La Lexicométrie Documentaire : Contribution à l'utilisation des Techniques Documentaires comme Méthodologie d'Etude en Sciences Sociales

Abdelkrim ABDOUN*

ABSTRACT

Lexicometry is one of discourse analysis approachs. It aims at getting highlights or meaning of a corpus of texts according to the conditions of communication. Its basis is concerned with exhaustive numbering of the keywords. It is known that documentary techniques are usually a mean of selecting the most significant words from a text, it is, then, possible to think about the possibility to use indexation and other types of documentary techniques in order to analyse the bodies of texts.

Considérations générales et problématiques :

Dans cette réflexion, nous essayerons de développer la problématique de l'utilisation des techniques documentaires-comme méthodologie d'analyse des phénomènes socio-historiques.

Autrement dit, une technique documentaire, peut-elle permettre l'interprétation, l'analyse et l'explication des phénomènes ressortant en principe, aux sciences sociales ?

Pour mieux comprendre le fondement de cette pensée, nous commencerons par quelques éléments de définition.

Qu'est-ce-que les sciences sociales ?

Les Sciences Sociales ont pour objet l'étude de l'homme dans son environnement, elles ont pour particularité, l'intérêt porté à la dimension historique.

Qu'est-ce-qu'une technique documentaire ?

En termes simplistes, il s'agit de toute méthode utilisée par les documentalistes-bibliothécaires pour la collecte, le traitement et la diffusion de l'information scientifique.

Ces explications tout en paraissant nous éloigner de notre problématique, nous en rapprochent au contraire.

Les techniques documentaires, dont la finalité est l'identification et la localisation des écrits transforment pour ce faire, l'information contenue dans les documents. Cela nécessite entre autres, la standardisation des éléments informatifs (mots-clés), pour représenter dans des catégories classificatoires, le contenu de nombreuses communications, qui traitent d'un sujet unique en l'exprimant de manières différents.

Ceci introduit pour nous les notions d'indexation et de classification, c'est à dire, de langage documentaire.

Nous soutenons, que le fait de représenter le contenu des documents par des catégories sémantiques (mots-clés), ou logiques (classes), à la fois manipulables et chiffrées (lexicométrie) permettrait de comprendre l'état et l'évolution de la pensée collective exprimée dans l'écrit, et ce, dans un espace temps bien déterminé.

Nous partons ici d'une hypothèse généralement admise, qu'il existe une corrélation entre l'évolution de

*Enseignant Chercheur au CERIST

l'écrit, support d'expression de la pensée collective, et l'environnement source et moteur de cette pensée.

On admettrait donc, que l'histoire d'une société donnée pourrait être appréhendée à travers ses publications, écrites et/ou audio-visuelles. L'information contenue dans ces publications serait le reflet de l'environnement générateur de ces publications. Cela relèverait de l'analyse de discours, méthode courante dans l'étude des relations entre l'écrit et les facteurs de production ; aussi, avant d'exposer la méthodologie que nous proposons pour l'étude des relations entre l'écrit et l'environnement, nous traiterons de la lexicométrie comme méthode socio-linguistique de l'analyse de discours. Ceci nous permettra dans un premier temps de nous démarquer de la linguistique pour nous confiner dans une méthodologie d'approche résolument documentaire, qui contribuera à l'aboutissement de cette brève réflexion.

La lexicométrie : Méthode linguistique

Peut-on parler de "la lexicométrie linguistique", ou plus précisemment "socio-linguistique" ? Cela ne resemble-t-il pas à un pléonasme, puisque la lexicométrie est par essence une méthode linguistique ?

Ce qualificatif (linguistique) est tout de même employé dans l'esprit de notre démarche tendant à séparer notre méthodologie de travail de celle utilisée par les linguistes. Cela ne saurait se faire sans une description préalable de la lexicométrie en tant que méthode linguistique.

Définition :

La lexicométrie est une des approches de l'analyse de discours, c'est à dire une problématique de la signification d'un corpus de textes par rapport aux "circonstances de communication qui en font un discours" (1).

Les circonstances de communication ou conditions de production constituent l'environnement stimulant et déterminant de toute production intellectuelle, à savoir, le cadre institutionnel, l'appareil idéologique, les rapports de force... (2).

Le but de la lexicométrie est donc l'étude des rapports et des influences entretenues entre le discours et les conditions générales de sa production. Elle chercherait donc à saisir la signification d'un corpus de textes par rapport au cadre spacio-temporel qui les a engendrés. Cela par le calcul des fréquences des unités sémantiques ; considérant que "le vocabulaire d'un texte, qui est un échantillon d'un lexique virtuel, obeït dans sa structure quantitative à des impulsions qui ne sont pas fortuites, et se construit suivant des lois complexes et mal connues encore" (3).

La lexicométrie est donc une des applications de la statistique, qui utilise pour ce faire des méthodes propres aux spécificités de l'objet lexique.

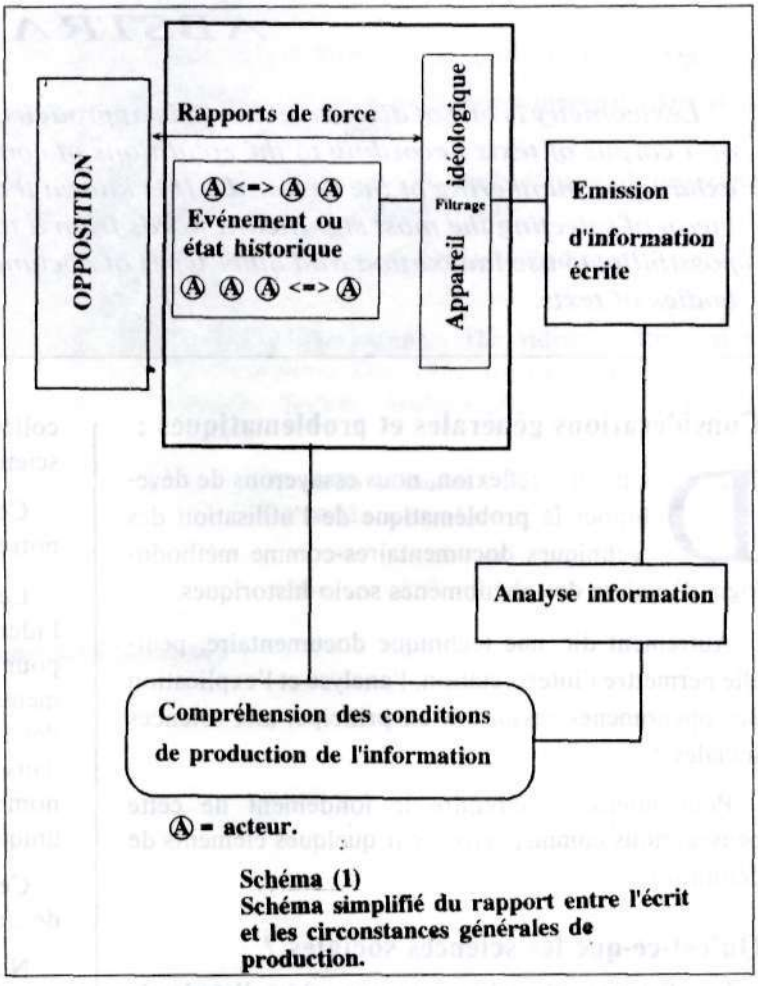


Schéma (1)
Schéma simplifié du rapport entre l'écrit et les circonstances générales de production.

Méthode :

Nous tentons ici une synthèse des principales constantes qui interviennent dans une étude lexicométrique :

- Choix du corpus :

Il n'existe pas de règle à ce sujet. Toutefois "pour travailler avec le maximum d'efficacité, mieux vaut

chercher à équilibrer entre discours et conditions de production pour que leur articulation soit la plus riche possible" (4).

Collecte des unités significatives :

Une fois déterminé le corpus, la première étape du travail consiste en la collecte des unités significatives. En cela, la lexicométrie se base sur l'exhaustivité des relevés, c'est à dire que chaque occurence est également significative, à l'exception des mots vides (mots qui ne sont ni substantifs, ni qualificatifs).

Choix du critère du dépouillement :

Il s'agit de préciser une forme unique pour l'ensemble des mots du corpus. On opte en général pour le lemme, ce qui signifie en français : l'infinitif pour le verbe, le masculin singulier pour les adjectifs, ...

Etablissement des tables de fréquences :

Cette étape comprend essentiellement :

- L'etude des répartitions des formes selon les émetteurs, et éventuellement les périodes;
- L'étude du fonctionnement des formes dans leur identité sur le plan syntagmatique ou co-occurences;
- L'établissement d'un index alphabétique des formes retenues avec indication de la fréquence de chacun;
- L'établissement d'un index hiérarchique classant les formes par ordre décroissant;
- L'établissement des tableaux de statistiques générales sur la structure du texte.

Calcul des fréquences :

Ce calcul utilise un grand nombre d'éléments de la statistique, et repose nécessairement sur des choix : choix statistiques et choix historiques essentiellement.

Critique :

En tant que méthode d'analyse de discours utilisant dans ses investigations l'outil statistique, la lexicométrie est sujette à certaines critiques qui relèvent essentiellement de la difficulté qu'a la statistique à rendre compte de tous les paramètres d'étude que peut fournir un corpus de textes.

La lexicométrie est réduite pour certains à un simple "comptage des mots" (5) puisque l'essence d'un texte ne peut être abordée par d'autres voies qu'intuitives.

Cette critique demeure quand même primaire, puisque l'outil statistique n'est pas un canevas, ou modèle figé, mais plutôt une procédure flexible et malléable selon l'initiative et les orientations du chercheur.

Vers une lexicométrie documentaire

Essai de dissociation : lexicométrie documentaire/lexicométrie linguistique :

La documentation peut-elle réellement contribuer par des méthodes propres qui permettent une véritable approche de la lexicométrie basée peu ou prou sur les techniques documentaires ?

Il faut d'abord que ces techniques représentent de manière objective le contenu des textes, il faut ensuite que ces représentations puissent être quantifiées.

Il existe à notre sens trois méthodes pouvant participer à la réalisation de ces deux conditions :

L'indexation;
La classification;
L'analyse documentaire.

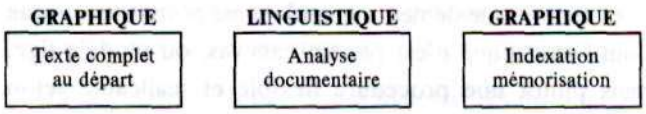
Le problème reste toutefois l'indépendance de la technique documentaire par rapport à la linguistique. Il existe en effet de nombreuses similitudes entre ces "deux disciplines" : d'une part, les techniques documentaires apparaissent comme une problématique de la réduction du contenu des documents, et comme une recherche de la signification ; deux opérations qui ressortent essentiellement à la linguistique, d'autre part, le fait de parler de langage documentaire équivaut à parler de langage naturel et des ses dérivées, donc de linguistique (6).

Il serait toutefois prématuré de classer la technique documentaire en sous-rubrique de la linguistique, puisque si des relations, de toute évidence existent entre ces deux disciplines, il est possible de dissocier l'une de l'autre, et de montrer par la même, la spécificité de la technique documentaire.

Examinons en exemple, la technique d'indexation : cette méthode met en cœxistence deux systèmes :

- Un système linguistique qui met en relation l'émetteur (documentaliste) et le récepteur (utilisateur) ;
- et un système graphique qui est celui des textes.

Ceci pourrait être représenté schématiquement de la manière suivante :



(Schéma n°2) Cœxistence des systèmes linguistique et graphique dans l'opération d'indexation (7).

Par ailleurs, si l'on se refère aux outils grammaticaux de l'analyse documentaire, une observation s'impose : c'est qu'ils sont plus proches des concepts logiques qu'ils ne le sont des concepts linguistiques (8).

Enfin, les techniques documentaires constituent une technique de travail spécifique de par :

Leur objet :

Le document d'une part (contenant), et le contenu du document, d'autre part.

Leur but :

Mémorisation et caractérisation du contenu documentaire.

Le choix des descripteurs :

Contrairement à la statistique linguistique qui se base sur les items formels (mots présents physiquement dans le texte), la technique documentaire repère les notions présentes dans le document et les transcrit en mots clés, qui peuvent ou non être présents dans le texte-

La relation entre les descripteurs :

Cette relation est établie à partir de facettes et rapports d'instrumentation, de localisation, de causalité, etc et non comme c'est le cas pour la linguistique, à partir d'éléments grammaticaux.

La lexicométrie documentaire :

Les niveaux d'analyse :

Précisons au départ qu'il n'existe pas de distinction dans le choix du corpus entre une lexicométrie sociolinguistique et une lexicométrie documentaire.

Pour appréhender un corpus de textes, deux niveaux d'analyse sont possibles :

- Le niveau des titres ;
- Le niveau des textes.

Le niveau des titres :

Le choix des mots clés du titre répond à plusieurs exigences : en s'interrogeant sur les rapports existant à l'intérieur de l'étude du livre entre les statistiques bibliographiques et linguistiques. ''Où commence l'une ? Où finit l'autre ?'' La réponse est claire... c'est le titre de la publication. Les mots clés du texte échappent comme tels à la statistique bibliographique et font l'objet de la statistique linguistique'' (9).

Cette affirmation bien que fondée, est déjà ancienne, puisque comme nous le verrons plus loin, le mot du texte peut également ressortir à la statistique bibliographique.

Le mot clé du titre apparait donc comme un quelconque indice de statistique bibliographique. Le problème posé est relatif à la valeur du mot clé du titre comme élément de description du contenu d'un document.

A priori, pour les œuvres techniques et scientifiques, ''aucun élément descriptif (mot clé du texte, résumé, référence...) ne semble réunir autant d'information que le titre d'un document'' (10). Mais, rien en pratique ne permet de confirmer cette valeur informative des titres, y compris ceux des articles de périodiques tenus pourtant de suivre des orientations normatives, qui précisent que ''le titre doit être significatif du contenu de l'article...'' (11). Cela demeure surtout valable pour les revues scientifiques et techniques, au contraire des journaux et revues diffusés pour le grand public, où le subjectif est parfois prédominant.

En l'absence de données résolument affirmatives, la responsabilité du chercheur demeure entière quant à la manipulation et au choix rigoureux des mots clés du titre. Choix et manipulations entrant dans une procédure de raffinement méthodologique progressif.

Le niveau des textes :

Le problème posé est le choix de la méthode la plus appropriée à l'extraction des mots clés du texte. Dans la pratique, cette méthode est l'indexation, systématique ou matières, c'est à dire ''l'opération qui consiste à décrire et à caractériser le contenu d'un document à l'aide de représentations des concepts contenus dans ce document, c'est à dire à transcrire en langage documentaire les concepts après les avoir extrait du document par une analyse. La transaction en langage documentaire se fait grâce à des outils d'indexation tels que thésaurus, classification, etc'' (12).

Méthode :

Nous ne proposons pas ici à proprement parler, une méthodologie de la "lexicométrie documentaire". D'ailleurs, si elle existait, une telle méthodologie resterait ouverte à l'initiative du chercheur.
Ce que nous proposons, c'est donc une classification des étapes d'une lexicométrie basée sur la technique documentaire.

Nous pouvons distinguer deux étapes essentielles :

1 - Composition des index ;
2 - Recherche des associations.

Composition des index :

On peut y énoncer les étapes suivantes :

Indexation :

Elle peut être faite à partir des mots clés du titre ou du texte. Dans ces deux cas, elle doit répondre à des règles rigoureuses à même d'éviter les erreurs fréquentes à ce niveau.

L'indexation établie à partir des mots clés du texte peut être :

- Exhaustive (relevant tous les concepts utiles) ou selective (ne relevant que les concepts essentiels) ;
- Libre (sans référence à un outil d'indexation), ou contrôlée (se référant aux outils d'indexation) ;
- Systématique (relevant plusieurs concepts du document), ou matière (ne relevant que le concept essentiel).

Par ailleurs, l'indexation peut se faire en même temps que l'analyse documentaire du texte, qui remplacera le titre dans les index KWIC et KWOC, comme il en sera fait référence plus loin.

Ecrémage :

Il s'agit d'éliminer tout mot clé non utile pour la recherche. Cette opération ne concerne que la première étape du travail (composition des index), puisque le mot clé à basse fréquence garde toute sa signification s'il est combiné à d'autres mots clés dans le cadre des classes de fréquence (étape 2).
L'écrémage se définit par la prise en compte uniquement des unités significatives, à l'exclusion des mots vides.

Est considéré comme mot vide :

- Un mot d'un poids sémantique faible ;
- Un mot appartenant à d'autres catégories que celles des substantifs et des qualificatifs (prépositions, conjonctions, pronoms...) ;
- Un mot porteur de subjectivèmes.

Uniformisation :

Au niveau de la "lexicométrie linguistique", nous avons pu voir que c'était le lemme qui était pris comme unité de base. Ici, et selon les usages adoptés en documentation (14), on retient comme unité de base, le substantif singulier-

Elaboration des index proprement dits :

Dans cette étape, nous envisageons quatre index essentiels :

1 - Index Kwic (Key words in Context) ou Index Kwoc (Key Words out of context) :

Ils permettent de classer les mots clés (alphabétiquement) dans leur contexte (résumé ou titre). Ils constituent pour toutes les opérations qui suivront un point de référence important.

Notons toutefois que ces index concernent surtout les mots clés du titre. Dans l'indexation des textes, ils sont remplacés par les index systématique ou matière.

Index alphabétique :

Dans lequel, les mots clés sont classés alphabétiquement, suivis de leur fréquence.

Index chronologique :

Classement par année des mots clés suivis de leur fréquence.

Index hiérarchique :

Classement des mots clés par ordre décroissant des fréquences.

Etude des fréquences :

L'étude des fréquences des mots clés dans chacun des index permet déjà une certaine représentation de la signification, mais l'on reste encore à un niveau quantitatif. L'analyse qualitative ne commencera qu'à partir de la deuxième phase.

Recherche des associations :

On peut considérer qu'il existe deux types d'associations :

- Associations logiques ;
- Associations linguistiques.

Associations logiques :

Il s'agit de répartir les mots clés par affinités logiques, à travers des classes, suivant les règles et les processus de la classification documentaire. L'étude de ces associations sur les plans synchronique et diachronique permet de mettre en exergue le schéma idéologique véhiculé par le corpus des textes.

Associations linguistiques :

Il s'agit de répartir les mots clés par affinités sémantiques selon les règles de l'établissement des thésauri.

Critique :

La critique dont nous ferons part ici, concerne la subjectivité de l'indexation comme méthode de choix des mots clés.

Il est en effet remarquable que des mots clés d'un même texte diffèrent selon qu'ils aient été extraits par tel ou autre indexateur. Les règles d'indexation sont trop vagues et la norme ne donne que des orientations générales sans apport sur le plan de l'uniformisation.

Il n'existe aucun moyen pour contourner cette critique, si ce n'est redoubler de prudence dans l'extraction des mots clés. Cela à deux niveaux :

A - Pendant l'indexation ;
B - Après l'indexation.

A) Pendant l'indexation :

Il existe un certain nombre de moyens pour réduire le taux de subjectivité :

La cohérence d'indexation :

Elle permet de limiter sérieusement les écarts probables. Cette méthode consiste à calculer la compatibilité existant entre les mots clés extraits d'un même texte par deux indexateurs, selon la formule :

$$Ci = \frac{ds}{dd}$$

ou ds représente le nombre de descripteurs semblables attribués à un même document par deux documentalistes, et dd le nombre total de descripteurs semblables ou dissemblables attribués au même document par deux documentalistes.

L'indexation exhaustive :

Relevé de toutes les notions présentant un intérêt :

La pondération :

Attribution aux mots clés, de niveaux de pertinence.

L'indexation contrôlée :

Permet d'éviter les doubles emplois par l'élimination des ambiguïtés du langage naturel.

B) Après l'indexation :

L'indexation conduit à l'élaboration d'outils de la recherche, c'est à dire en somme, à la réalisation d'un système documentaire. Il est donc possible d'appliquer à ce stade les méthodes d'évaluation d'un système documentaire qui en permettent la révision et l'optimisation, le cas échéant.

Nous citons parmi ces méthodes quelques ratios relatifs à l'évaluation de la recherche documentaire :

$$\text{Le taux de pertinence} = \frac{\text{documents pertinents retrouvés}}{\text{documents pertinents retrouvés + documents pertinents non retrouvés}}$$

$$\text{Le taux de bruit} = \frac{\text{documents non pertinents retrouvés}}{\text{documents pertinents retrouvés}}$$

$$\text{Le taux de silence} = \frac{\text{documents pertinents non retrouvés}}{\text{documents pertinents retrouvés}}$$

Conclusion :

Nous avons essayé au cours de cette brève réflexion de contribuer à une démonstration des applications possibles de la technique documentaire en dehors du cadre rigoureusement techniciste dans lequel elle évolue généralement.

Nous précisons toutefois que notre exposé est encore incomplet dans bon nombre de ses points, qui mériteraient un plus large développement au cours d'études que nous espérons incessament aborder.

Bibliographie :

(1) MAINGUENEAU, D. - Initiation aux méthodes de l'analyse de discours. - Paris : Hachette, 1976. p.13.

(2) Cf à ce sujet :
PECHEUX (M), FÜCHS (C) : - Typologie du discours politique. In : Langage, n°76, 1983.

(3) MÜLLER (Ch.) : - Initiation à la statistique linguistique - Paris : Larousse, 1968. p.165.

(4) MAINGUENEAU (D) Op. cité p.18.

(5) Ibid p.43.

(6) COYAUD (M) - Introduction à l'étude des langages documentaires. - Paris : Librairie C. Klinchsiech, 1968. - (Th. de 3e cycle).

(8) LONG (B) - Linguistique et indexation. In : Documentaliste, vol. 17, n°3, Mai - Juin 1980, p.100.

(9) GARDIN (J.C) Les analyses de discours. - Neuchatêl (Suisse) : Delacheux et Niestle S.A., 1974. - p.132.

(10) ESTIVALS (R), GAUDY (J.C), VERGZE (G) :
- L'Avant-garde : étude historique et scientifique des publications périodiques ayant pour titre l'avant-garde. - Paris : Bibliothèque nationale, 1968. - p.9.

(11) FONDIN (H) - Le titre comme élément de description du contenu d'un document. - In : Documentaliste, vol. 19, n°1, Janvier - Février 1982, p.3-15.

(12) ASSOCIATION FRANCAISE DE NORMALISATION. - NFZ 41-003, Janvier 1974. - Présentation des articles de périodiques. p.5.

(13) ASSOCIATION FRANCAISE DE NORMALISATION. - Principes généraux pour l'indexation des documents. - NFZ 47-102, Août 1978. p.2.

(14) ASSOCIATION FRANCAISE DE NORMALISATION. - Indexation matières. - NFZ 44-070, Août 1986.