Hakim Hessas
Université d'Alger 2

L'importance des corpus en linguistique

« Connaître c'est analyser. […] Si la connaissance est analyse, ce n'est tout de même pas pour en rester là. Décomposer, réduire, expliquer, identifier, mesurer, mettre en équations […] » [1]

Résumé

De nombreuses disciplines ont accepté la nécessité de travailler sur des corpus numériques. Pour les traiter et tirer parti des rapprochements de textes qui les constituent à chaque fois, les logiciels s'avèrent souvent indispensables et efficaces. Cependant, le chercheur se heurte souvent à des obstacles non négligeables : comment constituer ces corpus ? Et qu'est-ce qu'un corpus ? Une fois ces textes constitués et le logiciel bien choisi, les chaînes de caractères produites par ce dernier, ne pouvant pas seuls signifier, doivent être qualifiées par une indispensable interprétation sémantique. L'objectif de cet article est d'éclaircir ces points nodaux, de tenter de préciser le lien indispensable entre les méthodes quantitatives et qualitatives.

ملخص

لقد صار استعمال المدوّنات الرقمية أمرا مقبولا في حقول معرفية عديدة. و نحتاج لمعالجتها و الاستفادة منها استفادة فعّالة إلى برامج حاسوبية خاصة. و يتواجه الباحث في ذبك صعوبات جامة تتعلّق بتكوين هذه المدونات و ماهيتها و اختيار البرنامج المناسب بها. وبعد ذلك ينبغي أن نترجم نتائج التحليل الرقمي إلى دلالات لغوية, نريد في هذه الدراسة أن نعالج هذه النقاط المعضلة مع إبراز العلاقة الوطيدة التي تربط المنهجيات الكمية بالمنهجيات النوعية.

1 - Georges Canguilhem, La connaissance de la vie, Librairie Philosophique J. Vrin, 1980, p. 9.

1. Introduction :

Dans cet article, notre propos n'est pas de mettre en avant la seule pertinence des approches quantitatives, ni celle de la sémantique des textes par la construction de parcours interprétatifs, mais de préciser leur imbrication et leur interdépendance théorique et méthodologique dans les études descriptives et analytiques des phénomènes linguistiques.

Aucune discipline n'a la priorité sur une autre, du moment qu'elles prennent comme objet l'étude des textes oraux ou écrits2, dans une perspective qui met en avant l'importance de travailler sur des corpus numériques. Toutefois, si l'on admet que l'écriture d'un texte se fait à partir d'une langue, mais aussi en rapport à un genre, à un discours et à une pratique sociale, ce n'est pas pour les écarter lors de son interprétation. Autrement dit, l'interprétation d'un texte ne peut se faire judicieusement sans la connaissance de son genre, du discours auquel il appartient et de la pratique sociale qui le rend possible, car le système de la langue n'est pas constant en tout lieu, mais variable d'une pratique à une autre et donc des genres et des discours qu'elles renferment à chaque fois3.

Jusqu'à présent, de nombreux logiciels textométriques ont été développés dans le but d'assister les analyses de textes et corpus numériques. Mais, comme le suggère ce mot « assister », cette approche quantitative demeure insuffisante, car il faudrait pouvoir interpréter les résultats obtenus, c'est-à-dire donner un sens aux chaîn-

2 - F. Rastier, notamment dans Art et sciences du texte (2001), appelle à un remembrement des disciplines du texte (la rhétorique, l'herméneutique, la poétique, etc.).

3 - Chaque pratique sociale possède ses genres propres.

es de caractères recueillies, par la construction de parcours et des interprétations qualitatives. Après celle de la constitution de corpus, l'interprétabilité des résultats constitue la seconde difficulté de la linguistique de corpus.

2. L'importance des corpus :

La discipline linguistique est restée longtemps impuissante à rendre compte de certains phénomènes textuels, comme la solidarité entre les deux plans du langage (signifiant et signifié), que l'on nomme la sémiosis textuelle, et cela, en dépit des grands et divers progrès réalisés depuis Saussure, avec notamment Hjelmslev, mais aussi Greimas, Jakobson, Barthes, etc. L'obstacle épistémologique à ces stagnations se trouve dans l'objet d'étude lui-même ; on étudie le texte – lorsque l'on ne se limite pas aux signes isolés, aux exemples forgés – comme s'il était une grande phrase. Ainsi, on extrapole au premier (au texte) toutes les observations faites sur le second (la phrase).

En donnant la primauté à la phrase comme limite supérieure, la linguistique s'est désintéressée des textes et donc des corpus. Émile Benveniste avait constaté à plus forte raison que les notions de sémantique de l'époque étaient « vagues », « floues » et « inconsistantes » pour comprendre le « fonctionnement du sens dans la langue »4 ; dans cette perspective, disait-il, le sens était alors « fuyant » et « imprévisible ». Néanmoins, dans Problèmes de linguistique générale, au chapitre X consacré au niveau de l'analyse linguistique, il décrit la phrase comme une totalité : « Une phrase, écrit-il, constitue un tout, qui ne se réduit pas à la somme de ses parties […] »5. Mais encore, la phrase constitue « l'unité la plus haute »6. Pour lui, « […] l'unité sémiotique est le signe » et « […]

4 - On pensait d'ailleurs, écrivait Benveniste, que le domaine du sens était la propriété exclusive des psychologues et psycho-physiologistes (p. 216).

5 - Émile Benveniste, Problèmes de linguistique générale, Gallimard, Paris, t. I, 1966, p. 123.

6 - Ibid., p. 126.

l'expression sémantique par excellence est la phrase »7. Il faut peut-être rappeler qu'à cette époque le Cours de linguistique générale attribué faussement à Saussure était déjà bien connu8. On y décrivait la langue comme un système de signes9, « la langue en elle-même et pour elle-même », pour signifier que seule la langue relevait de la linguistique. Oswald Ducrot, pour sa part, avoue son incapacité totale à abandonner la notion de phrase, car elle représente pour lui la forme des organisations de mots la plus facile à saisir : « Ainsi donc, même si mon objectif ultime, admet-il, est d'arriver un jour à remplacer cette notion de phrase par une notion moins stricte de combinaison de mots, pour l'instant, je suis incapable dans la plupart des cas de réaliser ce projet » 10. Sans aucun doute, cette façon de considérer l'objet de la linguistique peut être considérée comme un obstacle épistémologique au sens bachelardien du terme.

Pour renverser cet obstacle et surmonter toutes les lenteurs qu'il engendre, la linguistique de corpus aborde ces problèmes autrement : en considérant d'abord le texte comme le palier primordial, elle part du corpus (ensembles de textes) vers le texte et du texte vers ses éléments, c'est-à-dire, le mot, la phrase, le paragraphe, le chapitre, etc. L'interprétation suit ces mêmes préceptes. D'où le principe épistémologique posé en sémantique interprétative par F. Rastier, à savoir que « […] la classe détermine l'élément, et […] le global détermine le local »11. Précisons que les classes sémantiques sont

7 - Émile Benveniste, Problèmes de linguistique générale, Gallimard, Paris, t. II, 1974, p. 224.

8 - Publié en 1916, après la mort de ce dernier, par Charles Bally et Albert Sechehaye.

9 - Benveniste adhérait à cette définition de la langue comme système de signe : « Nous dirons donc avec Saussure, à titre de première approximation, que la langue est un système de signe » (Benveniste, 1974, p. 219). Mais il voyait bien que cette manière de voir, le fait de considérer la langue comme un système de signe, était limité ; il voulait alors, comme il dit, aller « au-delà du point où Saussure s'est arrêté dans l'analyse de la langue comme système signifiant » (Ibid, p. 219).

10 - Cavadonga Lopez Alonso et Arlette Séré de Olmos, éd. Où en est la linguistique - Entretiens avec des linguistes, Paris, Didier Érudition, 1992, p. 66.

11 - François Rastier, Sémantique interprétative, Presses Universitaires de France, 1987, p. 12.

déterminées sur les deux niveaux syntagmatique et paradigmatique, par l'opération de substitution, pour le premier, et l'étude des cooccurrences, pour le second.

Un corpus est donc nécessaire, dans toute recherche qualitative ou quantitative, et sa spécification demeure indispensable, car certains phénomènes particuliers restent remarquablement sensibles aux variations des auteurs, des genres et des discours12. Mais, écrit E. Brunet, « […] un corpus est toujours artificiel. La nature n'en produit pas spontanément. » Il est le résultat d'une production personnelle ou collective à partir d'hypothèses pour des objectifs de recherche bien spécifiques (traitement documentaire, traitement linguistique ou statistique, etc.).

2.1. Ce que le corpus n'est pas. — Avec le développement des outils de numérisation et le foisonnement de textes numériques sur la Toile, l'élaboration de corpus peut sembler aller de soi, néanmoins il importe de savoir que toutes les compilations de textes ne peuvent pas être désignées ainsi. Il faudrait donc pouvoir distinguer, dans l'amas de documents numériques, les corpus d'étude des assemblages « naïfs » de textes. Pour cela, le principe épistémologique formulé par Saussure, à savoir que « c'est le point de vue qui seul fait la chose » est entièrement applicable dans le cas de la constitution du corpus. En le paraphrasant, on pourra dire que c'est le point de vue qui crée le corpus et le sous-corpus. La constitution du corpus constitue le point de départ du processus d'interprétation. Mais que faut-il entendre réellement par corpus ? Pour tenter de répondre à cette question, soulignons plutôt ce qu'il n'est pas :

a. Un vaste assemblage de mots ou d'exemples. — L'objet d'étude de la linguistique demeure les textes et non les mots ou les phrases isolés de leurs conditions de sémantisation, c'est-à-dire des textes eux-mêmes, du genre, du discours et de la pratique sociale auxquels

12 - François Rastier. Enjeux épistémologiques de la linguistique de corpus. Texto ! [en ligne], juin 2004. Rubrique Dits et inédits. Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html>.

ils renvoient. Ainsi, un corpus ne peut être réduit ni à des « fragments » de la langue, comme on les rencontre dans les dictionnaires, ni même à un ensemble de phrases comme les exemples qui sont généralement inventés par les chercheurs pour étayer leurs démonstrations. Ferdinand de Saussure, dans son Rapport sur la création d'une chaire de stylistique, parlait de la langue comme « dépôt passif » et désignait la parole comme « force active et origine véritable des phénomènes qui s'aperçoivent ensuite dans l'autre moitié du langage » 13. Il inscrivait déjà le signe dans la parole – les signes de parole14, écrivait-il. Des mots ou des phrases isolés proviennent donc d'une décontextualisation maximale, principale cause des divers problèmes sur la polysémie, la synonymie, les ambiguïtés15 syntaxiques et sémantiques, etc. D'où l'importance du contexte pour la description du sens d'une unité linguistique.

b. Un regroupement quelconque de textes. — Un regroupement de textes, sans hypothèses de départ ni aucun critère précis de sélection, peut renfermer des textes relativement éloignés les uns des autres, par l'éloignement de leurs genres et de leurs discours d'appartenance (littéraire, philosophique, linguistique, etc.). De pareilles bases de données restent sans intérêt immédiat pour la recherche, si ce n'est d'offrir des exemples, comme cela a été le cas de la banque textuelle Frantext pour le Trésor de la Langue Française, ou servir à l'étude spécifique de la langue, comme celle qui a été menée par E. Brunet16 sur le vocabulaire français moderne. Si l'objectif de mettre ensemble des textes est celui de pouvoir les contraster, pour tenter ainsi de saisir des propriétés qui ne se laissent pas entrevoir autrement, les distances dues à ces rassem-

13 - Ferdinand de Saussure, Écrits de linguistique générale, Édités par S. Bouquet et R. Engler, Paris, Gallimard, 2002, p. 273.

14 - Ibid., p. 265.

15 - Il suffit d'observer les exemples suivants : Les tartes aux (amandes fraîches) ou Les tartes (aux amandes) fraîches ; J'ai reçu (un vase de Chine) ou J'ai reçu (un vase) de Chine, etc.

16 - Dans cette étude, l'auteur a décrit les fluctuations du vocabulaire en fonction des différents genres, littéraires, philosophiques, scientifiques et techniques (Étienne Brunet, Le vocabulaire français de 1789 à nos jours, Genève-Paris, Slatkine-Champion, 1981, 3 vol).

blements sans principes desservent la comparaison : par exemple, ce contraste peut faire ressortir des éléments non pertinents qui appartiennent à tous les discours.

c. Un regroupement de textes incomplets. — Comme on le sait, le contexte c'est tout le texte. Les textes incomplets sont des textes « troués » auxquels il manque des parties qui peuvent contenir des instructions interprétatives importantes (comme les interprétants) nécessaires pour la construction de parcours interprétatifs. Dépourvus de certaines parties, ces textes restent difficilement interprétables, car en les fragmentant ils perdent du sens, de la même manière que des mots ou des phrases isolés. Ainsi, le British National Corpus (BNC), construit par des échantillonnages de textes oraux et écrits, peut difficilement être considéré comme un corpus, et cela malgré les diverses situations de communications et de genres représentés17.

d. Un recueil d'œuvres complètes. — On aurait tendance à penser qu'un ensemble composé des œuvres complètes d'un auteur représente le corpus idéal ; comme son nom l'indique, il comporte toutes les œuvres de l'auteur en question (jusqu'à ses traductions), en suivant une chronologie bien déterminée. On l'aborde alors comme s'il portait en lui, spontanément, les formulations et les questionnements de la recherche. Cependant, à cause de la diversité des genres qu'il renferme18, il regroupe en son sein des textes hétérogènes, inappropriés pour des explorations thématiques ou stylistiques. À moins de spécifier et de séparer ces différents genres dans le corpus, pour les prendre en compte par la suite lors de l'analyse comme une variable principale, ces œuvres seraient mieux appropriées pour l'étude des usages linguistiques (le système de la langue), car les distances entre les discours sont beaucoup plus importantes qu'entre les champs génériques ou entre les genres.

17 - Dans le but d'équilibrer toutes les parties des textes en taille et éviter ainsi que certaines soient sou- ou surreprésentées, on procède à un échantillonnage.

18 - Par exemple, dans les œuvres complètes de Voltaire, on retrouve du Théâtre, de la Poésie, de l'Histoire, des Romans, des Correspondances, etc.

2.2. La construction de corpus.– Comme on vient de le montrer par la négative à partir des différents exemples précédents, un corpus est un objet construit qui n'a rien de commun avec des ensembles naïvement assemblés ; il ne faut donc pas céder aux « divisions réelles » des textes, que l'on peut prendre facilement, sans aucun critère d'évaluation, un peu partout sur la Toile, ou des CD-ROM, comme l'archive, les banques textuelles, les œuvres complètes, etc. Ces dernières peuvent servir pour y distinguer des corpus à partir de points de vue bien clairs, sur la base d'un corps d'hypothèses. La multiplication d'exemples pris séparément dans un texte demeure de loin insuffisante, si l'on veut rendre compte de phénomènes régissant le texte. Ils ne sont définissables qu'au sein d'une problématique et d'une démarche. Relevant de choix théoriques et méthodologiques, un corpus est constitué en vue d'une exploitation bien déterminée. En reprenant la définition de F. Rastier, on peut dire qu'« 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. »19

En plus du fait qu'un corpus doit être construit et non préconstruits et que cette construction doit répondre à une problématique théorique, le fondateur de la sémantique interprétative, dans Arts et sciences du texte, énonce quatre critères20 à partir desquels on peut caractériser des corpus pertinents : la représentativité, l'homogénéité, la fermeture, et l'entretien.

La représentativité est la qualité d'un regroupement de textes constitué en corpus de façon à désigner un problème précis. Cependant, il faut souligner qu'aucun corpus ne peut être vraiment représentatif d'un problème et il est, par sa nature même, subjectif. La construction d'un corpus, qui est à chaque fois singulière, est toujours orientée, pour la bonne raison que celui-ci provient d'un corps d'hypothèses

19 - Rastier François, « Enjeux épistémologiques de la linguistique de corpus ». op. cit.

20 - Rastier François, Arts et sciences du texte, Presses Universitaires de France, 2001, p. 86.

bien déterminées. Les analyses linguistiques qui peuvent être menées sur ce corpus restent relatives à ce corpus, même s'il est possible ensuite d'en généraliser certains résultats. C'est dans ce sens que l'on parle de la représentativité comme principe épistémologique directeur. Par exemple, si l'on se propose d'étudier l'engagement politique de P. Bourdieu, un corpus « représentatif » de cette question pourra être constitué de l'ensemble de ses interventions politiques (Contre-Feux I, Contre-Feux II, etc.) et non de ses œuvres comme Langage et pouvoir symbolique, Science de la science et réflexivité, etc. Notons seulement qu'un travail critique qui tient compte des objectifs recherchés est toujours nécessaire pour élaborer ce corpus. Pour E. Brunet, « […] 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 a quo, un autre ad quem à délimiter. »21

L'homogénéité par laquelle on désigne un corpus est la propriété d'un ensemble de textes rassemblés en vue de représenter autant qu'il faut un problème déterminé, qui constitue en quelque sorte une certaine unité. Par exemple, il serait mal à propos de mettre ensemble des entretiens de télévision, des articles de journaux et des œuvres scientifiques en vue de les analyser ; on ne peut pas non plus contraster des articles juridiques et linguistiques. Cette règle d'homogénéité intéresse aussi « Dans un sens plus restreint, l'homogénéité pourra être fondée sur un choix d'éléments de même niveau, de relations de même type (Hjelmslev). »22 En fonction des objectifs de la recherche, celle-ci peut se réaliser, généralement, au niveau

21 - Brunet Étienne, Ce qui compte. Méthodes statistiques. Écrits choisis, tome II., Éditions Champion, Paris, 2011, p. 279.

22 - Greimas, A. J.& Courtés, J., Sémiotique. Dictionnaire raisonné de la théorie du langage, op.cit., p. 174.

des genres ou, spécifiquement, au niveau des discours, comme dans le cas des recherches sur le système de la langue. Mais pour F. Rastier, « […] l'homogénéité de genre doit être privilégiée par défaut, même pour les recherches stylistiques […] : en effet, un texte peut « perdre » du sens, s'il est placé parmi des textes oiseaux, car la comparaison avec eux ne permet pas de sélectionner d'oppositions pertinentes. »23 Il va sans dire, depuis au moins la découverte des manuscrits de Saussure en 1996, que la langue est variable et diachroniquement et synchroniquement. Elle présente également des modifications importantes dans un champ générique, c'est-à-dire dans un ensemble de genres s'opposant dans une même pratique. Par exemple, dans le discours littéraire constitué du théâtre, du récit, de la poésie, l'hétérogénéité des textes composés par la comédie, la tragédie, le drame, le récit, le roman, la nouvelle, etc. est aisément reconnaissable. Ainsi, si le sens provient de la différence entre des unités linguistiques et que l'avantage de construire un corpus est de pouvoir contraster ses textes, ce n'est pas pour comparer l'incomparable, c'est-à-dire des textes de genres ou de discours complètement éloignés les unes des autres.

La fermeture (ou l'ouverture) constitue un critère qui permet de distinguer un corpus d'étude fermé d'une base de données qui reçoit continuellement des textes ; si le contexte c'est tout le texte, et que le sens d'une unité linguistique locale est déterminé par le corpus global, ce dernier doit être délimité et clôturé, au moment de l'analyse, comme le serait une classe sémantique qui détermine le sens de ses éléments. Car chaque ensemble fermé (ici le corpus ou la classe) détermine des signifiés dont la valeur (au sens saussurien du terme) est fixe. « Au cours d'une recherche, le corpus de référence et le corpus de travail sont toujours fermés, car ils doivent être prédéfinis. En raison de sa méthodologie comparative, la linguistique ne peut d'ailleurs travailler utilement que sur des corpus définis. »24 Une fois les hypothèques qui ont présidé à la constitution du corpus ont

23 - Rastier François, Arts et sciences du texte, Presses Universitaires de France, 2001, p. 86.

24 - Ibid., p. 87.

été vérifiées, il est toujours possible d' « ouvrir » le corpus pour en retrancher ou en rajouter des textes, mais seulement à partir d'un nouveau corps d'hypothèses.

Ce que l'on désigne par l'entretien permet également de spécifier le corpus d'étude pour le distinguer d'un assemblage naïf de textes, qui peut facilement devenir obsolète et illisible si ses textes ne sont pas exploités, annotés, lus et interprétés ; au contraire, constitué conformément aux règles, un corpus entretenu ouvre une tradition interprétative qui rend ses textes lisibles et compréhensibles. « […] tout corpus, même fermé, qui ne fait pas l'objet d'une élaboration continue, se périme, et paradoxalement devient inutilisable s'il n'est pasutilisé25. »26

2.3. Définir le texte. — Avant d'évoquer l'importance des méthodes statistiques pour l'analyse des corpus, il importe de préciser rapidement ce que l'on entend par un texte.

Quand bien même certaines opérations interprétatives se feraient à des paliers textuels inférieurs (paliers lexicaux27), pour Hjelmeslev, dans sa théorie du langage, et pour F. Rastier, dans sa sémantique interprétative, le principal objet de l'investigation linguistique demeure le texte dans toute sa complexité.

Le linguiste Louis Hjelmslev a été le premier à donner au mot « texte » un contenu conceptuel bien spécifique en le dégageant du vocabulaire commun. La théorie du langage qu'il construit prend en considération, d'une manière explicite, le(s) texte(s) : au milieu d'un réseau conceptuel riche (langue, langage, analyse, expression, contenu, théorie, système, signe, sens, forme, etc.), le concept de texte tient une place prépondérante dans ses Prolégomènes à une théorie

25 - Il en va de même au plan de contenu : un texte qui cesse d'être lu peut devenir illisible, car il est coupé de sa tradition interprétative.

26 - Rastier François, Arts et sciences du texte, op. cit., p. 87.

27 - On distingue trois paliers de l'analyse linguistique : micro-, méso- et macrosémantique, auxquels correspondent successivement le mot ou le morphème, la phrase et le texte. L'incidence de ce dernier sur les autres paliers textuels est incontestable.

du langage, avec 119 occurrences28. On peut lire dans cet ouvrage de 1943 (traduction française de 1971) ceci : « La théorie du langage s'intéresse à des textes, et son but est d'indiquer un procédé permettant la reconnaissance d'un texte donné au moyen d'une description non contradictoire et exhaustive de ce texte. Mais elle doit aussi montrer comment on peut, de la même manière, reconnaître tout autre texte de la même nature supposée en nous fournissant les instruments utilisables pour de tels textes » 29.

Grâce à ses travaux, il engagera la linguistique dans un tournant décisif. Mais, pendant longtemps, les linguistes qui lui ont succédé ne le suivront pas sur cette voie ; au concept de texte, note M. Arrivé, ils préfèrent ceux de discours et d'énoncé30. Cette tendance se trouve clairement exprimée, de l'autre côté de l'Atlantique, par le linguiste américain Zellig Harris, pour qui le discours, objet de son analyse, désigne « […] une séquence de formes linguistiques disposées en phrases successives […] »31. Si Harris parle sans distinction du texte (ou de l'énoncé) et du discours, Michel Pêcheux, en France, les oppose. Cette mouvance domine au demeurant toute l'École française d'Analyse du discours, avec Maingueneau, Charaudeau, etc., jusqu'à Greimas et Courtés32 : « Considéré en tant qu'énoncé, le texte s'oppose au discours, d'après la substance de

28 - Pour un approfondissement du concept de texte chez cet auteur, nous renvoyons à l'article de Kyheng Rossitza, « Hjelmslev et le concept de texte en linguistique ». In Texto [en ligne], septembre 2005, vol. X, n°3. Disponible sur : <http://www.revue-texto.net/Inedits/Kyheng/Kyheng_Hjelmslev.html>.

29 - Hjelmslev, Louis. Prolégomènes à une théorie du langage, Éditions de Minuit, Paris, 1971, p. 26-27, cité par Kyheng Rossitza, Hjelmslev et le concept de texte en linguistique, op. cit.

30 - Arrivé et al., 1986, p. 670, cité par Rastier, François. Pour une sémantique des textes : questions d'épistémologie. Texto ! 1996 [en ligne]. Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_PourSdT.html>.

31- Harris Zellig S., Dubois-Charlier Françoise. Analyse du discours. In: Langages, 4e année, n° 13. . L'analyse du discours. pp. 8-45. url : http://www.persee.fr/web/revues/home/prescript/article/lgge_0458-726x_1969_num_4_13_2507.

32 - Rastier, François. Discours et texte. Texto ! juin 2005 [en ligne]. Disponible sur : <http://www.revue-texto.net/Reperes/Themes/Rastier_Discours.html>.

l'expression - graphique ou phonique - utilisée pour la manifestation du procès linguistique. Le texte serait alors un énoncé qui peut s'actualiser en discours. Autrement dit, le texte pourrait être considéré comme un produit, une substance (du côté de la langue) et non comme un processus »33.

Or on sait pertinemment que la séparation des plans du langage (signifiant et signifié ; expression et contenu), à partir de laquelle on oppose abusivement le texte au discours, est sans fondement. Aussi bien chez Hjelmslev que chez Saussure, les deux plans du langage se trouvent indissolublement liés l'un à l'autre, comme l'envers et l'endroit d'une feuille de papier. Dans les écrits retrouvés de Saussure, on peut lire qu' : « Il est donc entièrement illusoire d'opposer à aucun instant le signe à la signification. Ce sont deux formes du même concept de l'esprit, vu que la signification n'existerait pas sans un signe, et qu'elle n'est que l'expérience à rebours du signe, comme on ne peut pas découper une feuille de papier sans entamer l'envers et l'endroit de ce papier, du même coup de ciseaux » 34.

De cette façon, F. Rastier considère les textes et les discours exactement au même niveau ontologique35. Pour lui « un texte est une suite linguistique empirique attestée, produite dans une pratique sociale déterminée, et fixée sur un support quelconque » 36. Oral ou écrit (ou fixé sur d'autres supports), un texte se rapporte à un genre (comédie, tragédie, drame, roman, nouvelle, etc.) qui, à son tour, se rapporte à un discours (littéraire vs juridique vs politique vs scientifique, etc.) par l'intermédiaire d'un champ générique (théâtre, poésie, récit, etc.). Autrement dit, un discours particulier (littéraire, par exemple) présuppose l'existence d'un rassemblement hétéroclite de textes, largement différents les uns des autres dans leurs réseaux de relation (roman, nouvelle, comédie, tragédie, drame, etc.), mais à

33 - Greimas, A. J.& Courtés, J., Sémiotique. Dictionnaire raisonné de la théorie du langage, Hachette, Paris, 1979, p. 389, cité par Rastier, François. Discours et texte, op. cit.

34 - Saussure, Ferdinand de. Écrits de linguistique générale. Établis et édités par S. Bouquet et R. Engler. Paris : Gallimard, 2002, p. 96.

35 - Rastier, François. Discours et texte, op. cit.

36 - Rastier François, Arts et sciences du texte, op. cit., p. 21.

partir desquels il est possible d'organiser un corpus. Quel que soit le degré de son originalité, la production d'un nouveau texte procède de ce corpus de textes dont il hérite des traces sémantiques, lexicales, thématiques, etc., indélébiles. Considéré comme une expérimentation spontanée, opposée à l'exemple inventé par le chercheur pour les besoins de sa démonstration, un texte doit pouvoir être authentifié (date, auteur, lieu, genre, discours, etc.).

Toutes les disciplines des sciences humaines ont affaire à des textes. Un corpus est une compilation organisée de textes. Cependant, un texte ne doit pas être considéré comme une simple somme de ses parties, ou une simple chaîne de caractères (String), comme c'est le cas de nombreux langages informatiques. Si l'on continue de le considérer ainsi, si l'on n'omet de tenir compte de la globalité (du texte ou du corpus) qui donne aux phrases et aux unités textuelles des déterminations essentielles, on ne saura l'analyser ni le décrire.

Longuement ignorée – et à tort – par la linguistique, la corrélation indissoluble entre le plan du signifiant et le plan du signifié (sémiosis textuelle) ne concerne pas seulement le signe, mais également le texte. Définie déjà par Saussure par le concept de forme-sens, cette association des deux plans du langage doit intéresser, à tous les égards, la linguistique des textes.

3.2. Le contexte c'est tout le texte.— S'il est admis, depuis l'enseignement de Saussure, que la parole (au sens large) est l'origine véritable des phénomènes, toute forme qui s'y introduit ne peut être comprise ni interprétée en l'absence de son contexte. Saussure, dans ses manuscrits retrouvés dans l'orangerie de la famille, formule clairement cette idée lorsqu'il écrit justement que « La condition de tout fait linguistique est de se passer entre deux termes au minimum ; lesquels peuvent être successifs ou synchroniques »37. Dans ces conditions, un mot n'a de sens que pris dans un contexte, qui est tout le texte ; de la même manière, un texte ne peut être interprété que pris dans un regroupement structuré de textes entiers. Comme pour

37 - Saussure, Ferdinand de. Écrits de linguistique générale. Établis et édités par S. Bouquet et R. Engler. Paris : Gallimard, 2002, p. 123.

donner une première définition de cette notion, F. Rastier écrit que « le contexte c'est tout le texte, mais ce n'est pas tout dans le texte »38 ; pour une unité linguistique donnée, le contexte représente donc l'ensemble des unités qui entrent en relation d'incidence avec elle. « Plus précisément, le contexte passif d'un sémème est l'ensemble des sémèmes sur lesquels il a une incidence, et son contexte actif est l'ensemble des sémèmes qui ont une incidence sur lui. »39

Cette inséparabilité des unités linguistiques, qui présuppose l'existence d'autres unités, a permis à Saussure de poser la notion de valeur, qui découle, d'après S. Auroux, de la synonymique des Lumières40 ; en effet, on la retrouve chez l'abbé Girard mais dans un sens restreint. Cependant, en tant que concept novateur, c'est Saussure qui l'a utilisé en premier pour montrer qu'un élément ne peut être identifié que par la valeur que lui donne la collectivité. La langue n'existe alors que dans sa transmission, et ses signifiés ne sont définissables que par des relations d'oppositions. On parle alors de valeur en contexte ou de « valeur contextuelle »41. « On pourrait penser que les valeurs contextuelles ne font que modifier secondairement, par des nuances, la valeur en langue. Au contraire, la valeur en langue est surdéterminée par la valeur en contexte et n'importe quel trait sémantique défini en langue peut être annulé ou virtualisé par le contexte, local voire global. »42

On voit donc toute l'importance des corpus de textes dans la détermination des valeurs des mots ou des lexies. On comprend alors qu'une valeur n'a rien de commun avec le contenu d'un signe qui lui serait intrinsèque. « […] la langue consiste, non dans un système de valeurs absolues ou positives, mais dans un système de valeurs relatives et négatives, n'ayant d'existence que par l'effet de leur oppo-

38 - François Rastier, Sémantique interprétative, op. cit., p. 73.

39 - Ibid., p. 73.

40 - Rastier François, Sémantique et recherche cognitive, Presses Universitaires de France, 1991, p. 101.

41 - François Rastier. La Mesure et le Grain. Sémantique de corpus, Editions Honoré Champion, coll. «Lettres numériques» n°12, Paris, 2011, p. 30.

42 - Ibid., p. 30.

sition. »43 Le signifié d'un mot ne peut être obtenu sans le secours d'autres mots, sur les deux axes paradigmatique et syntagmatique. Les textes – et donc les contextes – constitués en corpus permettent l'identification des cooccurrents sémantiques44, une opération importante en linguistique de corpus pour l'élaboration des parcours interprétatifs, la détermination du sens et la construction des thèmes.

3. Mesure des quantités :

> *« 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 »45.*

Pour appréhender un phénomène linguistique, sociologique, historique ou économique donné, les analyses textométriques représentent le premier volet méthodologique de la recherche ; elles constituent une étape de description incontournable, notamment dans l'étude de grands corpus numériques, car de nombreux phénomènes ne se laissent pas appréhender facilement, à l'instar des thèmes qui ne sont pas manifestés par des lexèmes.

Par textométrie on désigne l'ensemble des techniques statistiques utilisées pour l'étude de textes et corpus numériques, que l'on retrouve également sous les termes de logométrie46, de statistique textuelle ou encore de lexicométrie47. De cette manière, « […] l'év-

43 - Saussure, Ferdinand de. Écrits de linguistique générale. op. cit., p. 80.

44 - Si l'hypothèse est solide sur un ensemble de cooccurrents donnés, si des relations sont établies entre eux, ces cooccurrents seront élevés au rang de corrélats sémantiques qui seront la base de la construction d'un thème.

45 - Muller Charles, La statistique lexicale, Langue française, 1969, n° 1, pp. 30-43, url : http://www.persee.fr.

46 - Mayaffre Damon, « Analyse du discours politique et Logométrie : point de vue pratique et théorique », Langage et société, 114 (2005) 91-121.

47 - Lebart Ludovic, Salem André, Statistique textuelle, Préface de Christian Baudelot,

olution de désignation de la « lexicométrie » en « textométrie » veut exprimer que l'analyse menée ne se cantonne pas à l'étude du lexique, mais s'intéresse avant tout à la description du texte, dans ses multiples dimensions » 48.

L'efficacité des logiciels d'analyse textuelle a souvent été exagérée et mal définie ; il n'a jamais été question, par la seule approche statistique, de faire des repérages thématiques, ni de rendre compte du sens d'un texte. Les traitements statistiques, dont l'efficacité revient aux régularités des phénomènes qui peuvent être observées et repérées quantitativement, n'auraient aucun sens si l'on oubliait que les textes sont dotés d'une structure régie par des associations aux deux plans du langage (contenu et expression) et aux différents paliers de complexité (morphème, lexie, chapitre, texte et corpus), en respectant le principe selon lequel le global détermine le local. D'après J.P. Bénzecri, c'est la conception purement déductive et mathématique de la langue qui a permis à l'auteur des Structures syntaxiques, N. Chomsky, d'affirmer injustement, « qu'il ne peut exister de procédures systématiques pour déterminer la grammaire d'une langue, ou plus généralement les structures linguistiques, à partir d'un ensemble de données tel qu'un recueil de textes que les linguistes nomment corpus » 49.

Ces différentes techniques statistiques sont donc offertes par un nombre important de logiciels qui se sont développés à la suite des travaux de Charles Muller et Jean-Paul Benzécri. En fonction des traitements offerts et du degré d'implication du chercheur, on distingue deux types de logiciels : i) les logiciels nécessitant une intervention constante du chercheur, et cela de la création des corpus et des sous-corpus aux différents calculs escomptés. SATO, MODALISA, ATLAS, NViVo, etc., sont de cette catégorie. ii) les logiciels ne nécessitant pas l'intervention du chercheur, sauf lors de la création

Dunod, 1994.

48 - Pincemin Bénédicte (2012) - « Sémantique interprétative et textométrie », Texto! Volume XVII, n°3, coordonné par Christophe Cusimano.

49 - Benzécri Jean-Paul, Histoire et préhistoire de l'analyse des données, Dunod, 1982, p. 102.

des corpus, accomplissent d'eux-mêmes, d'une manière automatique, l'intégralité des opérations et des calculs définis au préalable. De cette classe, on peut mentionner, à titre d'exemple, HYPERBASE, ALCESTE, LEXICO, Dtm-Vic, SPAD-T, SAS, 3AD, etc. Dotés de fonctionnalités rapides et efficaces, ces logiciels permettent en général une vaste exploration de corpus numériques. Sur la base d'hypothèses, ils sont capables de fournir une foule importante de données concernant la distribution et l'évolution du vocabulaire, de mettre en évidence des endroits spécifiques dans le corpus, de retrouver aisément les occurrences d'une forme. Par leurs calculs, ils permettent également la constitution de sous-corpus, pour enfin pouvoir les exploiter ou les comparer. Le logiciel SATO, par exemple, comme son nom l'indique, est un « Système d'Analyse de Textes par Ordinateur ». Il permet principalement la génération de lexiques50, la catégorisation et l'annotation de mots (en contexte et hors contexte), la constitution de sous-textes et leur comparaison sur la base de leurs lexiques correspondants, etc. À partir d'une base de données créée dans le disque dur de l'ordinateur, le logiciel HYPERBASE est capable de calculer les occurrences d'une forme donnée (et les hapax) dans le corpus et dans chaque partie du corpus, de suivre l'évolution du vocabulaire, de comparer les spécificités du corpus par rapport à celui de Frantext (spécificités positives et négatives). Il est également en mesure de faire une analyse factorielle des correspondances (AFC) à partir des formes les plus fréquentes dans le corpus, et d'estimer ainsi les distances lexicales entre les textes, etc.

Ces logiciels sont donc des instruments d'aide à l'analyse de corpus textuels, qui donnent des possibilités diverses et variées pour formuler et/ou vérifier des hypothèses (afin de les valider ou pas). Par divers calculs, ils permettent à certains faits, si l'on peut dire, de « sauter aux yeux ». « [...] certaines régularités, dispersées dans le corpus, observe F. Rastier, semblent « s'imposer d'elles-mêm-

50 - L'affichage des formes lexicales dépend du filtre choisi, selon nos hypothèses ; il faut donc le définir en sélectionnant les lexèmes à afficher. Par exemple, pour afficher les fréquences des formes commençant par queb-, en choisissant queb$, SATO affichera tous les éléments lexicaux qui commencent par queb, c'est-à-dire, quebec, quebecois, quebecoise, etc.

es » et prendre une valeur heuristique : par exemple, si le corpus de référence est soigneusement constitué, surligner les mots qui dépassent un seuil d'écart réduit permet à certains d'entre eux de « frapper à la porte » »51. Et selon l'expression bien trouvée de Viprey, l'analyse statistique permet « d'offrir des rives à l'intuition solitaire »52.

Cependant, il faut insister fortement sur quelques points essentiels :

1. Les questionnements préexistent à la constitution de corpus ;

2. Une stratégie de recherche prédéfinie, applicable comme une ritournelle à tous les types de corpus, n'existe pas ;

3. Le chercheur peut découvrir autre chose que ce qu'il recherche. Les explorations qui peuvent être menées sur les textes peuvent faire voir des phénomènes éloignés, sous tous les rapports, des interrogations de départ – comme pour les fouilles archéologiques.

4. Les logiciels de lexicométrie mobilisés pour le traitement de corpus ne sont que des instruments informatiques, des outils au service de l'analyse sémantique. Les réponses qu'ils livrent aux chercheurs nécessitent d'être interprétées, car le sens d'un texte ne peut être défini par ses seules chaînes de caractères. C'est l'un des problèmes essentiels qui se posent à la sémantique de corpus, il concerne le passage des identifications quantitatives aux qualitatives, autrement dit, de la mesure des quantités à la qualification des données.

4. Qualification des données :

Les analyses lexicométriques menées dans une recherche linguistique, ou autre, en manipulant des vocables, ne doivent pas faire croire que l'on s'enferme dans un chiffrage et un déchiffrage naïf

51 - Rastier François, Arts et sciences du texte, op. cit., p. 96.

52 - Jean-Marie Viprey, Dynamique du vocabulaire des Fleurs du mal, Préface d'Étienne Brunet, Éditions Champion, Paris, 1997, p. 65.

des signes ; qu'il suffise ainsi de repérer des « mots-vedettes » par de simples intuitions, pour y découvrir les sujets ou les thèmes correspondants. C'est le cas par exemple des systèmes de filtrage automatiques conventionnels des sites racistes, xénophobes et pédophiles, qui partent du seul et simple principe « qu'il y a des mots racistes et des mots qui ne le sont pas, sans considération pour leur mise en texte (ou condition d'énonciation). […] comme si le racisme était une langue de spécialité avec une terminologie stable et univoque. » 53

Au contraire, en considérant le texte – et non le signe – comme objet, les analyses et identifications thématiques doivent être menées par le biais d'opérations sémantiques interprétatives complexes. Il est à noter que la présence ou l'absence d'un mot dans un corpus ne dit rien – du moins, pas d'une manière systématique – sur la présence ou l'absence d'un thème. Ce dernier est construit autour de lexicalisations diverses par un ensemble toujours particulier de sèmes, qui différent selon des critères comme le genre, le nombre, etc. « Comme toutes les unités sémantiques, un thème est une construction, non une donnée ; aussi la thématique dépend de conditions herméneutiques : l'interprétation des données textuelles se place dans un cercle méthodologique dépendant du cercle herméneutique. »54

En revanche, les mesures ne sont pas à récuser, néanmoins elles doivent être prises avec prudence et d'incessantes correspondances dans le texte, car parfois ce sont ces mesures qui créent leurs propres démesures55. À la statistique (bonne ou moins bonne), écrit F. Simiand, on ne peut faire dire que ce qu'elle dit et dans les conditions

53 - Mathieu Valette, Natalia Grabar , « Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction se syntagme ? l'exemple du projet PRINCIP », Journées Internationales d'Analyse statistique des Données Textuelles, Louvain-la-Neuve : Belgique, 2004.

54 - Rastier François, Arts et sciences du texte, op. cit., p. 191.

55 - Le fait de trouver, par exemple, 892 occurrences du verbe "aimer" dans Corneille et seulement 316 chez Racine, « ne signifie pas pour autant que l'on aime deux fois plus chez corneille que chez Racine ! » (J. Emélina, cité par Sylvie Mollet et Marcel Vuillaume, Mots chiffrés et déchiffrés, Éditions Champion, Paris, 1998, p. 92.)

où elle le dit56. Il suffit de reprendre, à titre d'exemple, des tableaux statistiques loin de leurs contextes pour s'apercevoir que les chiffres qu'ils alignent ont peu de valeur en eux-mêmes, du moment qu'ils ne se rattachent à aucun texte ou phénomène qu'ils permettent de comprendre, d'interpréter ou de connaître. Ainsi, de la même manière que le global est au service du local, le quantitatif doit rester au service du qualitatif. « L'analyse statistique, explique C. Muller, pratiquée sur de grandes masses, ignore les nuances, mais peut suggérer des recherches plus précises et plus localisées où la sémantique et la philologie reprendraient tous leurs droits. »57

L'interrogation des textes numérisés par un logiciel reste donc insuffisante ; elle ne constitue qu'une étape d'un processus d'interprétation qui ne demande qu'à être poursuivi. Mais de l'autre côté, notamment dans le cas de grands corpus, une « autosuffisance sémantique » peut s'avérer elle aussi insatisfaisante, fragmentaire, lorsqu'elle fait l'économie du calcul statistique. Ainsi, que l'on soit dans le texte (même lors d'une simple lecture) ou que l'on s'y éloigne (au moment du calcul expérimental), il y a poursuite ou suspension de l'interprétation.

Alors que la « compréhension est immédiate et se suffit à elle-même58 », l'interprétation est médiate et nécessite le passage par d'autres éléments textuels, des plus proches, dans le paragraphe, aux plus larges, dans le texte et l'intertexte.

Si l'objectif visé est bien la recherche du sens, celui-ci n'est ni complètement dans le texte ni dans l'interprète 59, et encore moins dans les données statistiques. Il naît plutôt dans cette rencontre, dans

56 - Simiand François., Statistique et expérience, remarques de méthode, M. Rivière, Paris, 1922, p. 24.

57 - Charles Muller, Étude de statistique lexicale. Le Vocabulaire du Théâtre de Pierre Corneille, Slatkine Reprints, Genève, 1993, p. 134.

58 - Thouard Denis, Herméneutique contemporaine. Comprendre, interpréter, connaître, Paris, Vrin, « Textes clés », 2011, p. 9.

59 - Rastier François, Arts et sciences du texte, op. cit., notamment, p. 125.

ces va-et-vient entre l'interprétation sémantique et le calcul statistique. Cependant, comme nous l'avons signalé au départ, le problème demeure dans le passage du quantitatif au qualitatif, c'est-à-dire, des cooccurrents statistiques aux corrélats sémantiques. Ils ne se confondent d'aucune façon, contrairement à ce qu'affirmait C. Boudelot60 : la seule détermination des cooccurrents reste insuffisante ; il faudrait pouvoir distinguer dans ces signes des sèmes identiques, sur la base d'une présomption d'isotopie61, en tenant compte évidemment du corpus, du discours, du genre, etc., ce que l'on désigne habituellement par le niveau global.

Toute activité interprétative – de construction de parcours thématique et (ou) de compréhension d'un texte –, qu'elle s'appuie ou non sur des calculs statistiques, doit obéir ipso facto à trois principes62 sémantiques : 1) contextualité ; 2) intertextualité ; 3) architextualité.

On comprend par là, grosso modo, que le sens d'un texte – ou d'un élément de ce texte – est modifié à chaque fois qu'il « interfère » avec un autre texte, à travers les citations, les reprises et les reformulations63. La constitution de corpus est aussi une activité interprétative, lorsqu'elle met ensemble des textes ou des auteurs qui n'ont aucune chance de se rencontrer ordinairement. Donc, à l'instant même où deux passages (c'est-à-dire, le mot, la phrase, le paragraphe, le chapitre, etc.) sont mis côte à côte dans un texte donné, des traits sémantiques de chaque partie sont activés pour générer

60 - « L'une des contributions majeures de la statistique textuelle est précisément d'animer tous ces graphes en donnant la parole à chacun de ces individus. Grâce à Lebart et Salem, les fameux points-individus ne sont plus muets, ils parlent. Vole alors en éclats la traditionnelle, mais artificielle distinction entre le quantitatif et le qualitatif. » (Ludovic Lebart, André Salem, Statistique textuelle, Préface de Christian Baudelot, Dunod en 1994, p. V).

61- La présomption d'isotopie permet de mettre en place un processus interprétatif à travers lequel on tente de vérifier l'existence d'un effet de récurrence de certains sèmes.

62 - Pour plus de précisions, voir, Rastier François, Arts et sciences du texte, op. cit., notamment, p. 92-93 ; Sémantique interprétative, Presses Universitaires de France, 1987, p. 72-73.

63 - Rastier François, Arts et sciences du texte, op. cit., p. 92.

du sens (1) ; deux autres passages activeraient à l'évidence d'autres traits pour un autre sens, etc. La même chose se produit à un palier supérieur, c'est-à-dire, de texte à texte différent (2), ou d'un texte ou citation placée dans un corpus (3).

Cette « révolution numérique » modifie les préceptes de l'interprétation ; on rentre dans une nouvelle ère où certains problèmes de recherche et de construction de parcours interprétatifs et de données, qui échappent consciemment ou inconsciemment à l'œil nu dans une lecture linéaire d'un texte, s'y trouvent largement dépassés : par exemple, le calcul de la distance64 intertextuelle, que réalisent aujourd'hui de nombreux logiciels, permet de mesurer les ressemblances et les dissemblances entre plusieurs textes.

Le renouvellement méthodologique favorisé par la linguistique de corpus numériques engage ainsi un nouveau dispositif65, résumé dans le cycle suivant :

1. analyse de la tâche et production des hypothèses ;
2. constitution de corpus de travail et de référence ;
3. choix de la stratégie et du logiciel appropriés ;
4. traitement statistique du corpus ;
5. interprétation des résultats obtenus ;
6. validation de l'interprétation par un retour aux textes.

5. Conclusion. :

La linguistique de corpus, dont nous avons défini quelques principes, se distingue méthodologiquement de la linguistique spontanée. Avec l'étude de corpus, le mot cesse d'être le principal objet

64 - Cet indice est utilisé pour attribuer – ou non – certains textes « douteux » à des auteurs, une sorte de « certificats d'authenticité ou de paternité à la manière des empreintes digitales ou des séquences d'ADN » (Étienne Brunet, « Peut-on mesurer la distance entre deux textes ? », Corpus [En ligne], 2 | décembre 2003, mis en ligne le 15 décembre 2004, Consulté le 17 octobre 2012. URL : http://corpus.revues.org/index30.html).

65 - François Rastier. La Mesure et le Grain. Sémantique de corpus, op. cit., p. 13.

de l'analyse. Son sens n'est déterminable que dans un contexte, dans ses rapports aux autres mots de sa classe sémantique et de ses occurrences, en tenant compte du genre, du discours et de la pratique sociale. Le mot, la phrase, le texte et le corpus se trouvent également unifiés dans un seul et même objet, alors qu'habituellement ils sont séparément décrits par des disciplines diverses, comme la sémantique lexicale, la syntaxe et la pragmatique. Ainsi, ce nouveau terrain dans lequel s'engage la linguistique, depuis quelques années, avec l'exploration de corpus, permet des développements plus fermes. Avec les nouvelles conceptions du texte d'où elle tire sa force méthodologique, elle semble trouver une assise plus solide, qui ne laisse guère de place aux problèmes et approximations habituels (polysémie, ambiguïtés, etc.).

Si les statistiques servent des domaines divers, l'on ne peut continuer à ignorer leur utilité pour une discipline comme la linguistique de corpus. Leur efficacité revient aux logiciels rigoureusement construits (Hyperbase, Alceste, SATO, etc.), qui sont capables de « décomposer » les textes en leurs éléments. Les logiciels sont donc des outils permettant de frayer des accès au texte. Ils substituent à la lecture linéaire une lecture suggestive, en mettant en lumière une foule de détails importants. Ils traquent certains phénomènes textuels qui ne se laissent pas saisir autrement.

Cependant, on ne peut guère se satisfaire des données que fournissent les logiciels et conclure hâtivement, par exemple sur l'existence ou l'absence d'un thème, sur l'importance de la ponctuation dans un texte, sur l'usage de l'imparfait, etc. Au contraire, les problèmes posés par les données statistiques doivent être reformulés, car si le sens est fait de différence, il ne jaillit de nulle part ; il faudrait pouvoir le construire.

Il est donc nécessaire de ne pas séparer ces deux approches, statistiques et herméneutique des textes, alors complémentaires, mais plutôt de préciser davantage leur rapport plus que jamais étroit. La linguistique de corpus dépend des logiciels comme d'une dépendance théorique.

Bibliographie

• *Benveniste Émile, (1966), Problèmes de linguistique générale, Tome 1, Gallimard, Paris.*

• *Benzécri Jean-Paul, (1982), Histoire et préhistoire de l'analyse des données, Dunod, Paris..*

• *Brunet Étienne, (2011), Ce qui compte. Méthodes statistiques. Écrits choisis, tome II., Éditions Champion, Paris.*

• *Brunet Étienne, « Peut-on mesurer la distance entre deux textes ? », Corpus [En ligne], 2 | décembre 2003, mis en ligne le 15 décembre 2004, Consulté le 17 octobre 2012. URL : http://corpus.revues.org/index30.html).*

• *Brunet Étienne, (1981), Le vocabulaire français de 1789 à nos jours, Genève-Paris, Slatkine-Champion.*

• *Canguilhem Georges, (1980), La connaissance de la vie, Librairie Philosophique J. Vrin, Paris.*

• *Cavadonga Lopez Alonso et Arlette Séré de Olmos, éd. Où en est la linguistique - Entretiens avec des linguistes, Didier, paris ;.*

• *Ducrot Oswald, Schaeffer Jean-Marie, (1991), Dictionnaire encyclopédique des sciences du langage, Seuil, Paris..*

• *Greimas Algirdas Julien, Courtes Joseph, (1979), Sémiotique. Dictionnaire raisonné de la théorie du langage, Hachette, Paris.*

• *Harris Zellig S., Dubois-Charlier Françoise. Analyse du discours, In Langages, 4e année, n° 13. L'analyse du discours. pp.8-45. url : http://www.persee.fr/web/revues/home/prescript/article/lgge_0458-726x_1969_num_4_13_2507.*

• *Hjelmslev, Louis, (1971), Prolégomènes à une théorie du langage, Éditions de Minuit, Paris,.*

• *Kyheng Rossitza, « Hjelmslev et le concept de texte en linguistique ». In Texto [en ligne], septembre 2005, vol. X, n°3. Disponible sur : <http://www.revue-texto.net/Inedits/Kyheng/Kyheng_Hjelmslev.html>.*

• *Lebart Ludovic, Salem André, (1994), Statistique textuelle, Préface de Christian Baudelot, Dunod, Paris.*

• *Mayaffre Damon, (2005), « Analyse du discours politique et logométrie : point de vue pratique et théorique », Langage et société, 114 91-121.*

• *Mollet Sylvie et Vuillaume Marcel, (1999), Mots chiffrés et déchiffrés, Éditions Champion, Paris,.*

• *Muller Charles, (1969), La statistique lexicale in Langue française, n° 1, pp. 30-43, url : http://www.persee.fr.*

• *Muller Charles, (1993), Étude de statistique lexicale. Le Vocabulaire du Théâtre de Pierre Corneille, Slatkine Reprints, Genève.*

• *Pincemin Bénédicte (2012), « Sémantique interprétative et textométrie », in Texto, Volume XVII, n°3, coordonné par Christophe Cusimano.*

• *Rastier François, (1987), Sémantique interprétative, PUF, Paris..*

• *Rastier François, (1991), Sémantique et recherche cognitive, PUF, Paris.*

• *Rastier François, Enjeux épistémologiques de la linguistique de corpus in Texto [en ligne], juin 2004. Rubrique Dits et inédits. Disponible sur : <http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html>. (Consultée le ...).*

• *Rastier, François. Discours et texte in Texto, juin 2005 [en ligne]. Disponible sur :<http://www.revue-texto.net/Reperes/Themes/Rastier_Discours.html>.*

• *Rastier François, (2001), Arts et sciences du texte, PUF, Paris..*

• *Rastier François, (2011), La Mesure et le Grain. Sémantique de corpus, Editions Honoré Champion, n°12, Paris.*

• *Saussure Ferdinand De, (2002) Écrits de linguistique générale, établis et édités par S. Bouquet et R. Engler, Gallimard, Paris.*

• *Simiand François, (1922), Statistique et expérience, remarques de méthode, M. Rivière, Paris.*

• *Thouard Denis, (2011), Herméneutique contemporaine. Comprendre, interpréter, connaître, Vrin, Paris.*

• *Valette Mathieu, Natalia Grabar, (2004), « Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction se syntagme ? l'exemple du projet PRINCIP », Journées Internationales d'Analyse statistique des Données Textuelles, Louvain-la-Neuve : Belgique.*

• *Viprey Jean-Marie, (1997), Dynamique du vocabulaire des Fleurs du mal, préface d'Étienne Brunet, Éditions Champion, Paris.*