# METHODES QUANTITATIVE ET INFORMATIQUE DANS L'APPROCHE DES TEXTES: LE LOGICIEL HYPERBASE

**Yasmine ABAS KARA Chargé de cours à l'E.N.S.**

La décision de la mise en place d'un enseignement de l'informatique dans le domaine des lettres et sciences humaines montre l'intérêt que l'on porte à cet outil et surtout les avantages nombreux qu'il apporte aussi bien à l'élève qu'à l'enseignant. Les différentes utilisations pédagogiques de l'ordinateur peuvent être classées en trois grandes catégories:

- la gestion pédagogique (contrôle des absences, confection de bulletins de notes, renseignement sur le milieu socioculturel de l'élève).
- L'enseignement assisté par ordinateur. (Logiciels multimédia utilisés pour l'apprentissage des langues étrangères).
- La recherche.

Dans le domaine de la recherche en lettres et sciences humaines, introduire le langage et les méthodes de la mathématique relève d'un défi surtout lorsqu'elles sont exploitables uniquement par le biais de l'informatique. De nombreux linguistes ont relevé ce défit dont Etienne Brunet. Ce dernier a eu l'audace d'introduire de la rigueur et de la certitude en langue et littérature grâce aux méthodes de la linguistique quantitative et de l'informatique.

La statistique linguistique est née grâce à de nombreux travaux effectués d'abord sur les faits de langue. Un des premiers linguiste à s'intéresser à cette méthode est Pierre Guiraud. Dans son ouvrage, Problèmes et Méthodes, «il expose ses recherches sur le rapport qui pourrait exister entre les emprunts aux langues étrangères et le système phonologique de, la langue emprunteuse ; proposition hardie; puisque sa démonstration tend à prouver une recherche d'équilibre phonologique par des apports lexicaux, à introduire la notion de fréquence dans la description de ce système, et à déceler une finalité.»[(1)] Un tournant décisif va être marqué toujours par P. Guiraud et par R-L Wagner en sollicitant cette fois le texte littéraire, tous les genres sont abordés : le théâtre classique, la poésie et les textes littéraires français du XVIe-XXe siècles. Grâce à ces deux chercheurs et aux collaborateurs formés dans le laboratoire d'analyse lexicologique de Quemada, les procédés mis à l'épreuve se transposeront sur d'autres genres littéraires, d'autres époques et d'autres idiomes.

---

[1] Muller *CH.(1968). Initiation-&/à statistique linguistique, Larousse*, Paris, p6.

A la suite de ces travaux, l'apport de Charles Muller ne sera que très appréciable dans la mesure où il apportera aux linguistes non mathématiques des «connaissances utilisables en pratique» grâce à son ouvrage, Initiation à la statistique linguistique. Il s'agit comme le précise l'auteur, de «présenter les principes et non les résultats de la statistique linguistique, de décrire ses méthodes d'exploration et non ses découvertes ou ses conquêtes» et «de familiariser le linguiste avec le raisonnement statistique, de l'habituer ou de le réhabituer au langage algébrique qui sert de support à ce raisonnement.» (pp.6-5)

Cependant l'accès aux grandes bases de données documentaires et statistiques n'était pas une mince affaire, il deviendra de plus en plus facile et aisé grâce aux procédés informatiques de plus en plus performants, c'est à ce niveau que s'inscrit l'apport de Etienne Brunet. En effet, le logiciel Hyperbase réalisé au CNRS de Nice est utilisé pour créer des bases hypertextuelles, il permet le traitement documentaire et statistique des corpus textuels. Ce logiciel autorise deux types de navigation:

- navigation dans le dictionnaire d'un mot à l'autre d'une page à l'autre du dictionnaire au texte.

- navigation dans le texte. La seconde démarche propre à l'hypertexte est symétrique de la précédente.

Il existe plusieurs versions d'hyperbase (3.1, 4.0 février 1999) dont les fonctionnalités et la présentation diffèrent. Par exemple la dernière version windows offre l'étiquetage et la lemmatisation que l'on ne trouve pas dans la version Mac, car comme le précise cet auteur, très souvent on reproche au traitement documentaire son aveuglement aux faits de syntaxe. Ces remarques sont quelque peu fondées car si l'on s'intéresse qu'aux formes graphiques, on obtiendra des résultats impurs qu'il faille trier par exemple pour distinguer les homographes. Dans ce cas, les chiffres qui prennent en charge ce matériau conduiraient à l'imprécision voire l'erreur. C'est pourquoi « c'est sur le texte que le tri doit se faire et c'est la fonction du lemmatiseur ».

Le menu principal de cette version qui donne accès à plusieurs adresses notamment à la page grammaire, se présente sous cette forme. Nous donnons pour exemple le menu d'hypermammeri :

## LES DIFFÉRENTES FONCTIONS D'HYPERMAMMERI.

Hyperbase permet donc d'accompli un certain nombre de tâches (documentaires et statistiques) à partir d'un corpus préalablement scanné et traité selon les normes établies par Etienne Brunet.

L'exploitation documentaire permet d'assurer deux programmes essentiels CONCORDANCE et CONTEXTE qui obéissent aux même principes et ne se distinguent que par la présentation des résultats.

### LA FONCTION CONTEXTE.

En effet, si l'on met en œuvre le bouton CONTEXTE, chaque occurrence de la lexie recherchée est montrée dans le contexte naturel du paragraphe, cette lexie peut être convertie en capitale dans le paragraphe où elle est rencontrée. Le contexte restitué est très explicite : les références du passage sont livrées avec indication du texte et, de la page, et de la zone dans la page grâce à un code alphabétique. Par contre si l'on fait appel à la fonction CONCORDANCE du menu principal, on obtient un contexte étroit qui tient en une ligne et qui montre la forme ou l'expression cherchée, en position centrale. avec une demi-douzaine de mots à gauche et à droite. La fonction TRIER souligne grâce à une autre présentation la, résurgence de syntagmes répétitif, qui révèlent les tendances phraséologiques de l'auteur.

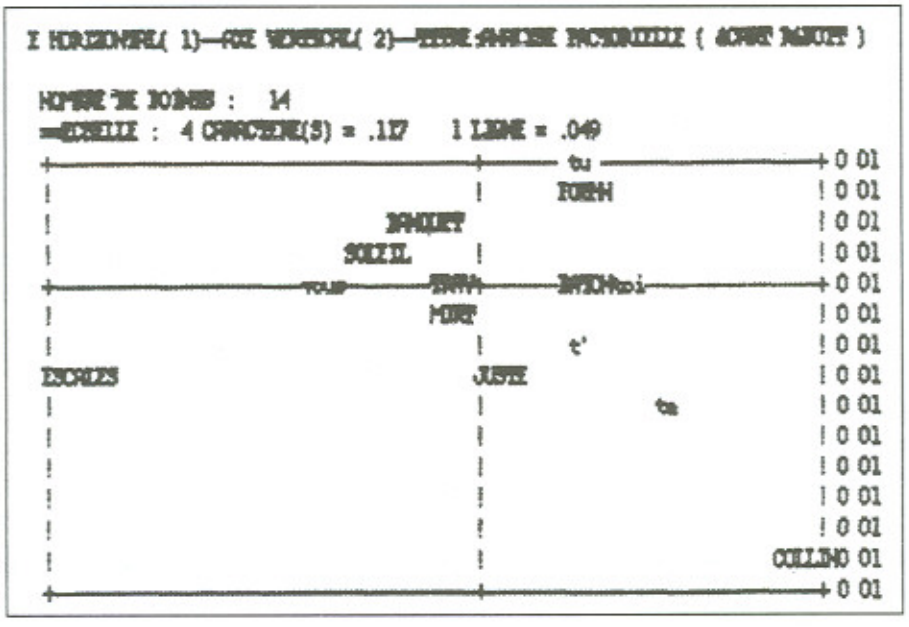
Outre les chiffres, ce logiciel permet l'utilisation des procédures statistiques plus synthétiques tels que des histogrammes et des analyses factorielles. On trouvera ci-dessous un exemple de ces figures réalisées sur le corpus de Mouloud Mammeri concernant la répartition des pronoms personnels (2è` personne).

### REPARTITION INTERNE DE *TU* ET DE *VOUS* DANS LE CORPUS MAMMERIEN.

Les pronoms personnels de la deuxième personne, notamment ceux qui établissent une relation familière, sont en excédent contrairement au pronom «vous». La relation noble est déficitaire sauf dans le théâtre. En effet, si l'on compare nos données à celles du TLF, on remarque que tous les pronoms de la seconde personne se détachent en relief, sans le moindre écart sauf pour vous. Le *tu (2585* occurrences, /73.6)l'emporte sur la forme atone *te* et sur *toi* mais surtout

sur le vous. Les relations familières exprimées par l'utilisation du *tu* dominent dans *Le foehn, L'opium et le bâton* et *Le banquet* alors que le vous est en excédent dans *Le banquet, La cité du soleil, La traversée,* amisi que le confirme l'analyse factorielle ci-dessous.

## ANALYSE FACTORIELLE DES PRONOMS PERSONNELS DE LA 2ème PERSONNE.




La figure ci-dessus confirme que les relations familières dominent dans les œuvres situées sur l'axe vertical, à droite et la relation noble sur l'axe gauche. Cette répartition est en relation avec le genre, nous avons à droite les romans et à gauche le théâtre.

On peut donc affirmer que le diagramme et l'analyse factorielle sont deux éléments complémentaires pour confirmer ce qui pourrait être au départ une hypothèse.

**Que conclure** ?

La statistique linguistique est un précieux auxiliaire dans l'analyse textuelle car elle permet une recherche approfondie. Le corpus issu des bases de données documentaires a l'avantage de posséder une structuration forte dont on peut tirer parti à condition d'avoir une idée précise de ce qu'on cherche, dans le cas contraire, on est vite submergé par une masse d'informations qui risquent de rester hors de notre portée. L'interprétation des résultats nécessite un retour au texte. Grâce à des

lectures plurielles et à des analyses qui prennent en charge la dimension énonciative, le texte s'ouvre, se transforme, se renouvelle et renaît de ses cendres.

**Références bibliographiques.**

**Abbés-Kara, AY.** (Février 2000). *Etude lexicologique, stylistique et pragmatique de l'oeuvre de Mouloud Mammeri*, Thèse de doctorat. Université de Sophia Antipolis. Nice.

**Brunet E.** (1978). *Le vocabulaire de Giraudoux. Structure et évolution* Slatkine, Genève. *(1982). Sodome et Gomorrhe*, édition critique, in Giraudoux,Théâtre, vol. 1, collection de la Pléiades, Gallimard.

(1980). *Index-concordance d'Emile ou de l'éducation.* Slatkine, Genève, tome 1, 585 p., tome 2.

(1981). *Le vocabulaire français de 1789* à *nos jours.* Slatkine

Champion, Genève Paris, tome 1, XIV-852 tome3,454p. Préface de Paul IMBS.

(1983). *Index des lettres écrites de la montagne, de J.J. Rousseau.*

*Slatkine-Champion,* Genève Paris, 364p. Postface de Nfichel Launay

(1983). *Le vocabulaire de Marcel Proust, avec l'indexe complet et synoptique de A la recherche du temps perdu.* Slatkine Champion, Genève Paris, tome 1, VI-261 p, tome 2 et 3, 1644 p. Préface de J.Y.Tadié.

(1985). *Le vocabulaire de Zola, avec l'index complet et synoptique des Rougon-Macquart.* Slatkine-Champion, Genève-

Paris, tome 1, étude quantitative, VI-472p., tome 2, Dictionnaire des fréquences, 646p., tome 3, index de Germinal et des Rougon Macquart, 357p. et 5500 p. sur microfiches. Préface de Henri Mitterrand.

(1985). *Hommage à Pierre Guiraud (ouvrage collectif).* Editions Les Belles Lettres, Paris, 343p.

(1986). Méthodes quantitatives et informatiques dans l'étude des textes (ouvrages collectif en hommage à Charles Muller).

Slatkine-Champion, Genève-Paris, 2 vol., 948p.

(1986). *Index des Considérations sur le gouvernement de Pologne et du Projet de constitution pour la Corse, de J.J. Rousseau.* Slatkine-Champion. Genève-Paris, 3177 p. En collaboration avec Léo Launay

(1986). *Index de l'oeuvre théâtrale et lyrique de JJ. Rousseau. Slatkine-Champion,* Genève-Paris, 507p. En collaboration avec Annick et Gilbert Fauconnier.

Muller, C.(1968). *Initiation* à *la statistique linguistique*, Paris, Larousse.