Université de Damiette.
Faculté des lettres .
Département de français.

Les statistiques lexicales et les logiciels lexicométriques

الاحصاءات اللفظية و برمجيات الاحصاء اللفظى

بحث علمي مقدم لنيل درجة الماجستير في الآداب

قسم اللغة الفرنسية وآدابها تخصص اللغويات

إعداد الباحث

تامر عبد الرحمن عبد الرحمن الرفاعى
معلم أول لغة فرنسية بالازهر الشريف

إشراف:
تحت إشراف الدكتور/هاني جورج فانوس.
أستاذ مساعد اللغويات بقسم اللغة الفرنسية جامعة دمياط.

TABLE DES MATIÈRES

1. L'approche lexicométrique :

1.1. Qu'est-ce que la lexicométrie ?

On désigne sous le vocable « lexicométrie » la discipline qui prend en charge l'analyse informatisée du discours et du lexique.
Paul A. Fortier [94] soulève que, « *bien que les textes soient composés de mots, leurs effets sont produits par des phénomènes d'un ordre supérieur et plus complexe* ». En revanche, il est convenu que les données issues d'analyses informatiques sont utiles pour d'autres analyses plus fines.

Cette jeune discipline est appelée également « analyse du discours assistée par ordinateur », ou encore « traitement automatisé du discours ». Néanmoins, les appellations proposées pour désigner « 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, Statistique textuelle » (André Salem), « Logométrie » (Damon Mayaffre), « Statistique lexicale (ou lexicostatistique) », « Statistique linguistique », etc. G. Mounin (1974 : 203), la définit comme le « Domaine de la lexicologie dans lequel les procédures de la statistique sont utilisées pour l'étude quantitative du lexique.».
Selon Leimdorfer et Salem (1995 : 133), 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 ». Enfin, pour Maingueneau, (2009 : 81), il s'agit d'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. »

La lexicométrie, d'après (LEBART et SALEM, 1994 : 315) c'est
«*l'ensemble de méthodes permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sur le vocabulaire d'un corpus de textes.* ».
Dans le même sens, (DUBOIS et al., 2002 : 282) préfèrent le terme « lexicostatistique ». Ces auteurs définissent ainsi cette approche:

[94]FORTIER Paul A. (1995) Categories, Theory, and Words in Literary Texts, Research in Humanities Computing, n°5, pp. 91-109

« *la lexicostatistique est 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* ».

G.Herdan définit la lexicométrie comme « *la quantification de la théorie saussurienne du langage* ».[95]

la lexicométrie est définie par Lowe et Matthews [96] comme "*consistant à appliquer des méthodes mathématiques afin d'extraire des données quantitatives d'un texte. Les données étudiées par la stylométrie sont les mots.*"

En revanche, si diverses et si nombreuses que soient les dénominations et les définitions de la jeune discipline, il faut retenir qu'elle s'intéresse au discours à travers l'analyse de son lexique.
Pour ce faire, elle fait appel à des méthodes et celle-ci des outils mathématiques et informatiques : des logiciels de statistique lexicale. Ceux-ci sont si nombreux, les plus connus dans l'univers francophone sont Hyperbase d'Etienne Brunet et Lexico d'André Salem. Citons aussi : Cordial, Sphinx, Alceste, Tropes, Sato, Weblex, Prospero, Astartex, etc. Mais il ne faut pas perdre de vue que cette discipline ouvre la voie à d'autres analyses, qualitatives notamment.

L'originalité de cette approche réside dans le fait qu'elle permet d'effectuer des analyses dont l'entreprise était jusque-là impossible. En effet, la conjugaison à l'analyse traditionnelle d'une analyse automatique assistée par ordinateur a introduit une rupture totale dans le champ de l'analyse des données. Grâce à la mathématisation de la recherche linguistique, on peut travailler sur des corpus plus étendus et assurer une certaine objectivité des traitements.
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. Avant l'apparition de la statistique lexicale, se posait encore 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'oeuvre complète d'un auteur sans se soucier du problème de la représentativité du corpus puisque l'analyse prend en charge l'intégralité de l'oeuvre. En effet, comment garantir, lorsqu'il s'agit par exemple d'aborder l'oeuvre 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

[95] LEBART L. et SALEM A. (1994) *Statistique textuelle*, Dunod, Paris, 342 p.
[96]LOWE David et MATTHEWS Robert (1995) Shakespeare Vs. Fletcher: A Stylometric Analysis by Radial Basis Functions, Computers and the Humanities 29, pp. 449-461

auteur, des oeuvres peuvent présenter des dissemblances discursives? Avec l'avènement de l'approche lexicométrique, ces obstacles sont levés. L'analyse automatique est donc une autre manière, inhabituelle et originale d'appréhender les discours.

L'apparition de l'outil informatique et des chercheurs qui s'intéressent à la statistique textuelle a eu un impact positif sur l'évolution de l'analyse de discours en tant que pratique qui s'appuie davantage sur le progrès de l'informatique. « *Tout ce qui est dit ou écrit est susceptible d'être soumis à une analyse de contenu* »[97]

L'analyse des occurrences des mots d'un texte permet de donner plus de sens au texte étudié, de mieux le comprendre, de focaliser l'attention des élèves sur les mots-clés en vue d'une lecture attentive.

La lexicométrie, s'intéresse au recensement des mots qui composent et qui structurent des énoncés, et qui devient de plus en plus une méthode d'analyse très fréquente à *des unités linguistiques jugées pertinentes.* [98] Elle permet d'ériger des comparaisons entre corpus, d'investir dans des méthodes s'inspirant des mathématiques et des statistiques en vue de proposer un modèle d'analyse efficace pour l'appréhension des textes. D'après Lebart et Salem, cet intérêt pour la statistique s'expliquerait ainsi :
« Les succès remportés par les applications de la méthode statistique dans de nombreux domaines des sciences de la nature (physique, biologie, etc.) mais aussi dans ceux des sciences humaines (psychologie, économie, etc.) et y compris dans des disciplines qui touchent à l'utilisation du langage finissent par attirer l'attention des spécialistes de l'étude du vocabulaire ».[99]

Nous pouvons ainsi dire que l'attractivité de la lexicométrie consiste dans le fait que la langue (voire la linguistique) est traitée mathématiquement, c'est-à-dire la science humaine est étudiée par une science exacte.
Cette *« démarche (la lexicométrie) se veut scientifique. C'est-à-dire qu'elle vise à créer, à systématiser un ensemble de connaissances, d'études d'une valeur universelle caractérisées par un objet et une méthode déterminés, fondées sur des relations objectives, vérifiables. Née du besoin profond ressenti par les spécialistes de l'étude des textes de dépasser les approches traditionnelles, jugées souvent trop subjectives, elle se propose d'apporter sur les textes un éclairage nouveau fondé sur le décompte et la localisation des formes qu'ils contiennent. »*[100]

[97] HENRY Georges (1975) Comment mesurer la lisibilité, Labor, Bruxelles, 173 p.
[98] Mayaffre, 2009.
[99] Ludovic LEBART et André SALEM, Statistique textuelle, Paris, Dunod, 1994, p.16.
[100] André Salem, Analyse factorielle et lexicométrie : synthèse de quelques expériences, p.148.

1.2. Débat autour de la démarche lexicométrique

On a souvent reproché à la lexicométrie son caractère purement matériel. Ainsi, selon ses détracteurs, 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 vouloir rendre compte du sens de ce matériel. Il serait donc, sans utilité scientifique d'appréhender des textes dans un sens graphique, visuel, matériel et informatique.
Pour ses défenseurs, la lexicométrie constitue, au contraire, un outil heuristique d'une grande utilité ouvrant la voie à des analyses lexico-syntaxiques. En effet, grâce au développement de certains outils qui ne s'arrêtent pas à la matérialité du lexique d'un texte mais s'enfoncent davantage dans sa structure, la question de la pertinence des analyses et de la fiabilité des résultats ne se pose presque plus

selon ses détracteurs,Cette lexicométrie a mis en place aussi une génération de chercheurs qui ne s'intéressent qu'à l'aspect extérieur de la langue, en ignorant que cette méthode est beaucoup plus un outil de description statistique qu'un outil d'analyse, qui cède sa place à un bilan parfois bâclé. Aucune méthode, originale soit-elle, ne peut négliger l'apport des écoles et les courants linguistiques.

Le génie du chercheur réside dans l'analyse des faits linguistiques ; objets de la recherche. Les compétences d'un linguiste se dégagent et se manifestent au moment où ce dernier s'intéresse à ce qui fait un texte un texte, ce qui fait un énoncé un ensemble d'idées qui apparaissent en un mode structuré.

Comme écrit Jean-Marc Leblanc dans sa thèse doctorale : « *[…] la fréquence d'un mot est révélatrice des préoccupations, des thématiques, du style d'un auteur, d'un locuteur. […] Les constats quantitatifs sont ainsi souvent pris comme indices, fondés ou illusoires, et utilisés en abondance par les commentateurs et les acteurs de la vie politique, mais aussi par chaque locuteur, dans le quotidien de ses pratiques langagières. L'inventaire et le décompte des mots et des choses sont aussi vieux que la pratique écrite, réfléchie ou mobilisatrice, de la langue* ».[101]

[101] Jean-Marc LEBLANC, Les voeux présidentiels sous la Cinquième République (1959-2001), Recherches et expérimentations lexicométriques à propos de l'ethos dans un genre discursif rituel, thèse de doctorat, Université de Paris 12 Val-de-Marne, 2005, p. 9.

1.3. Origine et évolution de la discipline

Du fait que le but de notre recherche consiste en l´analyse lexicométrique nous allons expliquer la notion de lexicométrie et nous allons mettre en évidence son origine , l´évolution et les principes de la démarche.

Si l'on fait remonter les premiers balbutiements de l'analyse automatique du discours au milieu du XX^{e} siècle, il n'en demeure pas moins que beaucoup de savants, à travers l'Histoire, ont effectué des mesures sur des textes et en ont calculé les mots, surtout dans le but d'étudier le style de leurs auteurs. Mais la discipline telle qu'on la connait 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, on ne peut nier le fait que le développement de cette branche de recherche ait toujours été tributaire de l'évolution de l'informatique, laquelle lui fournit les outils statistiques nécessaires.

Ce domaine linguistique est une jeune discipline scientifique dont les bases ne remontent qu´à la moitié du 20^{e} siècle. Comme le dit Philippe Galiana dans son article *« La lexicologie est l´étude scientifique du vocabulaire d´un texte. Lorsque cette étude scientifique d´un texte est faite avec l´outil informatique, on parle alors de lexicométrie. »*[102]

Au XXe siècle, les premiers projets d'analyse automatique du discours sont dus notamment à l'italien Roberto Busa [103] qui fut le premier à utiliser l'ordinateur dans l'analyse linguistique et littéraire (Busa, 1998).

L´origine de cette discipline est liée aux noms de Georges Kingsley Zipf (1902-1950) et Georges Udny Yule (1871-1951). Le premier nommé était le linguiste américain qui a étudié la statistique appliquée aux différentes langues et qui est l´auteur de la loi de Zipf expliquant la fréquence des mots dans un texte. Dans les années 1930, c'est-à-dire bien avant l'apparition des ordinateurs, le linguiste George Zipf a remarqué quelque chose d'étrange quant à la frequence à laquelle les gens utilisent les mots dans une langue donnée. Il a

[102] Philippe Galiana, Lexicométrie, Le bulletin de L´EPI (Association Enseignement Public& Informatique), n°63, 1991, p.111.

[103] Dès 1946 Roberto Busa perçut l'intérêt de l'informatique pour son projet d'*Index thomisticus*. En 1949, il rencontra Thomas J. Watson, fondateur d'IBM, et le persuada de commanditer la réalisation de ce projet. Ce projet s'est étendu sur 30 ans et a débouché sur la production de 56 volumes vers la fin des années 1970. En 1989, ce matériel a été porté sur CD-ROM. En 2005, le matériel a migré sur le Web grâce au financement de la Fundación Tomás de Aquino et du CAEL. Les aspects techniques du projet ont été pris en charge par E. Alarcón et E. Bernot, en collaboration avec Busa. https://fr.wikipedia.org/wiki/Roberto_Busa [Consulté le 24 janvier 2019]

constaté qu'un petit nombre de mots sont utilisés tout le temps, alors que la grande majorité est utilisée très rarement. Il a remarqué en classant les mots dans l'ordre de popularité que Le mot classé numéro un a toujours été utilisé deux fois plus souvent que le mot de deuxième rang et trois fois plus souvent que le troisième rang. Plus précisément, *« G.K. Zipf a montré qu'en classant les mots d'un texte par fréquence décroissante, alors, on observe que la fréquence d'utilisation d'un mot est inversement proportionnel à son rang. La loi de Zipf stipule que la fréquence du second mot le plus fréquent est la moitié de celle du premier, la fréquence du troisième mot le plus fréquent, son tiers, etc. »*[104]
Cette loi peut être écrite par la formule suivante :

$$r \times f = constante$$ [105]

La loi de Zipf est mise en pratique par la fonction Pareto dont dispose le logiciel Lexico 3 et elle nous sera utile lors de l´analyse de la richesse du vocabulaire contenu dans le manuel étudié .

En ce qui concerne George Udny Yule, statisticien écossais, il faut constater que ce personnage est également considéré comme le pionnier de la statistique moderne (soit de la statistique mathématique soit de la linguistique). De ses ouvrages, nous pouvons mentionner *Introduction to the Theory of Statistics, The Statistical Study of Literary vocabulary.* qui connut quatorze éditions.

L'autre personnage important qui contribue à l´évolution dans le domaine de la statistique lexicale est le linguiste français Pierre Guiraud (1912-1983). Dans son livre, LA STYLISTIQUE, il aborde la thématique de la stylistique et il met en évidence l´apport de la coopération de la stylistique et de la statistique. Il cite d´anciens linguistes selon lesquels le style est un écart par rapport à une norme, ainsi selon Guiraud la statistique est « *précisément la science des écarts; la méthode qui permet de les observer, de les mesurer et de les interpréter.* » [106]
Évidement, la statistique lexicale est un instrument à la fois nécessaire et efficace dans l´étude du style. Depuis cette époque, la statistique lexicale a beaucoup évoluée.

Jusqu'aux années 1960, les travaux étaient peu nombreux et menés par un nombre très restreint de chercheurs épars. En 1962, est créé le Centre de linguistique quantitative de Paris. Il donnera naissance peu après à la revue Langages qui fut, des décennies durant, un lieu de rencontre important pour les linguistes et les analystes du discours. Dans les années 1970 et 1980, une amélioration nette est à signaler. Les recherches ont fait un bond en avant grâce

[104] Encyclopaedia Universalis. *La loi de Zipf.* http://www.encyclopaedia-universalis.fr/ [Consulté le 25 janvier 2019]
[105] **r** réprésente le rang et la **f** la fréquence du mot donné.
[106] Pierre Guiraud, La Stylistique, Paris, Presses Universitaires de France, 1961, p.107.

à l'apparition du micro-ordinateur au début de la décennie 1970, car il faut savoir qu'au début, le matériel était très dispendieux, donc difficile à acquérir.

En France, les précurseurs sont notamment Charles Muller , Maurice Tournier et Pierre Guiraud. Le premier nommé linguiste a dirigé ses recherches et travaux surtout dans la discipline de la linguistique quantitative *(Principes et méthodes de statistique lexicale, Initiation aux méthodes de la statistique linguistique)*. Charles Muller a exploité dans sa thèse le dépouillement complet du théâtre de Corneille, Selon ses propres mots, la linguistique quantitative signifie que :

« *... l´objet étudié comporte certains caractères quantifiables, et que l´on juge bon d´isoler certains de ces caractères pour les soumettre aux opérations statistiques.* » [107] Les caractères peuvent être soit quantitatifs (par exemple âge, taille, poids), soit qualitatifs (sexe, couleur des yeux, des cheveux, profession, l´origine géographique ou sociale ...).

Dans le domaine de la quantification du langage, les caractères quantitatifs sont : la longueur du texte, le nombre de mots, de syllabes ou de phonèmes. Dans la catégorie des caractères qualitatifs appartiennent : catégories grammaticales ou sémantiques, différents niveau du discours, etc. Ce qui est important pour une analyse statistique c´est la stabilité des données qui doivent être traitées avec la responsabilité et l´attention et il ne faut pas déduire une conclusion hâtive et non vérifiable.

Il s´agit également de Maurice Tournier, linguiste et chercheur au CNRS, [108] directeur du Laboratoire de lexicométrie politique de CNRS-ENS de Saint-Cloud et de la revue *Mots. Les langages du politique*, qui a contribué au développement sur le plan de la définition de la linguistique quantitative.
Cette branche de la linguistique fait partie de la linguistique mathématique.
Le but de la linguistique quantitative consiste, selon Maurice Tournier,
en effort de donner un aspect scientifique aux hypothèses concernant le langage.
Ainsi, cette discipline prétend à devenir une discipline empirique qui pousse avant la connaissance de la langue à l´aide des théories vérifiables.

Quant à Pierre Guiraud,il développa une approche scientifique qui prend en charge l'examen du style d'un auteur en mettant en oeuvre une méthode statistique.

[107] Charles Muller, *Initiation aux méthodes de la statistique linguistique*, Paris, Classique Hachette,1973, p.5.

[108] Centre National de la Recherche Scientifique. http://www.cnrs.fr/ [Consulté le 02 février 2019]

Michel Pêcheux est lui aussi un des promoteurs de la discipline avec sa fameuse *Analyse Automatique du Discours 1969*. Le choix de Pêcheux pour l'automatisation de l'analyse du discours était motivé par
sa préoccupation de défendre les Sciences humaines contre ce qu'on appelle les « Sciences dures ».

A l'aide des recherches de Jean Dubois, il est possible de rendre compte qu'au cas où les statisticiens confortent les unités lexicales qui comportent plus d'un élément, ils préfèrent y voir une seule unité. A la place de traiter différemment au garçon et à la femme on peut traiter chacun de ces deux formes présentant une occurrence de l'article (le ou la). Il est à noter qu'en français comme dans les autres langues ; si deux linguistes dépouillent le même texte, ils ne peuvent pas trouver le même résultat et même si les linguistes dépouillent les textes différents, il est impossible de comparer les résultats.

Il faut retenir aussi et surtout le nom d'Etienne Brunet de l'Université de Nice qui a contribué fortement au développement de la démarche.
Ses recherches portèrent sur le lexique des grands auteurs classiques de la littérature française. On lui doit surtout le très remarquable logiciel Hyperbase.

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. C'est dire que l'équipement utilisé était peu développé et n'avait rien à voir avec les outils très sophistiqués de ces dernières années.

De nos jours, les obstacles auxquels la jeune discipline a fait face pendant les premières décennies de sa vie ont été en partie levés grâce à la disponibilité de bases de données contenant des textes numérisés en grandes quantités, disponibles sur le web. Il faut savoir que la numérisation des corpus prenait une grande partie du temps du chercheur puisqu'elle se faisait manuellement.

De la même manière, la disponibilité de logiciels assez perfectionnés a rendu possible le redémarrage de la discipline et renforcé sa fiabilité scientifique. La décennie 1990 est marquée par deux moments saillants: la publication en 1994 par Ludovic Lebart et André Salem de leur ouvrage intitulé Statistique textuelle et la création en 1997 par André Salem de la revue Lexicometrica.

Au début du XXIe siècle, on retient surtout le nom de Damon Mayaffre. Celui-ci propose une refonte de la discipline par une évolution de la lexicométrie traditionnelle vers une « logométrie » qui croise différents niveaux de traitements documentaires et statistiques du texte.

La logométrie étend son analyse à d'autres unités du discours en plus des formes graphiques. Elle prend en compte alors d'autres unités linguistiques du corpus analysé : les lemmes (formes canoniques des mots dans le dictionnaire : l'infinitif pour les verbes, le masculin et le singulier pour les noms et les adjectifs, etc.) ; les structures grammaticales; sémantiques et rhétoriques, etc.

Pour conclure, nous aimerions résumer et mettre en relief l´apport de la statistique linguistique ou bien de la lexicométrie aux linguistes. Évidemment, le traitement automatique des langues permet aux chercheurs de mettre en évidence, de la manière rapide et efficace, des données statistiques concernant les informations morphologiques, syntaxiques et lexicales. *« ... ces techniques, à travers une analyse plus fine des documents et requêtes prenant en compte différents niveaux de la langues, permettent d´extraire des informations plus riches... Ces connaissances favorisent une meilleure représentation du contenu informationnel et du besoin de l´utilisateur. »*[109]

[109] Fabienne Moreau, Pascale Sébillot, Contributions des techniques du traitement automatique des langues à la recherche d'information, INRIA (Institut National de Recherche en Informatique et en Automatique), Rennes, 2005, p. 29.

1.4. Notions de Lexicométrie

Voici quelques notions qu'il nous a semblé important de définir ici pour éviter toute ambiguïté concernant le sens que nous donnerons à ces concepts : [110]

Toute suite de lettres ou de caractères non séparées par un blanc et délimitée par deux blanc sera appelée une *forme* ou un *vocable.*[111]

Quand une forme revient N fois dans un texte, nous dirons qu'elle possède N *occurrences*. "*Toutes les fois qu'un élément linguistique (type) figure dans un texte, on parle d'occurrence (token). L'apparition du terme politique dans un texte analyse du point de vue linguistique sera une occurrence du mot politique.*"[112]

Par exemple, dans la phrase suivantes *:*
La cuisine traditionnelle e suivantes istiquerence dirons quce dirons qulle ée par deux blanc sera a.[113]
Cette phrase est composée de 17 occurrences, les formes *de* et *ses* ont chacune 2 occurrences :
La cuisine traditionnelle égyptienne est connue par le goût unique de ses plats et de ses pâtisseries.

On appelle donc *"occurrence d'un élément linguistique dans un corpus donné une instance d'utilisation de cet élément dans le corpus en question*."[114]

Un vocable qui est employé une seule fois dans tout le texte est dit un *hapax* ou bien le legomenon. *« ... la fréquence est égale à un dans le corpus (hapax du corpus) ou dans une de ses parties (hapax de la partie)*[115].
Dans l'exemple précédent, en dehors de *ses et de* toutes les formes sont des hapax.

Selon dictionnaire de linguistique Larousse "*On donne le nom d'hapax a une forme, un mot ou une expression dont il ne se rencontre qu'une occurrence dans un corpus donne, une oeuvre.*"[116] Le contraire de l´hapax, en d´autres

[110] http://www.tal.univ-paris3.fr/lexico/Lexico3doc0.pdf
[111] pour plus d'informations, vous pouvez consulter le premier chapitre.
[112] DICTIONNAIRE DE linguistique Larousse, 1994,333
[113] *Club@dos PLUS 3*,p 35
[114] POLGUÈRE Alain, *Notions de base en lexicologie*, Montréal: Université de Montréal, OLST, 2002,p. 82
[115] Ludovic LEBART et André SALEM, Statistique textuelle, Paris, Dunod, 1994, 314
[116] DICTIONNAIRE DE linguistique Larousse, 1994,230

termes le mot qui apparaît le plus fréquemment, dispose de la fréquence maximale.

L'ensemble des formes d'un texte est appelé *vocabulaire* de ce texte. Notre exemple comporte 15 formes. Les formes *de et ses* ne sont comptées qu'une seule fois.

La	*cuisine*	*traditionnelle*
égyptienne	*est*	*connue*
par	*le*	*goût*
unique	*de*	*ses*
plats	*et*	*pâtisseries*

Le nombre d'occurrences qui composent un texte est appelé la *taille* ou l'*étendue* ou la *longueur* de ce texte. La phrase de notre exemple possède une taille de 17 occurrences.

Grâce aux logiciels lexicométriques *« ces formes sont examinées sous le seul angle de leurs fréquences, de leurs co-fréquences, de leur répartition, de leurs localisations et distances dans le texte, c'est-à-dire de leurs caractéristiques mesurables. »* [117] Si nous envisageons d´identifier un lexème qui apparaît le plus fréquemment ou, au contraire, celui qui ne possède qu´une seule occurrence dans le texte, il est nécessaire de le compter et de marquer sa fréquence pour pouvoir confirmer ou révoquer, à l´aide des résultats mesurables et vérifiables, nos hypothèses.

La fréquence des mots est déterminée par le nombre de ses occurrences dans le texte. Puisque la fréquence forme la base de la statistique lexicale, nous allons mentionner plusieurs types de la fréquence avec lesquels nous pouvons nous rencontrer lors de la recherche.

Premièrement, nous nous concentrons sur l´explication de la distinction entre la fréquence absolue et relative.

La fréquence absolue, qui est parfois désignée un effectif, exprime le nombre total d´ occurrences contenues dans le corpus étudié ou dans une partie sélectionnée. Tandis que *la fréquence relative* met en évidence la fréquence rapportée à la taille du corpus. C´est-à-dire qu´il faut calculer le pourcentage des occurrences par rapport à la longueur du texte. Par conséquent, la fréquence *relative « ... se traduit alors par une expression numérique qui peut varier de 0*

[117]Maurice Tournier, D'où viennent les fréquences de vocabulaire ? La lexicométrie et ses modèles. In: Mots, Institut de la langue française, Saint Cloud, 1980, N°1. p. 190.
http://www.persee.fr/web/revues/home/prescript/article/mots_0243-6450_1980_num_1_1_1010
[Consulté le 03 Avril 2019]

à 1, ces bornes incluses ... » [118].Dans notre exemple la forme *de* a une fréquence absolue égale à 2 et une fréquence relative égale à 2 / 17.
Au cas où le but de la recherche résiderait dans la comparaison des deux ou plusieurs textes d´une longueur différente, il est plus pertinent de dénombrer la fréquence relative.

Toute suite de formes qui se répète dans un texte est appelé *segment répété***.** donc, segment répété "*suite de forme dont la fréquence est supérieure ou égale à 2 dans le corpus.*"[119] (voir la figure suivante)

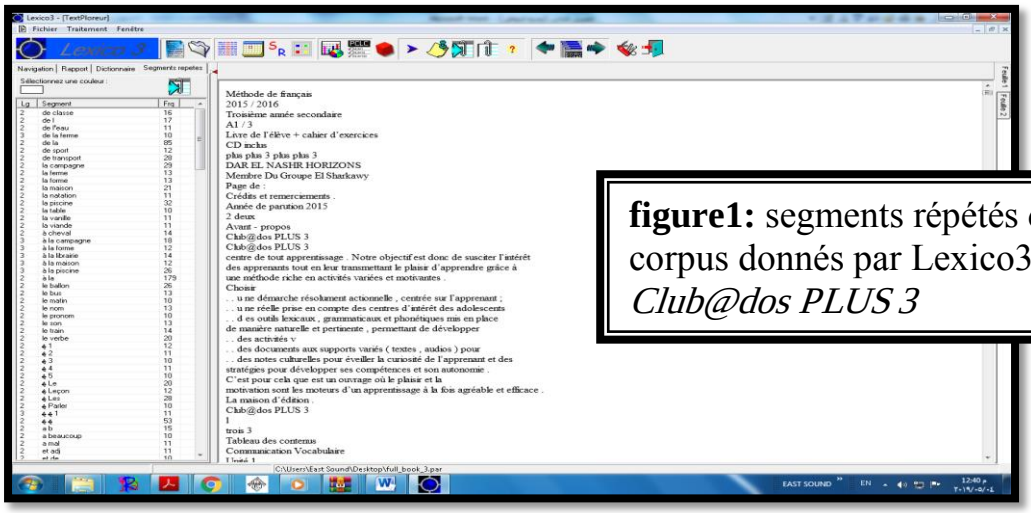

figure1: segments répétés de notre corpus donnés par Lexico3 *Club@dos PLUS 3*

L'index ou *Le dictionnaire* d'un texte est la liste des formes qui ont été employées dans ce texte. Elles se sont présentées généralement sous la forme d'un tableau à deux colonnes, la première contenant les vocables, la seconde, la fréquence absolue de chacun d'eux. " *Un index est, dans sa forme la plus standard, une table où tous les signifiants lexicaux du corpus sont énumérés, généralement accompagnés de leur nombre d'occurrences*."[120]
L'index contient aussi les signes de ponctuations ou autres caractères utilisés dans le texte accompagné et leur fréquence.

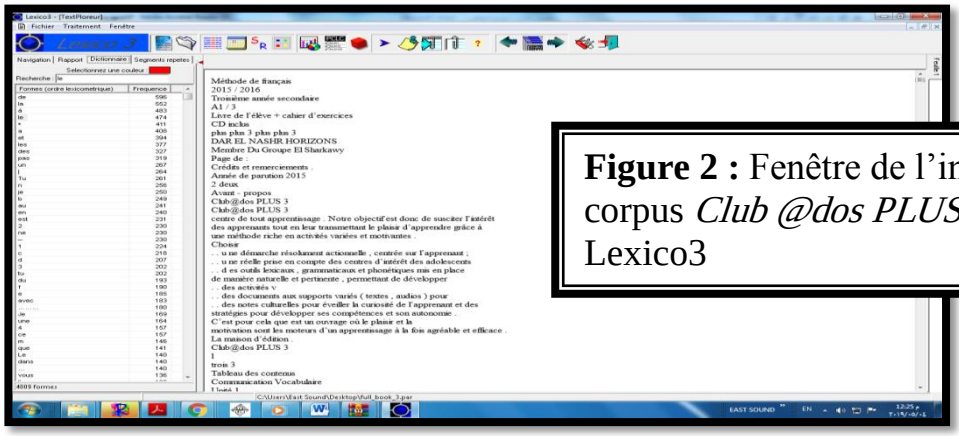

Figure 2 : Fenêtre de l'index de notre corpus *Club @dos PLUS 3* fournie par Lexico3

[118] Charles Muller, Principes et méthode de statistique lexicale, Paris, Honoré Champion Editeur‹ 1992, p. 47.
[119]http://www.tal.univ-paris3.fr/lexico/Lexico3doc0.pdf P.44
120 POLGUÈRE Alain, *Notions de base en lexicologie*, Montréal: Université de Montréal, OLST, 2002,p. 83

Une fois que le logiciel lexicométrie a terminé son travail de comptage de chacune des formes du texte nous obtenons des index et des concordanciers.

Un index n'est ni plus ni moins que la liste des formes du texte accompagnées d'un nombre représentant la fréquence d'apparition dans le texte.
« Les programmes statistiques proprement dits comparent la répartition des formes entre parties du corpus. Ils permettent d'établir d'une part diverses propriétés quantitatives générales des corpus, comme la répartition des fréquences relatives, ou la richesse lexicale des différentes parties. » [121]

Alors que l'index est une liste hors contexte, le concordancier, donne lui, les différents contextes dans lesquels on a utilisé une forme.En ce qui concerne par exemple une forme comme *classe* [122] , il est primordial de pourvoir la situer dans son contexte afin de la désambiguïser. Les index sont généralement de deux types, les index alphabétiques et les index hiérarchiques.

Les index et les concordanciers peuvent à leur tour donner lieu à des calculs statistiques et à des représentations graphiques.
comme exemple la forme dîner , le concordancier de cette forme, donné par le logiciel lexicométrie nous permet de déterminer dans quel cas du texte il s'agit du substantif " *Tu peux m'aider demain à préparer un dîner à mes amis?* " [123]
et dans quel autre cas du texte il s'agit du verbe *"dîner chez lui*"[124] .
C'est la désambiguïsation. Tous les contextes donnant lieu à une utilisation en tant que substantif sont alors répertoriés dans une base de données (hors de lexicométrie) et tous les contextes donnant lieu à une utilisation en tant que verbe sont répertoriés dans un autre fichier de la base de données. De là, peuvent être tirées des statistiques pour chaque emploi de la forme étudiée.

Les index et les concordanciers peuvent être globaux (portant sur tout le texte) ou sélectifs (portant sur un seul personnage ou une seule forme).
Ils peuvent nous faire apprécier l'importance d'un personnage parrapport aux autres personnages ou par rapport à la totalité de l'oeuvre.
Un personnage par exemple qui a lui seul représenterait 25% des formes de l'oeuvre est, sans nul doute un personnage qu'il vaudrait mieux étudier de près.

---

[121] Pierre Fiala, *L'interprétation en lexicométrie. Une approche quantitative des données lexicales,* Langue
française, ENS- Fontenay/Saint Cloud, 1994, p. 118.
[122] "chez une camarade de classe" *club @dos plus 1* p.106 classe ici un nom signifie un établissement scolaire
"classe-les dans un tableau" *club @dos plus 2* p.14 classe ici un verbe signifie mettre
"toutes les classes sociales" *club @dos plus 3* p.35 classe ici un nom signifie Ensemble d'individus défini en fonction d'un critère historique, sociologique, politique etc.
[123] *club @dos plus 3* p.22
[124]*club @dos plus 3* p.105

Les index permettent de confronter la richesse du vocabulaire du personnage A avec la richesse du vocabulaire du personnage B. De même, en examinant de près le vocabulaire d'un personnage, on peut tirer des renseignements intéressants sur son niveau socio-culturel.
Avec les index, nous pouvons aussi nous intéresser à l'étude des temps et des personnes verbales.
Bref. Il serait trop long ici de détailler ce qu'on fait avec ces index.

L'outil *Groupe de formes* permet de constituer des *types* rassemblant les occurrences de formes graphiques différentes liées par une propriété commune. On peut ainsi, moyennant certaine précaution, rassembler le pluriel et le singulier d'une même forme, les flexions d'un même verbe, des formes qui possèdent un lien sémantique.

1.4.1. Mesure de la richesse lexicale d'un corpus

En effet «*La statistique lexicale permet de résoudre, entre autres choses, des questions stylistiques sur « la richesse » objective d'un vocabulaire ; sur les oppositions stylistiques à l'intérieur d'un même texte ; sur les variations stylistiques chez un même écrivain ; sur l'individualisation lexicale des personnes que l'auteur fait parler ; sur la distance qui sépare deux oeuvres d'un même auteur ou même de deux auteurs différents etc.*»[125]

En effet *« Les programmes statistiques permettent d'établir d'une part diverses propriétés quantitatives générales des corpus, comme la répartition des fréquences relatives, ou la richesse lexicale des différentes parties. »* [126]

Si un corpus se réduit à un simple texte de longueur moyenne, on peut bien entendu en répertorier directement tout le vocabulaire. Cependant, la situation est rarement aussi simple. On peut notamment se retrouver devant un des trois cas de figure suivants :

1. Il arrive fréquemment que l'on veuille examiner le vocabulaire de corpus très vastes, ou même, de corpus dont la taille n'est pas fixe et qui continuent de croître. On peut donc être forcé de n'étudier en détail qu'une partie d'un corpus.

2. On peut vouloir déterminer à l'avance qu'elle devrait être la taille d'un corpus que l'on compte développer pour mener des études linguistiques d'un type donné Lexicologie et sémantique lexicale.

[125] » (Dubois et alii, 2002 :441).
[126] Pierre Fiala, *L'interprétation en lexicométrie. Une approche quantitative des données lexicales,* Langue française, ENS- Fontenay/Saint Cloud, 1994, p. 118.

3. On peut chercher à savoir quelle est la valeur d'un corpus dont on dispose, si on compte l'utiliser pour faire observations portant sur la langue en générale.

Chacun des trois cas qui viennent d'être mentionnés correspond à une situation où l'on doit être capable d'évaluer la représentativité linguistique de corpus Pour parvenir à faire ce type d'évaluation, on a été amené à examiner quel était l'accroissement du vocabulaire d'un corpus en fonction de l'accroissement de sa taille. Cela a permis de faire une série d'observations fort intéressantes. Notamment, on a constaté que cet accroissement présente la courbe caractéristique suivante[127], dans le cas de corpus relativement homogènes :

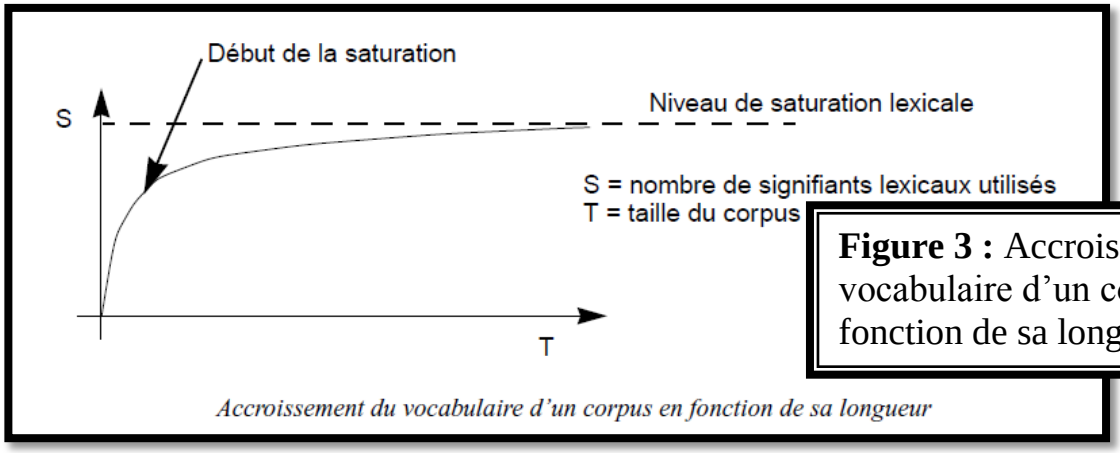

Figure 3 : Accroissement du vocabulaire d'un corpus en fonction de sa longueur

On considère ici que T, la taille du corpus, est mesurée en terme de nombre d'occurrences de signifiants lexicaux dans le corpus.

2.5. Logiciels lexicométriques

La lexicométrie a rendu l'analyse de discours accessible par le biais de différents outils et logiciels qui puisent davantage dans les progrès actuels dans le domaine des sciences informatiques.

Un logiciel de lexicométrie, comme celui du même nom de chez NATHAN[128],"*est un programme informatique plurilinguistique et pluridisciplinaire qui permet bien sûr de saisir un texte (le module de traitement de texte de lexicométrie étant assez peu puissant, nous lui préférons l'intégré WORKS dont nous utilisons également la base de données.), mais surtout de faire le décompte des formes (ou mots) employés dans ce texte. Pour chaque forme, le logiciel ne fait que compter le nombre de fois qu'elle apparaît.* "

[127] POLGUÈRE Alain, *Notions de base en lexicologie*, Montréal: Université de Montréal, OLST, 2002, p.89

[128]LEXICOMETRIE distribué par Cedic/Nathan et le C.A.R.F.I de Versailles. EGA/CGA Hercule

Le même travail pourrait se faire manuellement, mais il serait fastidieux voire impossible pour de longs textes ou de gros livres. Une fois les comptes terminés, l'utilisateur aura une somme d'informations avec laquelle il pourra au mieux étayer des hypothèses et diriger ses recherches.
Lorsqu'il s'agit de traiter de gros ouvrages, la saisie au clavier est hors de question et, dans ce cas, nous faisons appel au scanner.

Pour pouvoir bien traiter un corpus il faut avoir à la disposition un ou plusieurs logiciels qui possèdent les fonctions permettant une profonde analyse du corpus. L'utilisateur peut choisir de différents outils selon leur fonction et selon le but qu'il veut atteindre.Les logiciels sont eux-mêmes très nombreux, fruits d'un développement maintenant pluri-décennal. La principale difficulté peut-être pour les chercheurs est aujourd'hui la diversité des offres logicielles : trop nombreuses, elles finissent par décourager l'utilisateur qui ne sait quels outils choisir , Nous n'en mentionnerons que cinq [129] et sans doute ces cinq logiciels sont les plus célèbres dans le domaine de la lexicométrie et on les mentionne pour aider les autres chercheurs qui souhaitent traiter un corpus par un logiciel lexicométrique.

⊙ Quelques logiciels de lexicométrie

A. HYPERBASE : est un logiciel d'exploitation documentaire et statistique pour la création et L'exploitation de bases hypertextuelles crée par Étienne Brunet en 1989. [130] , à l'Université de Nice Sophia-Antipolis, HYPERBASE apparaît comme un logiciel complet. D'une conception ergonomique classique, il est sur le marché le logiciel qui s'articule le mieux aux deux lemmatiseurs sus-présentés : du texte brut aux sorties statistiques logométriques, l'utilisateur d'HYPERBASE se trouve pris par la main, pour un fonctionnement autonome. De plus, le logiciel offre une panoplie d'outils statistiques impressionnante expérimentée sur des corpus littéraires comme socio-politiques. Particulièrement adaptée au public non informaticien, l'utilisation d'HYPERBASE essaye d'être intuitive, tendue vers la description-interprétation des corpus textuels SHS.
L'intégration des données dans la base lexicométrique nécessite leur préparation, une mise en forme propre à chaque logiciel. Sous HYPERBASE, elle consiste à découper le corpus rassemblé au sein d'un fichier texte unique à l'aide de balises du type &&&text1&&&, propres à chaque niveau de découpage (texte, page, paragraphe). Un seul découpage des textes est possible

[129] d'autres logiciels à l'identité bien formée : ASTARTEXT, DTM, SATO, XAIRA, IRAMuTeQ,TXM, Sphinx Lexica, Alceste etc
[130] Étienne Brunet : est un linguiste français. Docteur d'État, il a été professeur de l'université de Nice Sophia Antipolis. Pionnier de la linguistique informatique et de la statistique textuelle française.

au sein du corpus, limité à 76 textes ; il faut créer une nouvelle base si l'on veut l'aborder avec un découpage différent [131] . HYPERBASE ne se préoccupe pas de la casse pour le découpage automatique des formes.
Le travail de création de la base est sensiblement plus long sous Hyperbase et monopolise l'activité sur le poste. Le logiciel effectue, dès ce moment, la quasi-intégralité des travaux et calculs (création du tableau lexical entier, spécificités, distance lexicale, etc.), à l'exception des AFC, ce qui rend le fichier exécutable assez lourd à transporter (5,8 Mo dans le cas du fichier saintjust.exe, contre 248 ko pour le fichier saintjust.num généré par Lexico3).
Lexico 3 est plus rapide, mais les calculs sont renvoyés à plus tard.
Ses interfaces sont en français mais il s'applique à toute langue qui utilise l'alphabet latin.

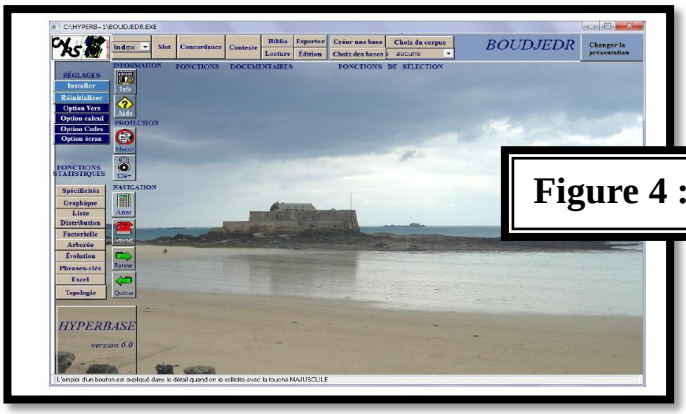

Figure 4 : l'interface d'Hyperbase

B. WEBLEX : Conçu par Serge Heiden, de l'ENS Lettres de Lyon, Weblex dispose sans doute du moteur de recherche le plus performant, permettant de faire des requêtes complexes (recherche d'expressions régulières, croisement des niveaux linguistiques, choix de l'empan ou de la fenêtre de recherche). Le traitement statistique est poussé, notamment dans le calcul et la représentation des co-occurrences (lexicogramme), mais WEBLEX n'offre que peu de possibilités de traitements synthétiques (AFC, ACP, Analyse arborée). Son principal inconvénient est de ne pas permettre à l'utilisateur de créer ses propres bases de textes sans passer par le concepteur même du logiciel.

C. LE CORDIAL : Il est né d'un correcteur orthographique intégrée en 1988 dans le logiciel de traitement de texte qui se mue en correcteur grammatical ; le mot cordial est un acronyme pour « CORrecteur D'Imprécision et analyseur Lexico-Syntaxique»

131 Le découpage des textes au sein du corpus est parfois quasi « naturel », il préexiste au travail de l'historien. C'est le cas ici des sept discours de Saint-Just, qui forment sept textes bien distincts. Mais dans le cas de l'étude d'un corpus d'éditoriaux quotidiens, par exemple, on peut vouloir découper le corpus en jours, en semaines, en mois, et considérer chacune de ces entités comme un texte dans les travaux statistiques.

Cordial est un logiciel de correction grammaticale et d'aide à la rédaction pour la langue française pour Microsoft Windows et Mac OS X.
Il existe en deux versions : standard et professionnelle. Les deux versions disposent du même correcteur mais la version professionnelle propose en plus le dictionnaire Littré, le Trésor de la langue française, quelques dictionnaires spécifiques (codes postaux, vrais / faux-amis, etc.) et des outils d'aide à la traduction (français, anglais, espagnol, allemand, italien, portugais), ainsi que la correction orthographique dans 70 langues, depuis la version 2012. Il est compatible avec les dernières nouveautés telles que Windows 8
Il corrige les textes, il fonctionne avec des logiciels. Il aide l'utilisateur à rédiger, il analyse les documents, il nous offert aussi une mise à jour permanente, c'est à dire nous pouvons bénéficier des mises à jour permanente, sur le modèle des mises à jour d'individus. Donc, le logiciel cordial est un outil pour accompagner et perfectionner la maitrise de la langue française pour écrire sans risque de commettre de fautes. [132]

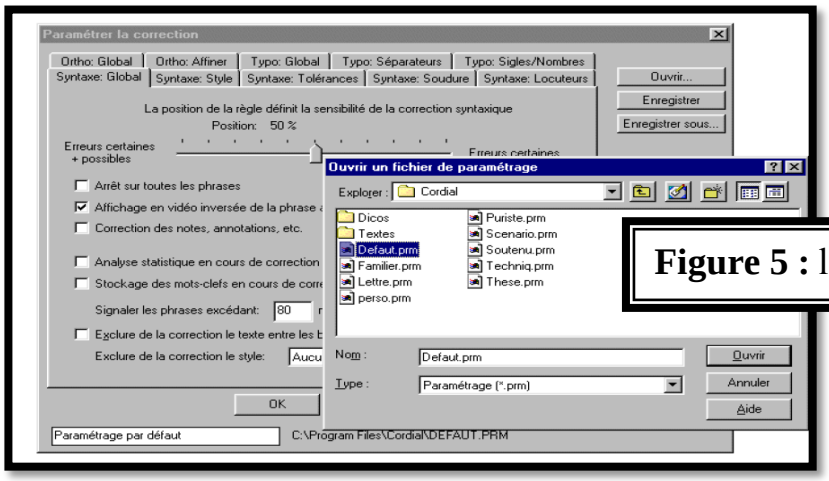

Figure 5 : l'interface du CORDIAL

D.Nooj: est un système de traitement de corpus, reprenant et améliorant les fonctionnalités D'INDEX[133], conçu pour l'enseignement des langues et de la linguistique. Nooj présente des fonctionnalités de TAL[134] qui paraissent prometteuses pour l'enseignement des langues et de la linguistique.
Son principale avantage est sa simplicité d'utilisation : il permet à la fois à l'enseignant de constituer des ressources linguistiques et de les paramétrer afin de constituer des projets pédagogiques destinés aux apprenants [135] .
Par ailleurs, à moyen terme, d'autres fonctionnalités plus avancées de Nooj, en particulier les transformations syntaxiques, pouvaient aussi être exploitées comme démonstrateurs d'opérations linguistiques.

[132] http://www. Je telecharge.com/Bureautique/2277. php/ Consulté le 10 février 2019
[133]La nouvelle version mouture du logiciel INTEX, appelée Nooj a été réécrite en particulier pour répondre aux besoins des utilisateurs pédagogiques.
[134] c'est l'abréviation de traitement automatique des langues
[135] http://alsic.revues.org/ consulté le11 février 2019

L'installation de Nooj peut être réalisée par un téléchargement gratuit à partir du site web Nooj : http://www.nooj4nlp.net.

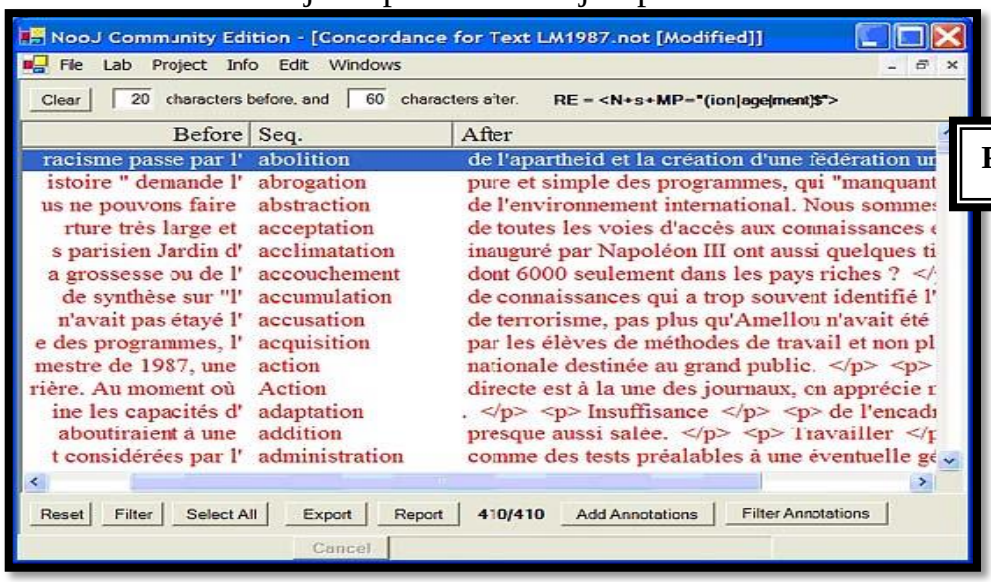

Figure 6 : l'interface du Nooj

E. LEXICO 3 : Conçu dès le début des années 1980, LEXICO représente peut-être le premier logiciel abouti de lexicométrie. Il est aujourd'hui développé, dans sa version 3, par André Salem et son équipe parisienne Syled-Cla2t. LEXICO dispose de toutes les fonctions essentielles de la discipline.
Sa première qualité est sa rapidité d'exécution puisqu'il ne faut pas plus de quelques secondes pour traiter de très gros corpus. Son deuxième avantage est son ergonomie moderne. Par un système de glisser-déposer (drag-and-drop), les unités (un mot du texte par exemple) sont tirés vers les fonctions pour faire apparaître un graphique, une concordance, une AFC. De plus, un système multifenêtrage rend la session de travail dynamique. La seule faiblesse de LEXICO 3 est son incapacité actuelle à traiter les sorties lemmatiseurs. On le favorisera donc seulement pour un traitement des formes graphiques et des segments répétés.
L'interface de lecture des textes sous lexico3 est performante, il invente une sorte de cartographie avec la représentation graphique colorée des occurrences d'une (ou plusieurs) forme (s), d'un groupe de formes, de segments répétés au sein du corpus. La lecture se fait ensuite par phrases, avec une navigation possible entre les différentes occurrences de la forme ou du segment. Ce système met immédiatement en valeur la distribution des formes au sein du corpus et invite donc dès la lecture à passer à l'interprétation.
Le logiciel rend compte de la longueur des textes, du nombre de formes, occurrences et hapax au sein de chaque partie du corpus. Le comptage sensible s'explique par la définition des délimiteurs de forme, et par la détermination adéquate des délimiteurs, les résultats permettent de préciser la forme la plus fréquente dans chaque texte. Ses fonctionnalités restent incontournables pour

dégager le sens, » *les mots de fréquences ont pour fonction de participer à la richesse lexicale d'un texte. »* [136]

L'analyse lexicométrique selon les balises introduites sur le corpus analyse de la structure vocabulaire en vue d'appréhender l'accroissement des vocabulaires, les hautes et les basses et leurs rapports avec les autres éléments textuels, les diagrammes de ventilation des mots par partition et par balises.

Le diagramme de Pareto offre la possibilité de classer les différents phénomènes linguistiques par ordre d'importance Il«*…fournit une représentation très synthétique des renseignements contenus dans la gamme des fréquences. Ce diagramme est constitué par un ensemble de points tracés dans un repère cartésien. Sur l'axe vertical, gradué selon une échelle logarithmique, on porte la fréquence de répétition F, qui varie donc de Fmax, la fréquence maximale du corpus. Sur l'axe horizontal, gradué selon la même échelle logarithmique, on porte, pour chacune des valeurs de la fréquenceF comprise entre 1 et Fmax, le nombre N(F) des formes répétées au moins F fois dans le corpus. La courbe obtenue est donc une courbe cumulée.*»[137]

La fonctionnalité de segments répétés permet de caractériser tous les éléments qui se répètent, ils sont aussi définis comme des *« des suites de formes dont la fréquence est supérieure ou égale à 2 dans le corpus. »* [138] , il faut noter que cette fonctionnalité permet de déterminer la nature des vocables employés propre au domaine étudié.

L'analyse factorielle des correspondances des tableaux croisés est un outil statistique qui permet de repérer la spécifiés et la fréquence des unités textuelles dans chacune des parties choisies. Cet outil met aussi en évidence les proximités qui peuvent apparaitre ou exister entre discours ou parties du discours, réparties selon des balises.

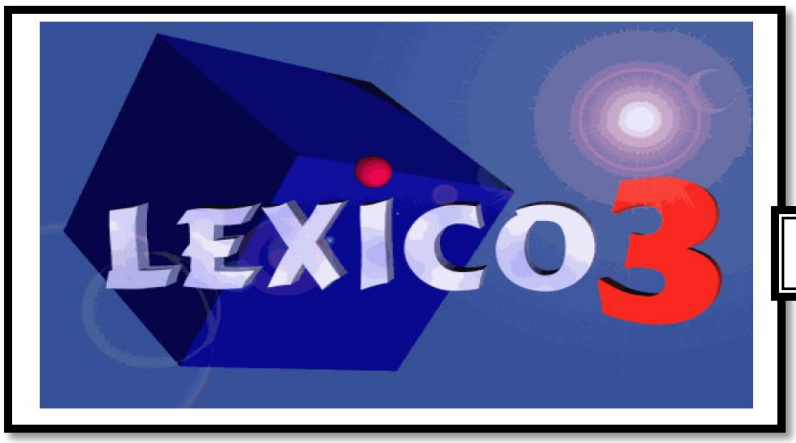

Figure 7 : le logo de lexico3

[136] Bensebia, 2011 p.162
[137] Lebart.et Salem., 1994, p. 47
[138] Lebart.et Salem., 1994, p. 60

Il reste à souligner qu'il existe d'autres logiciels qui cherchent à investir dans d'autres éléments comme des lemmatiseurs , la structure de vocabulaire, l'analyse arborée, l'analyse de co-occurrent et l'analyse de thèmes récurrents.

2.6. Présentation du logiciel «Lexico 3»

Parmi plusieurs logiciels utilisés à des fins lexicométriques, nous avons choisi le «Lexico 3».
Le «Lexico» est l'un des logiciels qui permettent le traitement lexicométrique de textes comportant plusieurs centaines de milliers d'occurrences.
Nous nous concentrerons maintenant, sur la présentation de ce logiciel et de ses fonctions dont nous nous servirons en analysant le corpus.

Au début, le projet Lexico a été conçu à l´École normale supérieure Fontenay Saint-Cloud à Lyon et il a été dirigé par Maurice Tournier et par André Salem. Ensuite, ce projet a été développé au sein du SYLED-CLA2T [139] à l'Université de la Sorbonne Nouvelle -Paris 3.

La première version du logiciel Lexico a été nommée *Lexicloud* et son apparition remonte à 1990. Plus tard, il a été surnommé *Lexico* et le logiciel *Lexico 3*, à l´aide duquel nous allons étudier et analyser notre corpus, La version que nous avons utilisée dans la présente recherche est (Lexico3.6). Il existe aussi d'autres versions antérieures (3.41,3.45) et une version récente qui a été mise sous test (Lexico5.beta). Lexico 5 est développé depuis le milieu des années 2010.

Lexico3 est développé sous le système d'exploitation commercial WINDOWS, vendu par MICROSOFT. Il utilise un découpage en balise du type < projet = 1>, qui permet cette fois- ci de superposer plusieurs niveau de découpages (par exemple par projet, séquence, partie…) sans avoir à modifier le fichier texte, et d'ingérer à nouveau les données.

Lexico3 permet de mener des analyses contrastives et chronologiques et offre les fonctionnalités suivantes [140] :

- D'obtenir la liste des formes présentes dans un document (nombre de formes, *hapax* , fréquence des mots, *etc*.)

[139] Système Linguistiques, Énonciation et Discursivité - Centre de Lexicométrie et d'Analyse Automatique des Textes, Responsable : Serge Fleury, l'Université de la Sorbonne Nouvelle - Paris 3.
[140] http://textopol.u-pec.fr/?tag=lexico3

- D’établir le dictionnaire de ces formes (classement lexicographique, C'est-à-dire par ordre alphabétique, ou lexicométrique, C'est-à-dire par fréquence décroissante)
- D’afficher la concordance d’un mot précis (en offrant la possibilité de trier le contexte à gauche ou à droite par ordre alphabétique, ce qui permet de trouver des expressions ou autres collocations).
- De rechercher des segments répétés dans le texte (suite de deux mots ou plus qui se répète au moins deux fois).
- D’établir des partitions du texte (en fonction des auteurs, de la chronologie, de la source, du genre, de la pagination originale, *etc*., ce qui permet d’effectuer par la suite des comparaisons entre parties).
- De déterminer des sections à travers lesquelles il est possible de naviguer.
- D’effectuer différents calculs statistiques.

Toutes les manipulations sont consignées dans un fichier généré automatiquement et tous les résultats peuvent être consignés dans un rapport disponible au format HTM.

Le logiciel Lexico3 peut être employé, par exemple, pour calculer la taille (ou longueur ou étendue) d’un texte, c’est-à-dire le nombre de mots qui le composent, dresser la liste de ces mots et correspondre à chacun d’eux sa fréquence absolue dans ce texte, c’est-à-dire le nombre de ses apparitions dans celui-ci. De cette liste pouvant être obtenue, soit par ordre alphabétique, soit par ordre de fréquence décroissante, nous pouvons dégager les vocables fréquemment utilisés, ceux qui le sont rarement et ceux de fréquence moyenne. Ce qui nous donnerait une première idée sur la teneur du vocabulaire employé par le scripteur. Les fréquences peuvent également diriger notre attention sur des phénomènes qui sont invisibles à l’oeil nu.

Lexico3 permet aussi de quantifier la ponctuation utilisée dans notre texte ou notre corpus. Ce sont des informations importantes pour une analyse textuelle. En effet, de celles-ci, il nous serait possible de déterminer et d’accéder rapidement à toutes les phrases composant le texte étudié. D’attribuer à ces dernières une longueur moyenne, ce qui est important dans une étude stylistique.

On pourrait également extraire tous les contextes d’un mot qui nous intéresse, en particulier, afin de déterminer les termes qui sont « trop » employés dans son voisinage, et ceux qui le sont « insuffisamment ». Ce qui nous révélerait entre autres le sens que donne le scripteur à ce mot.

Ce logiciel sert, par ailleurs, à comparer deux ou plusieurs textes entre eux ou avec un corpus numérisé, Il nous informe, par exemple, si un texte *T1* est plus riche en vocabulaire qu’un texte *T2*. C’est-à-dire quel est celui qui contient

plus de termes différents que l'autre. En d'autres termes, s'il y a répétition ou renouvellement des vocables utilisés. Il nous dit comment évolue l'accroissement du vocabulaire en chacun d'eux et détermine les moments de variation significatifs dans l'apport de ce vocabulaire.

En particulier, la statistique lexicale, par un simple découpage du texte en tranches, nous permet de déceler les parties qui ont connu un afflux de vocables nouveaux et nous indique la position où cette augmentation a eu lieu. Ce qui permet de savoir à quels moments de son discours le scripteur renouvelle son vocabulaire ou se répète. Le retour au texte nous révélerait alors les thèmes et les raisons de cette attitude.

Lexico3 nous donne, par ailleurs, des outils mathématiques pour mesurer la distance lexicale entre deux ou plusieurs textes, c'est-à-dire en gros le nombre de vocables qu'ils n'ont pas employé en commun.

Nous ajouterons que les chiffres et les graphes obtenus par celui-ci, s'ils sont riches en informations, ne sont pas suffisants. En effet, l'étape capitale en statistique est l'interprétation des données chiffrées.

Concernant les travaux effectués à l´aide du *Lexico 3*, il faut constater que, le plus souvent, les thématiques traitées touchent diverses analyses des discours politiques et de leurs évolutions diachroniques. Par exemple, dans sa thèse de doctorat, Jean-Marc Leblanc analyse les *Voeux présidentiels sous la cinquième République (1959-2001)* [141]."*Dans ses voeux du 31 décembre 1995, Jacques Chirac a prononcé sept fois le mot confiance,dans un contexte social agité, marqué par d'importants mouvements de protestations. Ceconstat quantitatif est alors abondamment commenté par la presse comme un signe d'apaisement donné par le chef de l'État."*

Parmi les autres travaux concernant cette problématique, nous pouvons mentionner l´analyse d´un discours royal espagnol de C. Pineira-Tresmontant et d'A. Salem [142], la recherche de F. Abbassi, *Discours théorique et discours d'action,* [143] qui étudie l'évolution de l´utilisation du vocabulaire de groupements

[141] Jean-Marc Leblanc, *Les voeux présidentiels sous la Cinquième République (1959-2001), Recherches et expérimentations lexicométriques à propos de l'ethos dans un genre discursif rituel,*Université de Paris 12 Val-de-Marne, 2005, p. 57

[142] C. Pineira-Tresmontant, A. Salem, *Discours royal espagnol,* Lexicometrica-revue électronique,2018.
http://www.cavi.univparis3.fr/ilpga/ilpga/tal/lexicoWWW/navigations/Gouv5.pdf
[Consulté le 13Mars 2019]

[143] Abbassi, F. *Discours théorique et discours d'action,* Lexicometrica-revue électronique, 2018.http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/navigations/SocPo2.pdf
[Consulté le 20avril 2019]

islamistes, entre autres. Il est à noter encore que Serge A. de Sousa a publié son essai Le discours de Fidel Castro [144] dans lequel il effectue l'exploration lexicométrique d'une série de discours de Fidel Castro qu´il a prononcé pendant les années 1959-2004. Lexico 3 est donc un outil important de repérage d'un langage totalitaire, par exemple.

Néanmoins, il ne s´agit pas seulement des textes politiques qui sont étudiés à l´aide du logiciel. Nous pouvons également trouver les analyses des oeuvres littéraires. dont nous pouvons mentionner la recherche des *Complaintes de Jules Laforgue* qui a été réalisée par le Centre de recherches Hubert de Phalèse dont *« la mission est de développer les études littéraires assistées par ordinateur et de diffuser ces nouveaux savoirs. »* [145] Cette équipe de recherche a été fondée en 1989 à l´Université Paris 3 Sorbonne Nouvelle (Lettres Modernes et Outils informatiques). Une autre étude concernant l´oeuvre poétique a été écrite par Giordano Righetti [146] de l´Université de Bologne (Italie). Ce travail présente la recherche du *Bateau Ivre* de Rimbaud et l´auteur se concentre surtout sur la structure et fréquence du vocabulaire utilisé dans le poème.

[144] Serge A. de Sousa, *Le discours de Fidel Castro, Essai de lexicométrie politique,* Lexicometrica-revue électronique, 2018.http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/navigations/STC2.pdf [Consulté le 13 mai 2019]

[145] Centre Hubert de Phalèse - littérature et informatique, Université Paris 3 Sorbonne Nouvelle.http://www.cavi.univ-paris3.fr/phalese/ [consulté le 03 Avril 2019]

[146] Giordano Righetti, *L'analyse textuelle du Bateau Ivre de Rimbaud avec le logiciel Lexico 3: ressources actuelles et possibilités de développement,* L'Università degli Studi di Bologna, http://www2.lingue.unibo.it/dese/didactique/travaux/Righetti/tesinainformaticarighetti.pdf [consulté le 28 Avril 2019]

RÉFÉRENCES BIBLIOGRAPHIQUES

- FORTIER Paul A. (1995) Categories, Theory, and Words in Literary Texts, Research in Humanities Computing, n°5.
- LEBART L. et SALEM A. (1994) Statistique textuelle, Dunod, Paris.
- LOWE David et MATTHEWS Robert (1995) Shakespeare Vs. Fletcher: A Stylometric Analysis by Radial Basis Functions, Computers and the Humanities.
- HENRY Georges (1975) Comment mesurer la lisibilité, Labor, Bruxelles
- André Salem, Analyse factorielle et lexicométrie : synthèse de quelques expériences, Dunod, Paris 1999
- Jean-Marc LEBLANC, Les voeux présidentiels sous la Cinquième République (1959-2001), Recherches et expérimentations lexicométriques à propos de l'ethos dans un genre discursif rituel, thèse de doctorat, Université de Paris 12 Val-de-Marne, 2005.
- Philippe Galiana, Lexicométrie, Le bulletin de L´EPI (Association Enseignement Public& Informatique), n°63, 1991.
- Pierre Guiraud, La Stylistique, Paris, Presses Universitaires de France, 1961.
- Charles Muller, Initiation aux méthodes de la statistique linguistique, Paris, Classique Hachette,1973.
- Fabienne Moreau, Pascale Sébillot, Contributions des techniques du traitement automatique des langues à la recherche d'information, INRIA (Institut National de Recherche en Informatique et en Automatique), Rennes, 2005.
- DICTIONNAIRE DE linguistique Larousse, 1994,333
- POLGUÈRE Alain, Notions de base en lexicologie, Montréal: Université de Montréal, OLST, 2002,p. 82
- Maurice Tournier, D'où viennent les fréquences de vocabulaire ? La lexicométrie et ses modèles. In:Mots, Institut de la langue française, Saint Cloud, 1980, N°1..
- Pierre Fiala, L'interprétation en lexicométrie. Une approche quantitative des données lexicales, Langue française, ENS- Fontenay/Saint Cloud, 1994.
- Charles Muller, Principes et méthode de statistique lexicale, Paris, Honoré Champion Editeur,1992.
- https://fr.wikipedia.org/wiki/Roberto_Busa [Consulté le 24 janvier 2019]

Encyclopaedia Universalis. La loi de Zipf. http://www.encyclopaedia universalis.fr/ [Consulté le 25 janvier 2019]

- Centre National de la Recherche Scientifique. http://www.cnrs.fr/ [Consulté le 02 février 2019]

- http://www.tal.univ-paris3.fr/lexico/Lexico3doc0.pdf
- http://www.persee.fr/web/revues/home/prescript/article/mots_0243-6450_1980_num_1_1_1010 [Consulté le 03 Avril 2019]
- http://www.tal.univ-paris3.fr/lexico/Lexico3doc0.pdf P.44
- http://www. Je telecharge.com/Bureautique/2277. php/ Consulté le 10 février 2019
- http://alsic.revues.org/ consulté le11 février 2019
- http://textopol.u-pec.fr/?tag=lexico3
- http://www.cavi.univparis3.fr/ilpga/ilpga/tal/lexicoWWW/navigations/Gouv5.pdf [Consulté le 13Mars 2019]
- Abbassi, F. Discours théorique et discours d'action, Lexicometrica-revue électronique, 2018.http://www.cavi.univ
- paris3.fr/ilpga/ilpga/tal/lexicoWWW/navigations/SocPo2.pdf [Consulté le 20avril 2019]
- Serge A. de Sousa, Le discours de Fidel Castro, Essai de lexicométrie politique, Lexicometrica-revue électronique, 2018.http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/navigations/STC2.pdf [Consulté le 13 mai 2019]
- Centre Hubert de Phalèse - littérature et informatique, Université Paris 3 Sorbonne Nouvelle.http://www.cavi.univ-paris3.fr/phalese/ [consulté le 03 Avril 2019]
- Giordano Righetti, L'analyse textuelle du Bateau Ivre de Rimbaud avec le logiciel Lexico 3: ressources actuelles et possibilités de développement, L'Università degli Studi di Bologna, http://www2.lingue.unibo.it/dese/didactique/travaux/Righetti/tesinainformaticarighetti.pdf [consulté le 28 Avril 2019]