L'article est publié dans trois versions :

[version en espagnol] Pincemin Bénédicte (2010) - « Semántica interpretativa y textometría », *in* Duteil-Mougel Carine & Cárdenas Viviana (éds), Semántica e interpretación, *Tópicos del Seminario*, 23, Enero-junio 2010, pp. 15-55. (ISSN 1665-1200 ; trad. Sebastián Giorgi).

[version française abrégée] Pincemin Bénédicte (2011) - « Sémantique interprétative et textométrie », *Corpus*, 10, 259-269.

[version française complète] Pincemin Bénédicte (2012) - « Sémantique interprétative et textométrie », *Texto! Volume XVII, n°3,* coordonné par Christophe Cusimano.

Sémantique interprétative et textométrie

Bénédicte Pincemin, CNRS
Laboratoire ICAR, Université de Lyon,
ENS, 15 parvis René Descartes, BP7000, 69342 Lyon cedex 07

Résumé

La textométrie propose une approche et des outils pour analyser les corpus numériques. À première vue, elle semble condamnée par une représentation trop fruste du texte : sac de mots (eux-mêmes simples chaînes de caractères), élimination éventuelle des hapax (qui pourraient être des lieux de singularités significatives), traitement statistique quantitatif. Et pourtant, les chercheurs en sémantique interprétative ont expérimenté et précisé comment la cooccurrence mesurée par la textométrie pouvait être mise à profit pour la description thématique, et comment une approche quantitative pouvait se révéler efficace pour la caractérisation des textes et des genres textuels. Pour éclairer ces réussites, on entreprend donc ici de repérer des adéquations essentielles entre la théorie linguistique de la sémantique interprétative, et les principes fondateurs de l'approche textométrique. Et les connivences sont nombreuses : la place centrale des textes à toutes les étapes de l'analyse, le souci de rester au plus proche du texte et d'éviter toute préconception réductrice, le rôle déterminant du contexte global construit par le corpus de référence, le fonctionnement différentiel des calculs contrastifs comme des tris... Enfin, la textométrie suscite actuellement de nouvelles propositions et de nouveaux développements : sans doute la sémantique interprétative peut-elle nourrir la réflexion théorique sous-jacente et participer tant à la conception de fonctionnalités pertinentes qu'à l'élaboration de repères méthodologiques et à la mise au point d'interfaces. Car, pour la textométrie, au plan technologique et méthodologique, comme pour la sémantique interprétative, au plan théorique et linguistique, c'est bien dans une activité de constitution et de parcours d'un corpus que se construit et s'affermit peu à peu un sens.

Résumé court

La textométrie propose une approche et des outils pour analyser les corpus numériques. Les chercheurs en sémantique interprétative ont expérimenté et précisé comment la cooccurrence mesurée par la textométrie pouvait être mise à profit pour la description thématique, et comment une approche quantitative pouvait se révéler efficace pour la caractérisation des textes et des genres textuels. Pour éclairer ces réussites, on entreprend donc ici de repérer des adéquations essentielles entre la théorie linguistique de la sémantique interprétative, et les principes fondateurs de l'approche textométrique. Dans le contexte de renouveau actuel des logiciels textométriques, la sémantique interprétative est appelée à nourrir la réflexion théorique sous-jacente.

1. Introduction

Cet article fait le point d'une réflexion, de discussions et d'expériences sur la pertinence de l'approche logicielle textométrique au regard des principes de la sémantique interprétative.

La *textométrie* s'appelle aussi logométrie ou statistique textuelle, c'est la forme actuelle de la lexicométrie (Lebart & Salem 1994). Elle propose des procédures de tris et de calculs statistiques pour l'étude d'un corpus de textes numérisés. À ces procédures quantitatives la textométrie articule fortement des moyens de parcours et d'interprétation qualitatifs, déterminants quant aux affinités possibles avec une théorie linguistique telle que la sémantique interprétative.

La textométrie ne se confond pas avec la linguistique de corpus. Toutes deux fondent leurs investigations sur un corpus numérique, dont la constitution est déterminante. Comme son nom l'indique, la linguistique de corpus poursuit un objectif de description et de modélisation de la langue. La textométrie, centrée sur le texte, a pu être mobilisée par diverses sciences humaines (histoire, littérature, sciences politiques...). Développée au sein d'une communauté scientifique s'intéressant à l'analyse des données textuelles (ADT)[1], elle se caractérise notamment par certains calculs fondateurs, statistiques (les spécificités, les cooccurrences) ou non (les segments répétés, les concordances), et accorde une place fondamentale au « retour au texte » (bien outillé dans les logiciels) pour interpréter les unités (généralement des mots) sélectionnées par les calculs. Une étude qui utilise une approche et des outils textométriques, et qui vise à observer et décrire des phénomènes linguistiques en corpus, peut donc relever à la fois de la linguistique de corpus et de la textométrie, sans pour autant qu'aucun de ces deux courants ne subsume l'autre.

La *sémantique interprétative*, théorie linguistique développée par François Rastier (1987, 2001) est *a priori* connue des lecteurs de ce numéro. Cet article nous amènera de toutes façons tout naturellement à en rappeler les principes fondamentaux, pour les confronter aux caractéristiques de la textométrie.

Rastier a suggéré des affinités de sa théorie avec des technologies de traitement actuelles (Rastier 1991, 2001, Rastier *et al*. 1994), il explore et expérimente aussi, dans le cadre de collaborations ou de la direction de travaux de recherche, certains outils logiciels sur corpus. Mais il passe le relais pour le prolongement, l'approfondissement et la systématisation de la réflexion sur les formes de technologies appropriées à une analyse interprétative. Le mémoire de Prié (1995) est une des premières réflexions générales sur ce sujet. En ce qui concerne l'approche textométrique[2], nous précèdent des publications mettant au point des usages textométriques dans un contexte de sémantique interprétative, généralement étayés par des réflexions théoriques tout à fait pertinentes pour notre sujet -et ces expériences nous seront très utiles ici. Mais peu d'investigations sont focalisées sur les liens entre les fondements de la théorie de Rastier et les principes de l'approche textométrique et proposant une réflexion d'ensemble, une synthèse : c'est la question que nous voulons aborder ici.

Notre démarche sera en quatre temps. D'entrée de jeu, il faut dépasser des *a priori* négatifs sur la textométrie quant à sa compatibilité avec une approche linguistique. Puis l'article invite à partager

1 Les actes des *Journées internationales d'Analyse statistique des Données Textuelles* (JADT) publient de nombreuses communications sur des questions de théorie textométrique et sur des exemples d'application variés, ainsi que des communications relevant d'autres approches d'analyse textuelle outillée. Une édition en ligne de ces actes est accessible depuis le site *Lexicometrica* (http://www.cavi.univ-paris3.fr/lexicometrica/), plus spécialement dédié à la textométrie.

2 D'autres technologies peuvent outiller avec complémentarité des analyses sur corpus en sémantique interprétative (cf. note 31), et d'autres théories linguistiques peuvent inspirer et guider bénéfiquement des pratiques textométriques.

l'enthousiasme des premières découvertes, premières rencontres lumineuses de la sémantique interprétative avec la textométrie, et leur consolidation en une proposition méthodologique pour l'analyse thématique. Le troisième temps est le cœur de cet article : tenter de comprendre pleinement les affinités de la théorie sémantique et de l'approche outillée en revenant à leurs principes fondamentaux, et prendre ainsi du recul avec une vue englobante. Sur cette base, le quatrième et dernier temps peut ouvrir quelques perspectives prometteuses.

2. Discussion de réductions à première vue compromettantes

2.1. Le texte, un « sac de mots » ?

Pour la textométrie, le corpus est segmenté en unités (habituellement de l'ordre des mots) : cela est lié à la procédure technique d'indexation, nécessaire pour mettre en œuvre efficacement tant les fonctionnalités de recherche de motifs que de calcul statistique. Le corpus peut être structuré en parties, typiquement les textes. Dans certains logiciels[3], il est également possible de définir des voisinages locaux utilisables pour les recherches ou calculs de cooccurences. Première remarque positive donc, le texte n'est pas aussi complètement déstructuré que pourrait le laisser entendre l'image du « sac de mots » : la textométrie mobilise *(i)* une contextualisation globale, *(ii)* une contextualisation locale, et enregistre *(iii)* la relation d'ordre des mots selon la linéarité du texte.

En fait, il s'agit de ne pas attribuer à l'ensemble de *l'analyse* textométrique ce qui est momentanément requis pour *un calcul* : ainsi, un calcul statistique de spécificités mobilise fortement les contextualisations globales mais ignore les relations de succession immédiate des mots[4] ; un autre calcul, de cooccurrences par exemple, peut ne considérer que les voisinages locaux et ignorer l'enchaînement exact des mots aussi bien que leur contextualisation globale ; un troisième, de segments répétés, ne considérera quant à lui que les enchaînements de mots, sans considération pour leur contextualisation globale. Pour autant, les résultats de ces calculs ne doivent s'interpréter qu'à l'aide de méthodiques retours au texte, qui permettent de lire celui-ci et d'observer directement les voisinages, successions et localisations complètes des mots.

Les recherches actuelles en textométrie sont de plus particulièrement sensibles au développement de nouveaux traitements intégrant la dimension syntagmatique du texte : « topologie » basée sur une modélisation du déroulement en tranches successives, « topographie » avec le succès[5] de la carte des sections implémentée dans *Lexico 3* (Mellet & Salem 2009).

De plus, ce n'est pas parce que à un moment donné le calcul s'appuie sur *une* segmentation en unités et *un* découpage en parties englobant ces unités, que le choix de ces unités et ces parties en textométrie est unique et prédéfini[6]. Si l'option la plus courante consiste à étudier le corpus à travers

3 Les phrases dans *Weblex* (exploitées par le moteur de recherche *CQP*), les sections de *Lexico 3*,...

4 Le découpage en parties peut aussi être un moyen de restituer, en première approximation, le déroulement syntagmatique du texte (la « tactique », dans les termes de la sémantique interprétative), ou le déroulement chronologique du corpus. Par exemple, Bourion (2001) propose une étude du *Père Goriot* structuré par son découpage en chapitres.

5 Il n'y a pas de palmarès scientifique ici, mais on peut tout simplement constater que les cartes de sections sont par exemple copieusement mobilisées dans le recueil (Salem & Fleury 2008), rassemblant des exemples d'analyses diversifiées réalisées avec le logiciel *Lexico 3*.

6 Même si dans certains logiciels l'importation du corpus suppose le choix d'une seule segmentation en « mots » (ex. *Lexico 3*) ou d'un seul découpage en « parties » (ex. *Hyperbase*), rien n'empêche de créer autant de bases textométriques -autant de versions interrogeables du corpus- que l'on souhaite, en variant la définition des unités ou des parties.

son lexique, on peut tout aussi bien le voir par le biais d'autres descriptions (ex. catégories grammaticales) et d'autres paliers (le morphème ou son approximation par des tri-grams, le tour de parole,...). De même, la variation du choix de la partition (découpage du corpus en parties) est un moyen de rendre compte de catégories méta-textuelles et de caractérisations philologiques (comme le genre, l'auteur, la période historique), de variables situationnelles significatives de la production des textes, comme de structurations intratextuelles (ex. chapitres).

2.2. Simples chaînes de caractères

Le « sac de mots » aurait non seulement la faiblesse d'être un sac, mais aussi celle de ne rassembler généralement pas même des mots (définis de façon acceptable linguistiquement), simplement des chaînes de caractères extraites mécaniquement du corpus. Plus généralement, il s'agit là d'abord d'un constat pragmatique : en pratique, l'analyse se base sur le matériau textuel numérisé, alors même que l'enjeu est d'accéder à des observations sémantiques, « on manipule des chaînes de caractères pour étudier des signifiés » (Rastier 2001 p. 79).

La textométrie ne prétend même pas partir des signifiants lexicaux, elle se base sur une segmentation linguistiquement grossière, approximative[7]. Autrement dit, elle relativise la question du bon choix des unités initiales[8]. Ce faisant, elle adopte selon nous une position fortement en affinité avec la sémantique interprétative : car si elle tolère la possible grossièreté de la représentation initiale[9], c'est qu'elle considère que les véritables unités sont non pas à l'initiale mais à l'aboutissement des traitements. Par la vue globale qu'elles intègrent, les statistiques et autres décomptes dégagent les grandes régularités traversant le corpus et déterminent ainsi la redéfinition des unités locales. Concrètement par exemple, c'est bien ainsi que la technique des segments répétés a été pensée, dès les débuts de la textométrie (Lafon & Salem 1983) : rectifier et ajuster *a posteriori*, au vu du corpus, des segmentations initiales malvenues.

Des expériences vertigineuses ont aussi démontré par les faits l'étonnante robustesse des analyses statistiques. Brunet (2006b) imagine ainsi de décrire le corpus non seulement par ses graphies (les mots en tant que chaînes de caractères entre deux blancs ou autres caractères séparateurs) ou ses lemmes, mais aussi par ses étiquettes morphosyntaxiques seules (ex. « ncms » pour « nom commun masculin singulier », indépendamment de la forme de ce nom), par les séquences de catégories grammaticales entre deux ponctuations (ex. « pvdn » pour « pronom verbe déterminant nom »), par les mots décomposés en séquences de quatre caractères (ex. « fenêtre » étant représenté par « fenê », « enêt », « nêtr », « être »), par les graphies réécrites comme successions de consonnes et voyelles (ex. « tant » comme « sont », « dans », « sang »,... deviennent « CVCC »), etc. Or, globalement, une analyse factorielle ou arborée appliquée à ces textes, représentés de façons si diverses et quelquefois si réductrices[10], fait ressortir les mêmes configurations de proximités ou

7 Les segmentations classiquement proposées dans les logiciels de textométrie sont définies à même la chaîne de caractères, sur des critères typographiques certes précis et pertinents, mais ne concordant pas toujours avec les structures linguistiques. Bien sûr, si le corpus est enrichi et présente déjà une analyse en unités, celle-ci peut en principe être exploitée par l'analyse textométrique, en remplacement de, voire en complément à, une segmentation d'ordre typographique (cela dépend de l'implémentation logicielle).

8 Il est vrai malgré tout que la communauté textométrique a été longtemps traversée par le fameux débat sur la lemmatisation préalable des corpus, maintenant peu à peu dépassé avec la coexistence puis l'articulation de plusieurs descriptions (Mayaffre 2005 ; Brunet 2011 chapitre 9).

9 Toute analyse initiale n'est cependant pas nécessairement bonne à prendre (Poudat 2006), en particulier une analyse irrégulière, ou difficile à interpréter (opacité de son mécanisme de production ou de la signification effective des étiquettes), ou inadaptée au corpus et à l'objectif de recherche, compromet l'exploitation des calculs textométriques.

10 Jusqu'au sentiment d'une représentation dénaturée : « désincarnation » du texte, « données perverties », « perte […] irrémédiable » (Brunet 2006b).

d'oppositions entre eux. Autrement dit, aussi fruste et indigente que soit la description initiale, certaines régularités textuelles sont telles qu'elles peuvent être captées par un traitement qui exploite pleinement la dimension globale du corpus.

2.3. Elimination courante des hapax, qui pourraient être des lieux de singularités significatives

Pour alléger les traitements statistiques, un élagage par les fréquences est communément pratiqué : concrètement ici, les mots de faible fréquence -voire aussi les mots grammaticaux de haute fréquence- sont écartés du calcul. Le cas particulier des hapax (mots de fréquence 1) est même dans certains cas naturellement favorable à une telle mise à l'écart, car ils sont d'emblée non pertinents si le calcul prend appui sur les répétitions. L'élagage sur les fréquences permet des analyses avec un bon aperçu des structures d'ensemble ; en revanche, en bonne pratique, il ne devrait pas être la seule vue sur le corpus, et des explorations ou des calculs plus focalisés sont l'occasion de reconsidérer des unités écartées dans un premier temps.

Néanmoins, il n'est pas sûr que les singularités pertinentes pour la description linguistique correspondent nécessairement à des basses fréquences au plan statistique. Dans le projet *Princip* (Valette 2004), on porte l'attention sur des néologies, au départ rares car innovantes ; mais celles-ci sont en fait composées de morphèmes qui peuvent être au contraire très présents en corpus, et qui s'avèrent de meilleures unités de caractérisation que le lexique[11]. Ou encore, derrière la variété de manifestations d'un thème (Rastier 1995) -lexicalisations diverses, synthétiques ou diffuses, etc.-, on pourrait retrouver une modélisation unique et donc assez fréquemment sollicitée, sous la forme d'un ensemble de mots « isotopants » (manifestant ensemble un sème commun), et se réalisant par la cooccurrence de quelques-uns d'entre eux (cf. le concept de « communauté » dans Bommier-Pincemin 1999). Le recours à un dictionnaire sémique serait une technique inverse pour capter et amplifier les récurrences de sèmes (Reutenauer *et al*. 2009). Bref, les éléments sur lesquels sont fondés les descriptions ne sont peut-être pas tous si rares ni uniques qu'il peut paraître, même si les manifestations apparentes en corpus sont singulières.

2.4. Traitement quantitatif vs qualitatif

Les calculs textométriques sont bien sûr quantitatifs. Or la sémantique interprétative n'est pas une sémantique formelle, dans laquelle le sens se modéliserait comme un calcul. Pour autant, une approche quantitative peut trouver une pertinence, si le volume du corpus est conséquent[12].

Tout d'abord, la textométrie exploite notamment des modèles statistiques. On mesure l'écart entre une répartition aléatoire des mots, et leur comportement effectif. Le principe opératoire, c'est donc bien de considérer la langue, et plus précisément son usage manifesté en corpus, comme réglés par des contraintes linguistiques, par opposition au seul hasard (Lafon 1980). On observe en effet que les calculs font ressortir des liens lexicaux, syntaxiques, sémantiques (isotopies de la sémantique interprétative[13]), génériques (relevant du genre textuel), stylistiques...

---

11 Le travail au niveau morphologique est évidemment aussi très pertinent pour l'analyse de la terminologie scientifique, avec les procédés de conceptualisation et d'emprunts (Loiseau 2006, Valette 2006).

12 Le lecteur humain est évidemment le meilleur interprète d'un texte, par comparaison avec tout traitement automatique et machinal, qui n'est d'ailleurs jamais une véritable lecture. Les atouts de l'ordinateur sont sa vitesse de calcul et sa mémoire : l'intérêt est de les mettre à profit pour aider la lecture humaine, lui suggérer des points d'appui, des pistes d'investigation, lorsque le volume des textes dépasse les capacités cognitives.

13 « Le concept d'isotopie est [...] basé sur la notion de redondance de l'information, c'est-à-dire d'une certaine façon sur un élément quantitatif. Si les traits d'une isotopie ne sont pas directement observables, puisque ce sont des

Si les calculs sont quantitatifs, l'analyse textométrique intègre une démarche qualitative englobante. En amont d'un calcul, il s'agit tout d'abord de formuler une problématique de recherche, de construire un corpus pertinent, de déterminer un point d'entrée approprié et de varier éventuellement les sous-corpus de travail, de définir le type de traitement adapté, de l'ajuster le cas échéant : bref, autant d'opérations qualitatives déterminantes. Et en aval d'un calcul, il est bien entendu que ce que l'on obtient est un résultat, non une réponse[14] : il reste toute la part d'interprétation, de qualification éventuelle de certains phénomènes[15], et de la progression de l'analyse en élaborant un parcours interprétatif. Rien de cela n'est donné en soi, ni fourni par le calcul. L'automatisation du calcul ne condamne en rien à un usage machinal.

3. Des expériences positives révélatrices

Après avoir écarté les objections empêchant même de considérer l'approche textométrique, nous invitons le lecteur à partager les premières découvertes de la textométrie par la sémantique interprétative.

3.1. Le contraste d'un texte par rapport à un corpus de référence

Il faut commencer par mentionner l'expérience marquante, par Rastier lui-même, de voir le calcul de l'écart-réduit mettre en évidence, à même le texte d'une nouvelle de Maupassant, des formes clés pour son interprétation : « alors qu'il m'avait fallu dix années pour comprendre l'importance du nombre *dix* dans la nouvelle de Maupassant intitulée *Toine* (l'auteur, 1989, liv.II, chap.V), le test de l'écart réduit [appliqué dans le cadre d'un corpus de référence soigneusement constitué] me l'a instantanément mis sous les yeux, et m'a même permis de tirer parti d'une occurrence à la première ligne, qui m'avait, je l'avoue, échappé, bien qu'elle eût renforcé mon propos. » (Rastier 2001 p.96 note 1)

Le calcul avait été produit par Bourion, à l'aide des programmes qu'elle avait mis au point avec Maucourt.[16]

> ON le connaissait à DIX lieues aux environs le père Toine, le GROS Toine, Toine-ma-FINE, Antoine Mâcheblé, dit Brûlot, le CABARETIER de Tournevent.
>
> Il avait rendu célèbre le HAMEAU enfoncé dans un pli du vallon qui descendait vers la mer, pauvre HAMEAU PAYSAN composé de DIX MAISONS normandes entourées de fossés et d' arbres.
>
> Elles étaient là, ces MAISONS, blotties dans ce ravin couvert d' herbe et d' ajonc, [...]

Figure 1 : Début de la nouvelle *Toine* de Maupassant, avec mise en évidence des

éléments du signifié et non du signifiant, ce caractère quantitatif peut-être la base de son identification. Les concepts descriptifs de la sémantique interprétative ne sont donc pas développés pour un cadre méthodologique quantitatif mais offrent plusieurs points d'articulation pour l'interprétation des données quantitatives » (Loiseau 2006, p.30)

14 Le calcul fournit toujours un *résultat*, fût-il vide ; en revanche, s'il est mal conçu ou n'entre pas dans une démarche d'analyse méthodique qui lui donne sens, il n'apporte aucune *réponse*, simplement des relevés oiseux et généralement encombrants.

15 « le quantitatif et le qualitatif ne s'opposent aucunement : seule une analyse qualitative peut rendre significatifs des phénomènes quantitatifs remarquables » (Rastier 2001 p.214)

16 Bien qu'ouvrant la partie de l'article consacrée à l'histoire de la découverte de la textométrie dans le cadre de la sémantique interprétative, l'expérience de Rastier n'est donc pas à proprement parler inaugurale, puisque Bourion avait déjà significativement entrepris une réflexion et des expérimentations sur le sujet. Mais cette expérience peut être considérée comme une étape, stimulant et renforçant les recherches dans ce domaine.

graphies spécifiques (écart-réduit > 3, fréquence dans *Toine* > 2) par rapport à un corpus de référence de 350 romans et nouvelles de 1830 à 1970 issu de *Frantext* (corpus décrit dans Bourion 2001)

3.2. La cooccurrence au service de la description thématique et sémantique

Puis très vite est venue une mise en relation des concepts théoriques d'isotopie (récurrence d'un sème) et de molécule sémique (groupement stable de sèmes) avec un calcul de cooccurrence, expérimenté initialement dans le cadre d'études thématiques sur les sentiments dans le roman français (Rastier 1995), puis repris dans d'autres contextes (Deza 1999, Bourion 2001, Valette 2004, Poudat 2006, Loiseau 2006...). La démarche méthodologique est ainsi synthétisée :

« Résumons les principales étapes d'une recherche thématique assistée : *(i)* Choix des hypothèses, en fonction de l'objectif général de la recherche (une préanalyse statistique peut guider la recherche d'hypothèse, mais la fréquentation préalable du corpus reste indispensable pour guider les intuitions). *(ii)* Recherche de cooccurrents par la méthode statistique des écarts réduits ou hypergéométrique. *(iii)* Transformation interprétative des cooccurrents en corrélats, et constitution des réseaux thématiques (cette étape est facilitée si l'on a pratiqué une interrogation simultanée sur plusieurs cooccurrents[17] ; cf. Bourion, 1995, I.2). *(iv)* Validation des résultats, par croisement de l'analyse thématique avec l'analyse d'autres composantes du même corpus, par test sur un corpus de contrôle, ou par confrontation avec d'autres recherches thématiques. » (Rastier 2001 p.213)

Ce que produit le calcul textométrique, ce sont donc des cooccurrents, au plan des signifiants ; et ce qui est visé, c'est l'obtention de corrélats, au plan des signifiés. On passe des premiers aux seconds par une interprétation qui reconnaît la présence d'un trait sémantique commun entre le ou les mots servant d'amorce à la recherche, et le cooccurrent alors qualifiable comme corrélat.

Parmi les cooccurrents, et tout particulièrement les cooccurents à faible distance, se mêlent aux corrélats des mots qui sont en relation phraséologique : cela a été observé tout particulièrement pour le lexique des parties du corps (pour « cœur » : « avoir à cœur », « savoir par cœur »,... ; pour « pied » : « plain pied », « faire le pied de grue », « sur un pied d'égalité », etc.). Mais cela n'est peut-être pas si négatif : d'une part, les locutions sont d'autant plus facilement repérables et interprétables que le phénomène est maintenant bien connu ; plus subtilement, et plus particulièrement dans certains corpus, des défigements sont toujours possibles, qui remotivent sémantiquement ces composants détachés. Ceci étant, certains indices pourraient aider à faire la part des phraséologies et des candidats corrélats (fort score de corrélation en particulier pour les formes non lemmatisées, et positionnement orienté, cf. Bourion 2001 p.58) ; et la recherche des cooccurrents doit donc privilégier un contexte pas trop étroit, de l'ordre du paragraphe (Rastier 2001 p.212, Deza 1999 chap. 5).

Dans les recherches en sémantique interprétative, ces calculs de cooccurrents ont été outillés principalement par deux programmes : un programme informatique développé à l'INaLF par Maucourt, et la fonction *Thème* dans le logiciel *Hyperbase* de Brunet. Le logiciel *Hyperbase* (Brunet 2006a) est l'un des plus diffusés en textométrie, et, depuis son introduction, la fonction *Thème* a confirmé sa pertinence. Le programme de Maucourt mérite ici un exposé rapide, car il n'a pas connu la même diffusion, et surtout il a été mis au point dans le contexte de recherches explicitement menées dans le cadre de la sémantique interprétative, dans une collaboration de l'informaticien avec Bourion. Il est mobilisé à plusieurs reprises dans le volume (Rastier 1995), et

17 Les cooccurrents sont choisis en fonction d'hypothèses sur la variation de lexicalisation des traits à observer.

on trouve une présentation de sa forme la plus aboutie dans la thèse (Bourion, 2001). Ce programme détermine la liste des cooccurrents statistiquement significatifs d'un mot pôle, selon la mesure de l'écart-réduit[18]. Il présente le résultat sous forme de liste (organisée et triée en fonction du score d'écart-réduit et de la fréquence), mais aussi affiche les contextes de cooccurrence sous forme de concordance, en mettant en évidence typographiquement les cooccurrents, et en triant les lignes de contexte en fonction des cooccurrents.

ils s' ABATTIRENT , haletants , au	pied d'un BUISSON incendié par les rayons du soleil couchant	MAUP.cn1881
l' OMBRE s' ABATTAIT inerte et lourde au	pied des ORMES	FRAN.om97
lie dernière de la clientèle conquise , ABATTUE aux	pieds du tentateur , évoquait l' image fière et VENGERESSE de	ZOLA.BD1883
PRêTRE se baissa vers *suzanne , toujours ABATTUE au	pied du LIT , la releva , la MIT dans un	MAUP.cn1883
d' ACAJOU , assez profondes , supportées sur deux	pieds de BRONZE et remplies d' une foule de ces beaux	SUE.A-G1831
vers le soir , *kai - *koumou ACCOSTA au	pied des MONTAGNES dont les premiers CONTREFORTS TOMBAIENT à	VER.ecG1868
n' ACCOSTA le rivage qu' à plusieurs milliers de	pieds de l' ENDROIT qui faisait face au point d' où	VER.im1874
les vêtements étaient ACCROCHéS au	pied du LIT , sous la moustiquaire .	MALR.ch1933
. ACCROUPIS au	pied d' un MUR , trois hommes mangeaient du pain qu'	PER.c1965
. le gamin , ACCROUPI au	pied du PARAPET , s' affairait à manier sa pelle ,	MART.Te1940
le soleil ou sous la pluie , ACCROUPIS au	pied d' un SAULE , le cœur battant , l' âme	MAUP.cn1886
de la GRANDE CASE des douanes , ACCROUPIE au	pied d' un manguier , une négrillonne gémissait .	MILLE.B1908

Figure 2 : exemple de sortie du programme de Maucourt : contextes de « pied(s) de » contenant plusieurs cooccurrents sélectionnés par l'écart-réduit, et triés alphabétiquement sur le cooccurrent de gauche le plus proche (Bourion 2001, tome II, p.8)

Le travail de dépouillement de telles extractions de contextes s'organise en regroupant les contextes qui réalisent le même motif sémantique, le même thème, définissable abstraitement comme certains sèmes structurés de façon actancielle : « Quand on étudie les cooccurrents (plan de l'expression) pour les qualifier éventuellement de "corrélats" du thème (plan du contenu), on repère également les relations casuelles, ce qui permet de représenter le thème comme un graphe où les nœuds représentent des composants et les liens des primitives (ergatif, accusatif, attributif, datif, bénéfactif, instrumental, final, cf. Rastier F., 1989, pp. 62-65) » (Bourion 2001 p.116). La lecture des contextes descriptifs de personnages, et le repérage de traits caractéristiques communs, permet aussi d'abstraire leurs rôles sous la forme d'agonistes, soit donc un travail non seulement sur la composante thématique du texte mais aussi sur la composante dialectique, au sens de la sémantique interprétative.

Les cooccurrents, comme indices potentiels d'isotopies (Mayaffre 2008), peuvent également être mis à profit pour contraster sémantiquement des lexèmes *a priori* proches. Ainsi, Deza (1999) étudie les cooccurrents respectifs de « pitié », « commisération », « compassion » et « miséricorde » dans un corpus de romans pour mieux caractériser leur sens effectif dans ce contexte. Loiseau

18 L'écart-réduit produit une valeur numérique, mesurant le caractère non aléatoire de sa cooccurrence. On sélectionne alors une liste de cooccurrents significatifs en convenant d'un seuil sur la valeur absolue de l'écart-réduit.

(2006) montre également comment l'étude d'un concept par ses cooccurrents (par exemple « nature » chez le philosophe Deleuze) peut aider à cerner à la fois son unité de sens et ses diverses acceptions.

3.3. Vers le repérage de passages

Bourion (2001) suggère un prolongement au repérage de corrélats thématiques, par une sortie plus sélective et plus souple que les lignes de concordance du programme Maucourt : « il reste à concevoir des programmes qui recherchent automatiquement des parties de textes comportant un nombre important (et statistiquement pertinent) des mots de la peur (le champ lexical d'étude), et aussi de ceux que nous avons qualifiés comme corrélats : des "rafales" signaleraient des passages probablement indexés sur l'isotopie de la peur. » (Bourion 2001 p.106) Il existe bien un calcul textométrique appelé « rafales », mais c'est plutôt du côté de la fonction *Phrases-clés* d'*Hyperbase* (Brunet 2006a) qu'il faudrait trouver une première réalisation de cette idée. Méconnue des linguistes de la sémantique interprétative, ou peut-être révisable dans sa conception (le calcul a été mis au point comme une suite d'ajustements heuristiques, sans encore avoir été l'occasion de débat scientifique[19]), il y a cependant dans la mise au point d'une telle fonctionnalité un terrain de collaboration entre la sémantique interprétative et la textométrie.

Une autre forme de repérage de passages proposée par la textométrie consiste non pas en une extraction sélective, mais en une représentation graphique de l'ensemble du texte avec, au fil du texte, un indicateur de densité des corrélats. Ehrich (1995) dessine ainsi des graphes figurant les manifestations du thème « ambition » dans le *Père Goriot*. La textométrie a depuis mis au point une autre représentation adaptée à la perception de fortes densités d'occurrence ou de cooccurrence au sein d'un corpus : la « carte des sections » (Lamalle & Salem 2002). Mais la question de l'indicateur qui serait à construire pour mesurer cette densité (intégrant des considérations non seulement de fréquence mais aussi de diversité, de spécificité, etc.), et la question plus délicate de la délimitation d'un passage, restent à travailler, tant au plan théorique qu'au plan technique.

De fait, les développements ultérieurs de la sémantique interprétative ont précisé le concept de « passage » et souligné son importance fondamentale dans la description linguistique, en le reconnaissant comme signe, articulant les plans du signifiant et du signifié (Rastier 2007). Or le repérage des zones denses en cooccurrents pourraient être un point d'appui pour le repérage de passages : « Quand il s'appuie sur des corpus de textes appartenant au même genre et au même discours que le texte analysé, le test de l'écart réduit permet de repérer des groupements de cooccurrents qui sont de bons candidats pour la constitution de passages » (Rastier 2008).

3.4. La caractérisation de textes et de genres textuels

L'analyse thématique par les cooccurrents peut être mise au service de la caractérisation des textes. Dans le projet *Princip* (Valette 2004), l'enjeu est de repérer et discriminer les pages racistes et les pages anti-racistes sur internet. La textométrie a été mobilisée pour construire des thèmes spécifiques aux unes et aux autres : pour une lexie appartenant au fond isotopique commun, comme « immigration » ou « étranger », on recherche ses corrélats dans le sous-corpus raciste et dans celui anti-raciste. Les sous-corpus s'avèrent aussi caractérisables par des indices de toutes natures, non seulement lexicaux mais aussi morphologiques et sémiotiques.

La textométrie a aussi permis d'observer concrètement l'incidence sémantique du cadre générique, qui occupe une place déterminante en sémantique interprétative. On a par exemple mis en évidence que le mot « amour » n'attire pas les mêmes corrélats, et donc ne construit pas les mêmes isotopies

19 Il s'apparente néanmoins au calcul des réponses modales exposé dans (Lebart & Salem 1994).

ni les mêmes thèmes, selon qu'on le trouve dans des romans ou dans des poésies (Bourion 2001 p.42-47). Du côté du roman, on pointe des contextes d'usage tels que : amour « platonique », « inspirer » (de l'amour/un amour...), amour « exclusif » ; et encore, « passion », « passionné », « jalousie », « ambition », « chagrin », « orgueil », « oubli », « renoncement », « révulsion », « vanité », « égoïsme ». Et pour la poésie : « allégresse », « hymen », « hyménée », « igné », « incendier », « rets », « rossignol », « soupir », entre autres.

Les analyses factorielles sur les décomptes et mesures fournis par la société *Synapse* sur son corpus (Malrieu & Rastier 2001, Beauvisage 2000) ont également été des expériences probantes, confirmant la détermination du global sur le local et les interrelations transverses aux paliers de description. Ces expériences de linguistique de corpus sont néanmoins en marge de la textométrie car, même si l'analyse factorielle est un calcul couramment pratiqué en textométrie, l'absence ici de retour au texte outillé (la société *Synapse* ne donnant pas accès à ses sources) ne permettait pas à proprement parler une démarche textométrique. En revanche, dans le même esprit, mais cette fois-ci en contrôlant l'analyse par des retours au texte, Poudat (2006) procède à la description d'un genre textuel, celui de l'article scientifique de linguistique français, selon les principes de la sémantique interprétative et en recourant à des procédures textométriques disponibles dans le logiciel *DTM*[20]. La description s'appuie tant sur une analyse lexicale que sur une description morphosyntaxique, et elle explore différents paliers, infratextuels (comme la section) et supratextuels (le style d'auteur, le domaine...).

Pour décrire les textes et les genres, la sémantique interprétative propose de s'intéresser à la composante thématique, mais aussi à trois autres composantes (dialectique, dialogique, tactique). La prise en compte de ces différentes composantes est d'autant plus importante qu'elles fonctionnent en interaction. Les thèses de Loiseau (2006) et Poudat (2006) proposent ainsi de nouvelles manières de mobiliser des calculs textométriques dans l'esprit de la sémantique interprétative, notamment des diagrammes de distribution à différents paliers comme celui du texte ou du paragraphe (« diagrammes tactiques » et « gammes de densité », cf. Loiseau 2006, §12.F.).

4. Connivences de fond

Les expériences précédentes ont montré le caractère prometteur de l'approche textométrique pour une recherche sur corpus dans le cadre de la sémantique interprétative. Pour confirmer l'intuition, nous entreprenons maintenant de relever de façon plus complète des adéquations essentielles entre la théorie de la sémantique interprétative, et les principes fondateurs de l'approche textométrique.

4.1. Linguistique et sémantique

La sémantique interprétative s'intéresse au sens. Selon elle, le sens peut s'élaborer à partir d'indices morphologiques, syntaxiques, phonétiques, etc. Il peut être transversal aux catégories grammaticales, mobiliser au même titre une ponctuation, un aspect verbal, un morphème, un profil prosodique et rythmique, une typographie, une mise en page (Bourion 2001, Malrieu & Rastier 2001, Beaudouin 2002, Valette 2004, Loiseau 2006). La textométrie est en mesure de prendre en compte des descriptions du texte de toutes natures, pour peu qu'elles soient explicitées par un codage du corpus.

La démarche sémantique proposée par la textométrie est tout à fait en accord avec la demande de

20 Le logiciel DTM est conçu et développé par Lebart, et est diffusé à l'adresse http://www.dtm-vic.com/. À dominante statistique, il est spécialisé dans les procédures d'analyse des données (analyse factorielle, classification...) et dans les techniques mathématiques d'aide à l'interprétation des résultats.

« dé-ontologie » formulée par Rastier[21]. Il s'agit bien d'éviter toute préconception réductrice, on veut surtout rester au plus proche du texte et ne pas commencer par l'étudier à travers le prisme d'une ontologie. Ainsi, dès ses débuts, la textométrie s'est fait une spécialité du traitement des questions ouvertes dans les enquêtes, pour éviter le postcodage des réponses (entre l'enquête et l'analyse) qui efface des variations d'expressions potentiellement significatives (Lebart & Salem 1994)[22]. De même, les logiciels d'analyse textuelle qui proposent des traitements statistiques ou quantitatifs quelquefois très proches des traitements textométriques, mais qui commencent par remplacer le texte par une représentation en termes de catégories prédéfinies (en projetant le texte sur une ontologie), sortent clairement de l'approche textométrique[23].

Le souci de fidélité au texte s'est même vivement exprimé dans le débat traversant la communauté textométrique et concernant la lemmatisation : faut-il vraiment segmenter le texte en formes graphiques « telles quelles », ou bien n'est-il pas plus juste d'appliquer une pré-analyse purement morphosyntaxique, qui assimile toutes les formes fléchies d'un même mot à l'entrée de dictionnaire correspondante ? Autrement dit, choisit-on de compter et analyser indépendamment « fleur » et « fleurs », « est », « étions » et « serai », ou bien préfère-t-on ne reconnaître ici que les lemmes « fleur » et « être » ? La question est longtemps restée ouverte, car si la lemmatisation était séduisante pour désambiguïser efficacement de nombreux homographes (ex. « un parti politique » *vs* « je suis parti »), l'attention au texte avait aussi révélé que bien souvent les différentes flexions étaient porteuses d'une sémantique différente, typiquement les pluriels étaient plus concrets que les singuliers (ex. « le travail » *vs* « les travaux ») (Geoffroy, Lafon, Tournier 1974). La sémantique interprétative partage cette sensibilité à ne pas écraser ces distinctions, comme le montre l'étude de Bourion sur « au pied de » vs « aux pieds de » dans un corpus littéraire : le singulier renvoie à des descriptions de localisation, avec un sème de /verticalité/ (« au pied de la montagne », etc.), alors que la forme plurielle correspond à des scènes d'imploration (« se jeter aux pieds de quelqu'un »), appelant cette fois-ci des sèmes /humain/ et /sentiment/ (Bourion 2001 p.62).

4.2. Sémantique des textes

4.2.1. La place centrale des textes à toutes les étapes de l'analyse textométrique

L'objet empirique de la linguistique n'est pas d'abord la phrase ou la proposition, mais le texte. La réalité observée est bien d'abord celle de textes, situés dans des pratiques, et non des phrases reçues indépendamment du texte dont elles sont issues. Dans les termes de la sémantique interprétative, développée ensuite en sémantique des textes, le global détermine le local, si bien que l'analyse d'une phrase, pour être juste et complète, requiert la prise en compte de son contexte textuel, voire

21 Au plan expérimental, Deza (1999) montre notamment comment la canonicité qui s'exprime dans le corpus est en décalage avec une représentation purement référentielle du monde, avec l'exemple de l'âge des personnages dans le roman français.

22 Un exemple donné par (Lebart & Salem 1994, p. 169 et 188) : à la question « Quelles sont les raisons qui, selon vous, peuvent faire hésiter une femme ou un couple à avoir un enfant ? », les réponses « manque d'argent » et « raisons financières » (ou encore « le travail de la femme » et « la femme travaille ») seraient *a priori* assimilées à la même réponse par un post codage ; or, une étude textométrique préservant ces formulations montre que ces manières de s'exprimer sont significativement corrélées à des répondants de catégories socio-professionnelles très contrastées, et d'y reconnaître finalement des nuances importantes.

23 Parmi les logiciels proposant une analyse textuelle via une réduction du texte à des catégories, et non en travaillant sur le texte lui-même, on peut citer Tropes (qui pourtant implémente des calculs d'origine textométrique comme les rafales) ou l'analyse sémantique livrée par Cordial (alors que la composante analyseur morpho-syntaxique du même logiciel peut tout à fait préparer un corpus pour une analyse par un logiciel textométrique). Le cas de Prospero est intermédiaire, au sens où les catégories sont construites par l'utilisateur -on n'est pas soumis à un dictionnaire « universel » prédéfini.

intertextuel -le texte est ainsi l'unité minimale de l'analyse.

La textométrie elle aussi affirme son attachement à l'unité texte, ne serait-ce que par son nom : l'évolution 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. Depuis son origine, la textométrie travaille naturellement sur des corpus de textes intégraux, par opposition aux pratiques de corpus de phrases ou d'échantillonnages de textes. De fait, la textométrie est employée tout autant par des linguistes que par des chercheurs d'autres sciences humaines (littérature, politique, histoire, philologie etc.), dont l'objectif est de se donner des moyens de renouveler la lecture de leur corpus, dans le respect des données recueillies.

Le textomètre a une très bonne connaissance de son corpus, sinon l'acquiert. Il l'a déjà parcouru, dans certains cas peut l'avoir même déjà lu et relu, et, -source ou résultat de cette fréquentation assidue ?- il y est souvent attaché. Car l'approche textométrique est celle de la curiosité d'une lecture renouvelée par la mise en évidence de régularités non encore perçues[24]. La textométrie est ainsi bien complémentaire d'approches comme la recherche ou le filtrage d'informations, voire les systèmes de question-réponse, où le corpus est en quelque sorte un réservoir, quelquefois constitué à la volée par filtrage sur quelques critères comme la présence de certains mots-clés, corpus-réservoir donc dont la connaissance globale importe peu, et dont le rôle n'est que de pouvoir livrer quelques extraits au contenu « pertinent », souvent sans considération particulière pour la formulation employée et sa contextualisation complète. La textométrie se caractérise ici de façon très claire par son attachement aux textes composant son corpus, construit et étudié pour lui-même. En ce sens, Geffroy et Lafon (1982) avaient tenu à souligner, non sans humour, « l'insécurité dans les grands ensembles » : l'application des calculs textométriques à des corpus trop gros pour en avoir une première connaissance intérieure, non superficielle, comme à des corpus où les textes sont fondus dans quelques grandes catégories méta-textuelles occultant l'unité de chaque texte, peine à livrer des analyses significatives, car l'interprétation des résultats des calculs reste approximative et limitée, voire peut se fourvoyer.

L'importance des textes se matérialise dans la conception même des logiciels textométriques. Dans *Hyperbase*, l'hypertextualité, massivement employée, est systématiquement mise à profit pour revenir au texte et visualiser les occurrences dans leur contexte textuel. L'ergonomie des logiciels textométriques prévoit toujours soit un affichage de texte aux côtés de l'affichage de listes, de tableaux ou de représentations graphiques, soit une navigation hypertextuelle permettant un accès immédiat à des contextes d'occurrence ciblés[25].

4.2.2. La contextualisation comme principe d'analyse et le rôle déterminant du corpus de référence

La textométrie compte, situe, caractérise, des unités dans des contextes : c'est ainsi que l'on décèle des liens (morphologiques, lexicaux, syntaxiques, sémantiques...) entre unités, que l'on établit également des similarités entre contextes (typiquement entre textes), que l'on dresse des cartographies dessinant des typologies. Comme on l'a vu (en termes de modélisation), les contextes sont tant locaux (cooccurrences, concordances) que globaux (spécificités, cartographie par AFC).

« Pour la problématique herméneutique, [le texte] est l'unité minimale (bien que non élémentaire). Un texte ne peut se lire que dans un corpus » (Rastier 2008). « La compréhension du texte [...]

24 D'où les connivences notamment avec les études littéraires (travaux sur les *Fleurs du Mal* de Viprey, sur le théâtre de Giraudoux de Brunet, etc.), la philologie (ex. interface de consultation de la *Base de Français Médiéval*) ou l'exégèse (cf. l'intérêt du *Centre Informatique et Bible* de l'abbaye de Maredsous pour un logiciel comme *Hyperbase*).

25 Voir par exemple (Heiden 2004) pour le logiciel *Weblex*.

procède par contextualisation et intertextualisation » (Rastier 2001 p.93). Cela se déploie à tous les paliers, se reformulant en autant de principes (Rastier 2001 p.92) : le principe de *contextualité* (« deux signes ou deux passages d'un même texte mis côte à côte sélectionnent réciproquement des éléments de signification (sèmes) [...] »...), le principe d'*intertextualité* (analogue pour deux passages de textes différents), et le principe d'*architextualité*, selon lequel tout texte plongé dans un corpus en reçoit des déterminations sémantiques, et modifie potentiellement le sens de chacun des textes qui le composent. En outillant une lecture non linéaire (par l'extraction de contextes, la génération de concordances), la textométrie joue fondamentalement sur les principes de contextualité et d'intertextualité. Les cooccurrences ont pu être comprises également comme une forme de contextualisation minimale, dans l'esprit de la sémantique interprétative (Mayaffre 2008). Quant au principe d'architextualité, le fonctionnement statistique du corpus de référence en est une concrétisation : en effet, tout texte inclus dans le corpus apporte sa contribution aux fréquences globales[26], et réciproquement, se trouve caractérisé par rapport à ces fréquences globales. Le choix du corpus de référence est déterminant pour l'analyse et conditionne complètement l'interprétation, la textométrie donne à voir un texte de façon tout à fait différente selon le corpus sur lequel on le profile. Le texte n'a donc pas ici un contenu à extraire, une seule « bonne » lecture, mais autant d'éclairages possibles que de contextualisations pertinentes en corpus. Par le biais du corpus de référence, le global détermine le local, et l'analyse est la mise en évidence de formes qui se détachent sur un fond (Rastier 2001 p.42 sq.)

4.3. Sémantique interprétative

A une approche ontologique, qui s'apparente à une forme de contemplation (de ce qui « est »), la sémantique interprétative oppose une conception dynamique du sens, une praxéologie, relative à des pratiques contextualisées. Le sens se construit au fil de la lecture, qui s'apparente à une reconnaissance de formes, peu à peu cernées, et même évolutives. Il s'agit d'une activité perceptive, le geste interprétatif s'ajuste en fonction des contraintes linguistiques reçues du texte. Ainsi se conçoit la richesse de sens d'un texte, mais aussi sa significativité, non arbitraire, les contraintes linguistiques empêchant de faire dire tout et n'importe quoi à un texte.

L'analyse textométrique procède également d'une démarche construite : on ne peut pas fournir un corpus, « faire tourner » le logiciel, et récupérer le résultat comme produit fini. Du fait de son importance déterminante, la constitution du corpus est une première étape engageant des choix interprétatifs : les données ne sont justement pas données (Rastier 2008). Le codage de ce corpus, et sa mise en correspondance avec la structure de données textométrique (pour définir les unités, les typages, les partitions, etc.) procèdent également de choix correspondant à des hypothèses, des attentes interprétatives. Puis il s'agira de trouver un bon point d'entrée ; de lancer un calcul pertinent, en comprenant selon quels principes il fonctionne ; de qualifier les résultats quantitatifs, avec un cheminement qui toujours renvoie à des tâches de lecture, de compréhension (parcours des contextes, comparaison, regroupements). La dynamique de l'interprétation se traduit encore par le choix d'un nouveau point d'entrée, d'un nouveau contexte, ou d'un nouveau calcul, qui bien souvent est en fait un ajustement du traitement précédent -et l'on retrouve très concrètement l'interprétation comme action et comme geste qui affine dynamiquement sa trajectoire.

On peut reconnaître dans certains processus textométriques des principes herméneutiques traditionnels, comme celui des « passages parallèles », qui consiste à s'aider, pour la compréhension

26 On peut aussi travailler avec un corpus de référence qui ne contienne pas le corpus de travail (cf. les spécificités exogènes dans *Hyperbase*, calculées par rapport au corpus littéraire *Frantext*), mais c'est un cas particulier plus rare (quasiment pas implémenté par les logiciels de textométrie) et souvent plus discutable (statut et qualité du corpus de référence, connaissance précise de sa composition et accès aux textes, adéquation et possibilités d'ajustement au corpus de travail).

d'un passage obscur, de la lecture d'un autre passage du même texte (ou d'un texte lié) abordant le même sujet. La délinéarisation et les réorganisations du texte facilitées par la numérisation outillent efficacement cette technique herméneutique des passages parallèles (Bourion 2001 p.116, Pincemin 2006).

4.4. Sémantique différentielle

La sémantique interprétative se définit comme une sémantique différentielle, par opposition à des sémantiques inférentielle ou référentielle. Elle est ainsi pleinement linguistique, car elle n'est pas fondée sur une réalité extérieure, physique ou psychique -même si elle permet de comprendre ensuite l'articulation du texte avec ces réalités d'un autre ordre[27].

4.4.1. Le fonctionnement différentiel des représentations et calculs textométriques

Formellement, au plan des décomptes de fréquences et autres calculs qui s'ensuivent, la textométrie suppose de convenir d'un typage des unités, qui fixe ce qui est reprise du même (et se cumule en termes de fréquences) et ce qui est différent (et participe donc au nombre de types). On a donc déjà fondamentalement une modélisation qui, à un moment donné, choisit d'assimiler certaines unités et d'en dissimiler d'autres. Le typage, qui règle ce jeu d'identification et d'opposition, est relatif au point de vue choisi, qui peut varier au fil de l'analyse -de même que, linguistiquement, les rapprochements et les différenciations évoluent selon l'activation, l'afférence ou l'inhibition de sèmes, dépendants eux-mêmes notamment des contextualisations.

Ensuite, les statistiques contrastives proposées en textométrie ont bien été comprises comme un mécanisme différentiel : car ces statistiques mettent en évidence ce qui s'écarte de la norme (définie par le corpus de référence), elles repèrent les contrastes dans un tout qui fait système (Bourion 2001 p.18, Rastier 2001 p. 86).

Mais aussi, toujours dans les procédures textométriques, et sans aller jusqu'aux procédures mathématiques élaborées, un simple tri alphabétique d'une liste du vocabulaire, ou un tri sur le contexte gauche ou droit d'une concordance, font aussi entrer en jeu heuristiquement dans leur lecture une perception différentielle : l'œil repère les motifs récurrents, rapprochés par le tri, et les variations à l'intérieur ou aux frontières de ces regroupements (Pincemin 2006).

4.4.2. L'attention à la structure et la dominance du qualitatif

La sémantique interprétative relève d'une approche structurale de la langue, dans la lignée des travaux de Hjelmslev, Greimas, Pottier, Coseriu. Il s'agit de situer les unités linguistiques les unes par rapport aux autres, à l'intérieur du système qu'elles forment, et non par une qualification ou une évaluation renvoyant à une réalité externe à la langue. Cela vaut aussi pour l'attribution d'une interprétation : la sémantique interprétative ne vise pas à associer à un signe, une proposition ou un texte, son interprétation ; mais elle recherche les contraintes posées par la langue et par son usage dans le texte considéré pour hiérarchiser des lectures possibles.

De même, certaines des mesures textométriques correspondent à des probabilités ou à des valeurs statistiques permettant une qualification en termes de significativité statistique ; d'autres sont de simples décomptes de fréquence, avec certaines valeurs particulières (comme la fréquence 1 des hapax) ; ces chiffres sont utilisées pour le seuillage des résultats et la détermination d'une sélection (de mots, de cooccurrents, etc.). Mais aussi et surtout, ils sont utilisés pour le classement qu'ils

27 On renvoie ici au concept de « pôles intrinsèques » du texte et à leur participation à la définition des genres textuels (Rastier 2001 p.17 sq.)

induisent, le « tri hiérarchique » qu'ils permettent d'opérer. La pratique textométrique consiste alors, après la génération d'une liste délimitée et ordonnée sur des critères quantitatifs, à travailler plus qualitativement sur des zones de la liste : la tête de liste donnant les éléments dominants, comme des zones intermédiaires (pour observer des phénomènes plus nuancés), voire proches du seuil (pour ajuster la sélection). Autrement dit, le quantitatif guide l'analyse, mais c'est un examen qualitatif (avec retour au texte, etc.), qui détermine l'interprétation.

4.5. Sémantique unifiée

La sémantique interprétative est unifiée, au sens où les principes différentiels et interprétatifs s'observent à différents paliers, typiquement ceux de la lexie, de la période, et du texte. Peut-être peut-on aussi ranger sous cette préoccupation unifiante le rejet des divisions disciplinaires qui séparent des points de vue pourtant complémentaires et intimement articulés : syntaxe, sémantique et pragmatique -une sémantique bien faite intègre des considérations syntaxiques et doit savoir décrire les phénomènes relégués à la pragmatique-, philologie et herméneutique, et plus généralement les « arts et sciences du texte » sur lesquels Rastier (2001) propose d'adopter un point de vue embrassant, du fait de leur objet commun, le texte.

La textométrie s'accommode très naturellement de la multiplicité des paliers et de leurs analogies de fonctionnement. De fait, formellement, la technique n'exige que de lui fournir des unités ou « contenus », réparties (mieux, contextualisées) dans des « contenants ». Peu lui importe la nature de ces contenus et contenants, au linguiste revient de déterminer les unités pertinentes, et de les faire varier comme bon lui semble. Les unités ne sont pas nécessairement des mots, les entités englobantes, pas nécessairement des textes.

Comme nous l'avons vu, les unités utilisées pour le calcul n'ont pas de prétention linguistique, les unités linguistiques/herméneutiques sont construites, et ce que l'on veut souligner ici c'est que ces unités construites peuvent relever d'un autre palier que les unités ayant servi au calcul. Clairement, la textométrie se range ainsi du côté des approches textuelles, elle n'est pas simple jonglage avec des signes prédéfinis. Comme le montrent les expériences de Brunet (2006b)[28], si signe il y a, il n'est pas considéré en tant que tel, il est saisi au vol pour un objectif plus global, celui de faire ressortir des lignes de forces, de dégager des formes signifiantes. Un calcul basé sur des mots peut ainsi conduire à repérer des sèmes (infralexicaux, au sens où un mot est *a priori* porteur de plusieurs sèmes) et à construire des molécules sémiques représentant un thème (supra-lexical, au sens où sa manifestation peut être diffuse sur tout un passage voire tout un texte, et où il se prête à des lexicalisations multiples). Rastier (2001 p.206) souligne comment, même en travaillant en apparence au niveau lexical du corpus, on saisit en fait des réalités d'un autre ordre : « Pour progresser, la thématique doit donc dépasser l'analyse lexicale. [...] Le mot à partir duquel peut commencer la recherche thématique n'en est pas l'objet, à la différence d'un mot-vedette qui ferait l'objet d'une recherche lexicographique. On va certes chercher, en utilisant les moyens d'assistance informatisés, d'autres mots et expressions qui sont cooccurrents. Une fois interprétés, les cooccurrents pour lesquels on aura identifié une relation sémantique seront considérés comme des corrélats, c'est-à-dire comme des lexicalisations complémentaires de la même molécule sémique. Le réseau des corrélats relie les manifestations lexicales du thème. Mais il faut pouvoir discerner les meilleurs points d'entrée dans ce réseau : la "vedette" n'est qu'un de ces points d'entrée, présumé lexicaliser synthétiquement le thème que l'on cherche à décrire. »

Les expériences de caractérisation de textes ou de genres, basées explicitement sur des unités non lexicales (mesures morphosyntaxiques du corpus *Synapse*, indices sémiotiques du corpus *Princip*, etc.) ont montré que les techniques statistiques permettent de capter des régularités significatives

28 Et d'autres avant lui, Salem par exemple avait fait des expériences similaires.

bien qu'imperceptibles pour une lecture non outillée. La textométrie ferait donc partie des techniques capables de plonger dans des dimensions profondes du matériau textuel. « Enfin, l'opposition humboldtienne entre la forme intérieure et la forme extérieure des textes, qui a fait couler tant d'encre chez les stylisticiens, pourrait recevoir une nouvelle formulation qui la relativise : la forme intérieure, loin d'être un mystère esthétique, est constituée par les régularités jusqu'à présent imperceptibles de la forme extérieure, celle de l'expression, que les moyens théoriques et techniques de la linguistique de corpus permettent à présent de mettre en évidence. En d'autres termes, le contenu d'un texte ne se réduit certes pas à une mystérieuse représentation mentale : un texte est fait de deux plans, celui des formes sémantiques et celui des formes expressives, dont le genre notamment norme la mise en corrélation. Au sein de chaque plan s'établissent des relations forme / fond, de type gestaltiste, qui permettent la perception sémantique et phonologique. » (Rastier 2005) La distinction des deux plans réaffirme que la textométrie peut donner à percevoir à des régularités expressives participant à la construction de formes sémantiques, sans pour autant livrer directement un sens, « extraire le contenu ». L'interprétation reste partie intégrante de l'analyse textométrique.

5. Perspectives d'apports mutuels

La textométrie suscite actuellement de nouvelles propositions et de nouveaux développements, tout particulièrement autour de la réalisation collaborative d'une plateforme logicielle ouverte, fédérant les recherches et les développements informatiques des principales équipes du domaine[29].

5.1. Modélisation des textes et des corpus

Le point de vue de la sémantique interprétative encourage à préserver et développer l'exploitation des corpus structurés et étiquetés, pour la pluralité des segmentations et des descriptions qu'ils rendent possibles (Loiseau 2006). Dans le même esprit, la logométrie (Mayaffre 2005), autre désignation récente de la lexicométrie (au même titre que la textométrie), affirme la pertinence d'une textométrie capable de travailler sur de multiples niveaux linguistiques, et la sémantique interprétative fait partie de ses fondements linguistiques forts.

La sémantique interprétative insiste aussi sur la possibilité de définir des sous-corpus « à pertinence enrichie ». Elle rejoint les retours des utilisateurs des logiciels de textométrie, qui constatent le besoin de pouvoir ajuster et redéfinir un sous-corpus au fil des analyses. Cette dynamique du corpus, déjà en partie assumée dans le logiciel *Lexico 3*, est inscrite au cahier des charges des nouvelles applications textométriques.

La discussion s'est engagée sur la redéfinition d'une modélisation du texte. Le modèle textométrique traditionnel est basé sur une segmentation de référence sans nécessairement de valorisation théorique associée, mais malgré tout par rapport à laquelle toutes les autres segmentations doivent se définir. Le point de vue de la sémantique interprétative soutient la proposition d'un modèle différentiel, dans lequel le texte se déploie au croisement des dimensions descriptives adoptées, affirmant ainsi l'ouverture de la modélisation (jamais complète), et son caractère différentiel et unifié (qui ne nécessite pas d'introduire de hiérarchie entre les dimensions descriptives) (Pincemin 2008).

29 Projet *Textométrie* et plateforme logicielle *TXM* (http://textometrie.ens-lyon.fr/), initiés avec un financement ANR (ANR-06-CORP-029, pour 2007-2010).

5.2. Conception de fonctionnalités

La sémantique interprétative peut proposer des repères pour comprendre les fonctionnalités textométriques existantes, chacune pour elle-même (ex. l'analyse des concordances dans (Pincemin 2006)). Elle peut renouveler aussi la manière de les envisager dans leur ensemble : par exemple la conception morphosémantique, qui s'appuie sur l'opposition entre points réguliers et points singuliers (Rastier 2001 p.45), pourrait suggérer une analyse globale des fonctionnalités textométriques entre celles qui captent des régularités -notamment des répétitions à l'identique- et celles qui cernent des singularités.

Les propositions de Bourion (2001), concrétisées dans les programmes réalisés par Maucourt, ont participé à l'amélioration de certaines fonctionnalités (croisement des concordances avec des coocurrences) et à la formulation d'innovations, comme les tableaux synoptiques (qui restent à intégrer dans les logiciels actuels). Les recherches de Malrieu (2006) explorent la mise au point de jeux d'indicateurs adaptés aux textes et de calculs appropriés à la description de structures intratextuelles. Loiseau (2007) développe aux plans théorique et logiciel l'exploitation des corpus multi-annotés, pour des analyses prenant en compte des unités linguistiques de taille et de natures diverses, selon l'idée rastiérienne de sémantique unifiée, avec de multiples possibilités de contextualisation. Son logiciel *CorpusReader* opère pour le moment des décomptes pas spécifiquement textométriques, mais il explore une voie dans laquelle la textométrie pourrait s'étendre.

Enfin, il est probable que le terrain de l'annotation dynamique de corpus, c'est-à-dire l'enrichissement du corpus par des interprétations pouvant servir d'appui à des analyses ultérieures, puisse intéresser la sémantique interprétative. Une collaboration sur cet aspect serait d'autant mieux venue que la mise au point d'une telle fonctionnalité, et des usages associés, conditionne encore complètement sa pertinence (mal définie ou mal mise en œuvre, une telle fonctionnalité peut rendre totalement ingérable et ininterprétable le corpus).

5.3. Elaboration de repères méthodologiques

Le travail autour de l'analyse thématique (Rastier 1995) a permis, on l'a vu, l'élaboration d'une méthode de repérage de corrélats pour la construction de molécules sémiques représentatives de thèmes. Les thèses de Bourion (2001) et Deza (1999) ont poursuivi et précisé cette réflexion méthodologique sur l'accès sémantique aux banques textuelles. Celles de Poudat (2006) et de Loiseau (2006) l'ont étendue en prenant en compte les possibilités d'enrichissement linguistique des corpus et en considérant plus systématiquement les quatre composantes de la description textuelle selon la sémantique interprétative -thématique, dialectique, dialogique et tactique. En continuant à développer leur expérience et leur pratique des procédures textométriques, les linguistes d'inspiration rastiérienne pourraient contribuer à mettre au point un ensemble consistant d'éléments méthodologiques[30] fondés sur un cadre théorique linguistique fort.

Toutes les étapes de la démarche textométrique peuvent être éclairées par une telle mise en perspective théorique et méthodologique : les considérations philologiques liées à la constitution d'un corpus initial et à son codage, l'interprétation sémantique des fonctionnalités (concordances, cooccurrences – cf. ci-dessus), les méthodes de dépouillement (comme l'organisation des contextes de concordance en fonction des sèmes actualisés), l'enchaînement de traitements. Ces apports méthodologiques peuvent être associés à la mise au point d'interfaces.

30 Par exemple, l'observation de rythmes sémantiques suggérée par (Bourion 2001, vol.I p.47 et vol.II p.18-19).

6. Conclusion : enjeux pratiques et théoriques

D'autres techniques, d'autres calculs, pourront certainement montrer leur pertinence pour l'approche de la sémantique interprétative : si l'argumentaire de cet article est enthousiaste, il ne veut néanmoins prétendre ni à l'exclusivité de la textométrie comme proposition de réponse logicielle à la théorie rastiérienne[31], ni au caractère idéal de la textométrie actuelle, qui est de fait encore en pleine évolution.

Ceci étant, il reste frappant de trouver tant de connivences de fond entre les deux approches : texte, contextualisation, intertextualité et corpus, sémantique différentielle, interprétation et dynamique de la construction d'un sens. Si bien que l'on pressent tout l'intérêt de poursuivre et approfondir les collaborations.

La textométrie pourrait être à la base d'un environnement de lecture « SAAS » (Système d'Aide à l'Analyse Sémantique), pour reprendre les termes de Bourion (2001). L'enjeu est de tirer parti des possibilités du numérique pour se doter d'outils logiciels renouvelant les parcours de lecture et les points d'appui interprétatifs. En ce sens, l'étude du potentiel de la textométrie pour accompagner une approche rastiérienne des textes relève pleinement d'une réflexion sur sémantique et interprétation : elle l'explore sur un plan concret, précis et révélateur. L'expérimentation de ces nouvelles formes de lecture et d'interprétation n'est pas une simple application, qui plus est inévitablement réductrice, de la réflexion théorique sémantique : elle est en mesure de la relancer et de la renouveler en lui dévoilant des réalités textuelles ou herméneutiques pas encore perçues ou oubliées.

Ce texte participe à la réflexion menée dans le projet Textométrie ANR-06-CORP-029.

Je remercie beaucoup Evelyne Bourion, Carine Duteil-Mougel, Serge Heiden, Sylvain Loiseau, Damon Mayaffre, Céline Poudat, et Mathieu Valette, pour leurs relectures attentives et constructives, et pour les nuances importantes et les précisions qu'ils m'ont permis d'apporter à l'article.

Bibliographie

Ablali, Driss, Poudat, Céline (2009) - « Sémantique de corpus . Concepts fondamentaux et dialogue avec l'ADT », *Ecole thématique CNRS Méthodes Informatiques et Statistiques en Analyse de Textes*, Besançon, 15-19 juin 2009 [présentation support de cours, non publiée].

Beaudouin, Valérie (2002) – *Mètre et rythme du vers classique -Corneille et Racine*, Paris : Champion, coll. « Lettres numériques », 2, 620 pages.

Beauvisage, Thomas (2000) – *Exploiter des données morphosyntaxiques pour l'étude statistique des genres - Application au roman policier*, Mémoire de DESS, Centre de Recherche en Ingénierie Multilingue, INaLCO, Paris, 73 pages.

31 On trouvera des exemples diversifiés de perspectives et d'applications logicielles en lien avec la Sémantique interprétative sur le site *Texto!* (http://www.revue-texto.net/), notamment dans les rubriques *Dits et inédits* et *Corpus et trucs*. La plupart ont des connivences avec la textométrie, par leur usage de statistiques textuelles (Rossignol, Mauceri, Reutenauer...) ou par la place centrale donnée à l'analyse qualitative, semi-automatisée, tpyiquement des applications d'annotation et de visualisation de sèmes et de parcours (Beust, Tanguy, Thlivitis, Bénel, Perlerin, Roy...). Mais sont aussi évoquées d'autres voies moins proches, comme le connexionisme ou la programmation logique par contraintes qui ont retenu l'attention par leur affinité avec le caractère perceptif de la sémantique, (Prié 1995, Rastier *et al*. 1994).

Bommier-Pincemin, Bénédicte (1999) - *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat, Linguistique, Université Paris IV (Sorbonne), 6 avril 1999, n°99PA040027, 806 pages.

Bourion, Evelyne (1995) – « Le réseau associatif de la peur », *in* François Rastier (éd.), *L'analyse thématique des données textuelles : l'exemple des sentiments*, Paris : Didier, collection Études de sémantique lexicale, 107-145.
http://www.revue-texto.net/1996-2007/Parutions/Analyse-thematique/Bourion.pdf

Bourion, Evelyne (2001) – *L'aide à l'interprétation des textes électroniques*, Thèse de doctorat en Sciences du langage, Université de Nancy II, soutenue le 14 décembre 2001.
http://www.revue-texto.net/Corpus/Publications/Bourion/Bourion_Aide.html

Brunet, Etienne (2006a) - *Hyperbase, logiciel documentaire et statistique pour la création et l'exploitation de bases hypertextuelles. Manuel de référence. Version 6.0 (mai 2006).* Laboratoire Bases, Corpus et Langage, UFR Lettres, Université de Nice.

Brunet, Etienne (2006b) - « Le corpus conçu comme une boule », *Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation, Actes du XVIIe Colloque d'Albi Langages et Signification*, Albi, 10-14 juillet 2006, Carine Duteil-Mougel & Baptiste Foulquié (éds), ISBN 2-907955-12-18, 85-94, et Texto!, juin 2006, vol. XI, n°2. Texte également réédité dans (Brunet 2011), chapitre 14.
http://www.revue-texto.net/1996-2007/Parutions/Livres-E/Albi-2006/Brunet.pdf ISSN 1773-0120

Brunet, Etienne (2011) - *Ce qui compte. Ecrits choisis, tome II, Méthodes statistiques*, Poudat, Céline (éd.), Paris : Honoré Champion, collection Lettres numériques n°11, 373 pages.

Erlich, David (1995) – « Une méthode d'analyse thématique. Exemples de l'ennui et de l'ambition », *in* François Rastier (éd.), *L'analyse thématique des données textuelles : l'exemple des sentiments*, Paris : Didier, collection Études de sémantique lexicale, 85-103.
http://www.revue-texto.net/1996-2007/Parutions/Analyse-thematique/Erlich.pdf

Geoffroy, Annie, Lafon, Pierre, Tournier, Maurice (1974) - « L'indexation minimale - Plaidoyer pour une non-lemmatisation », E.N.S. de Saint-Cloud, 30 pages - Communication au Colloque sur l'*Analyse des corpus linguistiques : Problèmes et méthodes de l'indexation maximale*, Strasbourg, 21-23 mai 1973.

Geffroy, Annie, Lafon, Pierre (1982) - « L'insécurité dans les grands ensembles. Aperçu critique sur *Le vocabulaire français de 1789 à nos jours* d'Etienne Brunet », *MOTS*, 5, 129-141.

Heiden, Serge (2004) - « Interface hypertextuelle à un espace de cooccurrences : implémentation dans Weblex », *Actes des 7es Journées internationales d'analyse statistique des données textuelles (JADT 2004)*, Gérald Purnelle & *al*. (éds), Presse universitaires de Louvain, Louvain-la-Neuve (Belgium), 10-12 mars 2004, 577-588.
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_055.pdf

Lafon, Pierre (1980) - « Sur la variabilité de la fréquence des formes dans un corpus », *MOTS*, 1, 127-165.

Lafon, Pierre, Salem, André (1983) - « L'inventaire des segments répétés d'un texte », *MOTS*, 6, 161-177.

Lamalle, Cédric, Salem, André (2002) - « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels », *Actes des 6es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2002)*, Saint-Malo, 13-15 mars 2002, Annie Morin & Pascale Sébillot (éds), Rennes : IRISA, 403-411

http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/lamalle_salem.pdf

Lebart, Ludovic, Salem, André (1994) – *Statistique textuelle*, Dunod.

Loiseau, Sylvain (2006) - *Sémantique du discours philosophique : du corpus aux normes. Autour de G. Deleuze et des années 60*, Thèse de doctorat, Sciences du langage, Université de Paris X Nanterre, 1er décembre 2006.

Loiseau, Sylvain (2007) – « *CorpusReader* : un dispositif de codage pour articuler une pluralité d'interprétations », *Corpus*, 6, 153-186.
http://corpus.revues.org/index1282.html

Malrieu, Denise (2006) - « Familles narratologiques et balisage du roman contemporain », *Proceedings of the First International Conference of the Alliance of Digital Humanities Organisations*, Paris:Centre Cultures Anglophones et Technologies de l'information, Paris IV, 131-139.

Malrieu, Denise, Rastier, François (2001) - « Genres et variations morphosyntaxiques », *Traitement automatique des langues*, 42 (2), 547-577.

Mayaffre, Damon (2005) - « De la lexicométrie à la logométrie », *L'Astrolabe*.
http://www.uottawa.ca/academic/arts/astrolabe/articles/art0048/Logometrie.htm

Mayaffre, Damon (2008) - « De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie », *Syntaxe & Sémantique*, 9, 53-72.

Mellet, Sylvie, Salem, André (éds) (2009) – Topographie et topologie textuelles, *Lexicometrica*
http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/special9.htm

Mézaille, Thierry (1995) - « La couleur des sentiments chez Proust », *in* François Rastier (éd.), *L'analyse thématique des données textuelles : l'exemple des sentiments*, Paris : Didier, collection Études de sémantique lexicale, 201-219.
http://www.revue-texto.net/1996-2007/Parutions/Analyse-thematique/Mezaille.pdf

Pincemin, Bénédicte (2002) - « Sémantique interprétative et analyses automatiques de textes : que deviennent les sèmes ? », Benoît Habert (dir.), Dépasser les sens iniques dans l'accès automatisé aux textes, *Sémiotiques*, 17, décembre 1999, 71-120.

Pincemin Bénédicte (2006) - « Concordances et concordanciers -De l'art du bon KWAC », *Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation, Actes du XVIIe Colloque d'Albi Langages et Signification*, Albi, 10-14 juillet 2006, Carine Duteil-Mougel & Baptiste Foulquié (éds), ISBN 2-907955-12-18, 33-42, et Texto!, juin 2006, vol. XI, n°2.
http://www.revue-texto.net/1996-2007/Parutions/Livres-E/Albi-2006/pincemin.pdf

Pincemin Bénédicte (2008) - « Modélisation textométrique des textes », *Actes des 9es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2008)*, Lyon, 12-14 mars 2008, Serge Heiden & Bénédicte Pincemin (éds), Lyon : Presses Universitaires de Lyon, ISBN 978-2-7297-0810-8, vol. II, 949-960.
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2008/pdf/pincemin.pdf

Poudat, Céline (2006) - *Etude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres*, Thèse de doctorat, Sciences du langage, Université d'Orléans, 20 juin 2006.
http://www.revue-texto.net/1996-2007/Corpus/Publications/Poudat/Etude.html

Prié, Yannick (1995) - *Contribution à une clarification des rapports entre Sémantique Interprétative et Informatique*, Mémoire de DEA, Informatique, Université de Rennes 1
http://www.revue-texto.net/1996-2007/Inedits/Prie95.pdf

Rastier, François (1987) – *Sémantique interprétative*, Presses Universitaires de France.

Rastier, François (1991) – *Sémantique et recherches cognitives*, Presses Universitaires de France.

Rastier, François (éd.) (1995) - *L'analyse thématique des données textuelles : l'exemple des sentiments*, Paris : Didier, collection Études de sémantique lexicale.

Rastier, François (2001) – *Arts et sciences du texte*, Presses Universitaires de France.

Rastier, François (2005) - « Enjeux épistémologiques de la linguistique de corpus », *in* Geoffrey Williams (éd.), *La Linguistique de corpus*, Rennes : Presses Universitaires de Rennes, 31-46.
http://www.revue-texto.net/1996-2007/Inedits/Rastier/Rastier_Enjeux.html

Rastier, François (2007) - « Passages », *Corpus*, 6, 25-54.
http://corpus.revues.org/index832.html

Rastier, François (2008) - « Que cachent les "données textuelles" ? », *Actes des 9es Journées internationales d'Analyse statistique des Données Textuelles (JADT 2008)*, Lyon, 12-14 mars 2008, Serge Heiden & Bénédicte Pincemin (éds), Lyon : Presses Universitaires de Lyon, ISBN 978-2-7297-0810-8, vol. I, 13-26.
http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2008/pdf/rastier.pdf

Rastier, François, Cavazza, Marc, Abeillé, Anne (1994) – *Sémantique pour l'analyse*, Paris : Masson.

Reutenauer, Coralie, Valette, Mathieu, Jacquey, Evelyne (2009) - « De l'annotation sémique globae à l'interprétation locale : environnement et image sémiques d'"économie réelle" dans un corpus sur la crise financière », *Conférence ARCO « Interprétation et problématiques du sens »*, Rouen, 9-11 décembre 2009.

Salem, André, Fleury, Serge (éds.) (2008) – « Explorations textométriques », *Lexicometrica*.
http://www.cavi.univ-paris3.fr/lexicometrica/numspeciaux/special8.htm

Valceschini-Deza, Nathalie (1999) – *Accès sémantique aux bases de données textuelles*, Thèse de doctorat, Linguistique, Université de Nancy 2, 29 juin 1999, 380 pages.

Valette, Mathieu (2004) - « Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet », *in* Patrice Enjalbert & Mauro Gaio (éds) *Approches Sémantiques du Document Numérique, Actes du 7e Colloque International sur le Document Electronique*, 22-25 juin 2004, 215-230 ; et (version légèrement étendue) *Texto!*
http://www.revue-texto.net/1996-2007/Inedits/Valette/Valette_Princip.pdf

Valette, Mathieu (2006) - « Observations sur la nature et la fonction des emprunts conceptuels en sciences du langage », *Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation, Actes du XVIIe Colloque d'Albi Langages et Signification*, Albi, 10-14 juillet 2006, Carine Duteil-Mougel & Baptiste Foulquié (éds), ISBN 2-907955-12-18, 107-114, et *Texto!*, juin 2006, vol. XI, n°2.
http://www.revue-texto.net/1996-2007/Parutions/Livres-E/Albi-2006/Valette.pdf