La logométrie:
origines de la méthode et principes directeurs

علم تحليل الخطاب بمساعدة الحاسوب (اللوجوميتريكس):
أصول المنهجية والمبادئ الرئيسية

Dr. Sarah Chatti
Department of Foreign Applied Languages
French University in Egypt

د. سارة شطى
مدرس بقسم اللغات التطبيقية الأجنبية
الجامعة الفرنسية في مصر

Logometry: origins of the method and guiding principles

Abstract:

Logometry is a method of computer-assisted discourse analysis that allows the quantitative and qualitative processing of linguistic elements of large textual corpora. The aim of this article is to present the logometric method in order to make it more widely known. To do this, we will first review the origins of logometry, a term that has only recently appeared in the scientific landscape but dates back to practices inherited from recognized mathematicians, linguists, and lexicometers. Secondly, we will discuss the guiding principles of the logometric method, which have introduced significant innovations in how we approach texts and corpora.

Keywords: discourse analysis, logometry, quantitative method, qualitative method, interpretative approach.

علم تحليل الخطاب بمساعدة الحاسوب (اللوجوميتريكس): أصول المنهجية والمبادئ الرئيسية

الملخص:

اللوجوميتريكس (logométrie) هى طريقة لتحليل الخطاب بمساعدة الحاسوب تسمح بالمعالجة الكمية والنوعية للعناصر اللغوية الخاصة بالمدونات النصية الكبيرة. يهدف البحث إلى تقديم المنهجية المتبعة في هذا العلم من أجل التوعية بها على نطاق أوسع. من أجل ذلك، سنقوم أولاً بتحليل أصول اللوجوميتريكس(logométrie) ، وهو مصطلح ظهر مؤخرًا على الساحة العلمية، ولكنه في الواقع يعود إلى الممارسات الموروثة من علماء الرياضيات واللغويين وعلماء القياس المعجمي (lexicométrie) المعترف بهم. سنقوم بعد ذلك بتحليل المبادئ الرئيسية لهذه المنهجية، والتي أدخلت طرق جديدة للتعامل مع النصوص والمدونات.

الكلمات المفتاحية: تحليل الخطاب، تحليل الخطاب بمساعدة الحاسوب (اللوجوميتريكس)، الطريقة الكمية، الطريقة النوعية، النهج التفسيري.

La logométrie: origines de la méthode et principes directeurs

Méthode d'analyse du discours assistée par ordinateur, la logométrie permet de traiter de manière quantitative et qualitative les éléments linguistiques de grands corpus textuels. Le présent article vise ainsi à présenter la méthode logométrique afin de la faire connaître au plus grand nombre. Pour ce faire, nous reviendrons, dans un premier temps, sur les origines de la logométrie, terme nouvellement apparu dans le paysage scientifique, mais qui remonte en réalité à des pratiques héritées de mathématiciens, linguistes et lexicomètres reconnus. En effet, il est important, à nos yeux, de rappeler les racines de la logométrie, d'une part, en hommage aux chercheurs de l'époque qui ont développé ces méthodes, sans qui les jeunes chercheurs d'aujourd'hui n'auraient pas pu utiliser ces outils, et d'autre part, car nous sommes convaincus que ce n'est qu'en connaissant les origines d'une méthode et ses fondements, qu'il est possible de se l'approprier et de l'utiliser à bon escient. Nous aborderons, dans un second temps, les principes directeurs de la méthode, qui ont introduit des innovations majeures dans la manière d'aborder les textes et les corpus.

1. Présentation de la méthode logométrique : arrière-plan historique et épistémologique

La logométrie, fruit des humanités numériques, puise ses origines dans les méthodes statistiques françaises et américaines des années 1950 et 1960. C'est dans les années 1980 que la logométrie s'est *« solidement constituée en corps de doctrine. »* (Mayaffre, 2010, p.99). Mais comme le souligne Damon Mayaffre, à l'origine du terme logométrie, *« la méthode utilisée n'a pas d'âge.* […] *Les concordanciers par exemple sont presque aussi vieux que les livres et sont répandus en Occident dès le XIII*^{ème} *siècle. Les dictionnaires alphabétiques des formes remontent eux à l'Antiquité. »* (Mayaffre, 2010, p.98). Il s'agit donc d'une méthode ancienne, qui a évolué pour s'adapter aux contraintes modernes, notamment celles des grands corpus. Dans cette première partie, nous retracerons le cheminement épistémologique et historique de la méthode logométrique, en revenant sur les pas des pionniers de la statistique

textuelle et lexicale, en passant par la lexicométrie et la textométrie pour arriver à la logométrie.

1.1. La statistique linguistique : travaux fondateurs des précurseurs

La statistique linguistique émerge en France au milieu des années 1950 et au début des années 1960, notamment à travers les travaux de Pierre Guiraud dans ses ouvrages intitulés *Les Caractères statistiques du vocabulaire*, publié en 1954, et *Problèmes et méthodes de la statistique linguistique*, publié en 1960, au moment où la communauté scientifique prend conscience de l'intérêt de la statistique dans le calcul de la fréquence d'utilisation d'un mot, notamment à travers la loi Estoup-Zipf (1949), et au moment où le développement de l'informatique permet d'effectuer des calculs impossibles pour le cerveau humain. Mais cet intérêt pour la statistique n'est pas resté confiné à l'intérieur des frontières de l'hexagone. En effet, comme le souligne Étienne Brunet, *« le modèle anglais se nomme quantitative linguistics et son représentant le plus connu* [est] *Gustav Herdan »* (Brunet, 2014, p.14), qui a publié deux ouvrages majeurs, à savoir *Quantitative Linguistics*, en 1964, et *The advanced theory of Language as Choice and Chance*, en 1966. Au départ, ces applications statistiques innovantes sont peu répandues chez les linguistes en France. Guiraud, dans son ouvrage, mentionne d'ailleurs que *« la linguistique est la science statistique type ; les statisticiens le savent bien ; la plupart des linguistes l'ignorent encore »* (1960, p.15), car *« le langage est un phénomène essentiellement statistique ; c'est-à-dire soumis à des constantes et à des lois numériques et susceptible, à ce titre, de définitions et d'interprétations quantitatives »* (1960, p.16). Mais peu à peu, l'engouement pour la statistique linguistique se fait sentir et commence à émerger dans certains cercles de recherche. En 1958, George Gougenheim, fondateur du CREDIF [1] rattaché à Saint-Cloud, conçoit, grâce à l'application de la statistique à la fréquence des mots, un dictionnaire du français parlé intitulé *Dictionnaire fondamental de la langue française*. En Belgique, à l'Université de Liège, sous l'impulsion d'Étienne Evrard, est créé en 1961 le Laboratoire d'Analyse Statistique des Langues Anciennes (L.A.S.L.A), qui étudie *« les langues classiques – grecque et latine – en recourant aux technologies du traitement*

automatique de l'information » (ULiège – faculté de philosophie et lettres). Dans le sillage de la statistique linguistique, deux courants évoluent côte à côte avec, d'une part, la statistique lexicale, incarnée par Charles Muller et Étienne Brunet, et d'autre part, l'analyse statistique des données linguistiques ou textuelles, fondée par Jean-Paul Benzécri.

1.2. La statistique lexicale: l'étude du vocabulaire

La statistique lexicale apparaît tout d'abord à Strasbourg, autour de Charles Muller, puis à Nice, autour d'Étienne Brunet. À l'époque, Muller s'était donné pour tâche de poser les bases méthodologiques de la statistique lexicale, à travers ses deux thèses, publiées en 1964[2] et 1967[3], mais surtout à travers son ouvrage de référence intitulé *Initiation à la statistique linguistique* (1968), puis dans ses versions mises à jour, l'une en 1973, intitulée *Initiation aux méthodes de la statistique linguistique*, et l'autre en 1977, intitulée *Principes et méthodes de statistique lexicale*. Au départ, la statistique lexicale porte sur les textes littéraires, *« elle s'intéresse surtout à la structure du vocabulaire de tel ou tel ensemble de textes »* (Née et al., 2017, p.10), notamment à partir des ouvrages de Corneille, Giraudoux, Rousseau, Proust, Zola et Victor Hugo, dont l'étude statistique a été possible grâce au Trésor de la Langue Française (TLF) qui avait informatisé ces ouvrages. Comme l'indique Beaudouin :

> *« La démarche statistique adoptée revient à comparer les données observées aux données calculées à partir d'un modèle théorique. Implicitement, il y a l'idée que le texte analysé est un échantillon représentatif de la langue et que par l'étude de ce corpus, on pourra inférer des informations sur la langue. »* (Beaudouin, 2000)

Les recherches portaient donc avant tout sur les aspects quantitatifs, dans leur dimension paradigmatique (hors contexte), tels que la fréquence, la richesse lexicale, les spécificités et l'évolution du vocabulaire, mais également sur les aspects qualitatifs, dans leur dimension syntagmatique (en contexte), telle que l'étude des cooccurrents (Maingueneau, 2009, p.172).

1.3. La statistique textuelle : introduction de l'approche distributionnaliste

La statistique textuelle a, quant à elle, été développée par Benzécri, mathématicien orienté vers la linguistique et considéré comme le *« père de l'analyse des données à la française »* (Beaudouin, 2000). L'un des apports majeurs de Benzécri est certainement la représentation des associations à travers l'analyse factorielle des correspondances (AFC). Cette dernière *« repose sur une notion de linguistique, l'équivalence distributionnelle »* (Benzécri, 1981, p.4). En ce sens, Benzécri se rapproche des méthodes distributionalistes de Bloomfield et Harris, qui, à partir du début des années 1950, *« sont appliquées aux "discours" »* (Dubois, 1970, p.4). Les distributionnalistes de l'époque s'étaient donné pour objectif de réaliser une description objective de la langue. Pour ce faire, ils ont *« réun*[i] *à propos de la langue à décrire, un ensemble de phrases homogènes de cette langue (corpus) et* [ont cherché] *les règles qui régiss*[aient] *ce corpus sans tenir compte du sens des phrases »* (Moharram, s. d., p.6). Ils ont donc eu recours à deux notions essentielles : celle d'« environnement », entendue dans le sens de contexte, et celle de « distribution », qui correspond à l'ensemble des différents environnements dans le corpus (Moharram, s. d., p.6)[4]. Pour en revenir à l'AFC, elle a permis de *« visualise*[r] *les proximités entre les individus* [mots] *et les variables (sur deux axes factoriels) »* (Beaudouin, 2016, p.20) et ainsi de guider le chercheur lors du processus d'interprétation, car *« la projection ouvre à la construction du sens »* (Beaudouin, 2016, p.20). On dépasse donc la démarche hypothético-déductive, largement répandue chez les chercheurs anglo-saxons, notamment chez Chomsky et le courant générativiste, pour s'orienter davantage vers une réflexion et une interprétation inductive. Cette méthode de visualisation des données sur plan a non seulement permis d'offrir sa renommée à l'analyse des données « à la française », mais a également permis de diffuser les méthodes statistiques appliquées aux textes, notamment à travers l'ouvrage de référence de Benzécri *Pratique de l'Analyse des Données, Tome 3 : Linguistique et lexicologie*, publié en 1981.

1.4. L'Analyse automatique du discours (AAD) : vers la prise en compte de l'idéologie

À la fin des années 1960, Michel Pêcheux s'est penché sur la question des techniques automatisées d'analyse de discours et a développé, en 1969, l'Analyse automatique du discours (AAD69), inspirée du découpage syntaxique de Harris. L'AAD69 s'inscrivait *« dans l'espace du structuralisme philosophique des années 1960, autour de la question de l'idéologie, et en particulier celle de la lecture des discours idéologiques »* (Pêcheux, Léon, Bonnafous, Marandin, 1982, p.95). En effet, Pêcheux était particulièrement influencé par des philosophes tels qu'Althusser, Foucault, Barthes ou encore par la psychologie lacanienne. Son approche visait à *« traquer une sorte d'inconscient des textes »* (Maingueneau, 2009, p.141). Ainsi, comme le souligne Jacqueline Léon, *«* […] *les deux tournants linguistiques* [de l'époque] *: linguistique computationnelle, et théorie linguistique du discours, arriment l'œuvre de Michel Pêcheux »* (Léon, 2010, p.89). L'Analyse automatique du discours a donc provoqué un *« choc des "gestes de lecture" »* (Mazière, 2010, p.64). Pour Pêcheux, l'informatique était l'outil indispensable pour défendre, d'une part, sa position sur la langue, qui était pour lui une *« construction fictive de nature métalinguistique »* (Mazière, 2010, p.64) et pour défendre, d'autre part, les sciences humaines vis-à-vis des sciences « dures » ; *« c'était sa façon de garantir une scientificité à l'objet langue »* (Mazière, 2010, p.63). L'AAD69 était *« l'un des rares "programmes d'analyse de textes" (par opposition aux traitements numériques), opérationnels en France »* (Léon, 2010, p.97). En 1980, Pêcheux développe la nouvelle version de l'AAD69, cette fois-ci nommée AAD80, et fait la rencontre de Pierre Plante de l'Université du Québec à Trois-Rivières, qui développe le logiciel Déredec. Mais la mort soudaine de Pêcheux signe un coup d'arrêt au projet d'AAD.

1.5. La lexicométrie : l'ancrage lexical

En France, dans les années 1970 et 1980, se développe la lexicométrie, qui est le trait d'union entre les sciences du langage, la statistique et l'informatique (Labbé & Labbé, 2013). Bien que les bases méthodologiques aient été établies par les précurseurs de la statistique linguistique, lexicale et textuelle, c'est *« la lexicologie sociopolitique,*

pratiquée depuis plus de dix ans à Saint-Cloud, [qui] *avait inventé la lexicométrie »* (Mazière, 2010, p.65). En effet, même si la lexicométrie est *« proche de la statistique lexicale,* [elle] *s'en différencie par le fait qu'elle s'intéresse non pas aux particularités du style d'un auteur mais aux régularités d'un discours, en les mettant en relation avec des déterminations idéologiques ou des positionnements sociaux »* (Née et al., 2017, p.10). C'est au sein de l'ENS de Saint-Cloud, dans le Centre de lexicologie politique dirigé par Robert-Léon Wagner et Maurice Tournier, aux côtés de mathématiciens et d'informaticiens tels que Guibaud, puis un peu plus tard Pierre Lafon et André Salem, que se développe la méthode lexicométrique. Le Centre s'est ensuite organisé autour de plusieurs groupes d'étude élaborant des index du vocabulaire de Rousseau, Montesquieu et Voltaire. Il existait également un groupe de dépouillement des périodiques du XVIII^{ème} siècle. D'autres *« petites équipes* [se consacraient] *à la langue et aux écrivains politiques des XIX^{ème} et XX^{ème} siècles »* (Tournier, 1969, p.82). Le premier colloque de lexicologie politique, intitulé *« Formation et aspects du vocabulaire politique français, XVIème-XXème siècles »*, organisé en 1968 par le Centre, ainsi que les premiers numéros de la revue *Mots*[5], fondée en 1980 par Maurice Tournier, vont permettre de diffuser davantage la lexicométrie dans le paysage scientifique francophone. Très vite, la lexicométrie intéresse des chercheurs issus d'autres disciplines, notamment des historiens, tels que Régine Robin et Jacques Guilhaumou, qui ont travaillé côte à côte avec des linguistes, telle que Denise Maldidier.

La lexicométrie, comme l'indique Maurice Tournier dans l'article « lexicométrie » du dictionnaire d'Analyse du discours, *« n'est pas une théorie mais une méthodologie d'étude du discours, qui se veut exhaustive, systématique et automatisée »* (Charaudeau, Maingueneau, 2002, p.342). Dans son article, Lafon expose le positionnement théorique de la lexicométrie au sein de la linguistique en rappelant les trois principes fondamentaux de la méthode, à savoir : travailler sur des textes réels, dont la situation socio-historique a été clairement identifiée ; élaborer une linguistique du texte (et non une linguistique de la phrase et de la proposition) ; et orienter la recherche vers l'aspect quantitatif en

s'intéressant particulièrement aux variations du lexique (Lafon, (s. d.), p.1).

Au milieu des années 1980, Guilhaumou décrivait la lexicométrie comme étant *« un ensemble de méthodes qui permettent de décrire quantitativement les séquences textuelles constitutives d'un corpus »* (Guilhaumou, 1986, p.27). C'est ainsi que Leimdorfer et Salem classifient les méthodes lexicométriques en trois catégories, qui doivent être effectuées successivement pour que l'étude lexicométrique soit complète :

« les méthodes documentaires qui opèrent une simple réorganisation de la surface textuelle [telles que l'index alphabétique] *; les méthodes qui opèrent, pour chaque texte pris isolément, des comptages et des calculs d'indices statistiques* [telles que l'index hiérarchique, les concordances, les inventaires de segments répétés] *;* [et] *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* [telles que le calcul des spécificités]. » (Leimdorfer & Salem, 1995, p.133)

Par ailleurs, la lexicométrie a recours à un principe clé : celui de la « délinéarisation » du texte grâce à la « mise en série ». Ainsi, « *on ne l'étudie plus* [le texte] *phrase après phrase, paragraphe après paragraphe, mais on utilise des outils qui parcourent l'ensemble du corpus pour produire, le plus souvent, des résultats sous forme de listes »* (Née et al., 2017, p.14). Ces procédés techniques permettent ainsi de faire apparaître *« des récurrences ou des rapprochements souvent impossibles à détecter à l'œil nu, à plus forte raison sur des grandes masses de documents »* (Née et al., 2017, p.14). La lexicométrie est donc une *« étude statistique de l'usage des mots »* dont le principe consiste à *« laisser l'outil effectuer une série de calculs, qui va apporter de nouvelles possibilités d'exploration »* (Poudat & Landragin, 2017, p.25).

1.6. La textométrie et la logométrie : l'extension au texte et au discours

1.6.1. La textométrie : prise en compte du texte

Peu à peu, les textes littéraires ont cédé leur place et les chercheurs ont commencé à s'intéresser aux données commerciales,

sociologiques, psychologiques et politiques (Brunet, 2014, p.17-18). À cette période, on perçoit un changement à la fois terminologique et épistémologique dans les pratiques statistiques. C'est ainsi qu'au milieu des années 1990, la lexicométrie est renommée « textométrie » (Rastier, 2008, p.24). Dans leur ouvrage de référence, Céline Poudat et Frédéric Landragin définissent la textométrie comme :

« […] *l'extension de la lexicométrie au texte, c'est-à-dire à des phénomènes qui ne concernent pas seulement les mots mais tout élément textuel,* [ce qui] *amène à la textométrie et conduit à inclure dans l'exploration l'ensemble des méthodes quantitatives et des analyses statistiques s'appliquant sur un corpus. Exploration et calcul de statistiques sont ainsi liés. »* (Poudat & Landragin, 2017, p.17)

Face à cette nouvelle approche des textes par la statistique, l'offre des logiciels de lexicométrie évolue et tend davantage vers la prise en compte du texte. En effet, les logiciels de textométrie *« sont plus orientés vers l'exploration de corpus de recherche définis* […] *avec retour constant au contexte d'énonciation »* (Rizkallah, 2013, p.143). Sur le site de l'ENS de Lyon dédié au projet Textométrie, Bénédicte Pincemin et Serge Heinden (2008) font état de nouveaux modèles statistiques créés pour mettre au jour les traits saillants des textes, tels que :

« [les] *attirances contextuelles des mots (phraséologie, champs thématiques,…),* [la] *linéarité et organisation interne du texte (par exemple mots bien répartis au fil du texte ou au contraire apparaissant en "rafales"),* [les] *contrastes intertextuels (mesure statistique fiable du sur-emploi ou du sous-emploi d'un mot dans un texte, et repérage des mots et des phrases caractéristiques d'un texte),* [les] *indicateurs d'évolution lexicale (période caractéristique d'un terme, détection des ruptures significatives). »* (Pincemin & Heiden, 2008)

À travers ces nouveaux outils statistiques, de nouvelles approches des corpus numériques deviennent donc possibles, alliant une approche globale, via les calculs statistiques, et une approche locale, via les contextes d'emploi. On parle également de linguistique instrumentée ou outillée.

1.6.2. La logométrie: prise en compte du discours

Parallèlement à l'essor de l'appellation « textométrie », émerge au début des années 2000, sous la plume de Damon Mayaffre, un nouveau terme, celui de « logométrie ». En effet, dans ce contexte d'élargissement du champ de la lexicométrie, Mayaffre a proposé de substituer le radical « lexico », jugé comme trop réducteur au seul lexique, par le radical « logo », signifiant « discours », plus englobant :

« La lexicométrie de la seconde génération – que nous appellerons "logométrie" car elle ne se contente pas de traiter du lexique (lexi) pour étendre ses procédures à toutes les unités linguistiques jugées pertinentes du discours (logo*) : mots graphiques, lemmes, cooccurrents, codes grammaticaux, enchaînements syntaxiques, etc. – connaît aujourd'hui un essor certain »* (Mayaffre, 2007, p.153-154).

Avec la logométrie, on passe ainsi de l'analyse lexicométrique traditionnelle des formes graphiques et du lexique, à une analyse discursive étendue aux *« réalités linguistiques d'ordre grammatical, syntaxique, sémantique ou encore rhétorique qui composent, ensemble, le discours »* (Mayaffre, 2010, p.23). C'est notamment par la lemmatisation et l'étiquetage (morphosyntaxique) que la logométrie entend développer un traitement linguistique plus abouti.

Ainsi, le terme « logométrie » tend de plus en plus à rentrer dans l'usage. Comme l'indique Pierre Lafon : *«* […] *le vocable lexicométrie n'a plus semblé adéquat pour désigner la méthode et* […] *a été remplacé ici ou là, à juste titre, par textométrie ou logométrie »* (Lafon, (s. d.), p.1).

Les raisons de cette évolution terminologique et méthodologique sont multiples. Nous pouvons tout d'abord citer le développement technologique et informatique qui, par *«* [la] *puissance et* [la] *généralisation des machines domestiques »* (Mayaffre, 2007, p.154), *« les mémoires de masse illimitées et les vitesses de traitement vertigineuses ont radicalement bouleversé le paysage de l'informatique et facilité aux chercheurs la maîtrise des données textuelles »* (Lafon, (s. d.), p.1). Par ailleurs, l'amélioration des performances, de l'ergonomie et des capacités des logiciels (Mayaffre, 2005, p.1) a permis de travailler à partir de macro-corpus textuels de plusieurs millions d'occurrences, dépassant

ainsi les capacités mémorielles et calculatoires humaines. De plus, l'essor des logiciels d'annotation a permis d'attribuer aux occurrences d'un texte des informations d'ordre linguistique, aussi bien grammaticales, que morphologiques, syntaxiques ou encore sémantiques. Enfin, la disponibilité et le nombre exponentiel de textes ou de corpus numérisés rend l'analyse par l'outil informatique d'autant plus facile que les chercheurs ne doivent plus – ou en tout cas moins souvent – passer par l'étape assez rebutante de la saisie numérique des textes (Mayaffre, 2005, p.1). Ces derniers sont aujourd'hui disponibles sur la toile et peuvent être exploités, soit immédiatement, ce qui est assez rare, soit après un traitement informatique.

Quoi qu'il en soit, la logométrie est une méthode d'analyse du discours assistée par ordinateur permettant de traiter quantitativement et qualitativement les éléments linguistiques de grands corpus textuels. Mayaffre, à l'origine du terme, en donne sa définition :

« Ce que nous appelons Logométrie, c'est un ensemble de traitements documentaires et statistiques du texte qui ne s'interdit rien pour tout s'autoriser ; 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 finalement un traitement automatique global du texte dans toutes ses dimensions : graphiques, lemmatisées, grammaticalisées. L'analyse ainsi portera sur toutes les unités linguistiques, de la lettre aux isotopies, en passant par les n-grams, les mots, les lemmes, les codes grammaticaux, les bi-codes ou les enchaînements syntaxiques. » (Mayaffre, 2005, p.9)

De façon plus concise, Mayaffre définit également la logométrie comme *« une méthodologie, sinon une discipline, visant à prendre la mesure du discours en conjuguant approche qualitative et quantitative, sans jamais les séparer »* (Mayaffre, 2014, p.1). Il s'agit donc d'une méthode combinant l'approche lexicométrique traditionnelle, à travers notamment l'analyse des formes graphiques et des réseaux lexicaux, et l'approche textométrique, en s'orientant davantage vers la prise en

compte des traits significatifs des données textuelles au sein d'un grand corpus.

2. Les principes fondamentaux de la logométrie : entre continuité et innovation

Forte de son inscription historique, la logométrie a su tirer parti des éléments méthodologiques qui ont fait la renommée des méthodes qui l'ont précédée, mais s'est également dotée de nouveaux principes directeurs. En effet, développer et améliorer une méthode déjà existence nécessite forcément de procéder à des changements d'ordre épistémologiques et pratiques. C'est ce que nous abordons dans cette seconde partie, en nous intéressant, tout d'abord, à l'importance du dialogue entre les approches qualitatives et quantitatives dans la démarche logométrique, ainsi qu'en soulignant, dans un second temps, le renouvellement de la démarche interprétative dans le processus logométrique à travers une approche inductive.

2.1. Combiner les approches qualitatives et quantitatives

Le débat entre qualitatif et quantitatif n'est pas nouveau. En effet, comme le soulignent très justement Jules Duchastel et Danielle Laberge, il est important de retourner aux origines de la distinction entre approche quantitative et approche qualitative en rappelant la séparation établie par Wilhelm Dilthey à la fin du XIX^{ème} siècle entre sciences naturelles et sciences de l'esprit. Ainsi, *« cette séparation se caractérisait par une opposition ferme entre explication et compréhension. Selon cette conception, les sciences naturelles étaient toutes entières tournées vers l'identification de relations causales entre phénomènes, alors que les sciences humaines cherchaient à débusquer le sens de l'expérience vécue historiquement située »* (Duchastel & Laberge, 2014, p.6). C'est donc cette opposition entre, d'un côté, une approche froide, distante, objective et neutre menée par le chiffre, et d'un autre côté, une approche tournée vers l'interprétation et l'analyse en profondeur que reflète l'opposition paradigmatique entre approches quantitative et qualitative. Ces deux approches s'opposent donc sur deux plans : méthodologique et épistémologique. Sur le plan méthodologique, les études quantitatives ont recours au comptage, alors que les études qualitatives ne recourent pas aux chiffres. Sur le plan épistémologique, pour les études qualitatives, la

phase d'analyse et le processus interprétatif se font « à l'œil », « à la main », « à l'humain », pour reprendre les expressions de Marie-Anne Paveau (2014, p.2-3), tandis que pour les études quantitatives, ces étapes se font « au logiciel », « à l'instrument », pour reprendre l'expression de Habert et al. (2005).

Si l'approche qualitative a su se développer sans l'aide du quantitatif, l'inverse est moins sûr. En effet, le quantitatif ne peut exister seul ; il doit s'enrichir du qualitatif, car *« ce qui importe c'est d'accéder au sens du message à travers les nombres »* (Duchastel & Laberge, 2014, p.3). D'où l'importance d'avoir recours à une mixité des méthodes afin de dépasser l'opposition quantitatif/qualitatif en combinant les deux approches, car la langue, comme le disait déjà Guiraud en 1960, présente un aspect à la fois qualitatif et quantitatif : *« les signes constituent un système de formes porteuses de sens d'une part, d'autre part un système de combinaisons numériques »* (1960, p.25). Tout l'intérêt de la logométrie réside donc dans l'alliance entre le quantitatif et le qualitatif. Le traitement quantitatif du corpus est certes essentiel, mais la logométrie ne se limite pas à cela : il ne s'agit pas d'une « dictature par les chiffres ». En amont du traitement statistique, le qualitatif s'exprime dans le choix du sujet et de la problématique de recherche, ainsi que dans la construction du corpus selon des critères bien spécifiques. Le qualitatif s'exprime également lors du traitement statistique, lorsque le chercheur tente de trouver de nouveaux points d'entrée dans le corpus en fonction des données obtenues. Enfin, le qualitatif s'exprime dans la dernière phase, celle de l'analyse et de l'interprétation des sorties logicielles. Ainsi, quantitatif et qualitatif nourrissent le processus heuristique au cœur de la logométrie.

Par cet aller-retour entre le quantitatif et le qualitatif, la méthode logométrique oscille donc entre une vision macroscopique ou globale du corpus et une vision microscopique ou locale, entre une approche paradigmatique et une approche syntagmatique, entre décontextualisation et (re)contextualisation (Mayaffre 2005, p.92 ; Mayaffre, 2010, p.23, 35). Ainsi, *« la logométrie cherche à allier la rigueur mathématique à la posture philologique de l'analyse de texte. Elle cherche* […] *à concilier la métrique et le scriptural ; les chiffres et les mots »* (Mayaffre, 2007,

p.179). La logométrie propose donc une *« lecture alpha-numérique »* (Mayaffre, 2009) des textes et des corpus. L'un des grands principes de la logométrie est que *« le nombre fait sens »* pour reprendre la formule de Mayaffre. En effet, le nombre, sans que l'on n'y voue un culte, permet d'effectuer un déchiffrement quantitatif des saillances linguistiques du corpus et s'avère d'autant plus utile que les corpus s'agrandissent de plus en plus.

2.2. Renouveler la démarche interprétative

Comme nous venons de le voir, la méthode logométrique répond à une certaine rigueur en repérant de façon systématique tous les éléments linguistiques saillants, réguliers et récurrents du corpus. La phase interprétative vient ensuite. C'est à cette phase que nous allons nous intéresser ici.

Le texte et par extension le corpus textuel, à travers leur matérialité linguistique, constituent la porte d'entrée vers l'observation et l'interprétation. L'herméneutique, entendue ici au sens d'interprétation des textes et non en son sens philosophique d'interprétation du monde, *« est fondée sur un ensemble de procédures de description et d'analyse des unités matérielles du discours »* (Duchastel & Laberge, 2014, p.7). Ainsi, un texte ne pourrait être interprété sans une mise en sens. Comme le souligne Jean Grondin, *« les herméneutes contemporains ont beaucoup insisté sur l'idée que l'interprétation était une activité créatrice de sens »* (Grondin, 2004, p.6). La logométrie mène ainsi à une réflexion sur l'herméneutique. Comme le souligne Mayaffre (2010), cette réflexion trouve ses origines dans l'herméneutique littéraire de Peter Szondi, qui en donne sa définition : *« nous entendons sous le terme d'herméneutique littéraire une théorie de l'interprétation qui, sans être non philologique, réconcilie l'esthétique et l'apprentissage de l'interprétation »* (Szondi, 1989, p.18). L'herméneutique de Szondi, en tant que science de l'interprétation des textes, se veut donc à la fois « matérielle » et « critique » (Berner, 2013, p.31). Le discours devient ainsi l'objet de l'herméneutique. Étant donné que la logométrie a recours aux technologies numériques à travers des corpus numérisés et des logiciels de plus en plus performants, on parle désormais d'*« herméneutique numérique »* (Mayaffre, 2010), car *«* [elle] *s'appuie sur la forme pour*

espérer atteindre le fond ; elle étudie l'expression brute du texte pour accéder à son contenu ; elle appréhende le mot en espérant atteindre le sens » (Mayaffre, 2010, p.26).

Ainsi, comme le souligne l'auteur, cette entrée dans l'herméneutique se fait par le bas, c'est-à-dire par le corpus, par la matérialité des textes (Mayaffre, 2010, p.38). La logométrie se doit donc *«* [de] *proposer des chemins interprétatifs et non des tunnels ; elle doit proposer un parcours herméneutique ordonné et non un cercle qui confond le commencement et la fin »* (Mayaffre, 2010, p.26). L'objectif initial de la logométrie est donc de *« mettre en place un protocole de lecture pour baliser, au sein des corpus* […], *des parcours interprétatifs contrôlables »* (Mayaffre, 2010, p.23).

Pour se faire, la logométrie entend renouveler le processus interprétatif. En effet, alors que traditionnellement les sciences humaines et sociales ont recours à la démarche hypothético-déductive ou descendante (*« top-down »*), la logométrie tend à privilégier la démarche inductive ou ascendante (*« bottom-up »*) : on passe ainsi d'une approche fondée sur le corpus *(« corpus-based »*) à une approche guidée par le corpus (*« corpus-driven »*), terme forgé par Elena Tognini-Bonelli. De ce fait, *« c'est le texte – dans toutes ses unités et sans sélection ou censure – qui interroge le chercheur et non le chercheur – avec sa part d'aveuglement et de parti pris – qui interroge partiellement et partialement le texte »* (Mayaffre, 2007, p.174). Cette démarche positiviste-inductive permet d'éviter deux dangers posés par les questionnements a priori ou exogènes de la démarche hypothético-déductive, à savoir : d'une part, le risque de trouver uniquement ce que l'on cherche en essayant de prouver une hypothèse en se basant sur des données chiffrées ; d'autre part, le danger de passer à côté d'hypothèses de lecture qui seraient peut-être plus pertinentes que celles posées a priori (Mayaffre, 2002, p.158). Ainsi, l'intérêt majeur de la logométrie réside dans son approche exploratoire des grands corpus textuels.

Cette exploration s'effectue à travers les logiciels, qui renvoient des données, dont l'interprétation relève de l'analyste. Car, comme le souligne Thierry Guilbert, *« utilisé sans précautions méthodologiques et*

sans souci des présupposés épistémologiques des sciences du langage et de l'AD, le logiciel semble acquérir une sorte de vertu magique : il dirait de lui-même et par lui-même de quoi est fait le discours » (Guilbert, 2014, p.5). Or, les sorties logicielles n'ont rien de magique, ce ne sont pas des réponses ; elles relèvent d'un parcours interprétatif encadré et prennent la forme d'interrogations. Comme le précise très justement Mayaffre, *« l'outil informatique et les algorithmes statistiques de décryptage servent ici de médiation afin d'objectiver autant que faire se peut l'interprétation des textes »* (Mayaffre, 2007, p.171). Il s'agit donc de retarder autant que possible l'entrée dans la subjectivité inhérente à toute interprétation. Auparavant l'analyse des textes et des corpus se faisaient en deux temps : lecture puis interprétation. Avec la logométrie s'ajoute une troisième étape : lecture, traitement logométrique, interprétation contrôlée (Mayaffre, 2007, p.171-172). Les sciences du texte ont avant tout pour vocation de contrôler l'interprétation plutôt que de fournir des preuves. En effet, la statistique n'a rien de probatoire, mais elle ouvre la voie à des questionnements, à de nouvelles hypothèses. Le sens ne se prouve pas ; il s'interprète. La logométrie est là pour objectiver non pas le sens, mais les parcours interprétatifs de lecture.

Pour pouvoir mener ce processus interprétatif, il faudra passer par un retour au texte, par une (re)contextualisation des données obtenues, car, pour reprendre la formule chère à Rastier et reprise par Mayaffre, *« le sens naît en/du (con)texte »* (Mayaffre, 2007, p.179). En effet, comme le souligne Rastier, *« l'activité interprétative procède principalement par contextualisation »* (Rastier, 2001, p.92) et ce processus s'effectue aux différents paliers. Ainsi, *« tout mot doit être contextualisé dans sa phrase ; toute phrase doit être replacée dans son paragraphe ; tout paragraphe situé dans son texte »* (Mayaffre, 2007, p.179) et tout texte doit être rapporté au corpus. Ainsi, comme le souligne Thierry Guilbert, *« le retour au texte* […] *consiste à rapporter, grâce aux résultats obtenus par l'analyse quantitative, l'objet (le discours) à sa formation discursive »* (Guilbert, 2014, p.8). Car les mots ne prennent pleinement leur sens que lorsqu'ils sont recontextualisés dans leur formation discursive.

La logométrie procède ainsi à une déconstruction des textes, en effectuant un traitement statistique des traits saillants du discours, en

rapportant ensuite les données à leur co(n)texte d'origine, pour enfin redonner sens aux données en menant un processus interprétatif. Nous proposons donc de résumer la méthode logométrique par le schéma suivant :

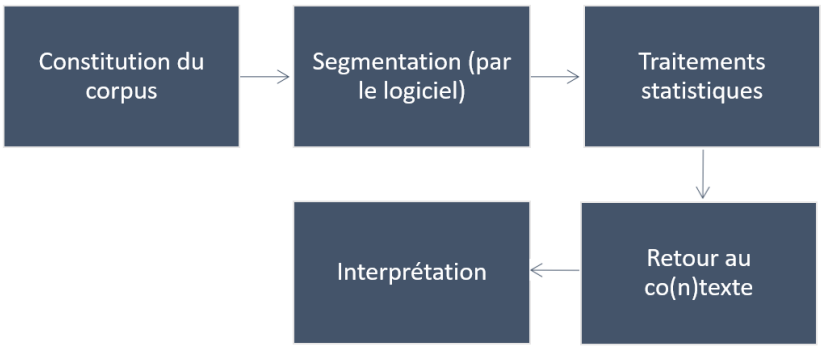

Figure 1 : Étapes de la méthode logométrique

Nous venons ainsi de retracer les origines de la méthode logométrique pour en décrire les éléments essentiels. Alors qu'auparavant la linguistique, et plus particulièrement l'AD, était considérée comme une « boîte à outils », il s'avère que, dans un renversement de situation, c'est désormais l'informatique, à travers la logométrie, qui sert d'outillage à l'AD. Par ailleurs, la méthode logométrique, en s'insiprant des méthodes qui l'ont précédée entend se doter de nouveaux principes directeurs. En effet, la logométrie aspire, d'une part, à combiner approches qualitatives et quantitatives et, d'autre part, à renouveler l'herméneutique traditionnelle en la « matérialisant » et en la « numérisant », privilégiant de ce fait une pratique interprétative inductive et émergentiste.

Il nous paraît donc intéressant de faire connaître cette méthode logométrique au plus grand nombre afin de l'intégrer dans les pratiques de recherche et dans les cursus universitaires en sciences humaines, notamment dans les sciences du langage, mais également dans les sciences politiques ou encore les sciences sociales. En effet, cette méthode constitue un enrichissement considérable tant la force heuristique de la logométrie, en reposant sur une pratique interprétative inductive et émergentiste, permet d'explorer les corpus en combinant les approches quantitatives et qualitatives. Le chercheur est ainsi guidé, non pas par ses a priori, mais par le corpus et les sorties logiciels, qui

constituent des points d'entrée dans le corpus et de nouvelles pistes exploratoires. La machine se trouve donc au service de l'humain, en traitant une quantité phénoménale d'occurrences en un temps record, mais l'humain continue de jouer un rôle fondamental à travers toutes les étapes dites qualitatives de la méthode logométrique.

Bibliographie :

BEAUDOUIN, Valérie (2000). Statistique textuelle : une approche empirique du sens à base d'analyse distributionnelle, *Texto !* [en ligne]. URL : http://www.revue-texto.net/Inedits/Beaudouin_Statistique.html>

BEAUDOUIN, Valérie (2016). Retour aux origines de la statistique textuelle : Benzécri et l'école française d'analyse des données, *Actes des JADT 2016*, p. 17-27.

BENZÉCRI, Jean-Paul et coll. (1981). *Pratiques de l'analyse des données. Linguistique et lexicologie*, Tome 3. Paris : Dunod, 565 p.

BERNER, Christian (2013). L'herméneutique dans son histoire. À propos de Peter Szondi, *Revue germanique internationale* [en ligne], n°17. URL : https://doi.org/10.4000/rgi.1373

BRUNET, Étienne (2014). La lexicométrie française : naissance, évolution et perspectives, *Revue de l'Université de Moncton*, vol. 45 (n°1-2), p. 13-33.

CHARAUDEAU, Patrick, MAINGUENEAU, Dominique (dir.) (2002), *Dictionnaire d'analyse du discours*, Paris : Seuil, 661 p.

CHATTI, Sarah (2022). *La question environnementale et les organisations internationales (1992-2018). Analyse logométrique des rapports d'activité du Programme des Nations unies pour l'environnement (PNUE), de la Banque mondiale, et de l'Organisation des Nations unies pour l'éducation, la science et la culture (UNESCO)*, Thèse de doctorat, Université libre de Bruxelles (ULB), Université Sorbonne Nouvelle (USN), 686 p.

CONEIN, Bernard, COURTINE, Jean-Jacques, GADET, Françoise, MARANDIN, Jean-Marie, PÊCHEUX, Michel (eds.) (1981), *Matérialités discursives*, Presses Universitaires de Lille, 216 p.

DUBOIS, Jean, DUBOIS-CHARLIER, Françoise (1970). Principes et méthode de l'analyse distributionnelle, *Langages*, n°20, p. 3-13.

DUCHASTEL, Jules, LABERGE Danielle (2014). Au-delà de l'opposition quantitatif/qualitatif. Convergence des opérations de la recherche en analyse du discours, *Corela* [en ligne], HS-15. URL : https://doi.org/10.4000/corela.3524

GRONDIN, Jean, (2004). Qu'est-ce que l'interprétation ?, *Philopsis* [en ligne]. URL : http://www.philopsis.fr/IMG/pdf/qu_est-ce_que_l_interpretation_.pdf

GUILBERT, Thierry (2014). Introduction : articuler les approches qualitatives et quantatives dans l'analyse de discours, *Corela* [en ligne], HS-15. URL : https://doi.org/10.4000/corela.3545

GUILHAUMOU, Jacques (1986). L'historien du discours et la lexicométrie. Étude d'une série chronologique : le « Père Duchesne » d'Hébert (Juillet 1793 - mars 1794), *Histoire & Mesure*, vol. 1, n°3-4, p. 27-46.

GUILHAUMOU, Jacques (1993). À propos de l'analyse de discours : les historiens et le « tournant linguistique », *Langage et société*, n°65, p. 5-38.

GUIRAUD, Pierre (1960). *Problèmes et méthodes de la statistique linguistique*. Paris : PUF.

HABERT, Benoît, NAZARENKO, Adeline, SALEM, André (1997). *Les linguistiques de corpus*, Paris : Armand Colin, 240 p.

LAFON, Pierre (s. d.). Statistique et lexicométrie : position des problèmes. URL : https://docplayer.fr/80784321-Statistique-et-lexicometrie-position-des-problemes.html (page consultée le 15/08/2019)

LEIMDORFER, François, SALEM André (1995). Usages de la lexicométrie en analyse de discours. Dans Daniel Barreteau (dir.), Traitement et emploi des langues : nouvelles techniques, nouvelles applications, *Cahiers des Sciences Humaines*, vol. 1, n°31, p. 131-143.

LÉON, Jacqueline (2010). AAD69 : archéologie d'une étrange machine, *Semen* [en ligne], n°29. URL : https://doi.org/10.4000/semen.8823

MAINGUENEAU, Dominique (2009). *Aborder la linguistique*. Paris : Points, coll. Points Essais, 177 p.

MAYAFFRE, Damon (2002). L'Herméneutique numérique, *L'Astrolabe*, Recherche littéraire et Informatique, p. 1-11.

MAYAFFRE, Damon (2005). De la lexicométrie à la logométrie, *Astrolabe*, p. 1-11.

MAYAFFRE, Damon (2007). Analyses logométriques et rhétorique du discours. Dans Stéphane Olivési (dir.), *Introduction à la recherche en SIC*, PUG, p. 153-180.

MAYAFFRE, Damon (2009). *L'analyse du discours assistée par ordinateur*. Formation, CNRS – UMR 6039 « Bases, corpus et Langage », Alexandrie.

MAYAFFRE, Damon (2010). *Vers une herméneutique matérielle numérique. Corpus textuels, Logométrie et Langage politique*, HDR, Histoire, Université Nice Sophia Antipolis.

MAYAFFRE, Damon (2014). « Ça suffit comme ça ! ». La fausse opposition quantitatif/qualitatif à l'épreuve du discours sarkozyste », *Corela* [en ligne], HS-15. URL : https://doi.org/10.4000/corela.3543

MAZIÈRE, Francine (2010). *L'analyse du discours. Histoire et pratiques*. Paris : PUF.

MOHARRAM, Sahar (s. d.). *Aperçu général de la linguistique* [document inédit], Université française d'Égypte.

NÉE, Émilie, (dir.) (2017). *Méthodes et outils informatiques pour l'analyse des discours*, Rennes : Presses universitaires de Rennes, coll. Didact méthodes, 250 p.

PAVEAU, Marie-Anne (2014). L'alternative quantitatif/qualitatif à l'épreuve des univers discursifs numériques, *Corela* [en ligne], HS-15. URL : https://doi.org/10.4000/corela.3598

PÊCHEUX, Michel, LÉON, Jacqueline, BONNAFOUS, Simone, MARANDIN, Jean-Marie (1982). Présentation de l'analyse automatique du discours (AAD69) : théories, procédures, résultats, perspectives, *Mots*, n°4, p. 95-123.

PINCEMIN, Bénédicte, HEIDEN, Serge (2008). *Qu'est-ce que la textométrie ? Présentation*, Site du projet Textométrie, URL : https://txm.gitpages.huma-num.fr/textometrie/Introduction/ (page consultée le 18/09/2019)

POUDAT, Céline, LANDRAGIN, Frédéric (2017). Explorer un corpus textuel. Méthodes - pratiques - outils, Louvain-la-Neuve : De Boeck Supérieur, 240 p.

RASTIER, François (2001). *Arts et sciences du texte*, Paris : PUF, 320 p.

RASTIER, François (2008). Que cachent les « données textuelles » ?, *Actes des JADT 2008*, p. 13-26.

RICŒUR, Paul (1986). Du texte à l'action, Essais d'herméneutique II, Paris : Seuil, 414 p.

RIZKALLAH, Élia (2013). L'analyse textuelle des discours assistée par ordinateur et les logiciels textométriques : réflexions critiques et prospectives à partir d'une

modélisation des procédés analytiques fondamentaux, *Cahiers de recherche sociologique*, n°54, p. 141-160.
SZONDI, Peter (1989). *Introduction à l'herméneutique littéraire* (traduit par Mayotte Bollack). Paris : Cerf, 154 p.
TOURNIER, Maurice (1969). Le centre de recherche de lexicologie politique de l'E.N.S. de Saint-Cloud, *Langue française*, n°2, Le lexique, p. 82-86.
Université de Liège, *Laboratoire d'Analyse Statistique des Langues Anciennes*. URL : http://web.philo.ulg.ac.be/lasla/ (page consultée le 17/01/2023)

[1] Centre de recherche et d'études pour la diffusion du français.
[2] MULLER, Charles (1964). *Essai de statistique lexicale, l'Illusion comique de P. Corneille*, Paris : Klincksieck. 204 p.
[3] MULLER, Charles (1967). *Essai de statistique lexicale, le vocabulaire du théâtre de P. Corneille*. Paris : Larousse, 380 p.
[4] Ces notions d'environnement et distribution, nous le verrons plus tard, seront fondamentales pour la logométrie.
[5] À l'époque, *Mots* signifiait *« "Mots-Ordinateurs-Textes-Sociétés" et non comme aujourd'hui "Mots / Les langages du politique" »* (Mazière, 2010 : 64).