Dr.Tabet Aoul Zoulikha
MCB USTO MB

17. La lexicométrie ou Lexico-statistique

«Le mot créé par un individu ne prend sa valeur que dans la mesure où il est accepté, repris, répété, aussi est-il finalement défini par la somme de ses emplois» Pierre Guirraud

Plan

1. Préliminaires

La somme des emplois étant un indice de pertinence dans les traitements de texte, l'outil informatique permet l'émergence de nouvelles catégories en matière d'analyse des textes. La taille et la longueur de certains corpus rendent nécessaire l'analyse quantitative, qui en plus de donner des dénombrements par les index, permet aussi de déceler des phénomènes non perceptibles à la lecture linéaire comme les associations singulières entre vocables, leurs répétitions, comme aussi les comparaisons faisant ressortir des spécificités à l'auteur, les hapax, les nullax.

2. L'approche lexico-statistique

L'objet de l'approche lexico-statistique est le mot « répété », ce sont donc les caractères quantitatifs du vocabulaire d'un texte à analyser.
C'est donc une forme de régularité descriptible et objectivable en ce qu'elle implique une trace matérielle avec une règle d'écriture reconnaissable par des grilles d'écriture. Cette régularité se fait sur le mode de la répétition systématique comme sur le mode de la répétition diversement modulées. La démarche lexico-statistique s'intéresse aux vocables du texte par leurs variantes quantitatives, par leurs associations et par les qualités de ces associations. L'environnement lexical du signe est un élément décisif dans la composante sémantique du signe lui-même.
Jean Peytard définit le texte comme :
« Un lieu de densification d'un réseau connotatif, réseau dense de relations entre des constituants, tel qu'une lecture nouvelle en est toujours possible »
En effet, des traits stylistiques distinctifs, notamment par le calcul les pourcentages de noms, verbes, adjectifs, adverbes, etc., à condition, bien sûr, qu'ils puissent être reconnus, peuvent être décelés. À partir, par exemple du décompte des pronoms, des analyses peuvent démontrer une dominance, un rejet, ou encore à partir d'un temps verbal privilégié peuvent être dégagées des représentations du monde.
Les signes de ponctuation sont également facilement repérables et analysables par ordinateur et révélateurs de certaines particularités d'écriture des auteurs. On peut même aller plus loin et voir la distribution des syllabes des mots, leurs longueurs, leurs fréquences qui sont autant d'indices. Un autre avantage de l'informatique est le gain de temps dans l'exploitation des corpus importants.

3. La lexico-statistique dans un corpus :

Le mot « répété » est l'objet de l'approche lexico-statistique basée sur les caractères quantitatifs et qualitatifs.

Les premières études en lexico-statistique ont été principalement réalisées sur des corpus scientifiques et des discours politiques : Jean Paul Benzecri et Charles Mûller sont les pionniers dans ce domaine.

En voulant transposer la méthode lexico-statistique aux textes littéraires les équipes de Saint-Cloud, de Guiraud, de Rastier, de Condé... se sont heurtées à ce qui constitue à la fois un atout et une difficulté du texte littéraire à savoir son caractère polysémique et hétérogène.

C'est l'étude sémiotique du signe qui va permettre le passage de l'univocité du signe scientifique à la plurivocité du signe du texte littéraire. L'environnement lexical du signe constitue un élément décisif dans la composante sémantique du signe lui-même.

La statistique lexicale informatisée combinée aux apports de l'analyse du discours et de la sémio-linguistique va permettre l'émergence de "points névralgiques" d'un parcours de sens et de voir aussi ces endroits où le sens apparaît par « entaille » discrètement dissimulé. La statistique lexicale informatisée, c'est aussi la possibilité de pratiquer une autre approche des textes multipliant en quelque sorte la capacité de lecture du chercheur, lui permettant d'aller au-delà de l'imprécision des lectures intuitives et de l'empirisme des interprétations herméneutiques. L'outil informatique combine précision du détail et vision d'ensemble.

Quelques définitions de la statistique lexicale peuvent éclairer sur cette démarche :

a) « *la statistique lexicale a pour objet les mots ; c'est l'étude quantitative des mots d'un texte ou d'un corpus en fonction d'un seul caractère : leur rattachement à un lexème. » (Mûller)*[81]

« la statistique lexicale qui se veut exhaustive, systématique et /ou automatisée a pour but le comptage des unités lexicales d'un texte ou corpus (Charaudeau). »[82]

« c'est l'ensemble de méthodes permettant d'opérer des réorganisations formelles de la séquence textuelle et des analyses statistiques portant sue le vocabulaire d'un corpus de textes. » (Salem)[83]

Il s'avère que la statistique lexicale appelée aussi linguistique quantitative (Miiller 1964), statistique linguistique (Guiraud 1960) et statistique textuelle (Salem 1987) n'est pas une théorie mais une méthode descriptive. Il faut avant toute manipulation du texte le « numériser ». Cette opération consiste à transformer une information signifiante pour l'homme en une série de codes lisibles par la machine. La numérisation « appauvrit » le texte d'une partie de sa signification (exemple des calligrammes d'Apollinaire) ; donc attention aux distorsions subies par les données, aux sauts de lignes, aux indentations, alinéas, majuscules et autres dispositifs typographiques voulus par l'auteur.

Le texte peut être numérisé soit par saisie manuelle, soit par reconnaissance des caractères. La première possibilité peut comporter des erreurs lorsque les corpus sont très longs. La deuxième manière utilise un scanner ou numériseur qui transfère à l'ordinateur l'image numérique du document. Puis un logiciel interprète cette image en reconnaissant la forme des caractères.

---

[81] Müller C, *Principes et méthodes de la statistique lexicale,* Champion, Paris, 1992.

[82] Charaudeau ,Maingueneau D,*Dictionnaire de l'analyse du discours*,Seuil, 2002.

[83] Lebart et Salem, *Statistique textuelle,* Dunod, Paris, 1994.

4. Les logiciels de traitement de texte

Une fois le texte numérisé, le choix d'un logiciel s'impose. Très nombreux sur le marché (Piste, Lexico 1, 2, 3, Hyperbase, Alceste, Lexplorer), ils offrent des possibilités de traitement de texte spécifiques. Une question délicate se pose : le fait de modifier le texte ne risque-t-il pas de réduire des ambiguïtés ou effets de sens essentiels dans le texte ?

C'est pour cela qu'il faut impérativement effectuer des retours au texte initial. Une fois le texte numérisé, le traitement par des logiciels appropriés peut débuter. Les logiciels d'analyse textuelle prennent en charge le découpage de la chaîne des caractères du texte en unités, la constitution d'un corpus et sa partition en texte puis les analyses statistiques pour fournir en sortie-machine des matériaux divers indexés, classés, hiérarchisés, sélectionnés. Ces matériaux sont triés (en unités spécifiques positives ou négatives), positionnés les uns par rapport aux autres (analyse factorielle des correspondances, arbres hiérarchiques), sériés entre eux (séries chronologiques ou grappes en évolution), articulés les aux autres (lexicogrammes des cooccurrences, graphes de connexion).

Pour les ateliers d'initiation, notre choix s'est souvent porté sur l'Hyperbase pour des raisons pratiques mais surtout pour ce qu'il offre comme performances. C'est un des logiciels des plus complets : outre les possibilités habituelles de tout logiciel d'analyse textuelle, il permet des particularités intéressantes : navigation en hypertexte entre les index et le texte, calcul des écarts réduits (et leur représentation graphique), analyse factorielle des correspondances, comparaison des fréquences du corpus avec d'autres bases de données. Rappelons que l'index est la liste fournie par l'ordinateur de toutes les formes classées par ordre alphabétique ou hiérarchique. L'indexation permet de lire pour chaque forme sa fréquence dans l'ensemble du corpus ou dans une partie seulement. Les formes classées par ordre décroissant de fréquence constituent l'index hiérarchique du texte. Le début de cette liste fait apparaître d'intéressants faits de répétitivité signalant les dominantes thématiques des textes mais aussi les tendances générales en matière de choix lexicaux et grammaticaux.

Comme dans presque tous les textes français, les formes les plus fréquentes sont à peu près toujours les mêmes : le, la, les, (catégorie des déterminants et prépositions). Ces formes sont appelées par Charles Millier les mots-outils ou formes vides (formes fonctionnelles). Ces informations peuvent être l'indice principal de richesse et de variété des structures énonciatives du texte.

Quant aux autres formes ou mots-pleins (appelées aussi en lexicométrie formes pôles ou formes pivots), ils peuvent être indicateurs de réseaux thématiques comme ils peuvent mettre en évidence certaines particularités d'une écriture. De plus, l'étude des spécificités est un précieux auxiliaire, elle permet notamment d'établir des comparaisons entre le lexique particulier de l'auteur et celui de ses contemporains, de découvrir l'évolution d'écriture d'un auteur, d'un mouvement, d'un concept sur le plan diachronique et synchronique.

Nous pouvons donc implémenter cette contribution en insistant sur la partie interprétative. En effet, et via la lexicométrie, la réception des textes devient active et dynamique de par l'effet de sens obtenu par les lectures numériques conjuguant les trois caractéristiques du texte numérique : l'interactivité, la multilinéarité et le syncrétisme.

Bibliographie

Béhar H., "*Un projet de banque de données d'histoire littéraire*" in *Méthodes quantitatives et informatiques dans l'étude des textes*, Genève, Slaktine, Paris : Champion 1986, p.43-54.

Bernard M, "*La banque de données d'histoire littéraire*", Bulletin de l'EPI, sept. 1988, N° 51, p.172-177.

Bernard M, "*La banque de données d'histoire littéraires*" in *Les banques de données littéraires comparatistes et francophones*, Limoges : Pulim, 1992, p.255.

Bernard M., *Introduction aux études littéraires assistées par ordinateur*, Paris, Puf, 1999. Charaudeau P. , Maingueneau D., *Dictionnaire de l'analyse du discours*,Seuil, 2002.

Lebart et Salem, *Statistique textuelle*, Dunod, Paris, 1994.

Müller C., *Principes et méthodes de la statistique lexicale*, Paris, Champion Collection Uni, 1992.

Müller C., *Initiation aux méthodes de la statistique lexicale*, Paris, Champion Collection Uni, 1992