Traitement instrumenté de corpus

Par **Hakim HESSAS**
Université ALGER 2

«Il ne viendrait à personne l'idée de publier une étude sur la population d'une ville ou sur les importations d'un pays en s'interdisant tout appel aux données quantitatives. Cela ne signifie certes pas que l'auteur d'une telle étude doive entreprendre de compter lui-même les habitants de la ville ou les marchandises qui passent les frontières du pays : l'état civil ou la douane se seront chargés de ces recensements et lui fourniront leurs statistiques détaillées»[149].

Résumé*:*
Si le sens est fait de différences, ce n'est pas pour contraster des mots ou des textes de genres ou de discours considérablement éloignés les uns des autres. Cependant, pour constituer des parcours d'interprétation intéressants, il faudrait d'abord constituer un corpus d'étude représentatif d'un problème déterminé, préalablement posé, à partir de textes aussi homogènes que possible et d'un point de vue bien clair. Dans le présent article, nous insisterons sur ces points qui revêtent une portée épistémologique importante pour toute recherche linguistique, sociologique ou autre.

1. Introduction*:*

Il est impossible de rendre compte d'un mot isolé; n'étant pas universel, un mot ne se laisse comprendre que par ses multiples emplois et ses rapports divers à d'autres mots (contexte et récurrence), en tenant compte de facteurs comme la pratique sociale, le discours, etc. Le prendre loin de son contexte, écrivait déjà Saussure[150], c'est tomber dans la figure vocale qui est du ressort de la physiologie et de l'acoustique. Car l'on sait parfaitement que de nombreuses formes de sons identiques (par exemple la donnée *pêche*) peuvent revêtir des sens différents :

-aller à la *pêche*,

-manger une *pêche*,

-avoir la *pêche*,

-recevoir une *pêche*, etc..

D'où l'importance des textes qu'il est possible d'organiser en corpus homogène, puisque tout texte relève d'un corpus. Cependant, une fois ce dernier bien constitué, pour saisir ses singularités, il doit d'abord être quantifié, une tâche que permettent aisément les logiciels d'interrogation: *accéder rapidement et facilement au corpus*, *rechercher des cooccurrents statistiques, suggérer des points d'entrée, etc*. Se rapportant à la philologie, ce premier mouvement constitue une étape nécessaire de l'analyse de grands corpus numériques, que cet article va tenter de montrer. Le second mouvement que nous n'aborderons pas ici, relève de l'herméneutique.

2. Linguistique et corpus*:*

Le traitement de corpus de textes par des logiciels a renouvelé de fond en comble la méthodologie de la recherche linguistique et des disciplines des sciences humaines, depuis l'accès numérique à l'écrit. La révolution tient principalement en deux points: le fait de travailler sur des corpus de textes numériques, au lieu des exemples forgés intuitivement par les

[149]. Muller Charles, *La statistique lexicale*, *Langue française*, 1969, n° 1, pp. 30-43, url : http://www.persee.fr.
[150] . Ferdinand de Saussure, *Écrits de linguistique générale*, Édités par S. Bouquet et R. Engler, Paris, Gallimard, 2002, p. 31.

linguistes[151]; l'assistance qu'apportent les logiciels pour le traitement des données statistiques recueillies.

Cette nouvelle approche de l'objet ne se fait pas sans se heurter à une difficulté : l'immensité et la variété des données qui peuvent être collectées, contrairement à celles que l'on peut récolter de la simple lecture linéaire. Par conséquent, il nous semble essentiel d'insister sur le fait que le traitement instrumenté de corpus n'est pas le traitement de corpus dans son entiéreté; il en constitue une partie nécessaire, mais non suffisante, car celle-ci doit être complété, à chaque fois, par une autre, aussi importante, à savoir l'interprétation des données: le passage des chaînes de caractères aux formes sémantiques, ou la représentation de la qualification des données en corrélats sémantiques.

Notons seulement que l'on se donne un corpus de textes comme objet scientifiquement traitable en le construisant en vue d'en constituer des données. On ne cherche pas à compiler des données pour éviter de lire les textes auxquels elles se rapportent, mais, au contraire, pour pouvoir mieux les lire et les interpréter, en construisant des parcours thématiques précis, à partir d'un corps d'hypothèses qui peuvent être construites lors de la conception de corpus – ou même avant, lors de la détermination des tâches. Ces hypothèses déterminent même le choix des textes qui entrent ensuite dans la composition du corpus. Comme il est impossible d'étudier tous les observables produits dans un corpus, il est préférable d'interroger des constructions particulières, à partir d'un point de vue déterminé.

Par conséquent, la manière d'aborder l'objet est cruciale. Nous aimerions donc insister sur quelques points qui revêtent une portée épistémologique importante pour toute recherche linguistique ou de sciences sociales et humaines: l'importance du corpus de textes; la constitution de corpus; l'importance du point de vue pour la constitution de corpus ; l'intérêt de la spécification des genres; l'efficacité des logiciels.

3. L'importance du corpus de textes:

La langue envisagée en elle-même et pour elle-même : tel est l'objet assigné à la linguistique depuis le *Cours de Linguistique Générale* (CLG) faussement attribué à Saussure. Ce point de vue a situé la linguistique dans un cadre restreint, autour de la langue et du système, négligeant par là la portée des corpus de textes. Dans le dictionnaire raisonné de la théorie du langage de A.J. Greimas et J. Courtés, on peut lire que « […] *d'une théorie à l'autre, […] le concept de parole a cessé, aujourd'hui, d'être opératoire.* »[152]

Si l'on insiste sur l'importance du corpus – non des mots ou des exemples confectionnés –, c'est pour signifier aussi celle des textes et en conséquence de la *parole* au sens saussurien du terme :

> «*force active et origine véritable des phénomènes qui s'aperçoivent ensuite dans l'autre moitié du langage [c'est-à-dire, la langue*]»[153].

La parole a longtemps été oubliée par une linguistique de la langue vue comme un système de signes fermé[154], négligeant par là même le caractère essentiellement culturel des langues, en oubliant également que la langue est diverse par la diversité des discours, des genres et des pratiques sociales.

---

[151] . En construisant ses exemples, le linguiste risque de construire avec eux les phénomènes qu'il recherche. De même, en plus de détacher les mots de leur contexte, il crée d'autres phénomènes qu'il qualifie ensuite par l'ambigüité, la polysémie, etc.

[152] . Greimas, A. J. & Courtés, J., *Sémiotique. Dictionnaire raisonné de la théorie du langage*, Hachette, Paris, 1993, p. 269.

[153] . Ferdinand de Saussure, *Écrits de linguistique générale*, *op.cit.*, p. 273.

[154] . Pour s'en convaincre, il suffit de lire dans le CLG attribué faussement à Saussure : « la langue en elle-même et pour elle-même».

Les manuscrits authentiques de Saussure, retrouvés en 1996, le montrent aujourd'hui amplement. En plus du système, il est indispensable de tenir compte non seulement du texte, mais des textes, objet empirique de la linguistique; il n'y a pas d'un côté la *langue* et de l'autre la *parole* et que la première préexiste à la seconde. La langue ne tombe pas du ciel; elle est d'abord produite dans des pratiques sociales diverses[155], reconstituée ensuite à partir de l'observation rationnelle des régularités qui se trouvent dans les corpus.

Pour Saussure, tout ce qui s'observe dans la langue a d'abord été expérimenté dans la parole, non par anticipation ou par préméditation, mais par improvisation. C'est ce qu'il affirme, par exemple, à propos des créations analogiques:

> *«Il faut <donc> se mettre en face de l'acte de parole pour comprendre <une> création analogique. La nouvelle forme ne se crée pas dans une assemblée de savants discutant sur le dictionnaire.*»[156]

Il faut donc rompre avec la langue comme système de signes pour l'envisager dans une dualité qui intègre les corpus de textes écrits ou oraux qui la caractérisent à chaque période historique. Elle ne saurait être autre chose que ce qui a été produit et confirmé pleinement dans la parole – et ses autres manifestations à savoir les textes et les corpus.

> *«Pour l'essentiel, une langue repose sur la dualité entre un système (condition nécessaire mais non suffisante pour produire et interpréter des textes) et des corpus de textes écrits ou oraux[157]. Non contradictoire, la dualité dynamique entre corpus et système constitue la langue dans son histoire.*»[158]

La langue ne se réduit pas à un trésor de mots, comme on les retrouve dans les dictionnaires, même s'ils demeurent nécessaires pour parler et écrire. De nombreuses disciplines, comme la psycholinguistique cognitive, prennent les mots (isolés) comme objet principal de leur investigation et malheureusement ne vont pas au-delà. Mais tout ce qui entre dans la parole (toujours au sens large) a besoin de son contexte pour être lu, compris et interprété. Comme pour le mot qui n'a de sens que dans un contexte, qui doit être tout le texte, le texte n'a de sens que dans un *corpus*, un «*regroupement structuré de textes entiers*».

4. La constitution de corpus:

Un mot n'a de sens que dans un contexte bien déterminé; son signifié, toujours particulier, se rapporte toujours à une langue, à un texte et à un genre particulier. *Le sens[159] est un réseau structuré de traits sémantiques*[160]; il est fait de différences entre les mots de la même classe. S'il est donc nécessaire d'avoir un contexte plus large pour saisir le sens d'une unité linguistique, l'objectif principal n'est pas d'obtenir une grande masse de textes (*du texte*) sans aucun critère de détermination; l'importance se trouve dans la qualité et la typologie des textes rassemblés (*des textes*) pour pouvoir les contraster, ce que l'on désigne par le concept de *corpus*.

Le corpus doit être construit à partir d'un point de vue bien clair, qui correspond à un choix théorique et méthodologique en vue d'une exploitation déterminée. Un corpus construit en fonction d'une problématique ne peut convenir véritablement à une autre.

---

155 . Toute production écrite ou orale procède d'un genre, d'un discours et d'une pratique sociale.

156 . (Cours I, Riedlinger, B, pp. 63-64)

157 . Dans le corpus d'une langue, les œuvres tiennent une place particulière parce qu'elles sont hautement valorisées : par exemple, l'italien est la langue de Dante au sens où son œuvre demeure le parangon historique qui a présidé à la formation de la langue italienne en tant que langue de culture.

158 . François Rastier. *La Mesure et le Grain. Sémantique de corpus*, Editions Honoré Champion, coll. "Lettres numériques" n°12, Paris, 2011, p. 14.

159 . Le sens d'un mot (ou d'un texte) n'est pas mathématiquement déterminable, en s'appuyant sur des règles figées, il est construit. Il n'est pas déductif, il est progressivement organisé par la découverte grandissante de ce que l'on appelle la logique des faits.

160 . Un trait sémantique (ou un sème) est ce qui note une différence entre deux mots. Il est toujours déterminé par un contexte et ensuite confirmé par des récurrences.

Ce serait une erreur de penser qu'il existe des corpus *naturels*. Ce qui existe ce sont des banques textuelles, des œuvres complètes d'un auteur, des textes disparates dans le Web, etc., qui ne peuvent, en tant que tels, être considérés comme des corpus en vue de l'analyse sémantique ou autre. Pour F. Rastier:

> *«Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés: (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications*.»[161]

Si l'on cherche des critères qui permettent de caractériser des corpus, F. Rastier[162] en a défini quatre : la *représentativité*, l'*homogénéité*, la f*ermeture*, et l'*entretien*.

La *représentativité* est la qualité d'un regroupement de textes constitués en corpus de façon à représenter un problème donné. Le corpus construit, qui n'est pas unique, est bien entendu orienté, puisqu'il est influencé par un corps d'hypothèses. E. Brunet écrit à ce propos:

> «*Tout d'abord un corpus est toujours artificiel. La nature n'en produit pas spontanément. C'est une création nécessairement subjective. Pire encore, la création est orientée, conditionnée par une hypothèse, par un objectif de recherche. Quelques précautions qu'on prenne pour affiner les critères de sélection, pour les justifier et pour les appliquer, il y a toujours des choix à décider, des doutes à faire taire, des contraintes à respecter, des compromis à négocier, un ordre à établir, un terminus ab quo, un autre ad quem à délimiter*.»[163]

L'*homogénéité* est la propriété d'un ensemble de textes, rassemblés en corpus pour représenter un problème déterminé, qui constitue en quelque sorte une unité. Pour Greimas et Courtès,

> *« […] l'homogénéité pourra être fondée sur un choix d'éléments de même niveau, de relations de même type (Hjelmslev).»*[164]

En fonction des objectifs de la recherche, celle-ci peut se réaliser, généralement, au niveau des genres ou, spécifiquement, au niveau des discours, comme dans le cas des recherches sur le système de la langue. L'on sait pertinemment, depuis les *Écrits* de Saussure, que la langue est variable diachroniquement et synchroniquement, d'un genre à un autre, comme ils «rivalisent» dans un champ générique ou dans un discours. Ainsi, dans le discours littéraire, par exemple, constitué du théâtre (comédie, tragédie, drame), du récit (roman, nouvelle), de la poésie, l'hétérogénéité des textes n'est pas difficile à observer. Donc, si le sens est fait de différences et que l'objectif de mettre ensemble des textes lors de la constitution de corpus est de pouvoir les contraster, ce n'est pas pour comparer l'incomparable, c'est-à-dire des genres ou des discours aussi éloignés les unes des autres.

4.1. L'importance du point de vue:

Si l'importance du corpus de textes est incontestable aujourd'hui en linguistique, sa construction doit partir d'un point de vue bien clair, dès le départ. Pour sa constitution comme, ensuite, pour son analyse, la question du point de vue est centrale.

Un point de vue n'est pas un simple point à partir duquel on tient à l'œil un ensemble de textes, mais un point à partir duquel on construit un corpus comme objet scientifique exploitable. Celui-ci n'est évidemment pas unique, comme le corpus de textes construit. Les bases de données, par exemple, comme tout autre ensemble de textes, n'ont guère l'initiative étant donné qu'ils ne

---

[161] . Rastier François, *Enjeux épistémologiques de la linguistique de corpus*. *Op. Cit.*
[162] . Rastier François, *Arts et sciences du texte*, Presses Universitaires de France, 2001, p. 86.
[163] . Brunet Étienne, *Ce qui compte. Méthodes statistiques. Écrits choisis, tome II*., Éditions Champion, Paris, 2011, p. 279.
[164] . Greimas, A. J. & Courtés, J., *Sémiotique. Dictionnaire raisonné de la théorie du langage, op.cit.*, p. 174.

fournissent des données que si on les interroge. Après avoir posé le problème, la question permet de chercher des relations entre les différents aspects des faits qui peuvent être observés.

> *«Pour un esprit scientifique, toute connaissance est une réponse à une question. S'il n'y a pas eu de question, il ne peut y avoir connaissance scientifique*.»[165]

On ne peut saisir complètement un fait de langage qu'en le considérant sous divers points de vue. C'est la multiplication des points de vue qui fait le fait de langage, comme l'affirme Saussure:

> *«Primordialement il existe des points de vue; sinon il est simplement impossible de saisir un fait de langage.»*[166] Mais encore*: «Rappelons-nous en effet que l'objet en linguistique n'existe pas pour commencer, n'est pas déterminé en lui-même. Dès lors parler d'un objet, nommer un objet, ce n'est pas autre chose que d'invoquer un point de vue A déterminé.»*[167]

Le point de vue, qui n'est pas unique, fait donc le corpus et le sous-corpus, et toutes les analyses et tous les parcours qui peuvent être menés sur lui doivent correspondre intimement à ce point de vue.

4.2. Spécification des genres:

On l'a bien vu, l'homogénéité est un critère important pour l'élaboration des corpus. Parmi les quatre niveaux de classification des textes répertoriés (c'est-à-dire, le discours, le champ générique, le genre et le sous-genre) ; le genre étant celui qui est retenu pour la caractérisation des textes, et non le discours. Ainsi, dans toute description linguistique, il est nécessaire de spécifier et de définir les genres[168] lors de la constitution de corpus, car la langue qui les caractérise est à chaque fois particulière. D'ailleurs, l'idée d'une langue générale est un leurre.

Il faut garder à l'esprit que la meilleure interprétation se fait au sein d'un corpus de textes qui partagent le même genre[169]. Cette homogénéisation permet d'écarter les ambigüités qui peuvent surgir lors de l'interprétation. L'on sait également qu'il existe des habitudes stylistiques, sémantiques, propres à chaque genre de texte qu'il est alors possible de comparer : vocabulaire, structure du texte, etc. De plus, les particularités spécifiques à chaque genre deviennent saillantes.

5. Traitement instrumenté de corpus:

Une rupture épistémologique s'est opérée en linguistique et dans toutes les disciplines des sciences humaines, par l'utilisation des méthodes statistiques et des logiciels pour traiter un nombre important de données. Néanmoins, si le sens est fait de différences, il demeure une construction – il n'est ni caché, comme le stipulent de nombreux auteurs, ni donné :

> *«Rien ne va de soi. Rien n'est donné. Tout est construit»*,

écrivait G. Bachelard dans *La formation de l'esprit scientifique*[170].

En conséquence, le traitement instrumenté de corpus n'est pas le traitement de corpus; il est nécessaire, notamment dans le cas de corpus de grandes masses, mais non suffisant. Si

---

[165] . Gaston Bachelard, *La formation de l'esprit scientifique. Contribution à une psychanalyse de la connaissance objective*, Librairie Philosophique J. VRIN, 1989, p.14.
[166] . Ferdinand de Saussure, *Écrits de linguistique générale*, *op. cit.* p. 19.
[167] . *Ibid*., p. 23.
[168] . On pense par exemple au travail de C. Muller sur les 32 pièces de théâtre de Corneille classées par genre (18 tragédies, 8 comédies, 6 pièces diverses).
[169] . Un corpus de champ générique (un groupe de genres qui rivalisent dans une pratique sociale) est un corpus non pertinent. Un corpus constitué par de textes incomplets est également inapproprié.
[170] . Gaston Bachelard, *La formation de l'esprit scientifique. Contribution à une psychanalyse de la connaissance objective*, *op. Cit*. p. 14.

l'interprétation se suffit parfois à elle-même, la *mesure* ne le pourrait pas: les chaînes de caractères que délivrent les logiciels, comme les mots, les signes de ponctuation, les paragraphes, etc. doivent être interprétées. *S'il est parfois possible d'interpréter sans mesurer, l'inverse n'est vrai.*

La mesure demeure cependant nécessaire pour indiquer des parcours interprétatifs intéressants, vérifier des hypothèses, faire une lecture rapide du corpus, repérer des zones textuelles thématiques (zones d'activités ou de changement thématiques), proposer des mots-candidats; les logiciels, par les différents calculs qu'ils opèrent, permettent de choisir des points d'entrée dans le corpus qu'il n'est guère possible de faire par la lecture linéaire – un point d'entrée intéressant est celui qui est solidement connecté à d'autres points (corrélations). Les points d'entrée dans des corpus, doivent être, à chaque fois, définis. Saussure affirme à propos des recherches linguistiques:

> *«Nous nous permettrons de remettre, jusqu'à trois et quatre fois sous différentes formes, la même idée sous les yeux du lecteur, parce qu'il n'existe réellement aucun point de départ plus indiqué qu'un autre pour y fonder la démonstration.»*[171]

Les analyses statistiques opérées sur des corpus numériques permettent de saisir la stéréotypie textuelle, de chercher les singularités dispersées dans le corpus, d'interroger le corpus sur de nombreux points, ce qui permet d'objectiver l'analyse.

Par l'exploration possible des corpus, les logiciels permettent de trouver ce que l'on ne cherche pas. La linguistique de corpus possède ainsi une valeur heuristique; elle permet de faire des découvertes. Comme première approche, il s'agit donc de soumettre le corpus à un traitement statistique par ordinateur[172] en vue d'en proposer une caractérisation globale. Au moyen de différentes techniques statistiques implantées dans des logiciels tel Hyperbase, les analyses chiffrées réalisées peuvent tenir lieu d'enquêtes et d'expérimentations sur la totalité du corpus, afin de rendre compte, synchroniquement et diachroniquement, des relations entre les éléments et les ensembles qui structurent l'ensemble des textes qui le composent (ordres, distances, rapprochements, ruptures, etc.).

Développé par E. Brunet, le logiciel Hyperbase permet aisément de tels traitements statistiques de corpus, pour son efficacité prouvée sur de nombreux types de corpus, pour les différents niveaux d'analyses sur lesquels il opère, pour ses diverses fonctions statistiques (spécificités, distribution, concordance, contexte, etc.) et graphiques (histogramme, factorielle, arborée, etc.) :

> *«Dans une première approche Hyperbase suit la méthode Jaccard qui ne se préoccupe pas de fréquences et pour un mot donné ne considère que sa présence – ou son absence – dans le texte considéré. Ou plus exactement, pour deux textes dont on cherche à apprécier la connexion, un mot contribue à rapprocher ces deux textes s'il est commun aux deux et à augmenter la distance s'il est privatif et ne se rencontre que dans un seul»*[173].
>
> «*Afin de situer le corpus par rapport à d'autres écrits, Hyperbase permet seul de calculer les spécificités externes du corpus par rapport à celui de Frantext ou à un découpage chronologique au sein de celui-ci.»*[174]

---

[171] Notes pour un livre sur la linguistique générale (Ms.fr.3951/9), présentation et édition : Kazuhiro Matsuzawa, (p. 319-322), *Saussure, Cahier de l'Herne*, Paris, Éditions de l'Herne, 2003.

[172] . «L'ordinateur, comme le dit Hubert de Phalèse, peut […] permettre de pratiquer dans les disciplines littéraires une véritable méthode expérimentale» (Hubert de Phalèse, *Les mots de Molière. Les quatre dernières pièces à travers les nouvelles technologies*, 1992, Pris, Nizet, p 6, cité par Sylvie Mollet et Marcel Vuillaume, *Mots chiffrés et déchiffrés*, Éditions Champion, Paris, 1998, p. 83.)

[173] . Étienne Brunet, *le corpus conçu comme une boule*, p. 70, http://www.revue-texto.net/1996-2007/Parutions/Livres-E/Albi-2006/Brunet.pdf.

[174] . **Emmanuel** Bonin **et Alain** Dallo, « Hyperbase et Lexico3, outils lexicométriques pour l'historien », *Histoire & mesure*, vol. XVIII – n°3/4, 2003, [En ligne], mis en ligne le 03 avril 2007. URL http://histoiremesure.revues.org/index840.html. Consulté le 26 avril 2009.

En partant d'hypothèses plus ou moins précises, il est possible de saisir, à différents niveaux, certaines spécificités du texte — ou d'une partie du texte —, dans ce qu'il met en avant, ou en arrière lorsqu'il évite, oublie ou au contraire emploie des formes textuelles bien déterminées. En premier lieu, l'on peut décrire le corpus en identifiant des ensembles plus au moins semblables, en recourant à l'analyse factorielle[175]. C'est particulièrement dans la phase exploratoire de départ que cette méthode se présente comme une technique pratique et efficace[176]; elle permet de saisir, de classer les différentes catégories du texte en les posant et les opposant les unes aux autres sur chacun des axes factoriels. Parce qu'elle fait abstraction du contenu des textes, cette étape sera essentiellement formelle.

Mais il ne faut pas perdre de vue que l'on travaille sur des textes, c'est-à-dire sur des éléments concrets. Cela requiert *ipso facto* la lecture, la description et l'interprétation des données. Dans ce cas, l'analyse travaillera essentiellement sur les éléments du lexique, et à chaque fois que cela est nécessaire elle combinera entre le contenu et les ensembles ou les formes identifiées.

> *«D'une part, il est nécessaire de découper des unités dans la chaîne textuelle pour réaliser des comptages utilisables par les analyses statistiques ultérieures. De l'autre, la chaîne textuelle ne peut être réduite à une succession d'unités n'ayant aucun lien les unes avec les autres, car beaucoup des effets de sens du texte résultent justement de la disposition relative des formes, de leurs juxtapositions ou de leurs cooccurrences éventuelles.»* (V. LEBART et SALEM)[177]

6. Conclusion:

La force théorique de la linguistique de corpus, comme on vient de le voir, ne consiste pas à offrir des instruments qui permettent de résoudre toutes les énigmes textuelles, mais de pouvoir donner à des hypothèses pertinemment posées des possibilités de confirmation ou d'infirmation.

Si la mesure ne se suffit pas à elle-même, l'interprétation des données ne peut se faire que si l'on prend connaissance du discours, du genre du texte en question, voire de la pratique sociale. La grammaticalité même d'une phrase suit les mêmes principes.

Les statistiques que permettent les logiciels n'ont donc pas de sens et de portée en elles-mêmes ; elles doivent être exploitées et surtout interprétées.

Références Bibliographiques:

1) BENVENISTE Émile., 1966*: Problèmes de linguistique générale*, Gallimard, Paris, t. I.
2) BENZECRI Jean-Paul., 1982*: Histoire et préhistoire de l'analyse des données*, Dunod.
3) BRUNET Étienne., 2011*: Ce qui compte. Méthodes statistiques. Écrits choisis*, tome II., Éditions Champion, Paris,
4) BRUNET **Étienne:** «Peut-on mesurer la distance entre deux textes?», *Corpus [En ligne], 2décembre 2003*, mis en ligne le 15 décembre 2004, Consulté le 17 octobre 2012. *http://corpus.revues.org/index30.html).*
5) BRUNET Étienne., 1981*: Le vocabulaire français de 1789 à nos jours*, Genève-Paris, Slatkine-Champion, 3 vol).
6) CANGUILHEM Georges., 1980*: La connaissance de la vie*, Librairie Philosophique J. Vrin.
7) CAVADONGA Lopez Alonso et ARLETTE Séré de Olmos*: Où en est la linguistique - Entretiens avec des linguistes*, Paris, Didier Érudition.

[175] . Cette méthode est aussi appelée analyse factorielle discriminante, comme de nombreuses autres méthodes de ce genre; elles sont dites aussi de *discrimination linéaire* par opposition à *discrimination globale*. (Ludovic Lebart & André Salem, *Statistique textuelle*, Editions Dunod, 1994, pp. 241-242)

[176] . Histoire & Mesure, 1997, XII-3-4, 271-298, Félicité des Nétumières. *Méthode de régression et analyse factorielle*, CNRS, Paris, 1997, p. 276.

[177] . Ludovic Lebart & André Salem, *Statistique textuelle*, Préface de Christian Baudelot, Éditions Dunod, 1994, p. 8.

8) DUCROT Oswald, SCHAEFFER Jean-Marie *: Dictionnaire encyclopédique des sciences du langage,* Point 1991.
9) GREIMAS Algirdas Julien, COURTES Joseph., 1993*: Sémiotique. Dictionnaire raisonné de la théorie du langage,* Paris, Ed. Hachette,
10) HARRIS Zellig, S. Dubois, CHARLIER Françoise*: Analyse du discours. In: Langages,* 4e année, n° 13. *L'analyse du discours.* pp. 8-45. http://www.persee.fr/web/revues/home/prescript/article/lgge_0458726x_1969_num_4_13_2507.
11) HJELMSLEV Louis., 1971*: Prolégomènes à une théorie du langage,* Éd de Minuit, Paris.
12) KYHENG Rossitza, «Hjelmslev et le concept de texte en linguistique». *In Texto [enligne],* septembre2005, vol. X, n°3. Disponible sur : <http://www.revue texto.net/Inedits/Kyheng/Kyheng_Hjelmslev.html>.
13) LEBART Ludovic et SALEM André., 1994*: Statistique textuelle, Préface de Christian* Baudelot, Dunod.
14) *MAYAFFRE Damon, « Analyse du discours politique et Logométrie : point de vue pratique et théorique », Langage et société,* 114 (2005) pp 91-121.
15) MOLLET Sylvie et VUILLAUME Marcel., 1998 *: Mots chiffrés et déchiffrés,* Éd Champion, Paris.
16) MULLER Charles 1969 : *La statistique lexicale, Langue française, n°* 1, pp. 30-43, url : http://www.persee.fr.
17) MULLER Charles., 1993 *Étude de statistique lexicale. Le Vocabulaire du Théâtre de Pierre Corneille,* Slatkine Reprints, Genève.
18) PINCEMIN Bénédicte 2012: « Sémantique interprétative et textométrie *», Texto!* Volume XVII, n°3, coordonné par Christophe Cusimano.
19) RASTIER François, 1987*: Sémantique interprétative,* Presses Universitaires de France.
20) RASTIER François., 2004*: Enjeux épistémologiques de la linguistique de corpus. Texto ! [en ligne],* Rubrique Dits et inédits. Disponible sur : <http://www.revue texto.net/Inedits/Rastier/Rastier_Enjeux.html>. (Consultée le ...).
21) **RASTIER François.2005***: Discours et texte. Texto ![en ligne]*. Disponible sur: <http://www.revue-texto.net/Reperes/Themes/Rastier_Discours.html>.
22) RASTIER François., 2001*: Arts et sciences du texte,* Presses Universitaires de France,
23) RASTIER François., 2011*: La Mesure et le Grain. Sémantique de corpus,* Editions Honoré Champion, coll. "Lettres numériques" n°12, Paris.
24) SAUSSURE Ferdinand de., 2002*: Écrits de linguistique générale.* Établis et édités par S. Bouquet et R. Engler. Paris, Gallimard,
25) SIMIAND François., 1922*: Statistique et expérience, remarques de méthode,* M. Rivière, Paris,
26) THOUARD Denis., 2011*: Herméneutique contemporaine. Comprendre, interpréter, connaître,* Paris, Vrin, «Textes clés»