Linx
Revue des linguistes de l'université Paris X Nanterre

43 | 2000
Linguistique de l'écrit, linguistique du texte

La Statistique de norme endogène dans l'analyse de textes

Endogenous normed statistics in text analysis

Jean-Marie Viprey

Édition électronique
URL : http://journals.openedition.org/linx/1099
DOI : 10.4000/linx.1099
ISSN : 2118-9692

Éditeur
Presses universitaires de Paris Nanterre

Édition imprimée
Date de publication : 1 décembre 2000
Pagination : 145-158
ISSN : 0246-8743

Référence électronique
Jean-Marie Viprey, « La Statistique de norme endogène dans l'analyse de textes », *Linx* [En ligne], 43 | 2000, mis en ligne le 11 juillet 2012, consulté le 01 mai 2019. URL : http://journals.openedition.org/linx/1099 ; DOI : 10.4000/linx.1099

Département de Sciences du langage, Université Paris Ouest

La Statistique de norme endogène dans l'analyse de textes

Endogenous normed statistics in text analysis

Jean-Marie Viprey

1 Cette communication développe certains axes d'une thèse, soutenue en 1996 et publiée par Champion, dans la collection « Travaux de linguistique quantitative » de Charles Muller, en 1997 (Viprey, 1997). Cette thèse elle-même avait pour objet de fixer une étape des travaux du Grelis, Équipe d'Accueil doctoral de l'Université de Franche-Comté, animée par les Pr. Claude Condé et Max Silberztein.

2 Jean Peytard, initiateur de nos travaux, a été l'un des pionniers de la diffusion en France des thèses de Mikhaïl Bakhtine, sur le dialogisme et l'interdiscursivité. Il faisait constamment référence à l'idée selon laquelle le sens langagier est un effet de la variation et de l'altération discursives, de la contextualisation dialogique. Il opposait tout aussi constamment à la sémantique de Greimas, dont il a montré les limites mécanistes, une organisation de la signification que résume à merveille la notion de *rhizome* chère à Deleuze.

3 Nous avons entretenu, sous et dans sa direction, une recherche sur l'automatisation des procédures lexicales : d'abord pour cerner la problématique des variantes manuscrites et éditoriales, puis (c'est le centre de ma thèse), pour améliorer la conception d'une statistique linguistique largement bloquée depuis les grands travaux de Pierre Guiraud, puis de Charles Muller.

4 L'objectif était et demeure de dégager les méthodes linguistiques, informatiques et statistiques d'explicitation et de modélisation des réseaux très complexes qui constituent la textualité.

5 En arrière-plan, nous entendons repérer des enjeux importants de la confluence des sciences de la littérature, du langage et de l'information.

1. Enjeux dans les champs disciplinaires

6 On peut, entre ces trois champs, cerner deux régions de débats qui nous concernent.

Discours et Texte

7 Nous ne pouvons prétendre désenchevêtrer définitivement ces deux notions. Il semble néanmoins nécessaire de travailler à leur distinction. Plutôt que de reprendre la discussion sur un plan théorique, considérons d'abord une classe de pratiques, très déterminée : l'Analyse de Discours telle que la définit l'École Française du même nom (Maingueneau, 1996 : 43), et plus précisément les méthodes statistiques illustrées par celle-ci.

8 Pour l'Analyse de Discours, un tableau de données lexicales sera du type suivant (fig. 1) :

		ÉNONCIATEURS				
		1	2	3	4	5
ITEMS LEXICAUX	a					
	b					
	c					
	d					
	e					
	f					
	g					
	h					

9 En colonnes, les descripteurs sont des instances discursives caractérisées socio-biographiquement (locuteur, lieu, temps). En lignes, ce sont des formes occurrentes : mots, séquences, phraséologies.

10 L'analyse multidimensionnelle de ces données produit des graphes où la projection des points-colonnes (énonciateurs) et des points-lignes (énoncés) présentent trois ordres d'interprétations :

- distribution des énonciateurs les uns par rapport aux autres (proximités et oppositions, groupements), tels que peuvent les caractériser des profils d'emploi des formes linguistiques : validation d'hypothèses sur les variations diachroniques (Congrès syndicaux, discours officiels...), sur les « véritables » affinités parfois masquées et que peut dévoiler cette procédure

- distribution des unités d'énoncés, qui peut servir à catégoriser idéologiquement des tendances d'emploi, des champs du lexique, en fonction des locuteurs qui y ont le plus fréquemment recours
- caractérisation croisée du vocabulaire et des positions discursives.

11 Dans les trois cas, la contrainte majeure est liée à la discursivité comme activité sociale, et dont on cherche à dévoiler les tendances foncières, masquées par la surface linéaire des énoncés. Les sociologues, les psycho-sociologues, les politologues, les sciences sociales en général ont intégré ces procédures à leur arsenal de base.

12 Ce constat nous amène à définir strictement le discours ainsi : ensemble des circonstances socio-biographiques d'une énonciation, acte et/ou champ d'activité par définition circonscrit.

13 Pourquoi ce parti-pris réductionniste ? Parce qu'on a cherché à appliquer les mêmes méthodes à des objets un peu vite considérés comme de même nature que le discours, parce que « faits de mots », à savoir les textes littéraires.

14 Dès ses premiers travaux, Guiraud veut contribuer à caractériser le style d'un auteur par les « choix » qu'il opère dans le stock lexical de sa langue : les mots-clés, plus tard rebaptisés spécificités (items surreprésentés dans un texte par rapport aux proportions d'un corpus de référence résolument exogène — le corpus du TLF, aujourd'hui noyau de la base textuelle *Frantext*). Rien n'affiche mieux la faiblesse heuristique de cette démarche que la publication en 1969, par P. Guiraud, d'*Essais de stylistique* où, sur les mêmes auteurs, il ne fait aucune mention des résultats de ses énormes calculs antérieurs.

15 De façon patente, cette application de la statistique lexicale renvoie à une conception de la stylistique fondée sur l'écart à la norme, en particulier sur ce que nous appelons aujourd'hui les *registres* de langue. Elle est en outre étroitement prisonnières des contraintes thématiques de l'énonciation littéraire : si tels et tels termes prédominent chez un auteur, comment déterminer si cela renvoie aux propriétés de sa langue, ou aux thèmes favoris qu'il aborde ?

16 Sans le dire très explicitement, Ch. Muller puis É. Brunet tirèrent un premier enseignement de cet échec, en préférant appliquer les spécificités aux rapports non plus d'un texte et d'un corpus hétérogène , mais d'un ensemble textuel et de ses parties organiques : pièces du théâtre de Corneille, romans de Zola.

17 Les résultats en furent bien sûr incomparablement plus probants, quant à la dynamique d'une oeuvre, aux contraintes croisées des campagnes d'écriture, à la pression des variétés de genres et de sous-genres.

18 Malgré tout, un aspect important demeurait inaperçu. Si dans le discours, la phraséologie isolée, l'impact d'un mot et surtout d'un segment, hors de toute syntaxe, peuvent prendre un relief déterminant, dans la textualité les relations syntagmatiques ne doivent plus être éludées ; les décomptes en masses ne peuvent être que des préludes à des analyses beaucoup plus fines, pour lesquelles la statistique multidimensionnelle semble avoir été précisément développée.

19 C'est d'ailleurs ce qui a historiquement détourné de ce domaine méthodologique beaucoup de spécialistes de la littérature.

20 Si le vocabulaire d'un discours peut encore se décrire par listes et fréquences, en y intégrant la problématique des *segments répétés* (cooccurrences les plus fréquentes), cela devient inopérant au niveau du texte. Le *vocable*, longtemps manipulé comme unité de

compte, doit désormais être pris en charge comme *champ*, comme voisinage dans un réseau hyperdimensionnel qui, en dernière analyse, est un modèle intéressant de ce qu'est la lecture sémiotique, ou littéraire.

Problèmes de la stylistique

21 Cela nous amène à l'autre discussion, qui porte précisément sur le style et la stylistique. Il y a un énorme malentendu, dans notre champ disciplinaire, que résume assez bien le terme de *stylométrie*. La prétention de « mesurer » le style, de quantifier ses constituants, doit pour le moins être tenue en lisière. Que peut signifier cette formulation ? Y a-t-il un programme scientifique possible derrière elle ?

22 Si l'on veut à tout prix quantifier, alors les grandes masses et proportions semblent bien les seuls critères accessibles. Mais aucun spécialiste du texte littéraire ne pourra prendre au sérieux de telles propositions, qui en effet ne lui apporteront pas d'aide appréciable et lui paraîtront seulement de nature à écraser les structures et propriétés réellement intéressantes, celles qui se jouent à l'échelle de la phrase et en-deçà.

23 Cependant, les choses en vont autrement si l'on retourne le problème et si, au lieu de se demander ce que peuvent bien apporter l'informatique et la statistique aux sciences des textes (car il faudrait admettre d'abord l'éventualité, largement défendue, d'une inadéquation radicale !), on examine les difficultés réelles de ces sciences, telles qu'elles se pratiquent couramment. L'une d'entre elles nous a retenu, qui concerne précisément style et stylistique. Le style peut se définir comme une propriété complexe, liée à un investissement particulier et irréductible de la langue, et se manifestant de manière diffuse à l'échelle d'un texte, d'une œuvre, mais dans son tissu le plus fin ; en d'autres termes, c'est une macro-propriété qui ne peut se manifester que comme micro-phénomène, un fait global n'ayant de réalité saisissable que locale.

24 Les faits stylistiques étant divers, et disséminés, leur saisie critique est infiniment difficile. La numérisation des textes a déjà apporté aux spécialistes la facilitation des recherches, en limitant le nombre et la dimension des lectures « la plume à la main », en accélérant la localisation et la collection des micro-éléments soulevés par une hypothèse en construction. Ce n'est pas là notre propos, même si nous serons par la suite amené à un retour critique et méthodologique sur cet aspect lui-même.

25 L'essentiel est du côté de la statistique. La statistique multidimensionnelle a pour objet précisément de dégager des macro-informations à partir d'ensembles de données hyper-complexes et engrénées ; ses procédures visent à extraire les faits saillants, nécessairement difficiles d'accès. Elle semble donc exactement adaptée au travail des hypothèses stylistiques.

26 Une autre difficulté des sciences textuelles réside dans la dichotomie des règnes de l'intuition et de la formalisation, avec pour toile de fond l'épineuse question de la *reproductibilité* des expériences, canon de scientificité universellement accepté. À ma droite l'approche érudite, le lecteur savant capable de citer exactement une source textuelle excitée par tel détour de sa recherche, l'intuition militante ; à ma gauche la fascination formaliste et machiniste, exacerbée par la montée en puissance des micro-ordinateurs, en quête des « secrets » du texte. Le compromis le plus courant de ces deux attitudes est l'idée d'une informatique totalement instrumentalisée, secrétant d'étranges

monstres conceptuels, comme le *Golem* d'Henri Béhar[1], qui indique : nous tenons la bête en laisse.

27 Et si les choses devaient être pensées en termes d'alternance, de dialogue ? Primo, l'informatique n'est pas d'abord un outil, mais un champ scientifique. Secundo, même considérée au plan de ses techniques et applications, elle ne saurait être un outil « comme un autre » : considérons le *temps*, même très bref, du déroulement d'un programme (même l'établissement d'une liste de formes) ; ce temps est celui de l'alternance hypothèse/validation. Plus le corpus est important et/ou l'algorithme complexe, plus ce temps est réellement perceptible dans le cours de la démarche ; il fait mûrir l'hypothèse, il suspend et relativise le système des attentes interprétatives et devrait favoriser la posture critique ; et aussi en favoriser l'apprentissage et la transmission, à condition de rompre à la fois avec une vision triomphaliste et avec une vision esclavagiste de l'ordinateur.

2. Vocable et vocabulaire : analyse distributionnelle et multidimensionnelle

28 Nous avons supposé que le vocabulaire d'un texte serait un candidat pertinent pour une première expérimentation, en raison de ses excellentes propriétés d'objet statistique (récurrence, dispersion, distribution), et aussi parce que nous serions d'emblée sur un terrain de dialogue avec les tentatives antérieures. Par différence avec ces dernières, nous proposons de considérer d'emblée le vocable, non pas comme un élément comptable caractérisé par son effectif, mais comme un élément de réseau, un fait relationnel, dont nous chercherons donc à calculer l'activité contextuelle, la collocation. Sa réalité mesurable sera le cotexte discontinu (Massonie, 1986A) de ses occurrences.

29 Nous considérons un ensemble textuel cohérent (*Les Fleurs du mal*), dans une dimension lemmatisée (laissons de côté les problèmes logiques de cette lemmatisation, qui justifieraient une autre communication : l'expérience a été menée, en 1996-97, à partir d'une lemmatisation fruste, largement emprique et fondée sur une conception très simpliste du *mot* : unité graphique simple renvoyant à une entrée lexicale unique).

30 L'*unité de contexte* est la mesure de l'environnement pris en compte autour de chaque occurrence, dans le relevé ; elle peut être mécanique (empan fixe en nombre de mots à gauche et/ou à droite), ou réglée textuellement (paragraphe, phrase, syntagme, strophe, vers...). C'est un paramètre à régler, puis à faire varier[2]. On peut dénombrer les items cooccurrents dans une unité de contexte définie : *cooccurrence brute*, puis rapporter cette mesure aux proportions du vocabulaire de l'ensemble, exprimées en termes de *cooccurrence théorique*, d'après l'hypothèse absurde d'une équirépartition intégrale ; ce rapport est positif (sur-représentation) ou négatif (sous-représentation), et il se calcule d'après la formule compensatoire de l'*écart-réduit*, décrit et expliqué par Muller (1992 : 69), que nous pouvons considérer comme un bon *indice de cooccurrence*.

31 Ces opérations nous permettent d'offrir à la consultation immédiate des listes de cooccurrents forts d'un vocable donné :

(fig. 2).

ciel		mer		monter		profond		vaste	
fond	4,39	vaste	7,219	mer	4,741	descendre	6,724	mer	7,219
bleu	3,865	profond	5,413	azur	2,946	mer	5,413	immense	6,015
soleil	3,596	vent	4,909	descendre	2,824	regard	4,461	feu	4,512
vaste	3,524	monter	4,741	clair	2,711	secret	4,461	ciel	3,524
matin	3,524	charmant	4,583	feu	2,711	riche	4,217	loin	3,356
triste	2,727	infini	4,43	blanc	2,606	odeur	3,42	couleur	3,337
enivrer	2,634	miroir	4,24	vie	2,508	puissant	3,254	deux	3,191
amoureux	2,47	immense	3,754	venir	2,383	fleur	3,072	clarté	3,188
azur	2,32	loin	3,706	doux	2,145	tombeau	2,714	clair	2,812
profond	2,244	mer	3,597	désir	2,101	vaste	2,714	profond	2,714
tomber	2,181	démon	3,293			doux	2,498	triste	2,705
rêver	2,052	amer	2,844			parfum	2,488	soleil	2,651
ouvrir	2,052	clair	2,717			ciel	2,244	esprit	2,599
clair	2,052	parfois	2,717			reine	2,039	bras	2,192
		soleil	2,361					sang	2,192
		main	2,11					grand	2,131

32 Une telle liste nous semble être un indicateur plus pertinent de l'activité et de la valeur du vocable dans l'ensemble textuel considéré, que la donnée de sa fréquence, voire de sa spécificité par rapport à un corpus de référence externe. MER est un mot-clé des *Fleurs du mal*, au sens que Guiraud donne à cette notion, c'est à dire qu'il y est significativement plus fréquent que dans la littérature du XIX° siècle prise comme norme. Cependant, il est bon de savoir à quels autres vocables du recueil l'écriture de Baudelaire l'associe préférentiellement.

33 Mais le plus intéressant dans cette direction consistera à livrer au calcul statistique, des matrices de grande dimension constituées sur le même principe, et dont la fig. 3 est un exemple réduit : il s'agit des matrices de cooccurrence (brute et corrigée) des 10 vocables les plus fréquents du recueil.

(fig. 3)

	âme	beau	ciel	cœur	dire	grand	noir	œil	plein	voir
âme	1	3	4	11	6	7	6	13	7	6
beau	3	3	9	20	14	6	5	19	3	4
ciel	4	9	1	11	4	8	9	13	1	11
cœur	11	20	11	10	20	14	13	15	13	7
dire	6	14	4	20	4	1	4	11	6	7
grand	7	6	8	14	1	1	8	13	6	5
noir	6	5	9	13	4	8	0	17	4	8
œil	13	19	13	15	11	13	17	12	15	14
plein	7	3	1	13	6	6	4	15	4	9
voir	6	4	11	7	7	5	8	14	9	5

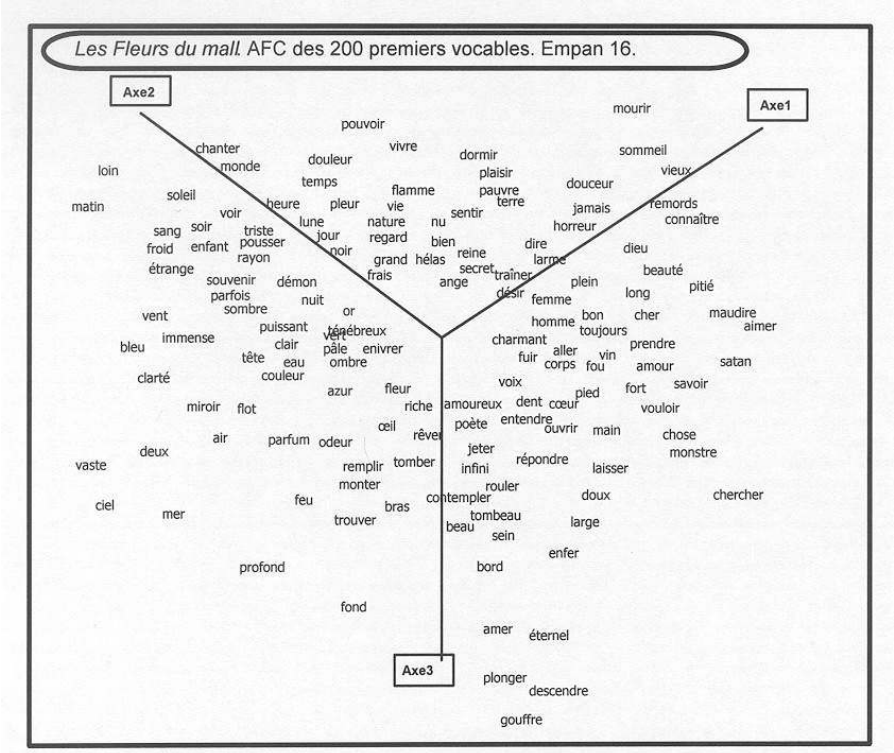
34 En première intention, ces matrices seront constituées à partir d'un critère de haute fréquence (les *n* plus fréquents vocables dans le texte), éventuellement croisé avec un critère de classe morphologique (les *n* plus fréquents verbes).

35 L'analyse Factorielle des Correspondances (AFC), mise au point par Benzécri dans les années '60, décrite notamment par Cibois (1983) et, dans les emplois que nous lui faisons assumer, par Massonie (1986A), est une procédure d'extraction de l'information pertinente dans une telle matrice, par paliers successifs (les « axes »). L'information pertinente est donc ici, en résumé, la différence entre le tableau des données constatées,

et un tableau « théorique » fondé sur l'hypothèse absurde de l'équirépartition. Cette différence, quand elle se creuse en négatif ou en positif, est justement ce que l'on nomme *correspondance*.

36 On notera que ces tableaux de données sont substantiellement différents de ceux qui ont été évoqués *supra* : ici les deux ordres de descripteurs sont des unités textuelles, homogènes. Il s'agit en effet de calculer des propriétés endogènes. L'Analyse Factorielle des Correspondances n'est que l'un des environnements logiques de la statistique multidimensionnelle, mais elle présente un avantage considérable pour ce qui nous préoccupe : elle organise ses résultats en vue d'une projection planaire orthogonale ou à 3 axes[3], ce qui permet d'obtenir un continuum graphique, ergonomique du point de vue de la poursuite de la recherche[4], dont voici un exemple :

(fig. 4)



37 Sur un tel graphique, la proximité de deux points indique que, dans les limites de la proportion d'information que la procédure a pu extraire, leurs profils contextuels sont proches. Cette proximité a d'autant plus de pertinence que les points sont loin du centre du graphe.

38 Soulignons qu'il s'agit d'un engrenage complexe de profils associatifs, qui offre donc une vue d'ensemble sur un aspect important de la structure du vocabulaire, donc sur un niveau essentiel de la structuration du texte à l'étude. Comme il s'agit d'un continuum, le repérage des tendances, groupements, se fait dans des conditions de plasticité et d'ouverture beaucoup plus grandes qu'avec des classifications dichotomiques. C'est une vue partielle, mais saillante, sur la *cooccurrence généralisée*.

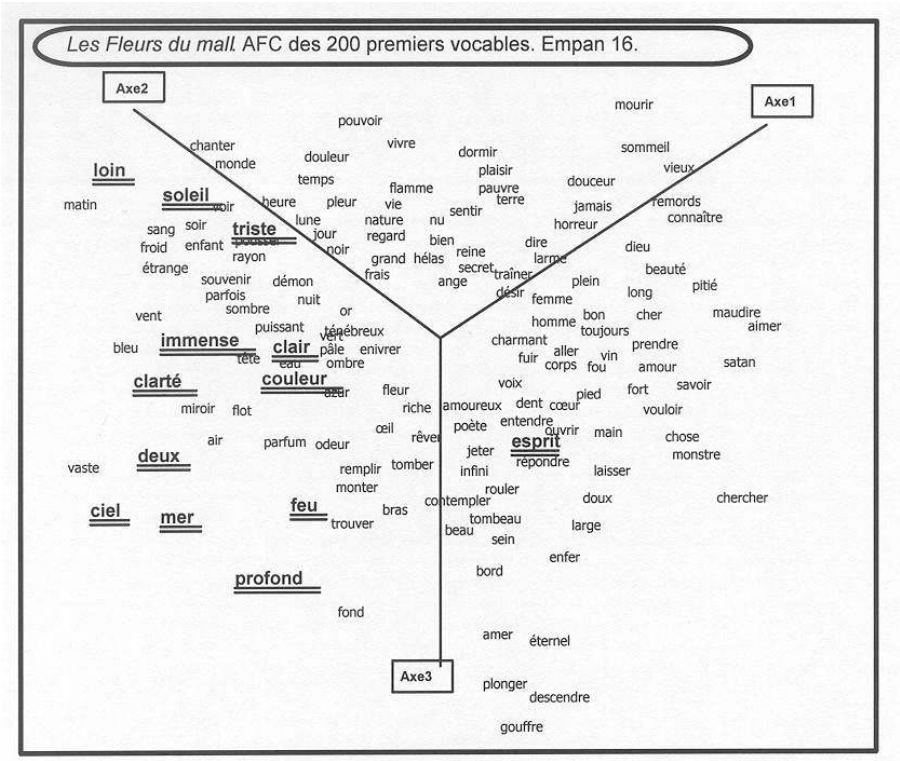
39 On vérifie que la procédure extrait de l'information significative, par des tests algébriques (Muller, 1992), mais aussi de manière plus empirique en constatant que ce sont bien les

vocables aux profils les plus accentués qui forment la périphérie du graphe, et que leurs cooccurrents les plus forts présentent une répartition significative dans l'ensemble :

40 la fig. 5 propose un repérage des cooccurrents de VASTE présentant un écart-réduit > 2,5. Si l'on descend la liste de la fig. 2, on voit le cercle s'élargir peu à peu.

41 La répartition manifeste à la fois une attraction commune vers VASTE, et une diversité des profils associatifs qui explique la dispersion des points dans toutes les directions. Nous appellerons les groupements plastiques rendus possibles dans la structure distributionnelle, des *isotropies* (gr. TREPEIN, *incliner vers*) ; cette notion est destinée à dialoguer avec celle d'*isotopie* selon notamment la sémantique greimassienne, qui sans ce dialogue tend à couvrir des pratiques exclusivement projectives (isotopies conçues *a priori*) qui écrasent de critères exogènes la lecture des textes particuliers.

(fig. 5)



42 Bien sûr, de tels résultats n'ont de sens que dans un dialogue poursuivi avec le texte. Nous envisagerons dans la dernière partie certaines conditions de ce dialogue, liées à l'hypertexte.

43 Envisageons tout d'abord quelques moyens simples. Nous pouvons notamment calculer quels sont les poèmes des *Fleurs du mal* les plus riches en occurrences des cinq vocables MER,VASTE,MONTER,CIEL,PROFOND.

44 En voici la liste (on indique l'occurrence brute et le volume en mots du poème) :

« Mœsta et errabunda »	cooccurrences	longueur du poème
« Le Balcon »	———————	———————
« Les Phares »	20	236
« Causerie »	12	248
« Les Bijoux »	11	342
	6	127
	6	273

45 Cette liste nous incite à retourner lire ces poèmes en regard les uns des autres, à y rechercher l'effet de l'occurrence particulière qui y est à l'oeuvre. Elle donne un relief thématique soutenu à ce quintil du *Balcon* :

> Ces serments, ces parfums, ces baisers infinis
> Renaîtront-ils d'un gouffre interdit à nos songes
> Comme *montent* au *ciel* les soleils rajeunis
> Après s'être lavés au fond des *mers profondes* ?
> — Ô serments ! ô parfums ! ô baisers infinis !

46 Il est enfin souhaitable de pouvoir s'intéresser à d'autres niveaux de structuration que le vocabulaire. Il est assez simple, par exemple, d'établir une liste des vocables fortement cooccurrents d'une forme non-lexicale, comme NOUS, ou comme le futur de l'indicatif, et d'indexer ces vocables sur le graphique, pour faire apparaître (ou non !) une ventilation significative de ces items. Dans les exemples ci-dessous, on constate une coïncidence manifeste, bien que les vocables concernés ne soient le plus souvent pas les mêmes. Cette observation nous invitera à des recherches plus poussées selon les critères ainsi mis en évidence.

47 Un travail analogue, trop complexe par les problèmes qu'il soulève pour être exposé ici, a été entrepris sur la corrélation vocable/allitération (Viprey, 1997 : 335 ssqq).

3. Application hypertextuelle

48 On l'aura compris, toutes les opérations sommairement décrites ici exigent un constant va-et-vient de la lecture aux calculs, des résultats graphiques aux contextes, d'un type de critère à l'autre. La statistique textuelle n'a de sens que si elle s'articule à un appareillage facilitant, voire autorisant tout simplement, cette circulation.

49 Je dois par exemple pouvoir obtenir, à partir du graphe de la fig. 4, une concordance organisée des contextes intéressés par les *n* vocables de mon choix ; ou encore, pouvoir confronter à ce graphe la liste des plus forts cooccurrents d'un vocable, de manière à pré-orienter la sélection de ces *n* vocables.

50 Je dois pouvoir relancer les calculs statistiques à partir d'une nouvelle hypothèse, et ce à tout moment. Par exemple, je peux vouloir repérer, dans la contrainte de la distribution des *n* premiers vocables par la fréquence, celle d'un « choix » de vocables déterminé par une hypothèse thématique (ce que l'on nomme souvent de manière un peu hasardeuse un *champ lexical*). Pour ce faire, je dois pouvoir soumettre une telle liste à un programme, qui formera la matrice de cooccurrence adéquate, en assurera l'analyse multidimensionnelle, et en affichera le graphique.

51 Ce graphique sera lui-même, plutôt qu'un résultat achevé, une simple base de départ pour de nouveaux retours au texte.

52 Nous nous proposons donc de systématiser ces besoins empiriquement définis. Ils nous posent tous le problème de l'hypertexte. Appliqué à la lecture critique des textes, ce dernier est aujourd'hui encore assez décevant. Avec des outils de recherche limités à la saisie clavier et aux clicks plein texte avec renvoi direct vers une concordance, la navigation devient vite fastidieuse et/ou vertigineuse si l'on ne veut pas se limiter à éditer de simples concordances.

53 La statistique multidimensionnelle, les graphes d'AFC, surtout si leur horizon est préférentiellement la norme de l'ensemble textuel lui-même (le texte saisi comme un ensemble de variations sur ses propres constantes), sont à notre disposition pour établir, au moins contribuer à établir, la voirie des ensembles hypertextuels littéraires ; une telle voirie est en fait une *viabilisation*, processus dynamique, endogène : les chemins apparaissent et disparaissent au fur et à mesure, seul l'historique en conserve une trace labile.

54 Le click plein texte, souvent irremplaçable mais qui souffre d'être opéré « au ras du sol », sera ainsi complété par le click sur un item du graphique, bénéficiant d'une meilleure vision panoramique et aussi de la possibilité d'être intégré à un choix multiple pertinent.

55 Dans une optique pluri-disciplinaire, que l'hypertexte stimule au premier chef, on pourra aussi considérer et théoriser ces graphiques comme une *cartographie* hypertexte. C'est de pleine actualité en sciences des textes, et du primordial intérêt de ces dernières. Les hypertextes en général, Internet en particulier, sont au centre des efforts actuels des cartographes.

56 L'informatique et la statistique n'ont pas pour cahier des charges de révéler le secret ultime des textes et des styles, ni de supplanter la lecture critique ; à l'inverse, elles ne peuvent être réduites au statut d'adjuvants inertes, de simples outils de facilitation. Comme tout ensemble de science et de technique, elles transforment les objets auxquels on les applique avec scrupule : les grandes bases textuelles, si nous voulons les rendre accessibles et opératoires, ne pourront être pleinement viabilisées que par une amélioration continue des procédures statistiques orientées vers la cartographie.

BIBLIOGRAPHIE

ADAM Jean-Michel (1998). – *Le Style dans la langue*. – Delachaux & Niestlé.

BALPE Jean-Pierre, LELU Alain, PAPY Fabrice, SALEH Imad (1996). – *Techniques avancées pour l'hypertexte*. – Hermès.

CIBOIS Philippe (1983). – *L'Analyse factorielle*. – P.U.F.

GUIRAUD Pierre (1954). – *Les Caractères statistiques du vocabulaire*. – P.U.F.

GUIRAUD Pierre (1969). – *Essais de stylistique*. – Klincksieck.

MASSONIE Jean-Philippe (1986A) – *Pratique de l'analyse de correspondances*. – Annales Littéraires de l'Université de Besançon, Les Belles-Lettres.

MASSONIE Jean-Philippe (1986B) – « Q-occurrences libres » in *Méthodes quantitatives et informatiques dans l'étude des textes*, éd. BRUNET Étienne (pp. 611-623). – Slatkine-Champion.

MULLER Charles (1992). – *Principes et méthodes de statistique lexicale*. – Champion.

PEYTARD Jean, GENOUVRIER Émile (1970). – *Linguistique et enseignement du français*. – Larousse.

VIPREY Jean-Marie (1997). – *Dynamique du vocabulaire des* Fleurs du mal. – Champion.

VIPREY Jean-Marie (1998). – « Une norme endogène pour le calcul stylistique du vocabulaire » in *4èmes Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 1998)*. – UNSA-CNRS (INaLF).

NOTES

1. Henri Béhar (1996) *La Littérature et son Golem*. – Champion.
2. Nous avons pu montrer, en comparant des résultats obtenus dans un empan fixe, à ceux obtenus dans le vers, un effet de *filtrage* de ce dernier (Viprey, 1997 : 202 ssqq), confirmant ainsi une hypothèse évidemment banale, mais dans l'optique surtout d'avancer vers la validation de la méthode elle-même.
3. C'est le cas de la projection ci-dessous ; il faut se représenter l'intersection des trois axes comme l'angle d'un cube, sur les faces duquel sont projetés les différents plans.
4. La plupart des autres méthodes présentent des classifications dichotomiques.

RÉSUMÉS

En linguistique textuelle, la statistique est appelée pour aider à la formalisation des niveaux de structuration, et à la liaison des macro- et des micro-analyses, notamment dans le domaine stylistique. Comme cette dernière, elle doit être reconçue en rupture avec la problématique des écarts exogènes, en faveur au contraire de la recherche des reliefs internes des œuvres : le texte se décrit et s'élabore dans l'analyse, comme un système d'écarts à sa propre "norme", notamment grâce aux techniques distributionnelles que l'on expérimente d'abord sur le vocabulaire. Les approches multidimensionnelles (benzécristes) sont privilégiées, mais aussi questionnées ; les résultats graphiques sont reçus comme de bons outils de repérage dans la perspective hypertextuelle.

In text linguistics, statistics are useful to help formalizing levels of structuration, and linking together macro- and micro-analysis, particularly in stylistic matters. As well as stylistics, statistics must be thought anew, far away from ideas of exogenous deviations. They would profitfully search about internal reliefs of works: one can describe and elaborate texts through analysis, as systems of deviations from its own "norm", provided that he uses distributional techniques, firstly upon vocabulary. Multidimensional approaches (benzecrist ones) are

emphasized and also critically tested; we show graphic outputs as good orientation and browsing tools in hypertextual outlooks.

AUTEUR

JEAN-MARIE VIPREY