans un type de journalisme appelé le *journalisme doctrinaire*.

# Ferhat Abbas a écrit dans des organes de presse différents. Depuis 1919, lorsqu’il publia son premier article[[99]](#footnote-100) dans le journal *L’Ikdam* de l’émir Khaled, jusqu’en 1960 lorsque parut dans *El Moudjahid*, journal du FLN, sa dernière interview, il n’a pas cessé de faire paraitre des écrits d’opinion dans de nombreux journaux.

Il convient de noter que nos données journalistiques sont le fruit d’une compilation que nous avons effectuée au niveau des Archives de la Wilaya de Constantine, dans celles de la Wilaya de Sétif, dans le fonds documentaire de la Bibliothèque Nationale d’Algérie (BNA), aux Archives Nationales d’Algérie et aux Archives de la Wilaya d’Alger.

# 2-1-Articles des années 1920 :

# Les articles des années 1920 ont paru dans les journaux suivants :

\**Le Trait d’Union* : journal francophone indigénophile qui paraissait à Alger. Il fut créé et animé par Victor Spielmann, un Européen d’Algérie. Ce dernier ouvrit les colonnes de son journal aux animateurs du mouvement Le Jeune Algérien dont la revendication principale était l’égalité entre les Algériens musulmans et européens.

\**Attakaddoum* (en français *Le Progrès*) **:** hebdomadaire bilingue créé en 1923 par le Docteur Belkacem Bentami. Il avait remplacé *L’Ikdam* de l’émir Khaled exilé en Syrie en 1922. Il défendait la même ligne éditoriale que l’*Ikdam*, c’est-à-dire la défense des revendications du mouvement Jeune Algérien. Il paraissait lui aussi à Alger etexista jusqu’en 1931.

# Voici les articles republiés dans *Le Jeune Algérien :*

**2-1-1-« Le service militaire des Indigènes algériens. »**: cet article fut écrit en novembre 1922 et publié le même mois dans *Le Trait d’Union*. Ferhat Abbas l’écrivit pendant qu’il effectuait son service militaire à Bône en réponse à un article paru dans la revue *L’Afrique latine*. Il y défend les conscrits musulmans.

**2-1-2-« L’exode des ouvriers algériens en France. » :** cet article fut écrit en avril 1924 à Guelmaet publié le même mois dans *Le Trait d’Union*. Ferhat Abbas l’écrivit pendant qu’il effectuait son service militaire. L’auteur proteste contre les mesures restrictives prises par l’Administration coloniale, sous la pression de la grosse colonisation, pour empêcher le départ des ouvriers musulmans en France.

# 2-1-3-« L'Intellectuel musulman en Algérie » : ce texte parut en septembre 1927 dans *Attakaddoum*. L’auteur, étudiant lui-même, défend les étudiants musulmans algériens. Ceux-ci avaient fait l’objet d’une attaque de la part des animateurs du mouvement algérianiste[[100]](#footnote-101).

# 2-1-4-« La Tragédie d'hier et l'incertitude de demain » : ce texte parut en juillet 1930. S’exprimant sur ce texte, l’auteur déclare dans *Le Jeune Algérien*:

# « *Après les fêtes du Centenaire de la Conquête de l’Algérie* […]*, j’ai écrit pour rappeler à la colonisation que si elle ne changeait pas de conception, elle se préparerait de sombres lendemains*»[[101]](#footnote-102)

# 2-1-5-« Justice et loyauté d'abord, politique après », article paru en 1930.

# Les trois derniers articles ont été écrits et publiés pendant que l’auteur effectuait ses études de pharmacie à la faculté de médecine d’Alger.

# 2-2-Écrits des années 1930 :

# Cette décennie fut prolifique pour l’auteur puisqu’il y connut ses vrais débuts en politique lorsqu’il devint collaborateur du Docteur Bendjelloul, président de la Fédération des Élus Musulmans du Constantinois et surtout en collaborant au journal de la Fédération, *L’Entente franco-musulmane.*

# Par ailleurs, Abbas créa, lorsqu’il présida l’Amicale des Etudiants Musulmans de l’Afrique du Nord (l’AEMAN), une revue illustrée, *Ettelmidh,* dans laquelle il publia de nombreux articles. Il publia également des articles dans la revue de l’AGEA (Association Générale des Etudiants Algériens) qui regroupait aussi bien les étudiants musulmans que les étudiants européens d’Algérie. Nous avons cherché en vain cette revue, appelée « *Alger-Etudiant* » Nous aurions souhaité inclure ses textes dans notre corpus, mais nos recherches se sont révélées infructueuses. Selon Guy Pervillé[[102]](#footnote-103), Ferhat Abbas avait publié des articles dans les numéros 92 (10-2-1929), 126 (14-1-1932), 129 (4-2-1932), 130 (18-2-1932), 131 (25-2-1932).

# Pourtant, dans *Le Jeune Algérien* (édition de 1981), Abbas ne cite pas *Ettelmidh* et *Alger-Etudiant : « Mes premiers articles parurent dans* L’Ikdam *de l’émir Khaled, dans* le Trait d’Union *de l’inoubliable Victor Spielmann, dans* le Ettakaddoum *du Dr Bentami puis dans* L’Entente  *du Dr Bendjelloul. »[[103]](#footnote-104)*

# 2-2-1-Ecrits parus dans *Ettelmidh*(1931-1933):

# *Ettelmidh* était une revue bilingue paraissant tous les mois et avait pour sous-titre « Bulletin mensuel de l’Association des Etudiants Musulmans de l’Afrique du Nord » Son siège était sis à 2, rue Marie-Lefèvre, Alger. Son premier numéro parut en 1931.

# Voici ci-contre l’image de la couverture de la revue :

# C:\Users\Samsung\Desktop\JOURNAL\DSCF6237.jpg C:\Users\Samsung\Desktop\JOURNAL\DSCF6182.jpg

# Les numéros que nous avons consultés à la Bibliothèque Nationale d’Algérie sont :

# *-La première année :*

# -Le n° 02 : décembre 1931.

# -Le n°03 : janvier 1932.

# -Le n°04 : février 1932.

# -Les nos 05 et 06 (numéro double): mars et avril 1932.

# -Les nos 07 et 08 (numéro double): mai et juin 1932.

# *-La deuxième année :*

# -Les nos 01 et 02 (numéro double) : novembre et décembre 1932.

# Parmi les douze numéros publiés, deux sont indisponibles, à savoir le n° 01 de la première année et les numéros 03 et 04 (numéro double) de la deuxième année (celui de janvier 1933). De ce fait, nous ne pouvons affirmer qu’Abbas y avait publié des articles. Dans les 10 numéros consultés, nous n’avons trouvé que deux textes publiés par Ferhat Abbas. Cela veut dire que ce dernier ne publiait pas régulièrement des articles dans cette revue même s’il en était le fondateur. Seuls les numéros 03 (janvier 1932) et 04 (février 1932) de la première année contiennent chacun un texte d’Abbas. Le premier contient un texte intitulé « Notre Local » et le second un texte intitulé « Nos Vœux au Congrès de l’UN ».

# 2-2-2-Ecrits parus dans *L’Entente Franco-Musulmane* (1935-1942)[[104]](#footnote-105):

# En 1933, Ferhat Abbas s’installa à Sétif où il exerça le métier de pharmacien et où il entama sa vraie carrière politique en devenant, tour à tour, conseiller général en 1934, conseiller municipal et délégué financier en 1935. Membre de la Fédération des Élus Musulmans du Département de Constantine, il devint très vite l’un des premiers collaborateurs de son président, le Docteur Bendjelloul, puis rédacteur en chef de son journal, *L’Entente franco-musulmane*, en 1937. Les écrits de ce journal constituent les principales publications de l’auteur pendant cette décennie.

# Présentation du journal :

# *L’Entente* était un hebdomadaire francophone qui paraissait sur quatre pages tous les jeudis. Il fut créé à Constantine[[105]](#footnote-106) le 29 août 1935 par la Fédération des Élus Musulmans du Département de Constantine. Si le siège de la rédaction, de l’administration et de l’abonnement était instable (à Constantine, Sétif et Alger), le siège social du journal resta à Constantine (à 08, Rue de Chabron) jusqu’à sa disparition. Il porta le titre « *L'Entente franco-musulmane*»*.* Il disparut en 1942. Son directeur politique était Mohamed Bendjelloul, président de la Fédération des Élus. Son rédacteur en chef était Mohammed El Azziz Kessous. Ferhat Abbas remplaça ce dernier en 1937.

# Ce journal connut plusieurs interruptions (notamment du 9 juin 1936 au 29 août 1937). *L’Entente* avait d'abord pour titre *L'Entente franco-musulmane* et pour sous-titre *Organe hebdomadaire d'union et de défense des Musulmans algériens.* Quand Ferhat Abbas prit les relais de Mohammed El Azziz Kessous à la tête de la rédaction, en 1937, le sous-titre changea et devint : *Organe de la fédération des élus musulmans et de l'Union populaire algérienne pour la conquête des droits de l'homme et du citoyen*.



Il faut souligner, en outre, que ce journal était un journal d’opinion. C’était le porte-voix de la Fédération des Elus puis de l’UPA (Union Populaire Algérienne, parti créé par Ferhat Abbas en 1938). Il défendait la ligne politique des deux partis. De ce fait, les différentes publications de Ferhat Abbas, notamment les éditoriaux, expriment des positions politiques. Sur ce point, les titres des articles sont plus que révélateurs : « Provocation et mensonge colonial*»*, « Arguments de concierge*»*, « L’Algérie devant le projet Violette*»,* « Les faux prophètes*»*, « Le Musulman algérien doit-il rester esclave ? *»,* etc.

# Quelques numéros de ce journal ne sont pas disponibles (04/157). Ces numéros sont: le 140, le 141, le 151 et le 153. Ils ne représentent que 2.54% de la totalité des numéros. Mais, du moment qu’ils ne sont pas disponibles, l’on ne peut savoir si Abbas y avait publié des textes.

# 2-3-Ecrits des années 1940 et 1950 :

# Ces deux décennies n’étaient pas moins prolifiques que la précédente puisqu’elles virent Ferhat Abbas devenir un acteur politique incontournable de la scène politique algérienne. D’abord, en rédigeant *Le Manifeste du peuple algérien*. La décennie 1950 est connue pour être celle pendant laquelle l’auteur mit fin à environ trente ans de politique légale pour rejoindre la lutte armée pour la libération nationale. Les organes dans lesquels les textes de Ferhat Abbas parurent pendant ces deux décennies sont principalement *Egalité*, *La République algérienne* et *El Moudjahid.*

# 2-3-1-Articles parus dans *Egalité* :

# Présentation du journal :

# Porte-parole des Amis du Manifeste et de la Liberté, ce journal francophone – excepté quelques numéros bilingues [[106]](#footnote-107) - fut créé en septembre 1944 par Abbas grâce à l’intervention de son ami Aziz Kessous auprès des autorités coloniales.[[107]](#footnote-108) Il porta le titre *Égalité des hommes, des peuples, des races* et le sous-titre *Organe du Manifeste Algérien*. Son premier numéro parut le 15 septembre 1944. Dans l’éditorial de ce numéro, Abbas explique la ligne éditoriale de son journal en ces termes :

# « *Ce journal est l’expression « indigène ». Il nait pour traduire l’opinion des Algériens. Il défendra en pleine clarté et sans réticence leurs intérêts et leurs légitimes aspirations. Il combattra pour ce qui est vrai et juste. »*

# La direction politique de *Egalité*était assurée par Ferhat Abbas tandis que la rédaction en chef était confiée à Azziz Kessous. Le siège social du journal ainsi que l’administration et la rédaction étaient à 06, Place Cardinal-Lavigerie, Alger. Ce journal fut très largement diffusé (130 000 exemplaires)[[108]](#footnote-109) Le mouvement dont *Egalité* exprimait les idées politiques connut grâce à son porte-voix un très grand succès. Egalité cessa de paraitre de mai 1945 à août 1946. Son directeur était en prison. Il le fit reparaître en août 1946. Ce journal connut un grand succès jusqu’en 1945 où il fut interdit.[[109]](#footnote-110)

# C:\Users\Samsung\Desktop\JOURNAL\DSCF1602.jpg

# En février 1948, *Egalité* changera de nom pour devenir *Egalité* *La République Algérienne.* Le mot « égalité » tombera ensuite. Ceci renseigne sur le changement de trajectoire politique d’Abbas qui revendique désormais la création d’une république algérienne dirigée par les Algériens.

# 2-3-2-Articles parus dans *La République Algérienne* :

# Présentation du journal :

En février 1948, Ferhat Abbas durcit ses positions en remplaçant *Égalité* par la *République Algérienne*[[110]](#footnote-111)*.* Ce journal était l’organe de l’Union Démocratique du Manifeste Algérien[[111]](#footnote-112). Il s’agit du même journal même si son titre a changé de « *Egalité »* à «*Egalité,*  *La République Algérienne ».* La numérotation des articles n’a pas changé.Ce journal n’est donc que le prolongement du précédent. Son directeur politique était toujours Ferhat Abbas. Son siège social et la rédaction et l’administration étaient à 2, Rue Arago, Alger. *La République Algérienne* disparut en décembre 1955.



Chose curieuse ! On peut constater que *L’Entente*, *Egalité* et *La République algérienne* constituaient tous le même journal bien que le nom changeât plusieurs fois. On sait que ces trois journaux étaient les organes de presse des partis créés par Ferhat Abbas. Ils exprimaient alors la ligne politique défendue par ce dernier.Il s’agit d’une rogression significative qui correspondait à la progression de la ligne politique de l’homme du Manifeste.

De *L’Entente* entre Français et Musulmans d’Algérie, on passe à l’*Egalité* entre ces derniers jusqu’à la création d’une *République Algérienne* fédérée à la France. Ce n’est là, parait-il, que le cheminement de la ligne politique de Ferhat Abbas.

Sur les 435 numéros parus, nous avons pu recueillir 394 numéros soit environ 90% du total. Les numéros manquants (42) ne sont disponibles ni aux Archives de la Wilaya de Constantine ni à la BNA ni aux Archives Nationales d’Algérie ni à la Bibliothèque Nationale de France.[[112]](#footnote-113) En revanche, du moment que ces numéros ne sont pas disponibles, nous ne pouvons affirmer que des articles d’Abbas y avaient paru.

# 2-3-3-Ecrits parus dans *El Moudjahid [[113]](#footnote-114) :*

# Présentation du journal :

# Journal francophone créé clandestinement à l’initiative d’Abane Ramdane en juin 1956, *El Moudjahid* (en français *Le Combattant*) était l’organe de presse du FLN. Il remplaça un autre journal arabophone intitulé « *al-Muqawama al-jaza’iriya* » (en français *La Résistance algérienne)*. Il était le porte-parole de la Révolution. À défaut d’un autre moyen de communication, le journal du FLN était le seul support médiatique qui rendait publiques les communiqués du FLN et les déclarations de ses dirigeants.

# C:\Users\Samsung\Desktop\100_33002546.jpg

# Les écrits parus dans *El Moudjahid* ne présentent pas tous les caractéristiques d’un article de presse. Ce journal a publié des déclarations du Président du GPRA, des appels lancés au peuple algérien, des correspondances diplomatiques, des discours, des mémorandums, des interviews accordées à des journaux étrangers (reprises), etc. À l’exception des interviews, ces écrits ne peuvent être considérés comme des écrits de presse.

# Les 47 écrits relevés dans ce journal, sont des messages diplomatiques (17), des appels au peuple algérien (10), des télégrammes (06), des déclarations faites devant un public de journalistes ou accordées, ou bien à *El Moudjahid* ou bien à d’autres organes de presse internationaux (09).

# Tous les numéros de ce journal sont disponibles aux Archives Nationales d’Algérie. Nous disposons donc de toutes les publications parues dans le journal du FLN.

# 3-Les autres publications :

# Elles sont constituées de rapports, de témoignages, d’appels au peuple algérien, de correspondances diplomatiques, de télégrammes, etc. Elles ont été ou bien publiées par des journaux, depuis *L’Entente* jusqu’à *El-Moudjahid* en passant par *Egalité* et *La République Algérienne* ou bien recueillies dans des livres d’Histoire*.*

# Pour une présentation détaillée de ces publications, nous renvoyons le lecteur aux annexes.

# Certains de ces textes ont marqué le parcours de Ferhat Abbas, et partant, l’histoire de l’Algérie colonisée ou indépendante. Nous pensons qu’ils méritent d’être présentés avec des détails :

# 3-1-Le Rapport au Maréchal Pétain (ou L’Algérie de demain) (1941):

# Ce rapport a été adressé par Abbas, le 10 avril 1941, au maréchal Pétain, chef de l’Etat français. Il constitue, selon son auteur, « *le dernier de [mes] écrits en faveur de l’égalité des droits, dans le cadre de la République française.* »[[114]](#footnote-115) Il le publia dans la deuxième édition du *Jeune Algérien* (édition de 1981).

# 3-2-Le Manifeste du Peuple Algérien (1943):

# Le Manifeste a été publié le 10 février 1943. C’est un des textes les plus célèbres de l’auteur. Ne l’appelle-t-on pas « l’Homme du Manifeste » ? Il marqua un tournant décisif dans le parcours politique de son auteur. Le Manifeste est reproduit dans presque tous les livres d’Histoire d’Algérie et de la France. Son auteur profita de la présence en Algérie des forces anglo-américaines depuis le débarquement du 8 novembre 1942 pour faire connaitre les aspirations de la classe politique algérienne. Avant la publication de son texte, Ferhat Abbas le fit signer par 28 personnalités musulmanes. Il y réclame la participation des Algériens aux décisions et aux institutions de leur pays.

# Le Manifeste s’articuleautour de« *deux pivots essentiels : d’une part la condamnation de la colonisation, et d’autre part, la volonté des Musulmans de participer sans délais et de façon effective, au gouvernement de leur pays*. »[[115]](#footnote-116)

# Le Manifeste doit sa réputation à la nouveauté des idées qu’il exprime. Il sonna le glas de la politique d’assimilation. En fait, Ferhat Abbas rompit avec ses revendications assimilationnistes et prôna, pour la première fois, la participation des Algériens à la gestion des affaires de leur pays en proclamant  *la nécessité de reconnaître la nationalité algérienne et l’État algérien*.[[116]](#footnote-117)

S’exprimant sur ce texte, Abbas déclare en 1948, dans un éditorial de *Egalité* :

« *Le Manifeste a établi le bilan d’un siècle de violences et de colonisation française dans un pays arabe. Il a mis en relief l’inanité d’une politique dite d’annexion et d’assimilation.*»[[117]](#footnote-118)

# 3-3-L’Additif au Manifeste du peuple algérien (1943):

# Après avoir remis Le Manifeste au gouverneur général d’Algérie, Peyrouton, il a été demandé à Abbas de rédiger un additif qu’il remettra à la Commission d’études économiques et sociales constituée par le Gouverneur lui-même. Ferhat Abbas rédigea le document entre avril et juin 1943.

# Ce texte est «*plus précis et même plus révolutionnaire que le Manifeste…Sur le plan politique, il va plus loin que le Manifeste. Il pose comme principe essentiel la reconnaissance de la nation algérienne, donc de la citoyenneté algérienne, et réclame l’autonomie politique de l’Algérie en tant que nation souveraine*»[[118]](#footnote-119)

# 3-4-Mon Testament politique (1946):

# Ce texte a été publié, la première fois, par le général Massu dans ses Mémoires. Abbas l’aurait écrit en prison, début 1946. Rappelons que Ferhat Abbas fut emprisonné le 08 mai 1945 suite aux évènements qui ont ébranlé l’Est algérien et causé la mort de milliers de victimes. L’historien et spécialiste de l’Algérie Charles-Robert Ageron l’aurait trouvé dans les Archives de l’Armée de Terre française. Il le publia en 1994 dans *La* *Revue Française d’histoire d’Outre-mer*.

# Charles-Robert Ageron déclare, à propos de ce texte :

# « *L’intérêt historique de ce document réside dans le fait qu’il se présente comme le testament politique d’une personnalité algérienne de premier plan qui continuera en réalité une longue carrière, mais qui, en 1946, déclare aspirer à la retraite. Abbas a écrit ce texte d’abord pour défendre son honneur et se justifier d’une accusation « arbitraire ».*»[[119]](#footnote-120)

# 3-5-La Lettre de démission de l'Assemblée nationale constituante (1963):

# En septembre 1962, Ferhat Abbas fut désigné Président de l’Assemblée nationale constituante. Quelques mois après, le 13 août 1963, il décida de se retirer de l’arène politique en adressant une lettre de démission à ses collègues de l’Assemblée. Le Président de la République, Ahmed Ben Bella, aurait réuni ses partisans dans un cinéma d’Alger pour élaborer la Constitution de la République algérienne au lieu de confier son élaboration à l’Assemblée constituante. Se sentant marginalisé et humilié, Ferhat Abbas décida alors de démissionner. Cette lettre est assortie d’un long texte intitulé : « Pourquoi je ne suis pas d’accord avec le projet de Constitution établi par le gouvernement et le bureau politique. »

# 3-6-L’Appel au peuple algérien (1976):

# En mars 1976, opposé à la politique du président Boumediene, Ferhat Abbas rédige avec Cheikh Kheireddine, ancien membre du Conseil National de la Révolution Algérienne, Benyoucef Ben Khedda, ancien président du GPRA et Hocine Lahoual, ancien secrétaire général du Parti du Peuple Algérien et ancien représentant du FLN à l’extérieur, un *Appel au peuple algérien* dans lequel ils désapprouvent le pouvoir personnel du président Boumediene. Aussitôt, les quatre personnalités sont arrêtées et emprisonnées.

CHAPITRE 3

Méthodologie et outils d’analyse

**1-Méthodologie :**

Pour effectuer l’analyse lexicométrique, nous procédons de la manière suivante :

Il faut savoir que notre corpus, qui englobe quelques centaines de textes, constitue une base lexicométrique assez importante. L’appréhension de ce matériel linguistique nécessite le respect de certaines règles.

Tout d’abord, il s’agit d’une étude comparative entre les différentes partitions du corpus que l’on va distribuer en fonction de plusieurs variables. En lexicométrie, les découpages du corpus se font habituellement en fonction des hypothèses de recherche.

Selon Leimdorfer et Salem[[120]](#footnote-121), les méthodes lexicométriques peuvent être grossièrement regroupées de la manière suivante :

-les méthodes documentaires qui opèrent une simple réorganisation de la surface textuelle ;

-les méthodes qui opèrent, pour chaque texte pris isolément, des comptages et des calculs d’indices statistiques;

-les méthodes statistiques « contrastives » qui produisent des résultats portant sur le vocabulaire de chacun des textes par rapport à l’ensemble des textes réunis dans un même corpus à des fins de comparaison.

C’est cette dernière approche que nous allons utiliser.

Pour Damon Mayaffre :

*« De manière générale, les segmentations exogènes des corpus, sur la base d’hypothèses de travail, apparaissent comme un moment critique de l’analyse lexicométrique où le risque de projection des hypothèses sur les résultats est important. »[[121]](#footnote-122)*

La distribution du corpus en sous-parties est donc indispensable pour l’analyse lexicométrique. L’objectif du découpage est de faire contraster entre elles ces différentes parties selon la variable mise en œuvre. Le découpage permet surtout de montrer la spécificité ou l'identité propre du discours de chacun des textes du corpus.

Si nous considérons de plus près le parcours politique de Ferhat Abbas, il est possible de distinguer plusieurs découpages :

**Le premier découpage** : quatre sous-parties sont possibles. Elles correspondent à quatre périodes différentes. Chaque période est déterminée par la position politique défendue par l’auteur et correspond à une partie du corpus qui contient un certain nombre de textes écrits par Abbas pendant cette période. Ce découpage est inspiré des jugements des historiens. (Voir plus haut la citation de Guy Pervillé)

Les sous-parties de ce premier découpage sont les suivantes :

1. De 1920 à 1943. Revendications principales : l’assimilation, l’intégration.
2. De 1943 à 1956. Revendication principale : le fédéralisme.
3. De 1956 à 1963 : Nationalisme et lutte indépendantiste.
4. De 1963 à 1985 : Lutte démocratique contre l’autoritarisme.

**Le deuxième découpage** : il s’agit d’une partition qui tient compte du genre des textes produits : articles de presse[[122]](#footnote-123) ou éditoriaux, textes littéraires, textes politiques. Le principal objectif de ce découpage est d’examiner les ressemblances ou dissemblances lexicales que pourraient engendrer les différences de genres.

Trois sous-parties se dessinent:

1. Textes journalistiques : éditoriaux, articles.
2. Textes littéraires : les Mémoires de l’auteur.
3. Textes politiques : rapports, discours, manifestes, correspondances, etc.
4. Les livres : *L’Indépendance confisquée* et *Demain se lèvera le jour*.

**Le troisième découpage** : selon qu’il soit en situation de « hors-gouvernance » ou de « gouvernance », les stratégies discursives employées par un homme politique sont différentes. Selon Patrick Charaudeau :

« *Que le sujet politique se trouve en situation d'énonciation de "hors gouvernance" (lorsqu'il s'agit, par exemple, d'accéder au pouvoir en étant candidat à une élection), ou de "gouvernance" (lorsqu'il est "aux affaires"), il lui faut agir et communiquer, communiquer et agir, mais avec des stratégies légèrement différentes, du fait qu'il se trouve dans des positions de légitimité différentes. Mais dans un cas comme dans l'autre, divers types de parole (ou stratégies discursives) sont à sa disposition : parole de* promesse*, parole de* décision*, parole de* justification*, parole de* dissimulation*.*»*[[123]](#footnote-124)*

L’on comprend donc que les stratégies discursives mises en œuvre par l’homme politique dépendent de la position qu’il occupe sur la scène politique.

De même, le sujet politique développe-t-il des stratégies différentes en fonction de sa présence ou son absence sur la scène politique ? S’agit-il d’un sujet politique actif qui participe à la gestion de la vie sociopolitique[[124]](#footnote-125)? Ou, par contre, ne joue-t-il que le rôle d’observateur du champ politique ?

Il faut préciser que Ferhat Abbas n’a été « aux affaires » que pendant une période très courte au lendemain de l’indépendance. Il a été donc toujours « hors-gouvernance ».

Ce que nous pouvons remarquer, c’est qu’on peut distinguer deux parties dans le parcours de Ferhat Abbas en tenant compte de ce qui vient d’être dit: selon qu’il soit acteur de la scène politique (Président de parti, délégué financier, député, représentant.) ou qu’il soit simple observateur de la scène politique de son pays comme ce fut le cas après l’indépendance. En fait, pendant la colonisation française, Abbas, qui était un autochtone, n’avait jamais occupé une place dans le système politique français qui dirigeait l’Algérie. Même lorsqu’il siégeait dans les assemblées élues de l’Algérie française ou dans le parlement français, ses projets politiques étaient à l’opposé de la politique coloniale.

Abbas avait occupé donc des « positions de légitimité » différentes : selon qu’il fût un sujet politique actif[[125]](#footnote-126), ou un simple observateur de la scène politique[[126]](#footnote-127). Employait-il, de ce fait, des types de paroles différents ? Le recul qu’il avait pris au lendemain de l’Indépendance lui permettait-il d’avoir un regard différent et par là même de développer des stratégies de communication différentes dans ses textes?

C’est ce que nous tentons de vérifier à travers ce découpage.

Deux sous-parties du corpus sont possibles :

**1-**Textes produits parFerhat Abbas en tant qu’acteur politique. Cette période commence aux années 1920 et se termine lorsqu’il quitte le champ politique en 1963, suite à sa démission de la présidence de l’Assemblée nationale constituante. Durant toute cette période, Ferhat Abbas a été un acteur incontournable de la scène politique algérienne. Ainsi, il a été, tour à tour, président de l’AEMAN, président de l’UPA, président des AML et de l’UDMA, membre du CNRA et du CCE, président du premier GPRA en septembre 1958 et du second GPRA en 1960, enfin Président de l’Assemblée Nationale Constituante (1962-1963).

**2-**Textes produits parFerhat Abbas en tant qu’observateur de la scène politique. Il se trouve dans une situation d’énonciation de « hors gouvernance », qui commence en septembre 1963 lorsqu’il quitte la vie politique, et se prolonge jusqu’à sa mort en 1985. Durant toute cette période – de 1963 à 1985-, il demeure dans l’opposition contre le système du Parti unique et écrit quelques livres.

Nous préférons ne pas présenter ici tous les découpages que nous avons effectués dans le corpus. Nous renvoyons le lecteur à la partie pratique où il trouvera plus d’informations sur les autres partitions, car nous ne présentons ici que les plus importantes.

Nous estimons que le plus important de ces découpages est le premier, c’est-à-dire celui qui s’appuie sur la variable chronologique. Les résultats permettent de voir les différences de langage entre les différentes parties du corpus, entre les différentes séquences chronologiques. Ces différences peuvent être expliquées et interprétées par des raisons sociohistoriques. Les convergences ou les divergences lexicales entre les sous-parties du corpus, qui correspondent chacune à une période chronologique déterminée, peuvent être expliquées par des déterminations contextuelles.

Pour effectuer l’analyse statistique, nous procédons de la manière suivante :

Dans un premier temps, il faut faire un enregistrement en ordinateur de tout le corpus. L’enregistrement se fait de deux manières différentes :

1-Lorsqu’il s’agit d’un texte en format papier, un livre par exemple, nous procédons à la numérisation de ce texte au scanner avant d’utiliser un logiciel de reconnaissance optique des caractères qui permet d’enregistrer le document sous format Word.

2-Quand il s’agit d’un texte dont nous ne pouvons obtenir la version papier, comme les articles de journaux disponibles uniquement dans les archives et dont le seul moyen de recueil est la photographie, nous effectuons la saisie manuelle au clavier.

Une fois rassemblés et numérisés tous les textes intégrés dans le corpus, notre base de données est constituée. Il faut ensuite effectuer les découpages nécessaires. Dans les découpages qui s’appuient sur la variable chronologique, le classement des textes se fait non pas selon la date de parution de la publication mais selon la date de sa rédaction (ou production) ; certaines parties du corpus n’ont pas été publiées dans la même année de leur rédaction. Dans ces découpages, le contexte et le temps étant étroitement liés, la division du corpus se fait selon une seule variable, la période pendant laquelle le texte a été produit.

Une fois obtenues toutes les versions du corpus, il faut créer les bases de données correspondantes dans le logiciel de traitement lexicométrique.[[127]](#footnote-128) Les variables selon lesquelles sont réalisés les découpages sont des «clés» qui divisent le corpus en « parties ». Le logiciel permet dès lors de savoir, grâce à des sorties-machines, si le vocabulaire se répartit de manière spécifique selon ces variables. Les résultats permettent de comparer le vocabulaire d’un texte par rapport à celui des autres textes de la même sous-partie ou par rapport à tout le corpus qui constitue la référence.

Nous obtenons donc des bases de données lexicométriques que nous pouvons explorer grâce aux fonctions documentaires du logiciel.

Relevons, comptons, comparons, nous interprèterons ultérieurement[[128]](#footnote-129)*,* écrit Maurice Tournier*.* C’est ce que nous nous proposons de faire dans la partie pratique.

**2-Quelques outils lexicométriques :**

Le développement de l’approche automatique du discours a été, depuis son avènement, tributaire de la maturité des outils fournis par une autre discipline, l’informatique. Les outils lexicométriques existants sont aussi nombreux que divers, par leurs programmes et leurs fonctionnalités. Nous nous limitons à présenter trois logiciels. Deux de ces logiciels, Hyperbase d’Etienne Brunet et Lexico d’André Salem, sont les plus connus, les plus performants et les plus utilisés dans l’univers francophone. Nous présentons aussi un troisième logiciel nécessaire pour les recherches thématiques, à savoir l’outil Tropes. Parmi la panoplie des outils existants, on peut citer Weblex[[129]](#footnote-130), Sphinx, Alceste, Sato, Weblex, Prospero, Astartex, etc.

Dans notre analyse, nous utilisons le logiciel Hyperbase.

**2-1-Le logiciel Hyperbase[[130]](#footnote-131) :**

Ce logiciel a été conçu et développé à l’Université de Nice (France) par Etienne Brunet et produit par l'UMR 6039 Bases, Corpus et Langage (CNRS-Université de Nice). C’est un logiciel hypertexte d’exploration des textes et de traitement statistique des données textuelles.

Cet outil, né en 1989, est actuellement le plus perfectionné des outils disponibles sur le marché eu égard à ses potentialités. Il permet un triple traitement du texte brut, du texte lemmatisé et du texte grammaticalement codé. Ce qui constitue une spécificité propre à Hyperbase et qu’on ne retrouve pas dans les autres logiciels.

Les fonctions d’Hyperbase sont: les index, les concordances, les contextes, les cooccurrences. Grâce à ce logiciel, on peut effectuer une comparaison entre la base de données constituée par notre corpus d’analyse avec le corpus du Trésor de la langue française (TLF) intégré au logiciel. Une autre comparaison est possible à l’intérieur de notre base de données que nous pouvons diviser en sous-parties pour rendre possible les études contrastives de vocabulaire. Hyperbase permet d’effectuer des analyses de spécificités, des analyses factorielles et des mesures diverses appréciant la richesse lexicale, l’évolution du vocabulaire, la distance ou connexion des textes, etc.[[131]](#footnote-132)

Parmi la floraison des outils existants, Hyperbase est, grâce aux évolutions récentes, le seul logiciel d’exploration et d’analyse statistique textuelle qui permet de travailler non seulement sur le lexique mais également sur la composition grammaticale (catégories et codes grammaticaux), aux structures syntaxiques, au style et à la sémantique. Cela est possible grâce à l’intégration à la version récente du logiciel d’un autre outil, un lemmatiseur, connu d’ailleurs pour sa fiabilité. Il s’agit de l’outil Cordial de la société Synapse. Alors, en plus de *la statistique lexicale* traditionnelle sur formes graphiques brutes, il propose *une statistique grammaticale et syntaxique* et ouvre la voie à l’interprétation sémantique.

Avant d’accéder au menu principal du logiciel, l’on doit parcourir un menu qui propose les bases lexicométriques intégrées au logiciel et permet de créer une nouvelle base ou d’exploiter une base autre que celles proposées sur l’écran du menu :



**Figure 01 : Menu principal du logiciel Hyperbase, Version 9.1, mars 2013.**

Les bases intégrées contiennent les œuvres des grands auteurs classiques de la littérature françaises ainsi que celles d’écrivains du XXe siècle et deux auteurs algériens, Mohammed Dib et Mouloud Mammeri.

Voici ci-contre l’image du menu de la dernière version du logiciel :



**Figure 02 : Le menu du logiciel Hyperbase[[132]](#footnote-133).**

L’on peut voir sur cette image que les fonctions documentaires sont proposées horizontalement en haut de l’écran. Les fonctions statistiques, quant à elles, sont proposées verticalement à gauche de l’écran.

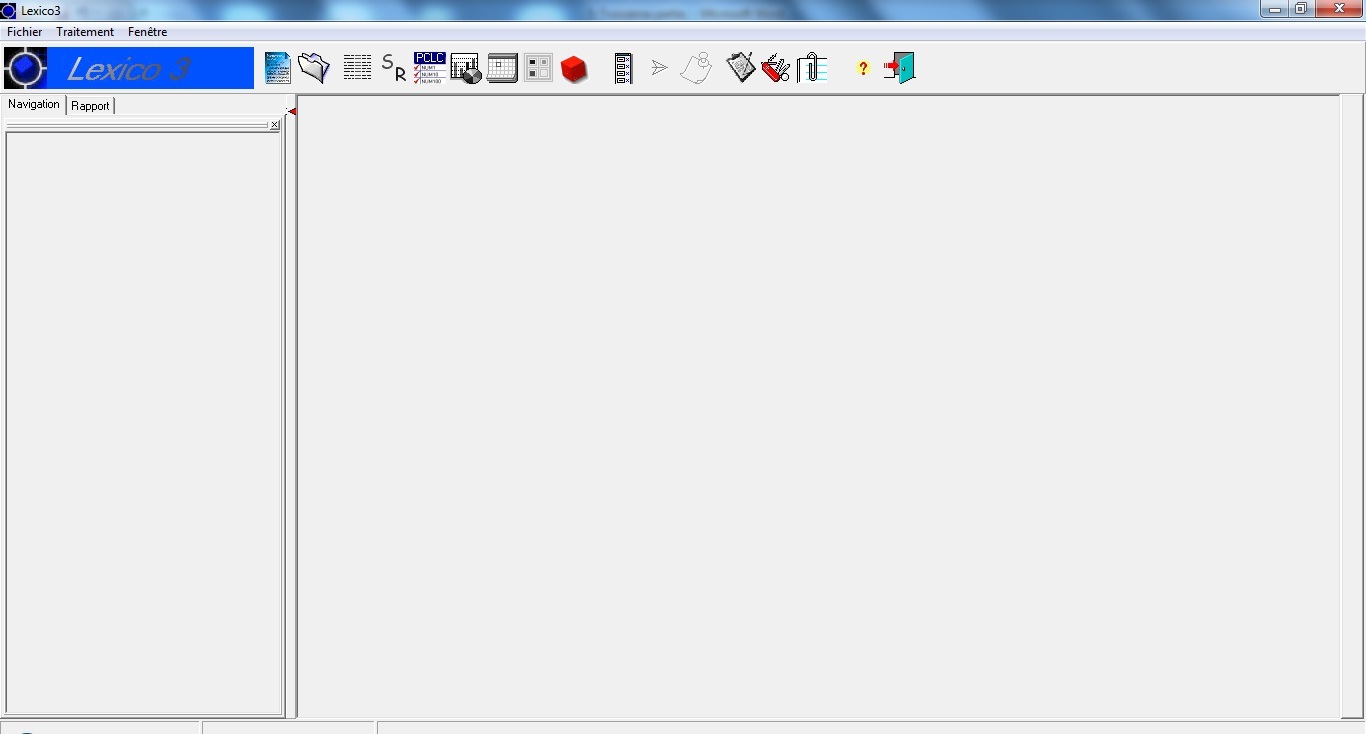
Soulignons enfin que nous allons utiliser dans notre analyse la version la plus récente du logiciel, la version 9.1 de mars 2013.

**2-2-Le logiciel Lexico :**

Moins connu et moins utilisé que le logiciel de Brunet, Lexico estsans conteste le logiciel le plus abouti de lexicométrie après Hyperbase. Mais il faut noter aussi que Lexico était, avant la parution de Hyerbase, le logiciel le plus abouti de lexicométrie. Il a été conçu et développé au fil des années par André Salem en collaboration avec Maurice Tournier à l’ENS de Fontenay Saint-Cloud avant d’être développé à l’Université de la Sorbonne nouvelle Paris 3[[133]](#footnote-134). Il est né au début des années 1980. La première version du logiciel conçue au laboratoire de Saint-Cloud a été nommée Lexicloud.

Les fonctionnalités de Lexico sont limitées. Par exemple, la version la plus récente du logiciel ne peut pas prendre en charge les sorties-machines des lemmatiseurs et ne permet donc pas la lemmatisation ni l’étude des co-occurrences. Elle se limite à un traitement lexicométrique sur les formes graphiques telles qu’elles apparaissent dans le corpus.

Hormis le fait qu’il ne permet pas la lemmatisation des corpus et l’étude des co-occurrences, Lexico offre pratiquement les mêmes fonctionnalités que les autres outils lexicométriques existants actuellement : segmentation, étude des spécificités, analyse factorielle des correspondances, concordances, segments répétés, distance ou connexion lexicales.



**Figure 03 : Le menu principal du logiciel Lexico 3.**

**2-3-Le logiciel Tropes :**

Tropes est un outil lexicométrique d’analyse sémantique de textes qui fonctionne sous Windows. Il a été développé la première fois en 1994 à l’Université de Paris VIII par Agnès Landré et Pierre Molette. Il s’agit d’un outil pluridisciplinaire dont le champ d’application couvre plusieurs domaines tels que la littérature, la sociologie, la linguistique et la psychologie.

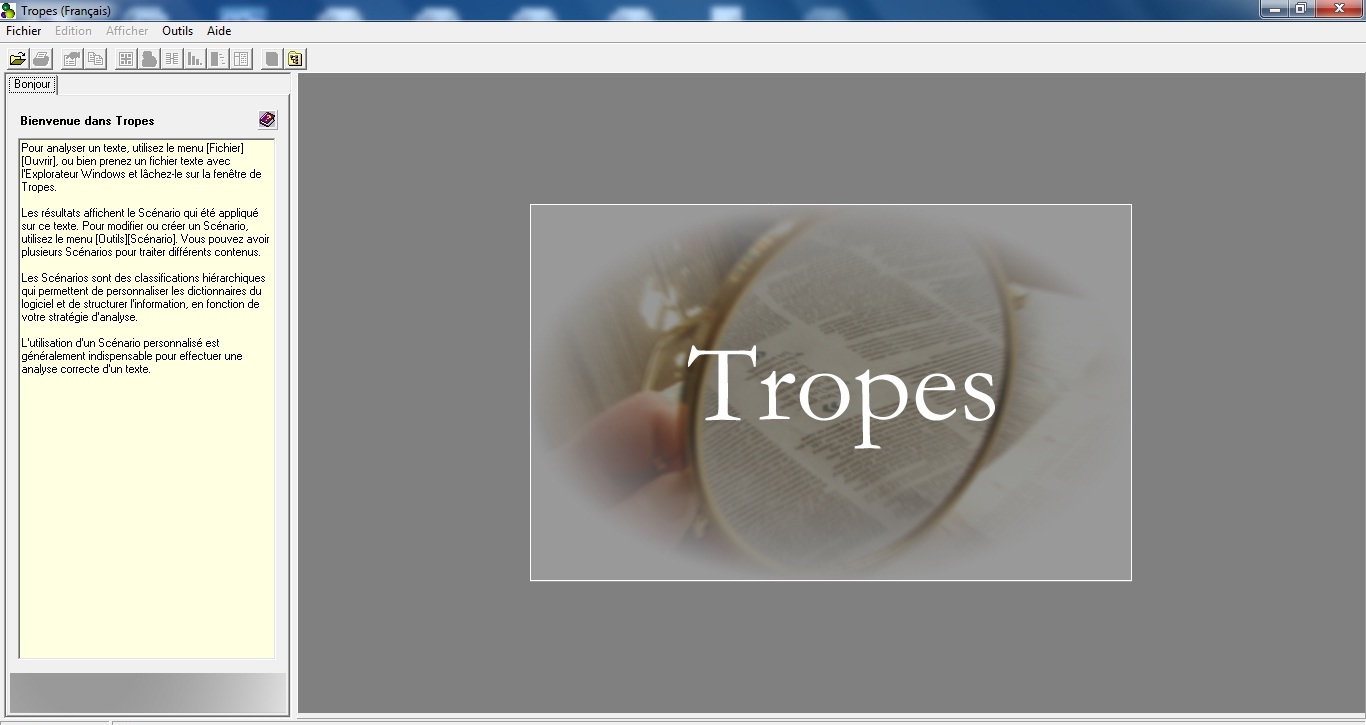
Les fonctions que recèle ce logiciel servent essentiellement à catégoriser les textes étudiés. Ces fonctions sont l’analyse du lexique, l’étude du style, les analyses morphosyntaxiques et les analyses sémantiques. Il détermine également les "Univers de référence" : les champs lexicaux qui aident à déterminer les thèmes du texte.

Tropes découpe le texte en plusieurs parties (épisodes) dont il donne (ou fournit) les caractéristiques linguistiques principales.

Pour analyser un texte, le logiciel procède tout d’abord à un découpage de ce texte propositions (découpage propositionnel). L’intérêt de ce découpage réside dans le fait qu’il permet l’étude des co-occurrences. Il procède ensuite à la désambigüisation sémantique et grammaticale des mots du texte (des homographes notamment). Ensuite, il effectue une classification grammaticale et sémantique des mots relevés. Cette classification est importante dans la mesure où elle facilite l’analyse qualitative du corpus.

L’étape suivante de l’analyse est l’analyse statistique. Le logiciel offre une palette de fonctions statistiques qui permettent de mieux aborder le texte en découvrant ses différents aspects. Elles permettent notamment l’étude du style, la chronologie du texte, l’analyse lexicale du discours. Enfin, le logiciel présente les résultats sous forme de graphes.

Tropes est directement téléchargeable sur le site : www.tropes.fr/download.htm.



**Figure 03 : Le menu principal du logiciel Tropes.**

1. Eu égard au nombre de recherches qui lui sont consacrées dans ce domaine. [↑](#footnote-ref-2)
2. Voir notamment Georges-Elia Sarfati, 2005, *Eléments d’analyse du discours*, Paris, Armand Colin, pp. 93-95. [↑](#footnote-ref-3)
3. Le linguiste prend plus de recul par rapport à l’objet de sa recherche. Son intervention n’est que partielle et sa tâche principale consiste en l’interprétation des sorties-machines. [↑](#footnote-ref-4)
4. François Leimdorfer et André Salem, 1995, op.cité, p. 139. [↑](#footnote-ref-5)
5. Cet exposé détaillé est à notre avis indispensable étant donné que la discipline et ses principes demeurent peu connus et nécessitent d’être présentés. [↑](#footnote-ref-6)
6. George Mounin (éd.), 1974, *Dictionnaire de la Linguistique,* PUF, Paris, p.203. [↑](#footnote-ref-7)
7. Ibidem [↑](#footnote-ref-8)
8. Jean Dubois (éd.), 2007, *Grand Dictionnaire Linguistique & Sciences du langage*, Paris, Larousse. [↑](#footnote-ref-9)
9. Patrick Charaudeau et Domnique Maingueneau (dir.), 2002, *Dictionnaire d’Analyse du discours*, Seuil, Paris, p. 342. [↑](#footnote-ref-10)
10. François Leimdorfer et André Salem, 1995, op.cité, p.133. [↑](#footnote-ref-11)
11. Dominique Maingueneau, 2009, *Les Termes clés de l’analyse de discours*, Seuil, Paris, p.81. [↑](#footnote-ref-12)
12. Simone Bonnafous et Maurice Tournier, 1995, op.cité, p.81. [↑](#footnote-ref-13)
13. Adresse : <http://www.cavi.univ-paris3.fr/lexicometrica/>  [↑](#footnote-ref-14)
14. Il faut noter la création, en 2000, de *L’Atrolabe*, revue électronique qui se fixe comme objectif d'explorer les relations entre littérature et informatique (<http://www.uottawa.ca/academic/arts/astrolabe/presentation.htm>). [↑](#footnote-ref-15)
15. Simone Bonnafous et Maurice Tournier, 1995, op.cité. p.69. [↑](#footnote-ref-16)
16. Outre la linguistique et la littérature. [↑](#footnote-ref-17)
17. Etude du style avec l’outil informatique. [↑](#footnote-ref-18)
18. Stéphane Sinclair, 2002, « Historique de l’analyse de texte informatisée. », in *L’Astrolabe*, http:/www.uottawa.ca/academic/arts/astrolabe/auteurs.htm. Consulté le 15 mars 2012. [↑](#footnote-ref-19)
19. Ibid. [↑](#footnote-ref-20)
20. Ibid. [↑](#footnote-ref-21)
21. Ibid. [↑](#footnote-ref-22)
22. Roberto Busa, 1998, « Dernières réflexions sur la statistique textuelle », in S. Mellet (éd.) *JADT 1998, 4e Journées Internationales d’analyse des données textuelles*, UNSA-CNRS, Nice, pp.179-183. [↑](#footnote-ref-23)
23. Par exemple, comment distinguer le vocable « marche », verbe fléchi, du mot « marche », nom commun féminin. [↑](#footnote-ref-24)
24. Sinclair, 2002, op.cité. [↑](#footnote-ref-25)
25. Il est l’auteur d’un ouvrage intitulé *La* *Stylistique*. [↑](#footnote-ref-26)
26. Citons: *Étude de* statistique lexicale. *Le vocabulaire du théâtre de Pierre Corneille*, Paris, Larousse, 1967, 379 p. [↑](#footnote-ref-27)
27. Francine Mazière, 2005, *L’Analyse du discours,* PUF, Paris. p.65. [↑](#footnote-ref-28)
28. Ibid. p.19. [↑](#footnote-ref-29)
29. Slatkine-Champion, Genève-Paris, 1824 p. [↑](#footnote-ref-30)
30. Charles Muller, 1973, *Initiation aux méthodes de la statistique linguistique,* Paris, Hachette. [↑](#footnote-ref-31)
31. André D. Robert, Annick Bouillaguet, 2007, *L’Analyse de contenu*, PUF, Paris, p.96. [↑](#footnote-ref-32)
32. Charles Muller, 1977, *Principes et méthode de la statistique lexicale,* Paris, Hachette. [↑](#footnote-ref-33)
33. André D. Robert, Annick Bouillaguet, 2007, op.cité, p. 96. [↑](#footnote-ref-34)
34. Ibidem. [↑](#footnote-ref-35)
35. Aux éditions Dunod. [↑](#footnote-ref-36)
36. 1997. [↑](#footnote-ref-37)
37. Chez Champion, Paris, 2004. [↑](#footnote-ref-38)
38. Sciences Po les presses, Paris, 2012. [↑](#footnote-ref-39)
39. Francis Chateauraynaud et Josquin Debaz, 2010, « Prodiges et vertiges de la lexicométrie. », <http://socioargu.hypothèses.org>. Consulté le 20 février 2011. [↑](#footnote-ref-40)
40. Voir sur ce point l’article d’Etienne Brunet, 2012, « Au fond du GOOFRE, un gisement de 44 milliards de mots. », accessible sur [www.jadt2012.ulg.ac.be/telechargements.html](http://www.jadt2012.ulg.ac.be/telechargements.html). Consulté le 04 avril 2013. [↑](#footnote-ref-41)
41. Francis Chateauraynaud et Josquin Debaz, 2010, op.cité. [↑](#footnote-ref-42)
42. Jean-Marc Leblanc, « Pascal Marchand, *L’analyse du discours assistée par ordinateur* », *Mots. Les langages du politique* [En ligne], 75 | 2004, mis en ligne le 23 avril 2008, consulté le 14 octobre 2012. URL : http://mots.revues.org/3643 [↑](#footnote-ref-43)
43. Chateauraynaud et Debaz, 2010, op.cité. [↑](#footnote-ref-44)
44. Ils se font de moins en moins nombreux ces dernières années étant l’évolution remarquable de la discipline. [↑](#footnote-ref-45)
45. Cette artificialité est reconnue par Pêcheux : « *Le caractère artificiel de cette lecture devient seulement plus évident quand le recours à l'informatique impose d'expliciter ces opérations à travers des algorithmes*», « Analyse de discours et informatique », in *Actes du Congrès international informatique et sciences humaines* - L.A.S.L.A. 1981, Université de Liège. [↑](#footnote-ref-46)
46. Damon Mayaffre, 2005, op.cité. [↑](#footnote-ref-47)
47. Serge Lusignan, *Quelques réflexions sur le statut épistémologique du texte électronique*.19 (1985), 209-212. Cité par Sinclair, 2002. [↑](#footnote-ref-48)
48. Al, B.F.F. «Possibilités et limites de la stylistique quantitative.» L'Ordinateur et les recherches littéraires et linguistiques: Actes de la XIe conférence internationale, 2-6 avril, 1984. Hamesse, Jacqueline et Antonio Zampoli (éd.). Paris-Genève: Champion-Slaktine, 1985, 1-11. Cité par Stéphane Sinclair, 2002, « Quelques obstacles historiques et épistémologiques dans le développement de l'analyse de texte informatisée », in *L’Astrolabe. Recherche littéraire et Informatique* (revue électronique) :

    <http://www.uottawa.ca/academic/arts/astrolabe/articles> Consulté le 15 mars 2012. [↑](#footnote-ref-49)
49. Sinclair. Ibid. [↑](#footnote-ref-50)
50. Ibid. [↑](#footnote-ref-51)
51. Mayaffre, 2007, « Analyses lexicométriques et logométrie », in OLIVESI, S. (dir.), Introduction à la recherche en SIC, Presses de l’Université de Grenoble, Grenoble, p.162. [↑](#footnote-ref-52)
52. Leblanc, 2004, op.cité, 143. [↑](#footnote-ref-53)
53. Simone Bonnafous et Maurice Tournier, 1995, op.cité, pp. 69-70. [↑](#footnote-ref-54)
54. Ibid.p.69. [↑](#footnote-ref-55)
55. Maurice Tournier, op.cité, 1996, p.183. [↑](#footnote-ref-56)
56. L’autre partie étant constituée de livres, donc plus facile à numériser au scanner. [↑](#footnote-ref-57)
57. Leimborfer et Salem, op.cité. p.132. [↑](#footnote-ref-58)
58. Bénédicte Pincemin, 2004, « Lexicométrie sur corpus étiquetés », *in Le poids des mots, Actes des 7es journées internationales d'analyse statistique des données textuelles (JADT 2004), G. Purnelle et al. (éds), vol. II,* p. 866. [↑](#footnote-ref-59)
59. Bonnafous, Tournier, op.cité. p.69. [↑](#footnote-ref-60)
60. L’hypothèse de recherche serait donc axée sur l’effet du temps sur le lexique de cette œuvre. [↑](#footnote-ref-61)
61. Leimdorfer et Salem, op. cité. p.132. [↑](#footnote-ref-62)
62. Damon Mayaffre, 2008, « L’entrelacement lexical des textes, co-occurrences et lexicométrie », revue électronique *Texte et corpus*, n°3 / août 2008, *Actes des Journées de la linguistique de Corpus 2007*, p. 91-102 (disponible sur <http://web.univ-ubs.fr/corpus/jlc5/ACTES/ACTES_JLC07_mayaffre.pdf>) page : 91. Téléchargé le 23 avril 2013. [↑](#footnote-ref-63)
63. Ibid. page. 95. [↑](#footnote-ref-64)
64. Ibidem, page : 92. [↑](#footnote-ref-65)
65. Mise au point au Laboratoire de lexicologie politique de Fonteney-Saint-Cloud. [↑](#footnote-ref-66)
66. Maurice Tournier, 1996, op.cité, p.185. [↑](#footnote-ref-67)
67. De suremploi ou de sous-emploi. [↑](#footnote-ref-68)
68. Tournier, 1996, op. cité. P. 186. [↑](#footnote-ref-69)
69. Pour qualifier des faits littéraires ou linguistiques, les logiciels de lexicométrie s’articulent sur des théories hypergéométriques mathématiques. [↑](#footnote-ref-70)
70. Mayaffre, 2005, op.cité. [↑](#footnote-ref-71)
71. Etienne Brunet, 2000, op.cité. [↑](#footnote-ref-72)
72. Sylvie Mellet et Gérald Purnelle, 2002, op.cité, p.529. [↑](#footnote-ref-73)
73. Le débat le plus connu est celui qui opposait deux grands noms de la discipline : Maurice Tournier (formaliste) et Charles Muller (lemmatiseur). [↑](#footnote-ref-74)
74. Etienne Brunet, 2000, op.cité, p.32. [↑](#footnote-ref-75)
75. Mayaffre, 2005, op.cité. [↑](#footnote-ref-76)
76. Benoît Lemaire, 2008, « Limites de la lemmatisation pour l’extraction de significations » in *Actes des 9es Journées internationales d’Analyse statistique des Données Textuelles*, UNSA-CNRS, Nice ,726. [↑](#footnote-ref-77)
77. Etienne Brunet, 2003, « Statistique et lemmatisation, l’exemple de Rabelais », *L’Astrolabe*, revue électronique. <http://www.uottawa.ca/academic/arts/astrolabe/articles>. Consulté le 05 août 2013. [↑](#footnote-ref-78)
78. Dominique Labbé indique que « *dans tout texte écrit en langue française, plus du tiers des mots sont «homographes»: une graphie et au moins deux entrées différentes dans le dictionnaire…* » cf, Labbé, 2002, « La lemmatisation des grandes bases de textes », in *L’Astrolabe*, revue électronique. <http://www.uottawa.ca/academic/arts/astrolabe/auteurs.htm> Consulté le 05 août 2013. [↑](#footnote-ref-79)
79. Mayaffre, 2007, op. cité, p. 160. [↑](#footnote-ref-80)
80. Dominique Labbé, 2002, op.cité. [↑](#footnote-ref-81)
81. Au lemmatiseur Tree Tagger également, version 9.0 de mars 2013. [↑](#footnote-ref-82)
82. Benoit Lemaire, op.cité. p.726. [↑](#footnote-ref-83)
83. Mayaffre, 2007, op.cité, p.160. [↑](#footnote-ref-84)
84. Sylvie Mellet, 2003, « Lemmatisation et encodage grammatical : un luxe inutile ? » in *Lexicometrica*, N° spécial. [↑](#footnote-ref-85)
85. Etienne Brunet, 2002, « Le lemme comme on l’aime », In *Actes des 6es Journées Internationales d’Analyse Statistique des Données Textuelles*, UNSA-CNRS, Nice, p.221. [↑](#footnote-ref-86)
86. Maurice Tournier, 1985, «Sur quoi pouvons-nous compter? Réponse à Charles Muller», *Etudes de philologie et de linguistique offertes à Hélène Nais*, *Verbum* (numéro spécial), Presses universitaires de Nancy. p. 487. Cité par Mayaffre, 2005, op. cité. [↑](#footnote-ref-87)
87. Charles Muller: 1984: IX, «De la lemmatisation», préface à Lafon P., *Dépouillements et statistiques en lexicométrie*, Paris-Genève, Slatkine-Champion.cité par Mayaffre, 2005, op.cité. [↑](#footnote-ref-88)
88. Mayaffre, 2005, op.cité. [↑](#footnote-ref-89)
89. Ibidem. [↑](#footnote-ref-90)
90. Ortega Y Gasset, *Difficultés du langage* (Diogène, n° 28) Cité par Boris de Schloezer, In Préface à : Léon Tolstoi, *La Guerre et la Paix*, Gallimard, 1972, Paris. Page 08. [↑](#footnote-ref-91)
91. *Le Jeune Algérien* a paru la première fois en 1931, aux Editions La Jeune Parque, Paris. Il a été revu et augmenté par l’auteur avant d’être réédité en 1981 chez Garnier Frères. L’auteur a joint à la deuxième édition *Le* *Rapport au maréchal Pétain* (avril 1941) [↑](#footnote-ref-92)
92. Page 17. [↑](#footnote-ref-93)
93. *Autopsie d’une guerre. L’Aurore*, avant-propos. Page 13. [↑](#footnote-ref-94)
94. Ibidem. [↑](#footnote-ref-95)
95. Page 20. [↑](#footnote-ref-96)
96. Page 25. [↑](#footnote-ref-97)
97. Page 91. [↑](#footnote-ref-98)
98. **Roselyne** Koren, 2004, « Argumentation, enjeux et pratique de l’«engagement neutre » : le cas de l’écriture de presse », Semen [En ligne], 17 | 2004, mis en ligne le 29 avril 2007, consulté le 15 novembre 2011. URL : http://semen.revues.org/2308 [↑](#footnote-ref-99)
99. Nous n’avons pas pu trouver cet article. Il ne fait pas partie de notre corpus. [↑](#footnote-ref-100)
100. Européens d’Algérie qui prétendaient être les légitimes et seuls Algériens et qui étaient groupés autour de la revue *L’Afrique Latine*. [↑](#footnote-ref-101)
101. *Le Jeune Algérien* suivi du *Rapport au Maréchal Pétain*, Alger-Livres-Editions, Alger, 2011, page.59. [↑](#footnote-ref-102)
102. Guy Pervillé, 2004, op.cité. [↑](#footnote-ref-103)
103. *Le Jeune Algérien*, p.28. [↑](#footnote-ref-104)
104. Documents recueillis aux Archives de la Wilaya de Constantine, à celles de la Wilaya d’Alger, à la BNA et aux Archives Nationales d’Algérie. [↑](#footnote-ref-105)
105. Charles-Rbert Ageron (*Histoire de l’Algérie contemporaine*, PUF, Paris, 1979, page 318) écrit que le journal parut à Alger puis à Constantine. Mais la collection que nous avons consultée aux Archives de la Wilaya de Constantine montre clairement qu’il fut créé à Constantine et non pas à Alger [↑](#footnote-ref-106)
106. Un supplément en arabe du même journal paraissait pendant l’année 1948 avant de disparaitre. Son titre était *El Watan.* [↑](#footnote-ref-107)
107. Abbas fut libéré en décembre 1943 de la résidence surveillée au Sud algérien. [↑](#footnote-ref-108)
108. Charles-Robert Ageron, 1979, op.cité. p. 568. [↑](#footnote-ref-109)
109. Christiane Souriau-Hoebrechts, 1975, *La Presse maghrébine,* Centre National de la Recherche Scientifique, Paris, p. 81. [↑](#footnote-ref-110)
110. Le numéro 115 du 27 février 1948 porte le titre *Egalité, La République Algérienne*. [↑](#footnote-ref-111)
111. Parti créé par Abbas en 1946 à sa sortie de prison. [↑](#footnote-ref-112)
112. Consultation du catalogue général faite en ligne, sur le site « www.bnf.fr » [↑](#footnote-ref-113)
113. Les numéros de ce journal ont été consultés aux Archives de la Wilaya de Constantine et aux Archives de la Wilaya de Sétif. [↑](#footnote-ref-114)
114. *Le Jeune Algérien*, 2011, page 137. [↑](#footnote-ref-115)
115. Youcef Beghoul, 2007, *Le Manifeste du Peuple Algérien,* Editions Dahlab, Alger, p. 29. [↑](#footnote-ref-116)
116. Guy Pervillé, 1987, « ABBAS, Ferhat (1899-1985) », in *Parcours, l’Algérie, les hommes et l’histoire, recherches pour un dictionnaire biographique de l’Algérie*, n°8, (novembre-décembre 1987), pp.5-16 [↑](#footnote-ref-117)
117. *Egalité*, n° 114, vendredi 6 février 1948. [↑](#footnote-ref-118)
118. Benjamin Stora et Zakya Daoud, 1995, op.cité, p. 123. [↑](#footnote-ref-119)
119. Charles-Robert Ageron, 1994, « Un Manuscrit inédit de Ferhat Abbas « Mon Testament politique » » in *Revue française d’histoire d’Outre-mer*, t, LXXXI (1994), n°303, p.181. [↑](#footnote-ref-120)
120. François Leimdorfer et André Salem, 1995, op.cité.p.133. [↑](#footnote-ref-121)
121. Damon Mayaffre, 2002, op.cité. [↑](#footnote-ref-122)
122. Dans son travail de thèse, Damon Mayaffre a inclu les articles de journaux des hommes politiques français. Il les considère donc comme des textes politiques. [↑](#footnote-ref-123)
123. Patrick Charaudeau, 2007, op.cité. [↑](#footnote-ref-124)
124. Même en étant dans l’opposition. [↑](#footnote-ref-125)
125. Pendant la colonisation. [↑](#footnote-ref-126)
126. Après l’indépendance, ou plus exactement, après son écartement de la scène politique à partir de 1964. [↑](#footnote-ref-127)
127. Chaque base lexicométrique correspond à un des découpages. [↑](#footnote-ref-128)
128. Maurice Tournier, 1996, op, cité, p.183. [↑](#footnote-ref-129)
129. Conçu par Serge Heiden de l’ENS de Lyon, France. [↑](#footnote-ref-130)
130. # Cette présentation s’inspire de la présentation du logiciel sur le site du Laboratoire « [UMR 6039 : Bases, Corpus, Langage](http://www.unice.fr/bcl) », <http://www.unice.fr/bcl/hyperbase?redirected_from=lolita.unice.fr/> Consulté le 15 octobre 2013.

     [↑](#footnote-ref-131)
131. Voir la présentation du logiciel par son concepteur et développeur: Etienne Brunet, 2001, « Le Logiciel Hyperbase », *L’Astrolabe*,

     <http://www.uottawa.ca/academic/arts/astrolabe/auteurs.htm>  Consulté le 20 décembre 2013. [↑](#footnote-ref-132)
132. L’image et le titre de la base sont personnalisés. Il s’agit de notre base de travail. [↑](#footnote-ref-133)
133. Lexico est réalisé par l’équipe universitaire CLE2T (SYLED) « Système Linguistiques, Énonciation et Discursivité - Centre de Lexicométrie et d'Analyse Automatique des Textes » de l’Université Paris 3 la Sorbonne nouvelle, sous la responsabilité d’André Salem. [↑](#footnote-ref-134)