Stylistique et statistiques Le corpus textuel et Hyperbase

Véronique Magri-Mourgues

Stylistique et statistiques

Le corpus textuel et Hyperbase

L'analyse de données textuelles est désormais admise comme partenaire de la linguistique textuelle. L'intrusion des modèles mathématiques dans l'explication des textes littéraires peut se manifester par exemple dans des journées qui lui sont entièrement consacrées, où se côtoient linguistes, stylisticiens et informaticiens, les Journées d'Analyse statistique des Données Textuelles, qui se tiennent une fois tous les deux ans.

L'avantage reconnu du traitement informatisé est double : il porte sur la taille des corpus susceptibles d'être étudiés et sur la fiabilité des critères de recherche qui peuvent leur être appliqués ; pouvoir soumettre les très grands corpus à un questionnement stable, même s'il ne peut toujours pas être revendiqué comme objectif, est en effet une des distinctions essentielles entre la recherche traditionnelle et la recherche informatisée.

Reste à déterminer la nature de ce questionnement dépendant des objectifs de la recherche qui évoluent eux-mêmes au fur et à mesure que se perfectionnent les logiciels d'analyse.

La lexicométrie s'est d'abord définie comme étude du vocabulaire, avant qu'on ne parle de logométrie – comme étude globale d'un discours – ou encore de textométrie, comme analyse d'un texte. Le terme de stylométrie[1], quant à lui, fonde sa spécificité dans la caractérisation d'une écriture. C'est sur lui que je m'arrêterai tout en posant aussi la question attendue et corollaire de la définition du style même. Deux acceptions du style sont à envisager, celle qui le définit comme recherche de la caractéristique ; celle qui le relie au sentiment esthétique de la littérarité. L'évaluation du corpus textuel évolue parallèlement à l'outil informatique : celui-ci est envisagé tantôt comme un ensemble de paradigmes, tantôt comme un ensemble de séquences.

1. Le corpus textuel comme un ensemble de paradigmes

1.1. Un corpus à géométrie variable

Un corpus de grande taille

La question du corpus jugé pertinent en stylistique se pose en préalable. Le corpus qui est un objet empirique, construit et structuré selon les objectifs de la recherche, influe sur la définition de la stylistique. La variation de la taille d'abord du corpus d'étude est un premier paramètre à prendre en compte. Si on envisage les relations entre statistique et stylistique, l'exercice atomiste universitaire est bien sûr à exclure : la statistique est inopérante avec les extraits de texte et s'accommode mieux de grands corpus, l'œuvre entier d'un écrivain, une époque particulière, un ensemble générique. Qu'elle soit individuelle, générique ou spécifique d'une période de l'histoire littéraire, la stylistique

[1] Le mot, dans le sens d'étude du style d'une œuvre, est attesté dès la fin XIXe siècle. Voir P. Tannery (1876). « La stylométrie : ses origines et son présent » *Revue philosophique de la France et de l'étranger*.

s'appuiera sur un corpus d'extension variable et modulable selon les enjeux de la recherche.

Un corpus contrastif : la norme et l'échantillon

Un second préalable est la démarche contrastive et différentielle, essentielle pour la caractérisation stylistique. Si le style est conçu comme « processus de singularisation » d'une œuvre[2], il implique la définition première d'un corpus de référence pertinent, susceptible de faire ressortir des patrons stylistiques de l'objet d'étude. Il faudra définir non seulement un corpus d'étude mais aussi un corpus de référence, ce dernier permettant de poser une norme-étalon ; une stylistique de l'écart retrouve ainsi une légitimité mais l'appréciation de la différence ne se fait pas par rapport à un usage supposé standard de la langue qui serait donné comme invariant pour tous les textes mais par rapport à un ensemble de textes englobant. Les contours de l'un et l'autre corpus sont à définir par le chercheur selon le point de vue adopté. Les protocoles d'analyse se règlent sur un objectif épistémologique premier : dès lors que l'exhaustivité est irréalisable, le choix de l'échantillon se fait en fonction de sa représentativité évaluée là encore par le chercheur.

On pourra alors choisir une norme externe, par exemple si on veut caractériser l'écriture d'un écrivain – en choisissant tout ou partie de son œuvre – par rapport à la langue de son époque. La base Frantext est alors un outil précieux pour la comparaison[3].

On peut, au contraire, choisir une norme endogène construite pour les besoins de l'analyse ; c'est le principe qui a prévalu à la réalisation d'une base de données *Auteurs* par É. Brunet : les œuvres complètes de soixante-dix écrivains distribués sur quatre siècles sont regroupées afin de rendre compte de l'usage comparé des mots chez les plus grands auteurs de la littérature française, dont l'œuvre de Baudelaire.

La notion de relativité et l'écart réduit

Le calcul de *l'écart réduit* est à la base de nombreuses évaluations différentielles ; il permet par exemple, par une mesure de la différence quantitative entre une fréquence théorique et une fréquence observée dans un corpus donné, de dresser le vocabulaire qui est dit spécifique d'une œuvre ou d'un ensemble d'œuvres par rapport à un autre ensemble. La lecture du corpus textuel est tabulaire : les textes et les mots sont croisés sur un tableau à deux entrées ; les mots se distribuent sur les lignes, les textes sur les colonnes. On obtient alors le genre de liste qui suit, établie par exemple pour les œuvres de Baudelaire dans la base *Auteurs* ; une confrontation est opérée entre le vocabulaire utilisé dans les œuvres de Baudelaire et celui utilisé dans les œuvres des autres écrivains répertoriés dans la base.

[2] A. Herschberg-Pierrot (2006), p. 31.
[3] Voir par exemple les travaux d'É. Brunet.

Vocabulaire spécifique Baudelaire

N°	Ecart	Corpus	Texte	Mot
35	120.52	375	132	opium
35	40.92	4548	169	couleur
35	40.73	3190	139	tableau
35	36.92	956	67	dessin
35	36.43	2334	106	peinture
35	32.55	2090	90	tableaux
35	32.08	627	47	tons
35	22.48	2507	71	artiste
35	22.28	1819	59	rêves
35	22.27	1993	62	peintre
35	21.76	1720	56	cerveau
35	21.61	1362	49	ivresse
35	21.43	166	16	célèbre
35	21.37	556	30	dessins
35	21.18	213	18	dose
35	19.77	296	20	romantisme
35	18.89	4301	83	manière
35	17.68	237	16	rubens
35	17.62	980	34	composition
35	16.75	796	29	portraits
35	16.72	1012	33	généralement
35	16.57	1203	36	artistes
35	16.55	154	12	chanvre
35	16.42	4608	77	vin
35	16.39	376	19	sculpture
35	16.17	216	14	compositions
35	15.76	881	29	tempérament
35	15.71	886	29	jouissance
35	15.71	446	20	voluptés
35	15.69	170	12	rembrandt
35	15.64	2695	54	ténèbres
35	15.58	1602	40	lecteur
35	15.48	1547	39	énergie
35	15.30	1299	35	parfum
35	15.28	2699	53	aspect
35	15.27	17969	0	comme
35	15.26	1440	37	atmosphère
35	15.26	426	19	originalité
35	15.19	995	30	parfums
35	15.15	1387	36	peint
35	15.13	7491	97	beauté
35	15.09	1757	41	paysage
35	15.01	156	11	doués
35	14.98	484	20	romantique
35	14.96	157	11	maint
35	14.95	216	13	confessions
35	14.84	1880	42	harmonie
35	14.47	1794	40	singulier
35	14.45	13549	137	ô
35	14.32	430	18	analogue
35	14.24	434	18	ténébreux
35	14.14	690	23	jouissances
35	13.89	43155	1870	une
35	13.76	1097	29	rêverie
35	13.71	217	12	béatitude
35	13.70	27686	215	esprit
35	13.37	303	14	violon
35	13.30	1302	31	volupté
35	13.19	12183	120	tes
35	13.18	2710	47	portrait
35	13.18	1474	33	idéal
35	13.17	775	23	peintres
35	12.61	292	13	miroirs
35	12.56	971	25	facultés
35	12.29	3979	56	ange
35	12.26	869	23	satan
35	12.11	157	9	caïn
35	12.10	2322	40	douleurs
35	12.09	230	11	analogues
35	12.07	158	9	hallucination

Cette liste fournit, pour chaque forme retenue, sa fréquence dans le corpus qui sert de norme de référence et sa fréquence dans le texte, autrement dit le sous-corpus baudelairien.

La stylistique comme recherche de la caractéristique se fonde ici sur les formes-occurrences pour ainsi dire brutes, qui constituent le corpus textuel et dont on compare les fréquences. Ces formes-occurrences répondent au double impératif d'être *observables et quantifiables*, pour pouvoir être évaluées par les modèles mathématiques. Mais ce ne sont pas les seules variables

à pouvoir être envisagées. La lecture du corpus textuel se module en fonction du choix des unités à étudier.

1.2. Les variables du corpus textuel

Quelles sont les unités qui peuvent encore être dénombrées et analysées les unes par rapport aux autres ?

Depuis quelques années déjà, le logiciel Hyperbase intègre un analyseur Cordial qui procède à l'étiquetage morpho-syntaxique des mots d'un texte : celui-ci fournit, pour chaque mot, la graphie, le lemme de rattachement - ou entrée lexicale du dictionnaire - le codage grammatical autrement dit sa catégorie, sa fonction dans la phrase, ainsi qu'une information d'ordre sémantique qui classe le mot dans un champ lexical. Hyperbase redistribue ces données dans les champs idoines et en effectue le dénombrement. Dès lors, la finesse du codage est telle qu'elle permet de travailler soit sur les graphies (le texte est constitué alors d'une suite de graphèmes séparés par un blanc ou un signe de ponctuation), soit sur les lemmes[4] (le texte est envisagé comme un extrait de dictionnaire), soit sur les codes grammaticaux[5], soit enfin sur les structures syntaxiques. Celles-ci correspondent à des séquences ordonnées de codes grammaticaux dont les limites sont fixées par les signes de ponctuation. Le texte équivaut alors à un ensemble de paradigmes qui privilégient tantôt le versant lexical du texte, tantôt le versant morpho-syntaxique.

Un aperçu du travail d'étiquetage opéré par le logiciel Cordial :

Chacun veut saisir sa malle ou ses paquets (Stendhal, *Mémoires d'un touriste*)	chacun_5vouloir_1saisir_1 son_7malle_2ou_8son_7paquet_2	[Pi-msnS] [Vmip3sV] [Vmn-V] [Ds3fssD] [Nc-fs-D] [Cc] [Ds3ps-D] [Nc-mp-D]

Chacun : pronom indéfini, masculin, singulier, sujet
Veut : verbe principal, indicatif présent, 3e personne du singulier, base de proposition
Saisir : verbe principal infinitif, base de proposition
Sa : déterminant, 3e personne, féminin, singulier, groupe objet
Malle : substantif, nom commun, féminin, singulier, groupe objet
Ou : conjonction de coordination
Ses : déterminant, 3e personne, pluriel, singulier, groupe objet
Paquets : substantif, nom commun, masculin pluriel, groupe objet

Les analyses qui portent sur les lemmes seront *a priori* plus sensibles à la thématique des textes que celles qui portent sur les codes grammaticaux par exemple. Celles-ci mettront davantage en valeur la structure morpho-syntaxique d'un corpus et les procédés d'écriture des différents auteurs.

1.3. Les variations remarquables

Si le corpus textuel est envisagé comme un ensemble de paradigmes, se pose encore la question de ce qui est remarquable dans les séries de paradigmes établies, et qui peut par conséquent faire l'objet d'un commentaire pour servir de point de départ à une interprétation

[4] Le lemme ou vocable correspond à une entrée du dictionnaire : les formes fléchies d'un verbe sont par exemple regroupées sous un seul vocable, le verbe à l'infinitif. Le lemme est la forme canonique.

[5] La version d'Hyperbase qui traite des corpus lemmatisés permet un codage des différentes formes d'un texte selon la catégorie grammaticale. La connexion entre les textes repose alors sur l'observation des parties du discours.

stylistique. L'association de trois critères valide la recherche statistique : *la répétition, la variation et la sériation.*

Le paramètre de la répétition peut être vérifié à différents niveaux : la répétition d'un mot, d'un syntagme, d'un lemme, d'un code est évaluée toujours de manière relative. La répétition est relative en ce sens qu'elle n'est signifiante que par rapport à une norme établie ; elle est indissociable du paramètre de la variation qui permet l'évaluation contrastive des corpus. Enfin, des paramètres récurrents et convergents identifiés peuvent être mis en série pour former l'armature d'un système. La démarche inductive peut ainsi à partir du corpus-échantillon valider les constantes pour un modèle – générique, linguistique ou autre.

Un genre littéraire, par exemple, peut se définir par des constantes observables à un moment donné, par des « régularités singulières »[6] ; une fois le modèle établi, on peut tabler sur le caractère prédictible des critères distinctifs et évaluer un texte par rapport à eux. Un texte particulier peut se conformer au patron stylistique ou, au contraire, s'en éloigner et s'ériger en exemple atypique avec des divergences plus ou moins importantes : la caractérisation stylistique de chaque oeuvre se fera comme une approche graduelle autour d'un noyau-prototype générique et engagera vers la caractérisation du style d'auteur.

Un exemple vient illustrer ces procédés d'évaluation de l'écriture fondée sur le dénombrement. En me plaçant dans la perspective de la stylistique générique, j'ai confronté deux sous-bases, l'une constituée de récits de voyage, l'autre de textes fictionnels écrits par douze écrivains depuis Chateaubriand jusqu'à Loti, et l'ensemble des vingt-quatre textes sert de norme de référence. On peut proposer, pour commencer, une image de ce qu'il est convenu d'appeler la distance intertextuelle.

Le calcul de la « distance intertextuelle »[7] permet de confronter les textes entre eux et d'obtenir une première image de leurs proximités.

[6] D. Malrieu et F. Rastier (2001).
[7] *Corpus*, n°2, *La Distance intertextuelle* (2003).

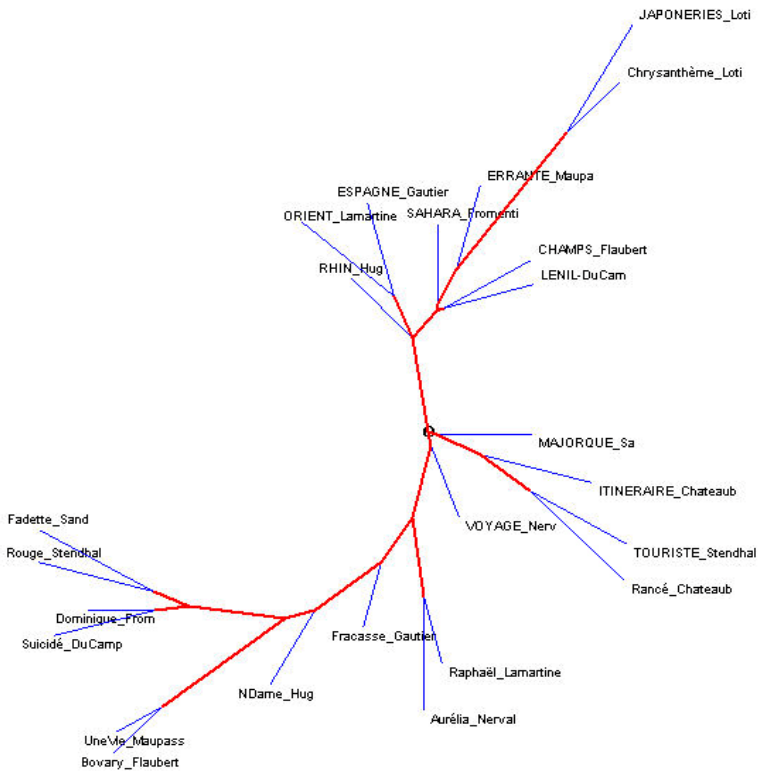

Figure 1 - La distance intertextuelle sur les lemmes. Calcul fondé sur la fréquence (méthode Labbé).

La méthode de calcul est celle de Labbé[8] qui repose sur un algorithme évaluant la distribution réelle des fréquences dans les textes confrontés. Plus les textes font un emploi quantitatif similaire des mots, plus ils se rapprochent sur le graphique et inversement. Le dépouillement du texte sous forme de lemmes permet le regroupement des formes qui se rattachent à la même entrée du dictionnaire en neutralisant les variations grammaticales ; il permet de mettre en évidence la structure thématique des œuvres[9].

La représentation graphique proposée est celle de l'analyse développée par Xuan Luong, qui permet non seulement de représenter les distances entre les textes de manière exactement proportionnelle, mais encore de mettre en évidence les nœuds de regroupement des textes au sein d'une structure arborée[10] particulièrement fiable et stable. Celle-ci permet de présenter une structuration classificatoire des textes en ajoutant la contrainte de la proximité entre les classes.

[8] Voir C. Labbé et D. Labbé (2003).

[9] Pour l'étude de la connexion entre deux textes, un autre protocole d'analyse peut être proposé par le logiciel Hyperbase. Il repose sur la méthode dite Jaccard qui se fonde sur le critère de la présence ou de l'absence d'un mot donné dans les deux textes considérés sans se préoccuper de sa fréquence. Si un mot est commun aux deux textes, il tend à les rapprocher ; la distance entre les deux textes augmente au contraire si le mot ne se rencontre que dans un seul. Tous les mots et tous les appariements des textes deux à deux sont envisagés. La version d'Hyperbase utilisée rapproche encore les textes par l'exclusion du même vocabulaire : les mots du corpus total qui ne se trouvent dans aucun des deux textes confrontés sont également pris en compte. La connexion intertextuelle se fonde par conséquent à la fois sur les mots utilisés par les deux textes et sur les mots également rejetés.

[10] La disposition dans l'espace du graphe, l'orientation ou les directions des branches importent peu pour l'évaluation de la distance intertextuelle ; seule compte la distance physique entre les points du graphique, ici les titres des œuvres, ainsi que les ramifications qui regroupent les points en bouquets ou au contraire les isolent ; les diverses bipartitions possibles de l'arbre manifestent la hiérarchie classificatoire.

On observe sur ce graphique la formation de trois ensembles : un groupe bien structuré rassemble la plupart des récits de voyage en haut du graphique, un autre un peu plus discordant réunit les récits fictionnels en bas du graphique. Certains « romans » se dissocient de l'harmonie d'ensemble : *Raphaël, Aurélia, Le Capitaine Fracasse* et *Notre-Dame de Paris* s'écartent en effet des deux ramifications principales du groupe fictionnel. Un autre groupe intermédiaire de quatre récits de voyage auquel s'associe *La Vie de Rancé* se détache au centre du graphe.

Et en haut à droite, on remarque le rapprochement des deux récits de Loti mais on sait les ambiguïtés génériques de ces récits, tantôt rangés par la critique parmi les romans, tantôt parmi les récits de voyage. La frontière est mince entre les deux et le logiciel met en évidence cette particularité. C'est le seul auteur, qui plus est, pour lequel le « roman » choisi est inspiré du voyage au Japon, à la source du récit de voyage. Les interactions tiennent peut-être à des convergences lexicales ou thématiques et brouillent le partage générique.

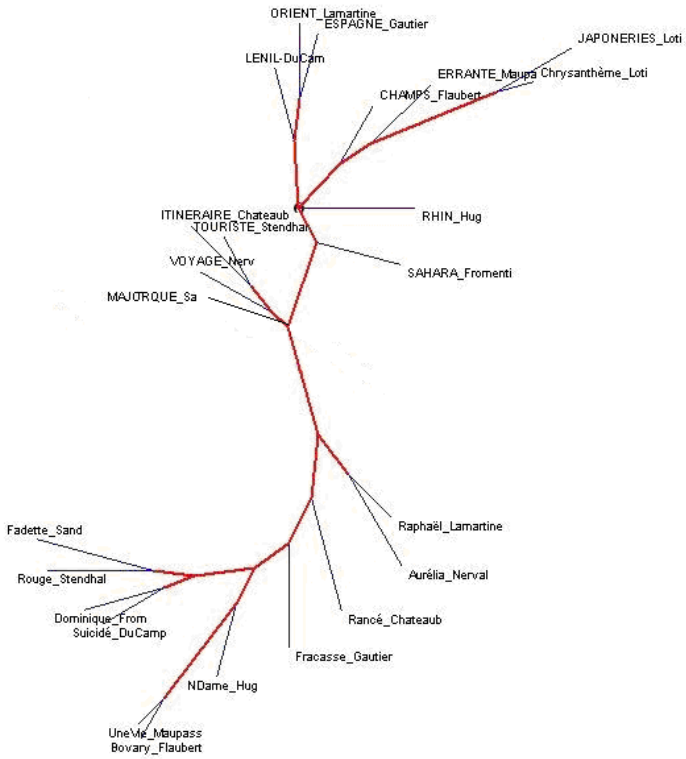

Figure 2 - La distance intertextuelle sur les codes. Calcul fondé sur la fréquence (méthode Labbé).

Le schéma est concordant avec le précédent permettant de montrer quels sont les textes qui s'attirent et quels sont ceux qui, au contraire, se séparent, selon l'usage qu'ils font des codes grammaticaux. Des groupes se forment, un axe d'opposition se dessine. Hormis pour Loti encore, les œuvres de chaque écrivain sont systématiquement dissociées l'une de l'autre. Les spécificités susceptibles de caractériser l'écriture de chaque écrivain sont ainsi éludées au profit de l'appartenance générique. En revanche, la distance grammaticale minimale qui se manifeste aussi entre les deux œuvres de Loti oblige à dépasser l'explication par la simple convergence thématique qui pouvait être proposée au vu du premier graphe ; c'est l'aspect grammatical qui est alors en jeu. L'écriture de l'écrivain se trouve caractérisée, indépendamment du lexique employé.

Le premier calcul qui porte sur les lemmes manifeste un rapprochement lexical des œuvres. Un vocabulaire similaire est utilisé dans chacun des deux groupes. En revanche, le schéma qui repose sur les codes grammaticaux met l'accent sur la structure grammaticale des œuvres.

2. Le corpus textuel comme un ensemble de séquences

2.1. La topologie – la dynamique du corpus textuel

On peut avoir cependant quelques réticences à traiter le texte comme une espèce d'urne, comme un ensemble de paradigmes variés parallèles qui déconstruisent le texte et menacent de le désintégrer. Après s'être organisée autour de l'observation des fréquences, l'analyse des données textuelles s'oriente davantage vers l'étude des séquences et retrouve ainsi l'ordre syntagmatique. Désormais, avec les nouvelles fonctions du logiciel Hyperbase, on se rapproche de la lecture naturelle d'un texte, qui suit son déroulement linéaire. Le texte est envisagé comme un espace où des réseaux se tissent entre les mots, entre les phrases, entre les paragraphes. Le texte se conçoit comme ensemble réticulaire[11].

Une fonction récente – celle de la topologie textuelle - dynamise l'étude du vocabulaire par une évaluation de la *répétition relative* au fil du corpus.

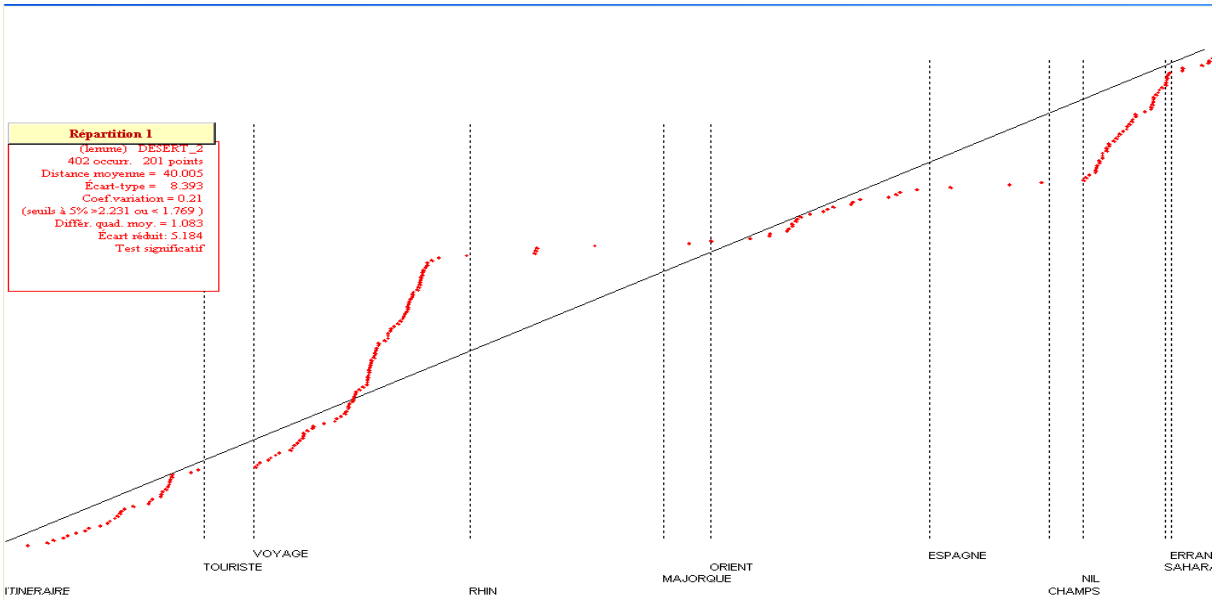

Figure 3 – La distribution du lemme « désert » dans onze récits de voyage.

L'étude porte sur onze récits de voyage du XIXe siècle qui forme un corpus textuel sans segmentation : le lemme étudié est le mot « désert » dont on observe la ventilation au cours du corpus. On constate facilement deux « rafales » de points – autrement dit une concentration d'occurrences du mot – matérialisées par une courbe qui se redresse pour le *Voyage en Orient* de Lamartine puis pour *Le Nil* de Du Camp et *Un Été dans le Sahara* de Fromentin. L'exemple ici choisi confirme simplement ce qu'on aurait pu aisément deviner dès le titre même des œuvres mais on peut multiplier les tests et parvenir à de véritables découvertes : la topologie textuelle, en corrélant des séquences qui présentent des caractéristiques communes, peut laisser percevoir des rapprochements inattendus entre deux passages ou deux textes et révéler l'architecture d'un texte ou d'un corpus. Les paramètres sont toujours les mêmes de la répétition et de la variation mais opèrent sur le texte, en quelque sorte en mouvement, décomposé en séquences successives.

[11] J.-M. Viprey (2005).

2.2. Approche sémantique du corpus

Les trois critères de la *répétition-variation-sériation* évaluent le corpus en termes quantitatifs et laissent au chercheur le soin de déduire une interprétation stylistique. Les limites de l'outil statistique seraient-elles atteintes quand on aborde le sémantique ? Des procédés littéraires qui jouent sur un détournement du sens, les tropes, paraissent – a priori – insaisissables par les logiciels d'analyse hypertextuelle.

Cependant, l'étude de l'environnement proche ou voisinage d'un mot peut amorcer de nouvelles directions d'étude : le calcul des corrélats lexicaux remet à l'honneur le cotexte étroit d'un lexème et rappelle la théorie contextuelle de la signification des linguistes anglo-saxons, selon laquelle les emplois d'un mot permettent d'en comprendre le sens.

> Nous avons vu qu'en sémantique lexicale, on peut appeler *signification* le contenu supposé invariant du mot et désigner par *sens* ses acceptions ou ses emplois en contexte : la signification est alors un type, constitué à partir des sens observés dans le discours, qui ont le statut d'occurrences.[12]

Un autre paramètre est alors à rajouter aux trois précédents énoncés, celui de la *distribution*.

Le principe de contextualité fait se rejoindre sémantique et distribution lexicale si on peut admettre que la somme des contextes dont un mot peut faire partie, de ses emplois possibles, en épuise la signification. Appliqué à un texte tout entier, le calcul des corrélats lexicaux ou des « associations privilégiées » est un premier test qui sert de base à d'autres calculs. L'analyse repose sur le choix préalable, effectué par le logiciel, des 400 substantifs les plus fréquents tout en évitant ceux dont la fréquence est trop élevée en raison de leur sens très usuel ; ils sont extraits du dictionnaire, sous la forme des lemmes. Le texte est exploré paragraphe par paragraphe et utilise le programme mis au point par L. Lebart (LX3AFC.EXE) intégré dans Hyperbase. Le seuil choisi pour que la cooccurrence soit retenue est de 3 ; le calcul des associations privilégiées repose sur un tableau général des cooccurrences fondé sur le calcul de la distance entre les mots de la liste, pris deux à deux.

Le programme analyse et trie le détail des associations deux à deux et propose une représentation sous forme de graphes des liens préférentiels qui tissent un réseau lexical autour d'un mot choisi pour pôle. Un parcours interprétatif, qui repose sur l'hypothèse initiale que « le contexte proche est structuré par des isotopies »[13], conduit de l'analyse lexicale à l'analyse thématique, des cooccurrents aux corrélats. Cette étude se veut purement expérimentale ; le choix du mot « homme » est arbitraire : il a été simplement senti *a priori* sensible à l'interprétation sémantique.

[12] F. Rastier (juin-septembre 2003).
[13] Rastier (1996).

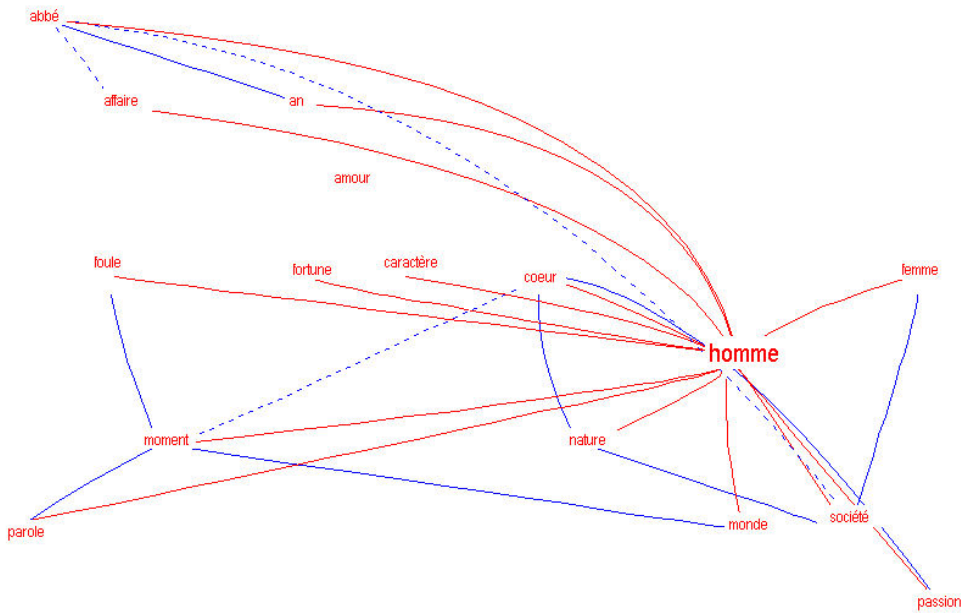

Figure 4 - Graphe des cooccurrences du mot-pôle « homme » dans le roman.

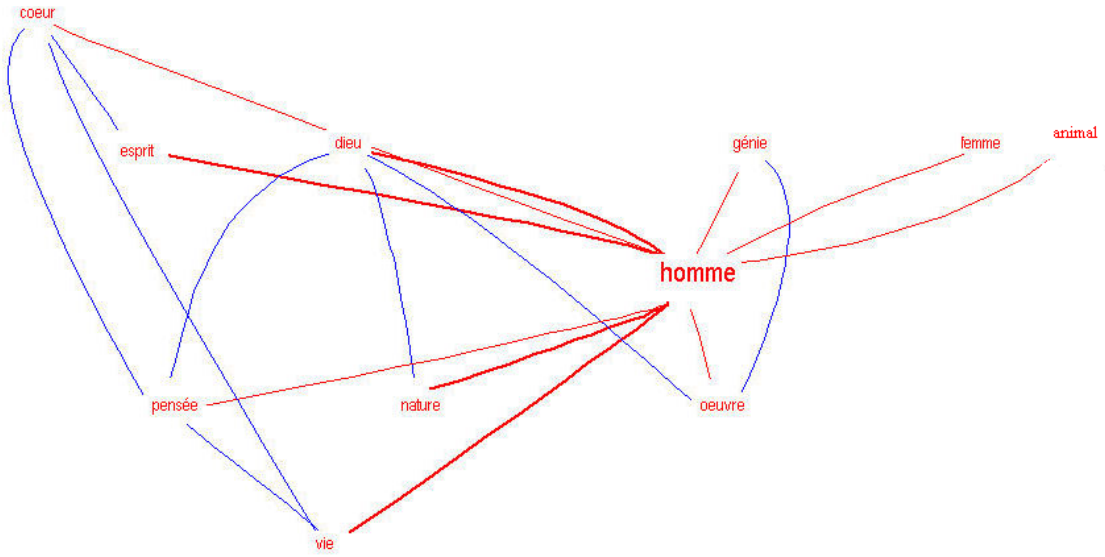

Figure 5 - Graphe des coocurrences du mot-pôle « homme » dans le récit de voyage

Sur les graphes, les conventions suivantes sont adoptées : les mots en rouge désignent les coocurrents directs du mot-pôle, reliés par un trait rouge, qui dessinent le premier cercle des cooccurrents. Les tracés en bleu matérialisent les relations des mots du premier cercle entre eux ; l'épaisseur du trait – trait en pointillés, trait maigre ou gras – correspond à la force de la liaison.

Dans le corpus fictionnel, le mot « homme » entre dans plusieurs réseaux lexicaux : un réseau affectif en premier lieu avec les mots « cœur » et « femme » et, dans une moindre mesure « passion » ; un réseau social et économique ensuite dessiné par les mots « société, affaire, fortune, monde, foule, parole » ; un réseau religieux enfin à peine esquissé toutefois par le terme « abbé ». Le texte factuel retrouve, quant à lui, le lien très fort entre l'homme et la nature et dessine davantage l'opposition entre le règne humain et animal ; ce couple « homme

– nature » paraît bien pouvoir fonctionner comme armature du récit de voyage dont l'enjeu est cette confrontation du voyageur – individu singulier et exemplaire représentatif – et des décors naturels traversés. Il affiche par ailleurs une visée plus didactique et philosophique avec des substantifs abstraits comme « esprit, génie, pensée, vie » et ébauche une réflexion métaphysique par le lien fort établi avec le mot « dieu », absent de l'environnement pour le corpus fictionnel.

Les affinités lexicales mises en évidence autour d'un mot choisi pour pôle permettent de cerner l'emploi de ce terme en contexte. On pourrait émettre cette hypothèse que le sens d'un mot se construit sur les termes qui se trouvent dans son « voisinage », dont l'étendue reste à définir. La problématique de l'interprétation est de pouvoir passer du niveau lexical au niveau sémantique : les cooccurrences doivent être interprétées comme des corrélats sémantiques pour pouvoir « être considérées comme des lexicalisations partielles d'un thème »[14].

Conclusion

La statistique impose des contraintes spécifiques au corpus textuel : ses dimensions d'abord mais aussi l'établissement d'une analyse différentielle et contrastive en vue d'une appréciation toujours relative des phénomènes. Elle suppose aussi des données observables et quantifiables qui pourront faire l'objet d'un commentaire si elles présentent des variations remarquables, c'est-à-dire répétées dans un texte ou au cours du texte, et qui peuvent être mises en série pour caractériser un corpus textuel.

Le corpus textuel peut être appréhendé de manière pour ainsi dire statique, comme un ensemble de paradigmes, ou de manière dynamique si on envisage sa décomposition en séquences de longueur variable. La lecture hypertextuelle se distingue toujours, quoi qu'il en soit, de la lecture linéaire.

Les outils statistiques remotivent la stylistique linguistique de Bally en s'appuyant sur des données concrètes, empiriques et observables et plaident en faveur d'une conception continuiste de la langue au style puisqu'ils servent une stylistique ancrée dans la matérialité linguistique, fondée sur les dénombrements qui se situe aux antipodes de la stylistique intuitive. La stylistique doit établir cette corrélation entre faits linguistiques microstructuraux et formes sémantiques. Le problème épistémologique qui se pose est d'articuler les observations quantitatives et l'interprétation qui cible le sens du texte. L'analyse qualitative demeure essentielle. La stylistique comme analyse de la caractéristique est pleinement légitimitée par l'investigation statistique ; comme recherche de la valeur littéraire, elle peut trouver sans doute une amorce dans l'analyse des environnements d'un mot, les corrélats lexicaux pouvant induire des inflexions sémantiques.

Bibliographie

Herschberg-Pierrot A. (2006). « Style, corpus et genèse ». *Corpus*, n° 5, p. 19-36.

Labbé C. et Labbé D. (2003). « La distance intertextuelle ». *Corpus*, 2, p. 95-118.

Malrieu D. et Rastier F. (2001). « Genres et variations morphosyntaxiques », *TAL Linguistique de corpus*, vol. 42, n°2, p. 548-577, repris dans http://www.revue-texto.net/Inedits/Malrieu_Rastier/Malrieu-Rastier_Genres.html.

Mayaffre D. (2007). « L'analyse de données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan. Retour sur les travaux actuels de topographie / topologie textuelle. *Lexicometrica*.

[14] F. Rastier (2001), chap. 8.

Mellet S., Barthélemy J.-P. (2007). « La topologie textuelle : légitimation d'une notion émergente ». *Lexicometrica*.

Rastier F. (1996). « La sémantique des thèmes ou le voyage sentimental », Texto [en ligne]. Disponible sur http://www.revue-texto.net/Inedits/Rastier/Rastier_Themes.html. (Consultée le 19 décembre 2007)

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

Rastier F. (juin-sept. 2003). « De la signification au sens. Pour une sémiotique sans ontologie ». *Texto !* [en ligne]. Disponible sur : http://www.revue-texto.net/Inedits/Rastier/Rastier_Semiotique-ontologie.html. (Consultée le 8 janvier 2008).

Viprey J.-M. (2005). « Philologie numérique et herméneutique intégrative », in Adam J.-M. et Heidmann U. (éds.), *Sciences du texte et analyse de discours*. Genève, Slatkine, 51-68.

Véronique Magri-Mourgues
Laboratoire BCL, Université de Nice Sophia-Antipolis, CNRS ;
MSH de Nice, 98 bd E. Herriot, 06200 Nice