1. POURQUOI CONCEVOIR UNE ASSISTANCE A L'INTERPRETATION DES TEXTES ELECTRONIQUES ?

Les nouvelles technologies permettent l'accès à des banques de données textuelles de plus en plus nombreuses et importantes, dont certaines autorisent le téléchargement gratuit.

1. 1. 1. Banques de données, textes numérisés et usagers

L'existence du *texte électronique* modifie l'activité même de lecture par rapport à son contexte classique, et on a vu apparaître, à partir de 1989[1], le concept de "lecture assistée par ordinateur", rendu nécessaire par le changement de support : il concerne "la suppléance aux limitations et inconvénients de ce mode d'existence du texte (par exemple : surmonter la perte de perception de la dimension "épaisseur", et, par le fait, celle des structurations typo-dispositionnelles à longue et moyenne portée)"[2]. Dans cet article, les auteurs considèrent que "cette sorte de lecture se distingue très clairement de la consultation de base de données textuelles", mais cet argument n'est pas convaincant, dans la mesure où, aujourd'hui on ne connaît que des embryons de "station de travail assistée par ordinateur", et où les modes de consultation des textes sur support informatique ne satisfont pas tous les besoins de l'usager.

De nombreux outils informatiques visant à "l'analyse" de texte sont proposés, mais ils portent la marque des besoins spécifiques pour lesquels ils ont été mis au point : recherches lexicales, documentaires, syntaxiques et problématique de l'édition scientifique, en particulier. Dans le domaine des textes littéraires, on sait que des logiciels comme Tact[3] permettent de renouveler l'approche des textes, mais cela n'est possible qu'après des heures de codage manuel fin, nécessitant de bonnes connaissances linguistiques, philologiques, littéraires, encyclopédiques, qu'il n'est pas possible d'automatiser, et ce travail est, en tout état de cause exclu pour les vastes corpus.

Les quatre "fonctionnalités de base de la lecture assistée par ordinateur" concernent :

- le marquage ou balisage des unités de la structure logico-linguistique du texte
- l'annotation, permettant à un lecteur particulier d'associer des commentaires à des passages
- "la prospection, par quoi on peut entendre l'ensemble des possibilités offertes pour réaliser des investigations fines du texte […] en termes lexicaux, syntaxiques, sémantiques, stylistiques, etc. La prospection gagne naturellement à s'appuyer sur des connaissances linguistiques variées et peut de surcroît exploiter diverses méthodes formelles (booléennes, statistiques, etc.).
- la structuration, ou organisation, ou classification de (segments ou unités de) textes, c'est-à-dire la composition d'entités textuelles issues de textes divers, en de nouveaux ensembles (corpus) arguments d'opérations textuelles. Cette fonctionnalité prend en compte le fait que

[1] "A l'occasion des études et travaux entrepris dans le contexte initial de la Bibliothèque de France", v. Virbel J., Maignien Y., 1999, p. 31-32.
[2] *Ibid. cf.* aussi : "le marquage ou balisage des structures logico-linguistiques des textes, introduites dans ou associées aux fichiers représentant ces textes, ainsi que diverses formes d'accès simulant des types de lectures de survol ou diagonales peuvent assumer cette fonction".
[3] V. Wooldridge T. R., 1991.

dans l'univers numérisé, toutes sortes de groupement/dégroupement de textes sont relativement aisées à exprimer et à exploiter. Elle doit reposer sur un système de catalogage, et au-delà, de représentations de connaissances encyclopédiques[4]".

On peut constater que ces fonctionnalités font partie des besoins de l'usager de bases textuelles, et nous verrons qu'une grande partie fait défaut dans l'utilisation de Frantext, l'importante banque de textes du XVI^e^ au XX^e^ siècle, à dominante littéraire, que l'INALF a mise à la disposition du public, depuis 1986 dans des stations d'interrogation et depuis 1996 sur l'internet. Même si on peut considérer que l'annotation n'est pas primordiale, puisque l'usager dispose en général de son espace de travail personnel où il peut rapatrier des résultats de recherche, sur les trois autres points, les bases textuelles que nous connaissons aujourd'hui sont loin d'être satisfaisantes. Par exemple, le British National Corpus, qui se présente comme "a unique snapshot of the English language" et comporte 100 millions de mots étiquetés, mêlant écrit et oral, est, à la différence de Frantext, documenté par une base d'en-têtes TEI, où l'on trouve éventuellement une classification rudimentaire des textes par "domaine" : pourtant, le fait que ce corpus ne propose pas de textes intégraux mais seulement des "échantillons" dans le but d'être représentatif des différents usages, justifierait un repérage précis des genres. De plus, l'option d'échantillons arbitrairement tronqués[5], ne permet pas à l'usager de disposer de la fonction "structuration-organisation" telle qu'elle est définie ci-dessus et rend problématique l'interprétation des résultats obtenus par la "prospection".

1. 1. 2. Lecture ou interprétation ?

Le terme de "lecture assistée par ordinateur" pose cependant problème car il ne rend pas compte du travail d'interprétation auquel est confronté l'usager, lié à ses questionnements mais aussi à l'état de structuration des données textuelles, aux outils (logiciels, etc.) et aux précisions sur les modes de fonctionnement de ces outils. En fait, dans ce domaine, force est de constater que l'emploi d'un mot comme *information* là où il y a opérations interprétatives d'un sujet, entraîne des ambiguïtés du même ordre, nous le verrons, que l'emploi de *langage* ou *langue*, quand il s'agit en fait d'un *texte* (c'est-à-dire d'un objet culturel bien particulier et situé).

Le contexte d'émergence de la théorie de l'information

Quelques aspects de l'histoire des idées d'une époque qui a bouleversé les méthodes d'accès aux textes méritent d'être soulignées, pour mieux comprendre certains choix. En effet, il n'est peut-être pas inutile de rappeler que la théorie de l'information assimilait le "message" à de la matière ou de l'énergie, dans un contexte intellectuel marqué par le monisme réductionniste (thèse de l'identité entre cerveau et esprit), où émergea la métaphore du cerveau semblable à l'ordinateur. La théorie de l'information de C. Shannon était partie intégrante de ce que Dupuy nomme "la première cybernétique" (c'est-à-dire le groupe pluridisciplinaire qui est à l'origine du nouveau champ du savoir nommé "sciences cognitives"[6]), où est née l'Intelligence Artificielle. "La cybernétique telle que l'a popularisée Wiener se présente comme la science des analogies maîtrisées entre organismes et machines. On peut résumer la position de McCulloch

[4] Virbel J., Maignien Y., *loc. cit*.
[5] Habert B. et *alii*, 1997, p. 146 ; cette habitude est courante dans le domaine anglo-saxon où elle est liée à la conception d'une langue une et universelle, qui néglige les systèmes de normes dont rend compte le genre. V. aussi *op. cit.*, p. 18.
[6] Dupuy J. P., 1999.

par cette citation extraite d'un texte de 1955 : "Plus nous apprenons de choses au sujet des organismes, plus nous sommes amenés à conclure qu'ils ne sont pas simplement analogues aux machines, mais qu'ils sont machines"[7]. Même si Wiener sortira la théorie de l'information du strict domaine de l'ingénieur des communications où voulait la cantonner Shannon, en la considérant comme une notion physique (et en particulier de la thermodynamique : l'information, c'est de l'entropie négative, "il est en vérité possible de concevoir tout ordre en termes de message"), "dans l'usage qui en est fait tout au long des conférences Macy, la "théorie de l'information" apparaît beaucoup moins comme la clef d'une nouvelle vision du monde que comme un outil familier, un simple moyen qu'on utilise à des fins très diverses"[8]. Et Dupuy cite la neurophysiologie (on s'interroge sur "l'énorme gaspillage d'information dont se paie le passage de la sensation à la perception"), la psychanalyse, la psycho-acoustique, et même la linguistique : "la théorie de l'information sert encore à évaluer la redondance d'une langue comme l'anglais et à s'interroger sur son éventuel caractère fonctionnel"[9]. Or, dans ce domaine, et malgré la remarque de von Foerster qu'il doit y avoir un optimum de redondance dans une langue : "plus la redondance est faible, plus la langue peut transmettre d'information ; mais plus elle est forte, mieux elle est structurée", on en viendra, l'idéologie interférant avec la science, à "juger que moins une langue est redondante (par exemple l'anglais !), plus elle est éloignée de l'"état primitif".

Bien sûr, comme la théorie de l'information bâtie par le mathématicien Shannon excluait toute référence à la signification, "il était inévitable que d'autres tentent de construire une théorie complémentaire dont l'objet serait l'information sémantique. Deux de ces tentatives furent présentées et discutées aux conférences Macy. Celle de Donald MacKay à la huitième et celle de Rudolf Carnap et de son disciple Yehoshua Bar-Hillel à la dixième"[10]. C'est donc une sémantique logique et véri-conditionnelle[11] qui prit en charge le problème de la signification : mais, comme nous allons le voir ci-dessous, cette approche ne constitue pas une prise en compte du contenu sémantique des textes, car elle se situe à la fois dans le cadre de la philosophie du langage (et non celui de la linguistique post saussurienne) et d'une "sémantique générale" (et non de la sémantique d'une langue particulière, qui est l'approche où nous nous situons).

D'ailleurs le problème du sens, et celui de l'interprète, évacués par Shannon reviennent en force, quand il présente ses recherches sur la redondance de l'anglais : "il s'agit de faire deviner à un sujet un texte qu'il ne connaît pas, lettre après lettre. Shannon a bien pris soin de rappeler d'emblée que la notion de quantité d'information est définie indépendamment du sens du message. [...] La tension est nette entre ceux qui prennent le point de vue de l'ingénieur, c'est-à-dire de celui qui décide souverainement de ce qui est signal et de ce qui est bruit, et ceux qui prennent le point de vue de l'organisation ou de l'organisme étudié, et par rapport auxquels une perturbation peut devenir sens, indépendamment de la liberté de l'observateur d'attribuer et

[7] *Ibid.* p. 42 ; *cf.* aussi : "dès le départ, pour ce qui est de Wiener et de ceux qui le suivent en tout cas, ils ont fait de l'information une grandeur *physique* (en ital.), l'arrachant au domaine des transmissions de signaux entre humains. Si tout organisme est environné d'informations, c'est tout simplement qu'il y a partout autour de lui de l'organisation, et que celle-ci, du fait même de sa différenciation, *contient* de l'information. L'information est dans la nature, et son existence est donc indépendante de l'activité de ces donneurs de sens que sont les interprètes humains." (*ibid.* p. 126).
[8] *Ibid.* p. 124.
[9] *Ibid.* pp. 125-128.
[10] *Ibid.* p. 131.
[11] *Cf.* p. 155 l'allusion au théorème de Tarski sur la vérité.

d'imposer des significations au système observé"[12]. Et en face de la simplification grossière de l'approche mathématique, les sciences de l'homme ont leur mot à dire : un psychologue gestaltiste, Klüver fait remarquer que "le non-sens absolu, c'est-à-dire un matériau totalement dépourvu de signification et ne donnant prise à aucune association, semble être demeuré un idéal qui n'a jamais été réalisé"[13].

Par ce bref rappel historique, nous avons voulu montrer qu'un ensemble de problèmes entoure la question du sens des suites linguistiques, de l'interprétation et de l'interprète, et que les réponses apportées varient selon les disciplines : dans l'actuel paradigme des sciences cognitives, le courant dit de "naturalisation du sens" considère que "l'information est une relation naturelle omniprésente, qui ne nécessite pas la présence d'un interprète : chaque fois qu'il existe une corrélation nomologique entre deux états de choses, l'état de chose postérieur indique, ou porte une information sur l'état de choses antérieur"[14].

L'approche logico-grammaticale du langage

Ce n'est pas le cadre, ici, de développer les raisons multiples qui ont contribué à ce qu'une théorie "naïve" de l'information considère les "données textuelles" comme des évidences, présentes dans l'environnement de "l'organisme" : certaines sont millénaires, bien ancrées donc dans les habitudes de pensée de la culture occidentale, et elles ont contribué à la méconnaissance du problème de l'interprétation que nous venons d'évoquer.

Nous rappellerons seulement, reprenant les termes de F. Rastier, que dans la conception aristotélicienne[15], les trois facteurs intervenant dans tout entretien sont les choses (*res*), les pensées (*intellectus*) et les paroles (*voces*), les choses et les états de l'âme signifiées par les paroles étant considérées comme identiques pour tout le monde, même si les paroles et les systèmes d'écriture diffèrent. Si cette triade subit quelques variations au fil du temps (suivant les écoles, chez les philosophes médiévaux ou les Messieurs de Port-Royal) elle reste remarquablement stable. Parallèlement, et depuis l'Antiquité, les trois disciplines que sont la grammaire, la rhétorique et la dialectique ont des affinités telles qu'elles seront regroupées (sous les noms de grammaire, logique et rhétorique) dans le *trivium*, la division inférieure et primordiale des "sept arts libéraux"[16] qui régissent la diffusion des connaissances pendant tout le Moyen Age, et sont le soubassement de notre système d'enseignement.

Or, cette conception du lien entre pensée et langage, qui fait d'une langue un simple véhicule au service de la pensée, était justement l'obstacle majeur à la naissance de la linguistique générale (qui étudie les langues et leur diversité) et plus particulièrement encore du secteur de la linguistique qui traite du sens, la sémantique[17], comme science dont l'objet n'est plus ni le concept d'ordre psychologique, ni le concept d'ordre logique, mais le signifié. La *sémantique différentielle*, selon laquelle le signifié prend sa valeur en contexte, et s'analyse en

[12] *Ibid.* p. 127-128.
[13] *Ibid.* p. 131.
[14] *Cf. Vocabulaire de sciences cognitives*, 1998, *s. v. intentionnalité* ; noter que "contrairement aux conceptions classiques qui, comme celle d'Edmund Husserl, identifient l'intentionnalité à la propriété d'un acte de représentation consciente, le concept d'intentionnalité est aujourd'hui généralement tenu pour indépendant de toute prise de conscience d'un contenu de pensée".
[15] Le texte concerné est le *Peri hermeneias*, diffusé par le célèbre commentaire de Boèce, v. Rastier F., 1990.
[16] Au Moyen Age, *les sept arts* désignent les disciplines enseignées en tant que méthodes et non en tant que connaissances abstraites, dans le *trivium* et le *quadrivium* (arithmétique, géométrie, histoire, musique -ces termes étant à prendre au sens médiéval) ; on notera que *libéral*, issu du latin, renvoie à la notion d'homme libre, par opposition aux esclaves (d'où : les activités auxquelles peut se livrer l'homme libre).
[17] Pour un exposé très clair et très documenté de la problématique du passage du concept au signifié, v. Rastier F., 1991, chapitre 3, pp. 73-114 ; v. aussi Rastier F., 1993.

termes de relations d'opposition, au sein d'une langue considérée comme système, s'est constituée depuis une vingtaine d'années seulement, sur la base des acquis de la sémantique structurale, principalement en Europe car les pays anglo-saxons restent sous l'influence de la philosophie du langage et de sa conception référentielle et logique de la signification. Dans le domaine de l'I.A., c'est cette conception logique de la philosophie du langage[18] qui a prévalu, pour des raisons de compatibilité d'approche et aussi d'importance institutionnelle, d'autant plus qu'elle rejoignait l'approche de l'autre courant dominant, la grammaire générative, qui semblait le cadre adapté à la modélisation des énoncés.

C'est dans cet ensemble de conceptions d'un héritage à la fois fort ancien et renforcé par des problématiques de recherche nouvelles, c'est dans ce cadre de pensée qu'a pu se développer la théorie de l'information comme "matière" d'une part, et d'autre part "évidente, faisant partie du monde", donc universelle et n'ayant pas à être objectivée par un sujet situé, au cours d'opérations interprétatives[19] qui lui sont propres. La constitution de la sémantique, secteur d'une linguistique qui se définit comme la sémiotique de la langue et des textes, permet de reprendre la question du sens en tenant compte des acquis des disciplines du texte, et dans une approche globale des différents paramètres de la situation[20]. Au sein du paradigme des sciences cognitives, d'ailleurs, et partant des recherches en biologie et neurophysiologie, sur la perception visuelle en particulier, on met l'accent à présent sur la relation complexe qui lie le sujet et son environnement dans ce qu'il est convenu d'appeler un *couplage*, où l'objet ne préexiste pas ; de même les découvertes vont de plus en plus dans le sens de la remise en question de la séparation entre le biologique et le mental, entre le corps et l'esprit.

Quand on utilise l'informatique pour chercher de "l'information", que fait-on ? On cherche si l'on peut s'approprier, sur la base de nos acquis et compte tenu de nos motivations, des "connaissances" nouvelles dans les textes, les suites linguistiques attestées que l'ordinateur peut mettre à notre disposition sur écran ou sur support papier, au terme de différentes opérations auxquelles on soumet ces textes. "On oppose traditionnellement la connaissance et l'action. Cette opposition est purement rhétorique, comme jadis la plume et l'épée. En effet, connaître

[18] Les conditions nécessaires et suffisantes, les valeurs de vérité, l'analyse en structures profonde et de surface, ont paru former le cadre tout trouvé pour l'extraction et la représentation des connaissances ; rappelons que la grammaire générative, du moins au début, a évacué la question de la sémantique. L'I.A. s'est ensuite tournée vers "le lexique" pour résoudre les problèmes du contenu ; la linguistique de corpus commence seulement à aborder la question du texte.

[19] "L'autre attitude extrême, celle des vitalistes, ou celle, plus importante aujourd'hui qui s'exprime dans le paradigme de l'ordinateur et qui considère l'information comme une chose, s'égare tout autant. *Cette attitude est intéressante dans la mesure où elle plonge ses racines dans le même terreau idéologique que les explications opérationnelles ; mais elle est appliquée ici à un domaine où elle ne peut être adéquate* [en ital.]. Il est significatif que l'ingénierie des systèmes et la science informatique rangent l'information et le traitement de l'information dans la même catégorie que la matière et l'énergie. La théorie des systèmes et la cybernétique se sont développées dans un univers technologique, conscient de l'insuffisance du paradigme purement causaliste […], mais inconscient de la nécessité d'*expliciter* le changement d'attitude accompli par la communauté des chercheurs. Pour autant que les sciences de l'ingénieur demeurent des sciences appliquées, ce genre d'erreurs épistémologiques n'empêche pas la réalisation d'un travail efficace. Mais ces erreurs deviennent inacceptables et inutiles, dès qu'on tente de les exporter au domaine de la description des systèmes naturels, des systèmes vivants ou des affaires humaines. Dans ces domaines, c'est un non-sens, me semble-t-il de considérer que l'information est une *chose* que l'on transmet, que les symboles sont des *choses* qui se réduisent à leur valeur nominale, ou que la finalité et les buts sont transparents au système lui-même, comme pour un programme. En fait, l'information n'existe pas indépendamment du contexte organisationnel qui engendre un domaine cognitif, ni d'une communauté d'observateurs qui choisit de qualifier certains éléments d'informationnels ou de symboliques. L'information *stricto sensu* n'existe pas (ni d'ailleurs les lois de la nature)". Varela F. J., 1989, p. 180.

[20] Dans son histoire de la cybernétique, Dupuy illustre fort bien sa thèse des "rendez-vous manqués" entre ce groupe (qui voulait édifier une science générale du fonctionnement de l'esprit) et les sciences de l'homme et du social : l'oubli des acquis de ces disciplines a appauvri leur capacité à aborder la causalité circulaire, la complexité et la question de l'autonomie des systèmes.

n'est pas une contemplation (malgré une longue tradition dont témoigne l'étymologie même du mot *theoria*). Connaître n'est rien d'autre qu'apprendre dans une pratique sociale. La pratique scientifique est l'une d'entre elles. Or, toute pratique sociale met en jeu des performances sémiotiques (verbales, gestuelles, musicales) comme par ailleurs des flux de conscience (ou des représentations dans la théorie du cognitivisme) et des interactions physiques" […] C'est dans la sémiotisation et par la sémiotisation que l'individu biologique se transforme en sujet humain situé"[21].

Il nous paraît important de reconnaître et problématiser l'activité interprétative de l'esprit alors même que le changement de support permet de soumettre le texte à des opérations de plus en plus lointaines des habitudes intellectuelles de l'usager : avec la profusion des "données" disponibles dans les réseaux internes et sur l'internet, la demande sociale est telle que les approches simplificatrices ne peuvent plus être de mise. La communauté du TAL (Traitement Automatique des Langues) prend en compte cette réalité après un demi-siècle de recherche qui a servi à "nous faire comprendre la difficulté extrême des règles du langage humain, et la complexité extraordinaire de l'intelligence qui le produit et le décode"[22].

1. 1. 3. Langage, langue, texte … et informatique

Les disciplines des textes, qui étudiaient les textes fondateurs, -les textes saints comme la Bible ou le Coran, ou profanes, Homère, Virgile, Dante, etc.-, puis la critique littéraire ont répandu l'habitude de dénombrer les faits langagiers, pour particulariser les auteurs, les œuvres, les courants : on appelle couramment "statistiques"[23] cet aspect quantitatif qui est devenu de plus en plus important quand on a pu passer des fiches manuelles de dénombrement à des cartes perforées, puis à des supports et outils performants dotés de "mémoire". Dans les débuts de l'informatique, et sous l'impulsion de la cybernétique, on voit se répandre des recherches sur le langage et la comparaison entre les langues.

La *Bibliographie critique de la statistique linguistique*, publiée à la demande du Comité de la Statistique Linguistique en 1954[24] par P. Guiraud est bien représentative de tous les secteurs où la linguistique "quantitative" et l'informatique entamèrent leur collaboration, parce que des habitudes de dénombrements et comparaisons quantitatives avaient été prises de longue date : phonétique, métrique et versification, concordances et indices[25], morphologie, syntaxe, problèmes philologiques et historico-littéraires (classement des langues, attribution de textes, influences), langage de l'enfant. "La vieille méthode des dénombrements en faveur en Allemagne au siècle dernier et si largement pratiquée dans les Universités américaines peut avoir l'ambition légitime de contribuer à un répertoire des faits de langue mais elle échoue devant leur interprétation. La statistique est précisément la méthode qui nous permet d'analyser

[21] Rastier F., 1996b, p. 221 ; en note : "il nous semble erroné de considérer la connaissance comme un objet susceptible d'une science. La connaissance est le mode herméneutique qui permet la constitution critique des objets, non un objet parmi d'autres". V. aussi pp. 230-231.
[22] Danlos L. et Véronis J. , 1997, p. 5.
[23] Cette dénomination se retrouve dans la fonction nommée "statistiques" de Word qui permet de comptabiliser un texte en pages, mots, caractères, paragraphes, etc.
[24] Guiraud P., 1954 ; l'auteur précise que peu de recherches parmi celles citées ici ont utilisé l'informatique : quant à la science statistique, elle est seulement en émergence alors.
[25] *Ibid.*, p. 30 : "C'est surtout pour les langues classiques et pour l'anglais que nous possédons un grand nombre d'indices. […] Il est vraiment regrettable que des compilations aussi longues et fastidieuses soient effectuées deux et même trois fois. Il existe cinq indices des comédies de Térence".

et interpréter les dénombrements[26]". On retrouve dans ce champ de recherche des mathématiciens comme Shannon, Wiener et Mandelbrot, qui s'intéressent aux "lois" mathématiques qui régissent la communication, et partant, pour eux, le langage (puisque leur approche n'est ni linguistique, ni textuelle), ou à la redondance et l'entropie des langues[27], comme on l'a vu.

C'est l'époque aussi des recherches sur la détermination d'un vocabulaire de base, qui pourrait contribuer à l'enseignement et à la diffusion de certaines langues[28]. En France on s'attacha à la recherche du vocabulaire fondamental[29] en suivant l'exemple des recherches qui avaient été menées par Ogden et Palmer, de 1923 à 1928 aux États-Unis et en Grande-Bretagne pour mettre au point le Basic English. A une première liste arrêtée à 850 mots en 1928, on a dû adjoindre des listes additionnelles, en particulier pour pouvoir traduire la Bible en Basic English pour les étrangers et aussi par besoin de mots scientifiques, ce qui fait qu'il comporte près de 2000 mots en 1933, quand il commence à être timidement employé pour les premiers essais d'enseignement au Japon, à l'École navale : il fut un moyen efficace de diffusion de la langue anglaise, mais en perdant de la rigidité qui lui venait de sa conception logique et universaliste[30].

La sémantique figure également dans la recension de P. Guiraud, mais il ne s'agit pas encore, et ce pour des raisons historiques comme nous l'avons vu, d'une approche du contenu en termes de sémantique différentielle : en fait, cette subdivision est conçue comme étudiant l'étymologie et l'emploi des figures et de certaines catégories de mots privilégiées chez certains auteurs[31] ou de la comparaison entre les mots les plus fréquents dans plusieurs langues, selon des listes de fréquence établies dans des corpus qui semblaient alors importants et représentatifs.

Ces éléments d'histoire du contexte historique de la collaboration entre linguistique et informatique montrent que des faits complexes ont contribué à négliger la dimension sémiotique de la langue et des textes et le problème de l'interprétation. Nous pensons qu'à présent, grâce aux acquis de la sémantique structurale et des disciplines du texte (comme la philologie, la rhétorique et la littérature comparée), nous pouvons penser les problèmes du texte en corpus électronique dans le cadre théorique proposé par F. Rastier, et dans une approche herméneutique renouvelée. La pluridisciplinarité permet d'aborder cette complexité : "[…] le caractère pluridisciplinaire du domaine […] contraint à collaborer entre eux des spécialistes de disciplines considérées comme bien éloignées les unes des autres, et traditionnellement rattachées à des modes de pensée qui ont tendance à s'exclure : sciences expérimentales, techniques d'ingénieur, humanités"[32].

---

[26] *Ibid.* p. 1.
[27] Par exemple : Shannon C. E. et Weaver W., The Mathematical Theory of Communication, 1949 ; Mandelbrot B., Mécanique statistique et théorie de l'information, 1951 ; Shannon C. E., Prediction and Entropy in printed English, sd ; Wiener N., Cybernetics, or control and communication in the animal and the machine, 1953, etc.
[28] Elles concernent surtout l'anglais, l'allemand, le français et l'espagnol.
[29] V. Gougenheim et *alii,* 1964.
[30] *Ibid.* p. 23 ; c'est le besoin d'une langue non ambiguë chez les juristes qui est à l'origine de l'influence qu'exerça Jérémie Bentham (1748-1832) sur les institutions anglaises. "En tant que jurisconsulte et législateur, il était indigné de voir circuler tant d'idées vagues et fausses. Il en voyait l'origine dans l'emploi de mots que nous avons accepté sans les contrôler et que nous continuons à employer par habitude. Nous sommes semblables, disait-il, à des douaniers qui, parce qu'ils ont apposé une fois leur cachet sur un ballot de marchandises, se croient dispenser d'en vérifier le contenu quand il repasse sous leurs yeux." *Ibid.* p. 25.
[31] V. Guiraud, *loc. cit.* p. 63-64 : ces études concernent Balzac et Baudelaire ou les pourcentages des parties du discours selon les genres.
[32] Laporte É. 1997, p. 65.

1. 2. LA SEMANTIQUE DIFFERENTIELLE ET LE CORPUS ELECTRONIQUE

L'un des principaux intérêts des corpus électroniques dans le domaine littéraire est de pouvoir pratiquer, à partir d'hypothèses de recherche, des comparaisons et contrastes entre genres textuels et œuvres d'un même genre, pour mettre en évidence des régularités sémantiques, des évolutions diachroniques, des faits d'ordre stylistique ou thématique ; la perspective du texte, de ses structures, et celle des trois types de littérarités[33] (la littérarité générale, générique ou singulière) demandent d'aller au-delà des fonctions de type "lexical", centrées sur le repérage de chaînes de caractères, ou de type "grammatical", opérant autour de "patrons syntaxiques" et "étiquetage morphosyntaxique", qui sont offertes généralement à l'utilisateur de banques de données textuelles. Que savons-nous d'un texte quand nous sommes devant la liste des vocables et des parties du discours dont il est "constitué" ?

1. 2. 1. La problématique du signe et celle du texte

Dans la pratique quotidienne d'un ou de quelques textes d'étude, dans le cadre d'une édition de texte, ou d'une monographie, on comptabilisait traditionnellement, à la main différents types d'unités (lemmes, vocables, mots rares, rimes, niveaux de langue, etc.) : mais ce travail servait d'une part à "s'approprier" le texte, à le découvrir et s'en imprégner, et d'autre part, ces dénombrements se faisaient selon des points de vue particuliers à la pratique. Aujourd'hui, devant des listes de fréquence "radiographiant" un texte, fournies par la machine, le spécialiste de l'auteur (ou de la période) et le lecteur novice ne sont pas à égalité : l'interprétation en est aisée pour le premier, qui a dans ses connaissances des pistes pour se forger des hypothèses interprétatives, alors que l'autre est totalement démuni et risque de tirer des conclusions à tout le moins hasardeuses. La raison en est que c'est le texte qui est l'unité minimale d'interprétation : cette approche du texte se situe dans la tradition herméneutique[34], attachant une importance toute particulière au contexte pour la détermination du sens. Le cercle philologique ou herméneutique s'exprime ainsi : le global détermine le local.

Il est symptomatique que des chercheurs impliqués dans le traitement automatique des langues en reviennent à privilégier cette problématique du texte, sous deux formes :

a) la négation de "l'existence *a priori* du lexique" : Sinclair J. 1996 intitule son article "the empty lexicon" car il recommande de partir du texte pour mettre au point le lexique adéquat, en acceptant la caractéristique propre à cette approche heuristique, le caractère provisoire, puisque ce "living lexicon", ne peut être ni complet ni absolu. Il nomme cette approche celle des Thespiens (du nom de Thespis, qui passe pour être le fondateur de la tragédie) par opposition à celle des Académiciens, tenants des dictionnaires, terminologies, et grammairiens de la phrase, à la recherche des "informations" du message coupé de ses conditions de production :

[33] *Cf.* G. Molinié, dans G. Molinié, A. Viala, 1993, p. 13.
[34] "L'herméneutique n'a jamais été une discipline autonome […] Dans notre tradition, elle fut d'abord un art d'expliquer les textes fondateurs, qu'ils soient littéraires, juridiques ou religieux. […] Il reviendra à Schleiermacher (1768-1834) de formuler un ambitieux programme général. D'une part, il étend le champ de l'herméneutique du religieux au littéraire, du littéraire à l'écrit, de l'écrit à l'oral, posant ainsi pour la première fois le problème herméneutique de la conversation. D'autre part, passant du général à l'universel, il trace le projet d'une herméneutique qui exposerait les principes universels de la compréhension". Rastier F., 1996a, pp. 23-24.

l' approche qu'il recommande est, pour lui la seule façon de se confronter au cœur, au noyau central du vocabulaire[35].

b) la nouvelle importance accordée aux caractéristiques linguistiques d'unités textuelles
- soit pour "profiler" les textes des corpus ou des parties de textes en les classant en groupes de données plus homogènes (Illouz et *alii* 1999, Habert B. et *alii* 2000)
- soit pour dépasser la méthode des mots-clés, en faisant de l'interrogation texte-texte, pour la recherche d'informations pertinentes (Bommier-Pincemin B., 1999).

Nous verrons, en particulier aux chapitres 2 et 3, l'importance des niveaux de structuration des textes : dans le cadre pluridisciplinaire de la philologie numérique, ce domaine suscite des collaborations nouvelles et permet à des communautés diverses de confronter savoirs théoriques et options pragmatiques.

"On peut distinguer voire opposer les problématiques du signe, comme modèles de la signification, *in abstracto* et hors contexte, à la problématique du texte, fondée sur l'analyse différentielle et qui définit le sens par l'interaction paradigmatique et syntagmatique des signes linguistiques, non seulement entre eux, mais avec le texte dans sa globalité"[36]. A côté des outils issus de la "problématique du signe", nous pensons que de nouveaux outils d'aide à l'interprétation peuvent être mis au point en réintroduisant la dimension du texte comme structure, pour renouveler les outils intellectuels qui ont fait leurs preuves dans les approches traditionnelles des textes.

1. 2. 2. Formes textuelles et cohésion

Nous présentons ici les principaux concepts de la sémantique différentielle utilisés dans cette étude, qui seront introduits et exemplifiés tout au long des chapitres[37].

L'accès au sens est considéré comme un processus de type "reconnaissance de formes" (et non comme un "calcul"[38]), que F. Rastier nomme "perception sémantique"[39]. Les concepts de la sémantique interprétative permettent de décrire les *parcours interprétatifs* et d'expliciter comment le linguistique impose des contraintes sur les représentations mentales : les faits sémantiques doivent être construits, et ce, en ne négligeant pas de prendre en compte la situation de communication -production et réception. Ainsi nous montrerons au chapitre 2 que certains énoncés du corpus examiné, bien que datant de la fin du XIX^{è}s. attestent des mots disparus de l'usage, pour lesquels un élève de français chercherait vainement une définition dans son dictionnaire préféré, parce que l'auteur préconise le retour à des formes et thèmes poétiques médiévaux[40] .

La cohésion textuelle est engendrée par la récurrence de traits sémantiques (ou sèmes), articulés en structures stables, propagés dans différentes zones de localité et dans l'ensemble du texte, grâce à des "mises en mots", des lexicalisations variées : l'effet de ces récurrences

[35] C'est ce que nous appelons expliciter la logique sémantique d'une forme, grâce au corpus, v. chapitre 2. L'évocation du fondateur de la tragédie rapproche les Thespiens de Sinclair de la tradition herméneutique-rhétorique, alors que les Académiciens nomment les tenants de l'approche logico-grammaticale, chez Rastier.
[36] Rastier F., 1996a, p. 15.
[37] Le glossaire à la fin du volume en donne les définitions.
[38] Ce qui est le cas de la conception logique et compositionnelle du sens.
[39] V. Rastier F., 1991, pp. 205-223 ; v. en annexes du chapitre 1 un exemple de dessins prêtant à interprétation différente suivant le rapport figure/fond (pour la perception visuelle).
[40] V. dans le glossaire la définition de *ordre herméneutique* ; sur le concept de "lecture" comme "texte à vocation métalinguistique", v. F. Rastier, 1987, pp. 231-246 et sur la prise en compte de l'émetteur et du récepteur dans la situation de communication : *Id.*, 1989, chapitre III.

produit des *isotopies sémantiques*[41]. Les *isotopies génériques* à l'œuvre dans un texte constituent des "fonds sémantiques" sur lesquels se détachent des "formes sémantiques", constituées d'ensembles de traits sémantiques, les *molécules sémiques* [42] : tout cela induit "l'illusion référentielle"[43] qui donne l'impression, par exemple, de "voir" tel personnage, tel décor, telle action en train de se dérouler. L'isotopie générique principale d'un texte est celle qui permet d'interpréter "le fond", c'est-à-dire "ce texte est une histoire d'amour, un recueil de recettes de cuisine, un traité de minéralogie" : sur ce fond vont se détacher au fil de l'interprétation des "formes" et des "configurations" particulières au texte.

Plutôt qu'une "lecture productive" qui "réinterprète le texte au gré du récepteur", F. Rastier propose une "lecture descriptive" répondant "à l'objectif modeste mais ambitieux de restituer le contenu du texte en reconstituant l'entour de la communication initiale."[44]

Les opérations interprétatives

Dans la description du signifié d'une unité lexicale en contexte, les notions d'*assimilation*, *dissimilation*, *virtualisation*, *actualisation* de traits sémantiques permettent d'expliciter comment une occurrence peut s'écarter du "type"[45], et s'appuient sur l'analyse des sèmes afférents propagés par le contexte : par exemple au chapitre 2, nous verrons des énoncés où le trait /animal/ définitoire de la sémie-type *biche* est virtualisé au profit du trait /humain/ dans des sémies-occurrences. En fait, c'est parce que nous avons l'habitude de consulter des dictionnaires qui décontextualisent le sens que nous pouvons considérer que le type est ainsi "déformé" dans l'occurrence, alors que c'est l'inverse : les définitions lexicographiques font abstraction de certains traits (ou y suppléent) dans le but de permettre l'application à toutes sortes de contextes, cela fait partie des contraintes de ce genre de discours particulier[46].

Les composantes textuelles

Au plan du texte, les concepts de *composantes textuelles, la thématique, la dialectique, la dialogique* et *la tactique* permettent de décrire les structures textuelles et de rendre compte de la cohésion des différents systèmes sémiotiques à l'œuvre dans tout texte : leurs modes d'interaction permettent aussi d'aborder autrement la question d'une typologie des textes, en la

[41] La sémantique opère au-dessous et au-dessus du niveau du mot, c'est-à-dire au niveau de la plus petite unité de signification, le sème ou trait sémantique, et au niveau des structures stables d'un texte, les isotopies, les thèmes, les fonctions narratives.

[42] Les isotopies génériques sont des classes sémantiques (comme l'eau, l'animalité -étudiées chez tel auteur) alors que les groupements de traits stables que sont les molécules sémiques ne sont liées à aucune classe déterminée, leur lexicalisation pouvant appartenir à des classes très diverses (comme le groupement stable des traits /jaune/ /visqueux/ /chaud/ et /néfaste/ repéré dans L'Assommoir, v. Rastier 89, p. 57).

[43] "L'illusion référentielle substitue à tort la réalité à sa représentation, et a à tort tendance à substituer la représentation à l'interprétation que nous sommes censés en faire" (M. Riffaterre dans Barthes R. et *alii*, 1982, p. 93). Pour F. Rastier (qui lui préfère *impression référentielle*), "les images mentales sont les corrélats psychologiques des signifiés linguistiques, et (…) la référenciation s'opère par appariement entre images mentales et percepts d'objets".

[44] "En tant que discipline scientifique, seule la sémantique interprétative peut y prétendre. Elle s'appuie nécessairement sur les résultats de la philologie ; mieux encore, les développements de ces deux secteurs de la linguistique se conditionnent mutuellement." Rastier 89, pp. 51-52 ; la note 42 précise : "Nous ne prétendons pas à la restitution d'un seul sens contrairement à ce que présumaient les philologues du siècle dernier, le contenu textuel peut être plurivoque."

[45] Le type (v. chapitre 2) correspond *grosso modo* à la "représentation en langue", à la description lexicographique du sémème (par ex. la subdivision I A de tel dictionnaire, correspondant à la synchronie du texte) ; dans un énoncé, un sémème peut perdre un sème inhérent, cf. ci-dessous au chapitre 2, l'exemple de *biche*, dans le signifié duquel le sème /animal/ est neutralisé et remplacé par /humain/. Sur la différence entre description statique et dynamique et les notions de type et occurrence, v. Rastier F., et *alii*, 1994, pp. 55-56, 89-91.

[46] La lexicographie ne donne pas "la vérité" sur les mots, malgré l'attente des utilisateurs : le dictionnaire est soumis à des contraintes de discours et de genre, de même que le glossaire d'édition de texte, ou la monographie sur un thème, etc.

fondant sur des critères linguistiques. Ces composantes sont des universaux de méthode, qui rendent compte de points de vue divers sur la textualité, mais elles ne sont ni ordonnées, ni hiérarchisées *a priori*. D'autre part, "chacune des composantes peut être régie par trois types de systématicité : le système fonctionnel de la langue ; les normes sociolectales (dont les normes de genre) ; les normes idiolectales (qui seules sont facultatives). Un continuum s'étend du premier au troisième. Généralement les prescriptions du premier l'emportent sur celles du deuxième, qui l'emportent à leur tour sur celles du troisième. Toutefois, dans des conditions favorables, les usages d'un individu peuvent devenir la norme d'un groupe ; et la norme d'un groupe, l'emporter sur la norme standard qu'on nomme la langue. Pour tous les constituants de la langue et jusqu'aux règles syntaxiques, les contradictions entre types de systématicité sont de puissants facteurs d'évolution"[47].

La philologie numérique

Données informatisées et pratiques sociales : une banque est un objet culturel

Frantext est un ensemble textuel issu des données qui furent saisies, sur bandes mécanographiques, à partir de 1965, pour constituer le fonds de référence du dictionnaire *Trésor de la langue française des XIX-XXè s.*, publié par le CNRS : cette banque offre actuellement une collection de textes électroniques du XVI^{è} au XX^{è}s. Le corpus initial comportait environ 70 millions de signes et 1000 unités textuelles des XIX^{è} et XX^{è}s. : actuellement Frantext recouvre les périodes XVI^{è}-XX^{è}s., surtout littéraires, et représente un ensemble de 2800 textes environ. D'autre part, un corpus de textes de la période du moyen-français -1330-1500- a été constitué pour la rédaction du *Dictionnaire du Moyen-Français*, en cours d'élaboration sous la direction de Mr R. Martin : quand il sera accessible au public, des recherches diachroniques larges, sur des textes complets, seront donc possibles pour le français[48].

Un ensemble de choix scientifiques et de limites d'ordre technique explique certains aspects de l'état actuel de la base Frantext. Les premières consignes de saisie datent de 1966 et sont tributaires des possibilités techniques de l'époque, en matière de saisie de fiches mécanographiques et d'informatisation sur l'ordinateur Bull Gamma 60 (par exemple la limite à 64 caractères typographiques). D'autre part, le but d'assistance à la lexicographie, et des possibilités de documentations complémentaires, ont introduit des disparates : par exemple, le recours aux fiches manuscrites du fonds "Inventaire Général de la Langue Française" de M. Roques et F. Lecoy, (dépouillé entre 1936 et 1969) explique des choix "historiquement justifiables" de ne pas saisir certains textes mais ces lacunes peuvent poser problème aujourd'hui dans un cadre où la communauté scientifique et le public doivent avoir accès légitimement à des œuvres complètes ou à des textes importants pour des périodes ou mouvements littéraires. Les limites techniques anciennes ont marqué les textes électroniques dont la qualité philologique doit cependant être reconnue et mérite l'effort d'amendement

[47] Rastier F., 1989, p. 105 ; un exemple de la norme d'un groupe qui se répand actuellement est l'usage de *grave* comme adverbe pour marquer l'intensité (*cf.* "j'hallucine grave", où le signifié du verbe est renouvelé également).
[48] Il faudra cependant disposer d'outils réglant certains problèmes comme la variation graphique, et des connaissances particulières sur des genres textuels disparus, par exemple ; et la question de l'enrichissement du corpus pour l'interrogation revêt des aspects encore plus importants dans le cas de la distance diachronique, car l'usager ne peut se fier à sa "conscience linguistique".

entrepris, en particulier au niveau du codage, surtout si l'on songe à l'état des textes qui se diffusent quotidiennement sur l'internet.

Reste ouverte la question de la constitution d'une banque "raisonnée", pour le français, qui serait représentative des emplois que les différents genres textuels font du système fonctionnel de la langue et des usages qui se sont imposés (pour l'écrit et l'oral transcrit). Cette banque devrait faire la part du patrimoine littéraire[49] (auteurs du premier et du second rayons), tout en permettant l'accès à des corpus de domaines variés des sciences et des techniques, des arts et loisirs, etc. : elle engage des coûts considérables et nécessite donc des partenariats multiples[50]. Les recommandations de la philologie électronique devraient être entendues pour ne pas multiplier des saisies inexploitables par manque de respect de règles déontologiques élémentaires. "Par un nouveau rapport empirique aux textes, comme par les problèmes nouveaux que posent leur codage et leur parcours, la numérisation conduit aussi à un renouvellement de l'herméneutique philologique. La grammaire y retrouve sa place : comme le travail sur la langue part des textes pour y revenir, elle demeure une discipline auxiliaire de l'entreprise philologique"[51].

Le corpus de référence

De l'ensemble textuel Frantext, nous avons choisi de constituer un corpus homogène en genre, comprenant environ 350 romans de 1830 à 1970, c'est-à-dire en commençant à une période où la langue se stabilise[52] et en ayant une représentation équivalente des tranches de temps[53]. Le corpus de référence comprend également quelques textes de nouvelles (le plus important étant le recueil de *Contes et Nouvelles* de Maupassant[54]) et quelques contes, ces deux genres fictionnels ayant été classés avec le roman par le système documentaire : la question s'est donc posée de savoir si on acceptait de mener la recherche sur ce corpus "hybride".

Il s'agissait d'inaugurer un type d'expériences de sémantique sur gros corpus (350 œuvres, représentant 40 millions de signes) et d'autre part, la problématique de départ était de vérifier si on pouvait mettre en évidence des régularités entre les textes pour la composante thématique : il nous a semblé possible d'accepter cette relative hétérogénéité, puisque sur le plan thématique, rien ne permet *a priori* de faire une démarcation nette entre le roman, le conte et la nouvelle, pour la période XIX-XX^{è}s. Les entreprises typologiques peinent à tracer la frontière entre ces genres, pour cette synchronie, et, en tout état de cause, il n'y a pas de critères linguistiques proposés pour cette distinction : compte tenu du fait qu'il n'était pas possible de constituer à part un corpus de nouvelles où le "poids" de Maupassant soit correctement relativisé, ni un

[49] Sur ces exigences, v. par exemple Émelina J.,1998 : "La fréquence inhabituelle d'un mot-clef dans une œuvre est toujours digne d'intérêt. Cet intérêt, plus encore, naît de la comparaison avec d'autres œuvres du même auteur. C'est ce que fait, par exemple, Jean Rohou (*L'évolution du tragique racinien*, Paris, Sedes, 1991, pp. 163-64) à propos des termes "fuir", "partir", "quitter", "sortir", "séparer" dont il relève un emploi exceptionnel dans *Bérénice*, par rapport aux autres tragédies de Racine. Plus intéressante encore serait, au sein de cette langue tragique aux allures si ressemblantes, une comparaison avec d'autres auteurs contemporains. Nul doute que notre connaissance du génie racinien serait améliorée si l'on confrontait le vocabulaire de sa *Phèdre* avec celui de *Phèdre et Hippolyte* de Pradon, celui de l'*Ariane* de Thomas Corneille ou du *Bellérophon* de Quinault. La statistique lexicale a encore d'immenses terres à défricher".

[50] C'est tout un consortium qui est à l'origine du BNC, v. Habert B. et *alii*, 1997, p. 158.

[51] Rastier F., 2000, pp. 120-121

[52] V. Brunot F. 1968, 530-539.

[53] A l'époque où le corpus de référence a été constitué, la tranche 1970-1990 ne comportait que peu d'œuvres et les besoins d'une équipe qui s'intéressait aux "marges" du français orientait le choix des textes saisis dans Frantext vers des romans "populaires" : ces différents éléments ne permettaient pas une bonne représentativité de la période contemporaine.

[54] La dénomination *contes* n'est pas due à Maupassant mais à son éditeur rouennais.

corpus suffisamment fourni de contes, pour pouvoir comparer roman et genres brefs, on a accepté cette particularité du corpus, qui ne nous a pas paru plus gênante, *a priori*, que les autres types de disparates par lesquelles s'expliquent les difficultés de classement des genres narratifs[55]. Mais on s'est donné pour consigne supplémentaire d'observer si d'éventuelles différences pouvaient être rapportées à ce critère. Par commodité, nous avons décidé de nommer *corpus Roman* ce corpus de genres narratifs où le roman est largement dominant[56] : dans les deux opérations de contraste du chapitre 2, il est opposé à un corpus technique et à la poésie, qui ne sont pas des genres narratifs.

Le corpus ne comporte qu'un texte du courant du "nouveau roman", *La Route des Flandres* de C. Simon ; mais d'autres, ne se réclamant pas de cette école, se démarquent du corpus par des faits linguistiques à rattacher à des techniques narratives modernes, comme *Le Chinois d'Afrique*[57] de R. Sabatier. Les études que nous avons menées confortent les conclusions des pionniers du traitement statistique des textes, comme P. Guiraud, Ch. Muller et É. Brunet, sur l'importance du genre (et plus généralement des quatre niveaux de structuration des textes proposés au chapitre 2) dans la caractérisation des œuvres : ce qui impose de "documenter" les textes électroniques selon les recommandations de la philologie électronique[58]. "[…] les normes linguistiques varient tout à la fois selon les discours (qui correspondent à des types de pratiques), les genres (qui correspondent à des situations typiques) et enfin les styles. Par ailleurs, en raison de la détermination du global sur le local, les variations sont déterminées par les discours, les genres et les styles individuels, et elles ne peuvent être ordonnées et comprises que si l'on ménage, dès le choix du corpus, les conditions de cette compréhension"[59].

1. 3. PERSPECTIVE SEMIOTIQUE ET TEST STATISTIQUE

Dans la perspective sémiotique qui est la nôtre, l'utilisation de l'approche statistique est quelque peu différente de celle qui a prévalu dans les nombreux travaux de statistique lexicale, lexicométrie ou statistique textuelle qui ont été menés depuis les années 70.

1. 3. 1. Statistique et corpus électronique

Une approche différente

En effet, puisque notre cadre théorique considère que le mot isolé n'a pas de sens, nous n'utilisons pas la statistique à la recherche de mots-clés ou de mots-thèmes, et si ce qu'il est convenu d'appeler "le vocabulaire caractéristique" d'un corpus ou ses "spécificités" nous intéresse, c'est que nous les considérons dans le cadre d'une théorie sémantique. Nous avons

[55] De fait, quand on mène une recherche thématique, on s'aperçoit que des textes "atypiques" comme les contes fantastiques de Nodier, sont très rarement sélectionnés avec des romans historiques, ou d'apprentissage, par exemple : ce sont bien des critères sémantiques et génériques qui rassemblent des passages parallèles, ce qui autorise à aller plus loin dans ce type d'expériences pour fonder une typologie des textes sur des critères "formels" en distinguant, à l'aide des méthodes statistiques, des faisceaux de "qualités génériques".

[56] Nous verrons que pour le plan des variables morphosyntaxiques, le roman et la nouvelle sont plus proches que le conte et le roman (Malrieu D. et Rastier F., à paraître).

[57] Ces deux textes ont fait l'objet d'une analyse contrastive avec *Le Père Goriot*, au plan du système des signes de ponctuation dans Bourion E. 1998, reproduit ici, *cf.* annexe du volume 1.

[58] Des catégories comme écrit/oral, pour convaincre/pour distraire (Biber D., 1992, 1993), prose/vers (Frantext) etc. ne permettent pas de rendre compte de la diversité des textes.

[59] Malrieu D. et Rastier F. à paraître.

voulu tester l'hypothèse que le test probabiliste permet l'accès à des éléments qui font que le corpus de travail est un "tout signifiant" : il met en évidence des faits qui sont à interpréter, en termes de traits et associations de traits, donc au niveau inférieur et supérieur au mot. Par sa nature, il calcule un écart (mesure issue de la comparaison entre un modèle théorique et une réalité observée) et cette démarche est tout à fait compatible avec celle de la sémantique différentielle, qui analyse les relations en termes d'opposition et contraste à l'intérieur d'un "système" : le texte, objet culturel, polysémiotique et signifiant dans ses différentes composantes : langue, genre, style, caractéristiques de la présentation matérielle, lignée et influences, etc. Il peut permettre ainsi de caractériser, selon les unités quantifiables, un "individu-roman" par rapport à une "population de romans", pour reprendre les termes de la statistique. L'interprétation des faits, le passage du quantitatif au qualitatif revient, bien sûr au spécialiste (linguiste, littéraire) qui a, de toutes façons, assuré la phase initiale de constitution d'hypothèses, de choix des caractères, population, individus à étudier. Pour que cette opération statistique soit valide, il faut d'une part que le corpus de travail fasse partie du corpus de référence, et d'autre part que les variables soient en grand nombre et indépendantes, ce qui est le cas dans les suites linguistiques (texte, partie d'un texte, ou parties d'un corpus centrées sur un mot-pôle) car de multiples déterminations sont à leur origine : complexité du système fonctionnel de la langue, normes de discours, de genre, préférences stylistiques, projet esthétique, lignées et influences, etc. Ch. Bernet, qui a pratiqué l'analyse assistée de textes classiques, reconnaît que "ce qui est significatif en termes de quantité découle de causes très diverses […] Ce qui provient des choix stylistiques, conscients ou inconscients, n'est pas toujours aisément identifiable, et souvent délicat à interpréter[60]". Cette opinion cependant est liée à l'approche "restreinte" qui est celle de la statistique *lexicale*, d'une part, et d'autre part au but de repérer grâce à la statistique des faits d'ordre *stylistique* : nous pensons que les écarts sur les variables calculés par le test probabiliste sont à interpréter au plan sémantique et textuel et qu'ils attirent l'attention sur toutes sortes de faits qui font de ce texte un tout signifiant, et pas seulement sur ceux étudiés traditionnellement par la stylistique. Comme nous le verrons au chapitre 2, la sémantique lexicale est une approche trop restreinte pour décrire les faits observés en corpus : dans l'interprétation, on ne peut éliminer le recours au contexte, à des zones contextuelles de plus en plus larges, qui dépassent le cadre de l'énoncé, pour aller du texte jusqu'au genre et au discours.

Si tous ceux qui ont étudié des textes littéraires à l'aide des outils de la statistique ont toujours été frappés de voir comment cet "idiot hypermnésique qu'est l'ordinateur"[61] peut les surprendre, c'est que le travail d'interprétation des faits quantitatifs peut les amener sur des chemins imprévisibles, liés à toutes sortes d'éléments concernant l'auteur, la période, le courant littéraire, etc. : "La recherche informatisée ne renouvelle pas la critique ; elle l'affine, l'infléchit, lui donne des assises plus solides qu'elle est seule capable de fonder. En architecture, les matériaux employés ne décident pas de la structure ni de la destination d'un édifice, mais de son aspect et de sa couleur. En littérature, les analyses quantitatives permettent enfin d'apprécier à leur juste valeur la "couleur", "le grain" et la "densité" d'un texte en fonction du matériel lexical choisi par l'auteur. Certes, rien ici n'a le caractère irréfutable et systématique d'une lecture scientifique de données. Le quantitatif n'est qu'un élément susceptible d'aider à

[60] Bernet Ch., 1983, p. 181.
[61] Bernard M., 1999, p. 85.

une meilleure évaluation du qualitatif, et le général doit rester au service du particulier. Mais, ces réserves admises, la statistique lexicale dont le domaine ne cesse de s'étendre, a le mérite d'offrir au littéraire d'immenses terres nouvelles où braconner"[62].

Dans ce cadre, le débat déjà ancien sur la préférence à accorder, pour la recherche des spécificités, au calcul hypergéométrique ou au test de l'écart réduit[63], comme celui qui porte sur la possibilité théorique de traiter le texte à partir de l'image du schéma de l'urne où l'on tire aléatoirement des boules blanches ou noires[64], sont sans objet. De même, la lemmatisation[65] : est une question épineuse qui a longtemps opposé des groupes de la communauté de la lexicométrie : mais des travaux récents sur de vastes corpus, utilisant l'analyse multifactorielle pour "radiographier" les textes sous l'angle de leurs graphies, de leurs lemmes ou des étiquettes morpho-syntaxiques de leurs composantes montrent des "images" très proches[66], parce que les facteurs explicatifs, de signifiance et de cohérence, se situent à d'autres niveaux (discours, genre, projet esthétique, etc.). Dans un récent travail, É. Brunet a enrichi un corpus de 26 textes, de même genre (roman), de 13 auteurs (en prenant des textes aux deux extrémités de leur production littéraire) avec l'étiquetage permis par l'analyseur morpho-syntaxique Cordial pour lui appliquer l'analyse multi-factorielle, dans le cadre de son Hyperbase[67] : les classements opérés, pour les parties du discours ou pour les variables "modes, temps, personne", calculant la distance entre les œuvres d'auteurs différents comme la distance intra-auteur (entre les deux textes séparés par un laps de temps important) sont très stables. Les résultats de ce type de recherche renforcent la conviction que des régularités sont à découvrir, à l'aide de ces outils, dans des corpus enrichis et permettant des partitions diverses, sous-tendues par des hypothèses sur la "texture" des textes (comme des paramètres de l'énonciation, contrastés dans des parties de dialogue et de description) : le changement d'échelle permet d'objectiver des faits que les spécialistes des textes littéraires avaient déjà perçus dans un cadre plus restreint.

La véritable question rejoint les points évoqués à propos de la philologie numérique : l'accès à des corpus suffisamment fiables, enrichis et pertinents pour l'hypothèse qu'on veut tester. D'autre part, comment disposer des éléments nécessaires pour l'interprétation (c'est-à-dire : à quoi renvoie ce contraste numérique ?) sachant que certains d'entre eux, sur l'histoire de la langue ou l'histoire littéraire, ou celle de faits de "civilisation" sont à rechercher dans des ouvrages spécifiques, ce qui pose la question d'une "station de travail" bien équipée. Il y a, de plus, un problème matériel et ergonomique d'accès aux éléments sur lesquels opèrent les calculs statistiques : quand nous avons abordé ce travail de contraste sur un corpus de 350 romans, nous avons été étonnée de constater qu'il n'existait pas de documents textuels spécifiques pour interpréter les résultats du test. C'est pourquoi nous nous sommes attachée à

[62] Émelina J., 1998, p. 100.
[63] *Cf*. Brunet É., 1983a, Cossette A., 1994, v. aussi Valceschini-Deza N., 1999, pp. 116-124 ; la question a été abordée également dans notre groupe de travail par B. Gaiffe et L. Romary, *cf. L'Accès sémantique aux banques textuelles*, 1996, pp. 50-59.
[64] Mellet S., 1990.
[65] Brunet introduit son article avec un rappel des pièces du dossier, *cf*. Brunet É., à paraître ; cette question sera évoquée ici à plusieurs reprises, puisqu'une lemmatisation satisfaisante au plan de la langue n'est pas réalisable automatiquement sur de vastes corpus, avec les outils dont nous disposons actuellement : mais dans la pratique interprétative, on arrive à jongler pour faire coïncider l'unité de l'ordinateur et celle de sa conscience linguistique.
[66] Malrieu D. et Rastier F., à paraître, Brunet É, à paraître.
[67] Ce "produit", tout à fait original, conçu par une des rares personnes qui allient culture informatique et statistique et formation littéraire et linguistique, offre à la fois des corpus et des outils statistiques, ainsi que la possibilité d'appliquer ces outils à ses propres corpus ; il a été conçu en 1989, et est constamment amélioré depuis. Dans les annexes du chapitre 6 figurent des exemples de "sorties Hyperbase".

concevoir des documents contextuels, qui reconstituent la "chair" du texte, à côté des listes qui l'éclatent et en représente une sorte d'ossature séquentialisée. Nous les avons élaborés avec notre point de vue sémiotique et sémantique, dans le but d'avoir en même temps sous les yeux le maximum d'éléments du plan local et du plan global, pour faciliter l'interprétation. En effet, nous pensons que les critiques à l'endroit des méthodes statistiques dans le domaine du texte, émises par les littéraires surtout, qui en sont même venus à demander "un moratoire"[68] sur leur pratique, proviennent du manque de documents adéquats pour l'interprétation : il nous a paru important de les concevoir sur la base de ceux qui ont fait leurs preuves dans les disciplines du texte depuis des siècles, les concordances et les tableaux synoptiques, en les enrichissant des résultats de la pondération statistique. "La recherche sur les textes à l'aide de l'ordinateur n'effectuera de véritables progrès que si elle se concentre sur la mise au point de modèles d'analyse qui soient propres au texte électronique ; il faut apprendre à extraire de celui-ci des informations qui conduisent à des interprétations significatives tout autant qu'imprévisibles à la lecture du texte imprimé[69]".

Avec l'aide de J. Maucourt, informaticien à l'Inalf, nous avons mis au point un programme d'aide à l'interprétation, le SAAS (Système d'Aide à l'Analyse Sémantique) pour tester l'emploi du test probabiliste sur vastes corpus et pour confronter les concepts descriptifs de la sémantique interprétative aux résultats de la mise en œuvre de cette méthode statistique.

Le modèle statistique

Le modèle statistique à partir duquel sont conçus les programmes du SAAS est celui qu'a mis au point Ch. Muller et qu'il a utilisé dans toutes sortes de travaux passionnants, qui offrent des pistes à parcourir à l'aide des nouvelles possibilités que vont offrir les corpus enrichis. On le retrouve également dans les travaux d'É. Brunet, et dans Hyperbase.

La notion de *hasard* est importante pour cette démarche : "(…) l'homme ne peut imiter le hasard. Si l'on invite des sujets à écrire une longue séquence de 0 et de 1 répartis "au hasard", on obtient des résultats qui sont plus ou moins irréguliers, ou plus réguliers que ceux que donnerait un tirage au sort dans une urne contenant des 0 et des 1 en quantités égales ; s'il s'agissait de mêler ainsi les dix chiffres de 0 à 9, l'échec serait encore plus flagrant"[70]. Elle conduit à formuler l'*hypothèse nulle* : "Le linguiste sait fort bien que, dans la langue qu'il étudie, la catégorie d'un phonème exerce une influence sur la catégorie du phonème suivant, sans pourtant la déterminer strictement. Si un phonème est une voyelle, la probabilité pour que le suivant soit aussi une voyelle est faible : pour que les deux suivants soient des voyelles, plus faible encore. Il y a là une très forte **contrainte**, qui n'a guère besoin d'être démontrée ; mais, pour décrire son action de façon objective et précise, et pour l'assortir d'une mesure (car il est possible et probable qu'elle varie d'un idiome à l'autre), il faudra d'abord savoir ce qui se passerait si cette contrainte n'existait pas, c'est-à-dire si consonnes et voyelles se succédaient comme si elles avaient été tirées au sort dans une urne qui contiendrait 43,5% de voyelles et 56,5% de consonnes. Hypothèse absurde, sans doute, mais indispensable pour mesurer la contrainte phonétique définie plus haut. C'est ce qu'on appelle une **hypothèse nulle**, fondée sur le libre jeu du hasard"[71].

[68] Lusignan S., 1985 ; *cf*. aussi Émelina J. 1998 et Bernard M., 1999.
[69] Lusignan S., 1985, p. 211.
[70] Muller Ch., 1992 [réimpr. de l'éd. de 1973], p. 22.
[71] *Ibid*. p. 48.

Nous employons ces notions selon la perspective sémiotique : étant donné un corpus homogène selon les critères de discours et de genre, il existe un certain nombre de faits langagiers communs aux textes du corpus (la partie centrale de la courbe de Gauss, ci-dessous), et d'autres qui sont particuliers à tel texte (ou tel corpus de travail formé de séquences entourant un mot pôle, où se diffusent des traits sémantiques spécifiques, qui lui donnent, par hypothèse, une densité sémantique particulière) : les cooccurrents sortis par le test, peuvent permettre d'interpréter ces faits, parce que la présence de ces signes dans cette zone de localité ne peut être due au hasard. Après analyse du plan de l'expression (cooccurrents sélectionnés) on peut qualifier comme *corrélats* ceux qui sont pertinents au plan du contenu[72], et qui ressortissent aux multiples "variables" dont est composée la "texture" du texte : c'est le cas des mots de l'ancienne langue sélectionnés dans un poème de Moréas ou de ceux qui sont répétés plusieurs fois chez Péguy, pour créer le rythme litanique, et ce, contrairement aux règles du "bien écrire" qui proscrivent les répétitions, surtout dans les textes littéraires, (*cf*. chapitre 2) .

"[…] on retiendra qu'une enquête statistique comporte en principe les phases suivantes :

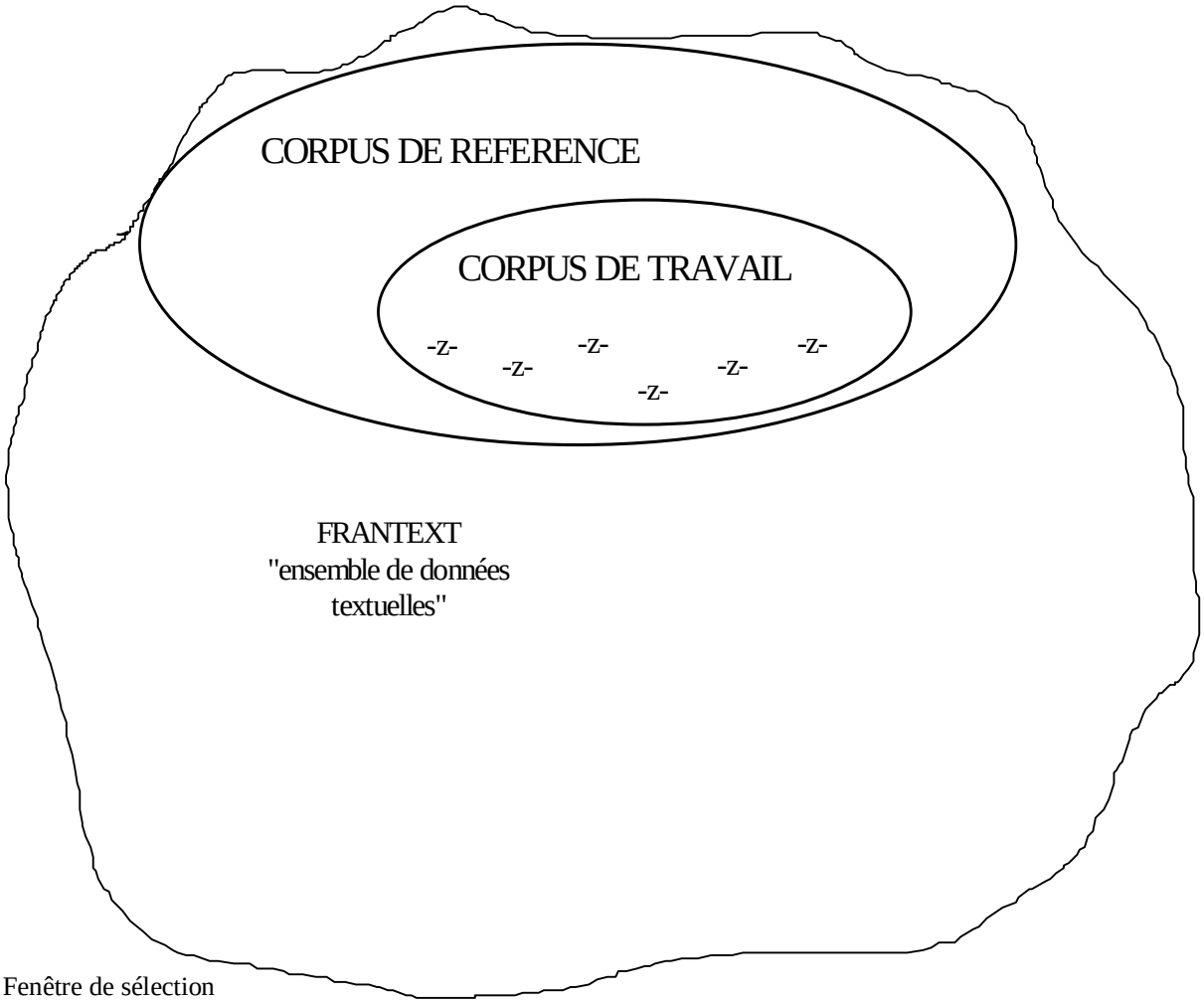
- la construction d'un modèle théorique ;
- l'observation de la répartition réelle, et des écarts qui existent entre celle-ci et le modèle théorique ;
- l'application à ces écarts d'un test statistique, qui les appréciera en probabilité ; si la probabilité est forte, les écarts sont déclarés non significatifs : l'hypothèse nulle ne peut être rejetée, et aucune conclusion linguistique ou stylistique ne peut être retirée de l'expérience ; si la probabilité est faible, les écarts ne peuvent être attribués au hasard : le fait linguistique ou stylistique existe ;
- l'interprétation des écarts entre modèle et observation, si ceux-ci ont été reconnus significatifs"[73].

La formule du test de l'écart réduit est celle-ci :

$$r = \frac{F_{obs} - F_{théor}}{\sqrt{F_{théor} \times (1-p)}}$$

où p est le rapport du nombre d'occurrences du corpus de travail (formé par exemple par l'ensemble des contextes d'une fenêtre de sélection autour d'un mot pôle), au nombre d'occurrences du corpus de référence (corpus Roman) : l'écart réduit permet d'apprécier la déviation d'une fréquence observée (Fobs.) par rapport à la fréquence théorique (Fthéor.) ($Fthéor. = Fabsolue * p$) qu'elle devrait avoir si l'on fait l'hypothèse que les mots se répartissent au hasard dans un texte (plus le score d'écart réduit est élevé, moins la cooccurrence s'explique par le hasard). On a retenu un seuil r=3 à partir duquel on considère que la cooccurrence est non aléatoire[74].

[72] Selon les concepts proposés par Hjelmslev ; nous verrons aux chapitres 2 et 3 que ces cooccurrents sélectionnés peuvent n'être que des parties du plan du contenu (par ex. une composante d'un syntagme), mais ils y donnent accès. Rappelons qu'il n'y a pas correspondance terme à terme entre les unités des deux plans du langage, celui de l'expression et celui du contenu.
[73] *Ibid.* p. 54.
[74] L'hypothèse nulle (…) est rejetée à partir du score 1.96 (*cf*. ci dessous la courbe de la loi normale).

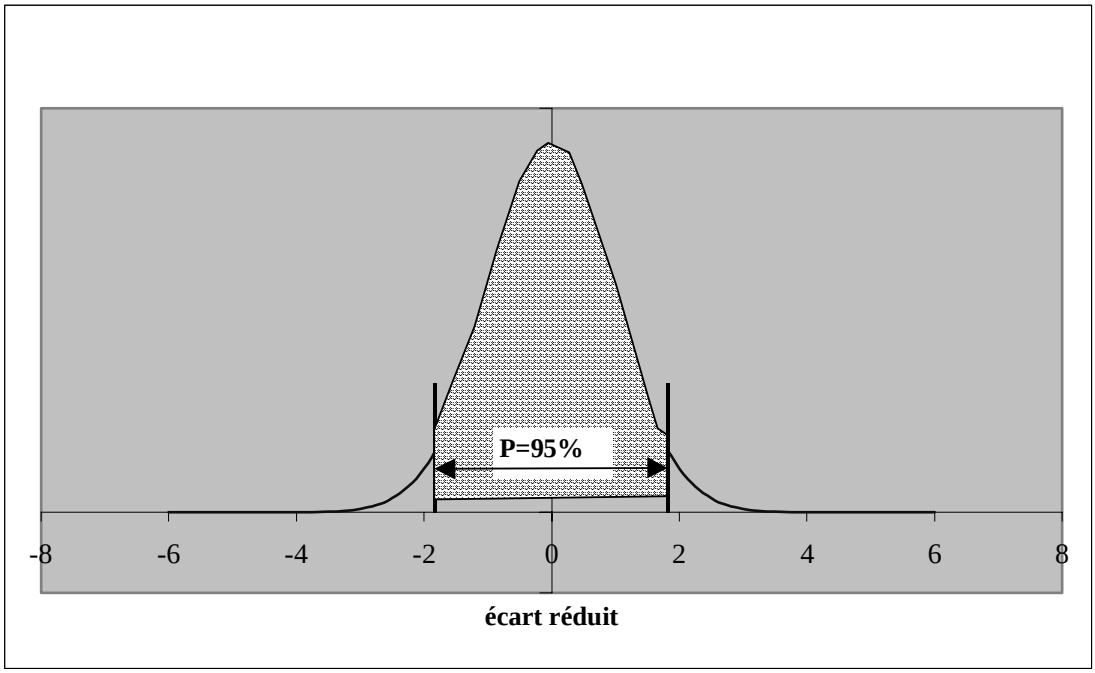


Les corpus de référence et de travail où opère la sélection par le test statistique

Le choix du seuil de rejet s'opère ainsi : "Dans les tests qui nous intéressent, on admet généralement un risque d'erreur de 0,05 ou 5%[75], ce qui met le seuil de rejet à deux écarts types (exactement 1,96) de la moyenne, de part et d'autre de celle-ci ; cela revient à dire : en rejetant l'hypothèse nulle et en admettant l'hypothèse contraire, j'ai une chance sur 20 de me tromper, au plus (il va sans dire que ceci est vrai si le résultat franchit à peine le seuil ; plus il le dépasse et plus le risque est petit). C'est un risque que l'on peut admettre avec sérénité ; car il est assez probable que parmi les hypothèses linguistiques ou philologiques proposées sans apparat statistique, il s'en trouve bien une sur 20 qui est fausse[76]. […] Si l'on veut être plus prudent, (c'est le cas dans d'autres sciences, par exemple en médecine quand on doit tester une

[75] Ce seuil est visible dans les graphiques issus d'Hyperbase, *cf.* annexes du chapitre 6.
[76] En note : ne pas oublier qu'un seuil a une largeur, dans l'acception courante du mot : il subsiste une marge entre l'intérieur de la maison et la rue, une zone indéterminée ; notion à transposer dans les tests d'hypothèse.

médication nouvelle), rien n'empêche de prendre un intervalle d'acceptation plus large, par exemple avec 2,5 écarts types ; au-delà de 3 écarts types, on peut admettre que le risque d'erreur devient pratiquement nul. […] Du reste, dans nos disciplines, les choses ne prennent pas en général l'allure d'un dilemme dramatique ; elles conduisent plutôt à un classement des faits, du plus significatif au moins significatif, qu'à une dichotomie stricte entre significatif et non significatif"[77].



Loi normale centrée réduite

"Les fréquences observées ne sont que le reflet, toujours imparfait, inconstant et aléatoire, d'une réalité constante (dans des limites à définir) qui est faite de probabilités. Celui qui n'observe et n'interprète que les fréquences ne voit que des apparences, comme les prisonniers qui, enchaînés dans la caverne mythique, ne voient que les ombres des objets et des êtres. Le calcul statistique et la connaissance des lois statistiques communes aux phénomènes les plus divers sont des moyens d'atteindre une réalité plus solide et plus simple. En linguistique, ce sont des moyens qui, derrière les apparences du discours, font entrevoir les réalités de la langue"[78]. Nous ajoutons : "des discours, des genres et des projets esthétiques", dont des études rigoureuses peuvent être menées sur des corpus enrichis et bien documentés. En effet, nous pensons que la question de l'hétérogénéité des scores obtenus, à l'origine de nombreuses critiques faites, à juste raison, aux résultats statistiques, proviennent de l'inadéquation entre corpus de travail et corpus de référence, que nous évoquerons dans ce travail.

Pour différentes raisons, nous avons testé notre méthode sur un corpus littéraire : or, le degré de complexité est maximal dans les textes littéraires, dans lesquels le niveau agonistique se superpose au niveau événementiel, comme nous le verrons au chapitre 6. Cette complexité reflète l'intrication des composantes textuelles. Cependant, la méthode peut être appliquée à toutes sortes de corpus enrichis, moyennant une réflexion préalable sur l'adéquation entre

[77] *Ibid.* p. 93.
[78] Muller Ch., 1992, p. 81.

corpus de référence et corpus de travail, en fonction de la tâche[79] : nos résultats montrent qu'on peut construire une description fine des acteurs, des interactions et des rôles, ce qui importe au premier chef dans toute interprétation du contenu, que les acteurs possèdent ou non le trait /humain/.

1. 3. 2. SAAS, le système d'aide à l'interprétation des textes électroniques

Pour l'aide à l'analyse sémantique, on dispose d'une batterie de documents obtenus à la suite de différentes opérations informatiques et statistiques, que nous détaillons ci-dessous : pour les spécimens, v. annexes du chapitre 1.

I. Cas de recherche thématique autour d'un mot-pôle

1) sélection du corpus de travail autour du mot pivot ; le point arrête la sélection
2) tri entre les "mots sémantiques" et les "mots grammaticaux"[80]
3) recherche des candidats cooccurrents : élimination des noms propres et des mots grammaticaux. Les candidats sont des mots sémantiques ou des signes de ponctuation (suivant option : lemmatisation à l'aide d'un dictionnaire de référence)
4) sélection des cooccurrents par le test de l'écart réduit : on élimine ceux dont la fréquence de cooccurrence avec le mot pôle est 1
5) résultats du test

Les résultats sont présentés sous diverses formes :
listes de cooccurrents sélectionnés par le test de l'écart réduit
documents contextuels (ensemble des séquences où voisine(nt) un/des cooccurrent(s) sélectionné(s) avec le mot-pôle)
tableaux synoptiques de résultats obtenus sur différents corpus.

a) Listes de cooccurrents sélectionnés

Huit sortes de listes proposent les cooccurrents sélectionnés par le test de l'écart réduit, en fonction de critères différents de présentation.
Dans chaque liste on donne, pour un *lemme*, 3 renseignements numériques :
- le score d'*écart réduit*, fourni avec des décimales.
- la fréquence du lemme dans le *corpus de travail*.
- la fréquence du lemme dans le *corpus de référence* : elle tient compte de toutes les formes du lemme (y compris celles qui ne sont pas rencontrées dans le corpus de travail).

Liste de cooccurrents : présentation par score d'écart réduit (spécificités positives)
Liste de cooccurrents : présentation par score de fréquence
Listes alphabétiques (sous-ensembles)

[79] Comme notre travail en "linguistique de corpus" se situe au carrefour de plusieurs disciplines, linguistique, littérature, informatique linguistique, nous présenterons dans les chapitres 2 à 5 :
- une partie descriptive : les faits observés dans le corpus
- une partie méthodologique, récapitulant les problèmes que nous avons rencontrés dans l'étude et faisant des propositions d'amélioration des données et des outils.

[80] En corpus multi-auteurs, pour la recherche thématique, on peut se dispenser du traitement des mots grammaticaux, qui sont étudiés de toutes façons sous l'angle de la diffusion des traits sémantiques avec les mots sémantiques sélectionnés.

- Présentation par liste alphabétique des cooccurrents qui ont satisfait aux contraintes "fortes" : score stat. = ou >4 et fréquence de cooccurrence = ou >4. Cette liste permet de cerner rapidement les mots les plus consensuels pour lexicaliser les traits sémantiques récurrents autour du mot pôle.
- Présentation par liste alphabétique des cooccurrents qui ont satisfait aux contraintes suivantes : 3= ou < score stat.<4 et fréquence de cooccurrence = ou >4.
- Présentation par liste alphabétique des cooccurrents qui ont satisfait aux contraintes suivantes : score stat. >4 et fréquence de cooccurrence <4.
- Présentation par liste alphabétique des cooccurrents qui ont satisfait aux contraintes suivantes : 3= ou <score stat. <4 et fréquence de cooccurrence = ou <4.

Liste de cooccurrents : présentation par score d'écart réduit (spécificités négatives).
Présentation par liste alphabétique des cooccurrents dont le score d'écart réduit est négatif.

Cette présentation en listes constituées sur la base de critères différents est destinée à favoriser la vérification d'hypothèses, de raisonner dans le sens langue --> notion, aussi bien que dans le sens notion --> langue (démarches sémasiologique et onomasiologique alternées). Elle permet en outre de susciter ou vérifier des éléments d'interprétation en rapport avec les composantes textuelles (thématique, dialectique, dialogique et tactique), leurs interactions et aussi les normes sociales ou idiolectales.

Pour favoriser l'émergence d'hypothèses, les cooccurrents de fréquence de cooccurrence 2 sont décalés sur la droite dans la liste par ordre de score, car ils peuvent ressortir à des normes idiolectales : soit qu'ils ne lexicalisent pas les traits sémantiques de la molécule sémique avec les "mots" qui sont "consensuels" chez les auteurs de l'ensemble étudié, soit parce qu'ils ont le statut de "point nodaux d'interprétation" (ils peuvent lexicaliser à la fois des traits d'isotopies spécifiques et génériques), soit que leur présence dans la sélection provienne d'une répétition. Dans les listes alphabétiques, l'astérisque signale les cooccurrents ne se trouvant que dans un texte du corpus.

b) Documents contextuels

Dans ce type de document, où apparaissent deux ou plusieurs lemmes sélectionnés par le test, on lit le mot-pôle au centre et le contexte qui l'environne est classé par ordre alphabétique des formes des lemmes, sur une ligne, en format paysage, du type des "concordances". Toutes les formes des différents lemmes sélectionnés par le test d'écart réduit apparaissent en majuscules, et font l'objet de deux classements, selon la présence à gauche ou à droite du mot pivot.

Les références d'auteur, titre, date figurent, de façon abrégée, au bout de la ligne.
D'autre part le fichier comporte deux sous-ensembles :
les contextes "denses" (c'est-à-dire comportant au moins deux cooccurrents sélectionnés)
les contextes pauvres (ne comportant qu'un seul cooccurrent)

Ce type de document contextuel permet un dépouillement optimal de gros corpus car il limite le bruit au maximum, et favorise la formulation d'hypothèses sur les interprétants des relations sémantiques : en mettant en évidence les types morpho-syntaxiques figés à côté des associations libres, il présente l'avantage d'autoriser le repérage des traits sémantiques dans des zones de localité comparables.

c) Tableaux synoptiques des résultats de la sélection par le test (r = ou >3)

Ces documents ont été élaborés pour les cas où la recherche sémantique porte sur un "champ lexical" que l'on suppose "thématiquement homogène" (ensemble de mots-pôles, substantifs, adjectifs, verbes contenant le trait /peur/, tels que *terreur, horrible, effrayer, faire peur*, etc., famille morphologique *-haïr, haine, haineux-*, ou mots qui possèdent des traits sémantiques communs et dont on veut comparer les univers sémantiques, mots réputés "synonymes", pour contraster un auteur par rapport à un corpus, ou pour observer de façon optimale les résultats de la sélection si on fait varier la taille de la fenêtre de recrutement des corrélats, etc.

II. Cas de contraste d'une œuvre sur un corpus de référence

1) sélection du corpus de travail (par exemple, le texte entier, chaque chapitre[81], *cf*. chapitre 6)
2) recherche des candidats cooccurrents : élimination des noms propres . Les candidats sont des mots sémantiques, des mots grammaticaux, ou des signes de ponctuation[82].
3) sélection des cooccurrents par le test de l'écart réduit ; on élimine les graphies de fréquence de cooccurrence 1
4) résultats du test

a) Listes des mots du texte sélectionnés par le test, les "mots spécifiques"

Comme dans le cas d'une recherche autour d'un mot pôle, différentes listes sont proposées :

- classement par score statistique
- classement par score de fréquence
- classement par combinaisons de critères, avec présentation par ordre alphabétique

b) Documents contextuels : concordances des "mots spécifiques"

Chacun des mots sélectionnés apparaît au milieu d'un contexte de type "concordances" en format paysage, les autres "mots spécifiques" figurant dans l'environnement étant en majuscules ; la pagination est précisée, pour le plan de la composante "tactique".

c) Tableau synoptique présentant les résultats du test par subdivisions (chapitres) et pour l'ensemble

d) Fichier comportant une version "spécificités" du texte : les mots sélectionnés sont en majuscules.

[81] En l'absence de balises spécifiques, on effectue ceci "à la main", dans un traitement de texte, après vérification de la pagination dans l'ouvrage-source.
[82] Pour ce cas, le traitement des graphies est préférable à celui des lemmes (*cf*. chapitre 6) ; à la différence du corpus multi-auteurs, on applique le test également aux mots grammaticaux, qui caractérisent certains aspects du texte comme unité (pour ceux qui sont très fréquents on peut refuser la sortie "concordances", puisqu'ils sont étudiés avec les mots sémantiques, mais l'usager grammairien peut en tirer profit).