SUR LES RÉPARTITIONS LEXICALES

CHARLES MULLER

Je me suis proposé de vous parler du problème posé par la répartition lexicale; mais il faut dire que si j'avais su que M. Etienne Brunet serait présent à ce colloque, j'aurais probablement choisi un autre sujet, car c'est lui plutôt qui devrait traiter ce sujet; mais j'espère qu'il prendra part à la discussion et qu'il comblera les lacunes de mon information par son expérience.

J'ai lu dans un article de M. André Robinet une phrase qui m'a accroché. Il écrit ceci: « Un commentaire de texte philosophique ou littéraire a tout avantage à partir de l'examen détaillé des résultats obtenus par les dépouillements automatisés avant de retomber dans les interprétations sémantiques ». J'adhère tout à fait à l'idée, mais dans cette phrase il y a un mot qui m'a choqué: « retomber », « retomber dans les interprétations sémantiques ». Je pense que, au contraire, quand on passe du recueil des données lexicales, statistiques, quantitatives, et qu'on passe même du traitement de ces données à l'interprétation sémantique, on ne tombe pas, on monte: on monte sur une pente assez glissante et assez dangereuse. L'idée que je voudrais développer et qui m'a amené à parler de ce sujet, c'est que autant nous sommes bien équipés en terminologie et en méthode dans les domaines de l'informatique et de la statistique, autant nous sommes démunis, terminologiquement et méthodologiquement parlant, dans le domaine de l'interprétation sémantique.

Je n'ai pas du tout l'intention de vous apporter des solutions, mais de poser le problème et d'appeler à des solutions que d'autres que moi pourront peut-être créer. Il s'agit donc du problème qui vous occupe depuis longtemps, à savoir de l'utilisation des moyens modernes pour l'interprétation des textes. Et on voit là-dedans trois phases, en gros. Il y a une phase informatique, qui consiste à enregistrer le texte, en accord avec celui qui est le spécialiste du texte et que j'appellerai le philologue (il peut être un philosophe, un historien, un linguiste, peu importe); le philologue, donc, demande que l'informatique vienne à son secours pour traiter un texte, et l'informaticien va le lui enregistrer avec un certain nombre de précautions que nous connaissons tous. Puis, quand l'informatique aura ainsi fait son office en enregistrant le texte, et l'aura reclassé par ordre (alphabétique, ou suivant d'autres critères), alors interviendra le statisticien qui, placé devant un grand nombre de résultats numériques, quantitatifs (tel vocable employé tant de fois, etc.), va appliquer certains tests, certains traitements; il apporte des indices qui permettent de calculer certaines choses, de synthétiser certaines données; et puis la dernière phase est celle du philologue, du spécialiste du texte, de celui qui le connaissait auparavant et en possède les données historiques, les données philologiques, et qui va se trouver devant les résultats statistiques et va devoir les interpréter. Je pense que c'est là le moment auquel nous sommes le moins bien préparés, et qu'il y aurait peut-être intérêt (ce serait là un bon sujet pour un colloque) à essayer de mettre sur pied une terminologie.

Cela dit, nous avons à parler de la répartition. Et je rappellerai d'abord que depuis qu'on parle de fréquence, on parle aussi de répartition.

Dès les premiers travaux (s'il s'agit des textes de langue française, je pense à ceux de Van der Beke, au français fondamental, aux dictionnaires de Juilland...), dès les premiers travaux, on a constaté que la fréquence devait être associée à la répartition,

c'est-à-dire que quand on avait établi combien d'occurrences avait une forme lexicale dans un corpus ou un texte, il était peut-être bon aussi de se demander comment ces occurrences se répartissaient dans l'étendue linéaire que représente le texte.

Seulement, dans ces premiers travaux sur fréquence et répartition, le but était toujours plus ou moins l'établissement de vocabulaires fondamentaux, ce qui n'est pas du tout notre objectif ici; on visait par conséquent la détermination de l'utilité du vocable, et, sa répartition intervenait alors comme une sorte de correctif, comme un critère d'élimination des vocables qui, ayant une fréquence élevée, étaient « mal répartis » (et déjà ce mot de « mal » nous met sur la voie de l'interprétation), qui étaient répartis de façon anormale, irrégulière, par exemple concentrés dans une petite partie du corpus; et c'était une raison de les éliminer ou de les déclasser; quiconque a suivi l'établissement du français fondamental ou de toute autre sélection de ce type connaît cette méthode et ses variantes.

Il s'agit ici de tout autre chose: d'utiliser la répartition, comme la fréquence, pour participer à la description d'un texte, à son analyse, à la recherche de ce que l'on peut savoir sur la façon dont il est composé, dont il a été composé. C'est donc un autre objectif. Il faut y inclure ce que j'appellerai (entre guillemets) le « comportement » d'un vocable donné, ou d'un groupe de vocables donnés: caractériser ce comportement, voir s'il est caractéristique de l'œuvre étudiée, ou de son auteur, ou si au contraire il est commun à tous les textes de même nature.

Et on constate aussitôt ceci: la fréquence a une expression simple; c'est un nombre, tout bonnement; quand je dis que tel vocable est employé 123 fois dans tel texte, il n'y a pas d'autre manière de le dire (à moins que Guilbaud en trouve de meilleures...?). En revanche pour la répartition nous sommes actuellement non pas devant une solution, ni même devant plusieurs solutions, mains dans un certain vide. On ne sait pas comment

caractériser, surtout par un chiffre, une répartition donnée. Donc les deux termes du couple « fréquence - répartition » ne sont pas au même niveau et sur le même plan: l'un est très facile à codifier, l'autre ne l'est pas.

Je me suis dit que le mieux était peut-être de prendre des exemples; des exemples qui n'apporteront aucune révélation, et n'obéissent qu'à le nécessité de faire du concret. J'ai donc pris un texte qui est enregistré sur support informatique, à savoir l'*Emile* de J.-J. Rousseau, dont notre ami E. Brunet vient de publier un index et une concordance (je ne m'occuperai ici que de l'index); je ne suis nullement spécialiste de Rousseau, et ce choix est de pure commodité.

Voici donc un texte dont il existe une bande enregistrée, contenant dans leur suite naturelle les 289 000 « mots » de cette œuvre (signes de ponctuation inclus), enregistrés dans un souci philologique parfait; on a respecté jusqu'à l'orthographe de la première édition (reproduite dans l'édition de la Pléiade). D'autre part nous avons cet index; puis l'inépuisable complaisance d'E. Brunet m'a fourni des listages dont je reparlerai. Voyons ce qu'on peut faire avec tout cela.

J'ai pris un vocable, à peu près au hasard, le vocable *cœur*, d'abord parce qu'il n'a pas d'homographe; en fait, je ne considérerai ici que la forme du singulier, bien que je reste farouchement partisan de lemmatiser; ce n'est que pour des raisons pratiques, mais quelques vérifications m'ont prouvé qu'en ajoutant les occurrences du pluriel *cœurs*, cela ne changeait rien aux résultats. Je constate tout de suite que *cœur* est employé 320 fois dans l'*Emile*; c'est sa fréquence. Si je la compare avec celles de vocables qui appartiennent au même champ sémantique, comme *esprit* ou *raison*, je constate qu'elles sont du même ordre de grandeur (265 et 277).

Mais si je m'intéresse à la répartition de ces 320 occurrences,

de quels moyens puis-je disposer? Eh bien, la première chose que j'observe, c'est que l'index me fournit non seulement cette fréquence, mais également les cinq sous-fréquences dans les cinq livres de l'*Emile*, à savoir 10, 23, 11, 158 et 118, et en outre, pour chacune des 320 occurrences, le numéro de la page où elle se trouve et une lettre (de *a* à *g*) qui indique la zone de la page, chaque page étant divisée en 7 zones d'environ 6 lignes; si bien que deux occurrences qui sont référencées 786 e et 786 f se signalent comme très proches. Voilà les données brutes sur lesquelles on peut travailler.

Je vais faire circuler un document dans lequel sont consignés les résultats numériques auxquels je suis arrivé, et qui vous permettra, je crois, de suivre plus commodément (v. p. après 78).

Le premier test auquel on pense, c'est évidemment d'étudier la manière dont ces occurrences se répartissent dans les cinq livres. Or ceux-ci sont très inégaux. La premier fait 9 % de l'ensemble, le second 20 %, le troisième 10 %, le quatrième 32 % (à l'intérieur duquel 1/3 environ est constitué par la *Profession de foi du Vicaire savoyard*), et le dernier 28 %. Tout le monde connaît le test de Pearson, le χ^2; c'est celui qui figure en haut de notre feuillet, à gauche. La première colonne donne les longueurs relatives des cinq livres; la seconde donne les produits de ces nombres par la fréquence totale, 320, donc des sous-fréquences théoriques. On constate que le résultat de ce test, 87,79, est énorme, c'est-à-dire que la répartition de ce vocable est très loin d'être aléatoire (permettez-moi ce terme non encore défini). Mais ce qui m'intéresse tout de suite, c'est de voir où sont les écarts; et on voit aussitôt un déficit très net dans les trois premiers livres, un excédent correspondant dans les deux derniers (le quatrième surtout); c'est-à-dire que ce vocable *cœur*, après avoir eu une fréquence relativement faible, passe tout d'un coup au premier plan du vocabulaire: Emile a pris de l'âge, il fait la connaissance de Sophie, et les sentiments prennent une importance énor-

me. Ce n'est pas là une découverte, et quiconque a lu l'*Emile* pouvait s'en douter.

Si je fais une opération semblable sur *raison*, qui forme plus ou moins un couple antithétique avec *cœur*, ce n'est plus du tout la même chose (en haut à droite); d'abord la répartition est encore très irrégulière, mais beaucoup moins que pour *cœur*, puisque nous arrivons à 19,26 (contre 87,79); mais en outre ce n'est plus une marche régulière: il y a d'abord un déficit, puis un excédent, puis un déficit, encore un excédent, et un déficit, l'excédent le plus notable étant celui du livre IV; donc les deux vocables sont majoritaires dans ce livre, mais *cœur* l'est également dans le livre V, où *raison* ne l'est plus. Je laisse aux spécialistes de Rousseau le soin d'interpréter ces faits.

Mais ma curiosité quantitative ne se satisfait pas de ce résultat; je me dis: *cœur* est réparti irrégulièrement, il est déficitaire dans les livres I à III, excédentaire dans les deux derniers; mais si je séparais ces deux sous-ensembles, si je considérais séparément les trois premiers livres, puis les deux derniers, que se passerait-il? On se demande alors si en prenant non plus les parties organiques de l'œuvre, voulues par Rousseau, mais des tranches artificielles, d'égale longueur, on retrouverait quelque chose de semblable. Pour cela, je prends tout naturellement, parce que j'en ai les moyens ici, la division en pages. Ayant sous les yeux les références, dans l'index, je constate que les occurrences du premier livre sont aux pages 243, 250, 259, 260, etc. Je n'ai qu'à suivre le défilé des références pour trouver combien de fois il se produit que deux références soient sur la même page, ou trois, ou quatre... J'observe ainsi très facilement que sur les 626 pages du texte (voir le second tableau de la colonne de gauche), il y en a 158 qui contiennent une seule occurrence, 53 qui en contiennent deux, 12 qui en ont trois, et 5 qui en ont quatre; par soustraction, on trouve que 398 des pages de l'*Emile* n'en ont aucune (là, petite invocation reconnaissante à la déesse Informatique:

quel travail absurde ce serait de relire *Emile* uniquement pour compter les pages sans *cœur!* Ne serait-ce pas humiliant pour l'esprit?).

Voilà donc une répartition par tranches artificielles, non voulues par l'écrivain, à laquelle je puis confronter une répartition calculée, celle que fournit la loi de Poisson; elle dit par exemple que si le hasard seul avait présidé à la répartition du vocable *cœur*, il y aurait 375 pages sans occurrence; or il s'en trouve 398, 23 de plus; ce qui entraîne évidemment un déficit dans les pages qui en contiennent, donc des groupements que le seul hasard n'explique pas: un peu plus de pages avec 3 ou 4 occurrences que le hasard ne le voudrait. Détaillons ce calcul:

	CALC.		OBS.		
N^bre d'occ.	N^bre de p.	Occ.	N^bre de p.	Occ.	Ecart
0	375,5	0	398	0	+ 22,5
1	191,9	191,9	158	158	— 33,9
2	49,0	98,0	53	106	+ 4,0
3	8,4	25,2	12	36	+ 3,6
4	1,2	4,8	5	20	+ 3,8
	626,0	319,9	626	320	0,0

Le test du χ^2 (tableau) donne 13,37, ce qui signifie que les écarts observés auraient moins de 1 chance sur 100 de résulter d'une répartition au hasard.

Mais on va voir ce qui se passe si l'on sépare les livres; suivez la première colonne de notre tableau: en considérant les trois premiers livres, les choses deviennent tout à fait normales: concordance de la répartition observée et du calcul, et un χ^2 minime (0,11): ainsi dans les 247 pages de cette partie, la loi de Poisson fait prévoir 206 ou 207 pages sans occurrence, et il y en a 206 exactement; de même 36,8 pages avec une occurrence, et il y en a 38. La correspondance est excellente, presque inquiétante.

Même opération pour le livre IV: là encore, excellente concordance et un χ^2 faible: 0,79. Avec le livre V, cela devient un peu plus irrégulier: on y constate des tendances aux groupements qu'on n'avait pas observées dans les livres précédents. Le χ^2, sans devenir vraiment significatif, monte à 4,30. On pourrait discuter sur les chiffres, mais cette expérience permet de dire que si le vocable que j'ai choisi est réparti de façon très anormale entre les cinq livres de l'*Emile*, il l'est d'une manière très normale à l'intérieur des trois premiers, également dans le quatrième, et de façon presque normale dans le dernier; donc une fréquence faible dans les trois premiers, une fréquence moyenne ou élevée ensuite, mais dans ces parties une répartition quasi aléatoire, sauf peut-être dans le dernier. Voilà. Cette expérience a pu se faire uniquement avec les données de l'index, et bien entendu l'édition de la Pléiade, pour aller voir les endroits du texte où il y a des concentrations.

On pourrait me dire: ce n'est pas très rigoureux; avez-vous tenu compte du fait qu'il y a des pages incomplètes (fins de chapitres, notes occupant plus d'une demi-page, etc.)? Certes, on pourrait chicaner là-dessus.

Il y a alors un moyen plus strict. Grâce à E. Brunet, j'ai pu étudier un certain nombre de vocables sur un listage dans lequel leurs occurrences sont numérotées d'après leur place dans le texte (encore un travail inhumain!): ainsi la première occurrence de *cœur* est le 1 205^{e} mot de l'œuvre, la seconde est le 3 513^{e}, et ainsi de suite. Seule l'informatique peut nous fournir cette précision.

Avec cet outil, il devient très facile d'imaginer des tranches rigoureusement égales, de 1000 mots par exemple, ou de 500, de 200, etc. J'ai fait l'expérience d'abord en considérant que les presque 289 000 mots du texte forment 289 tranches de 1000 mots (la dernière seule est incomplète). En suivant sur notre tableau la colonne du milieu, on verra d'abord que dans l'en-

semble du texte il y a 124 tranches sans occurrence, 80 avec une occurrence, 38 avec deux, etc. Là encore, la comparaison avec une distribution calculée fait apparaître des excédents parmi les tranches ayant 3 occurrences ou plus, et un χ^2 élevé: 36,49. Ensuite, en suivant cette colonne du milieu, on verra ce qui se passe pour le groupe des trois premiers livres, puis pour les deux derniers. En somme l'opération qui prend pour base la page et celle qui, d'une façon plus rigoureuse, se fonde sur le découpage en tranches de 1000 mots donnent des résultats concordants (heureusement, car je me demande bien comment on pourrait interpréter un désaccord!).

Et puis, ma curiosité n'étant pas tout à fait satisfaite, je me suis demandé ce qui se passerait si l'on prenait des tranches plus courtes. J'ai pris successivement des tranches de 500 mots (il y en a 577), puis de 100 mots (2885); les résultats figurent au tableau, en bas à gauche. Avec 500 mots, on trouve un χ^2 élevé (10,41), et ce sont les rencontres de 2 occurrences qui révèlent un excédent sensible. Avec 100 mots, les écarts deviennent négligeables, ce qui indique que la concentration d'occurrences ne crée pas, pour ce vocable, des répétitions à très courte distance; là aussi, la régularité est plus grande si l'on traite séparément les livres I à III et IV-V. Le tableau vous donne enfin dans sa colonne de droite des traitements semblables pour le vocable *raison*, qui révèle que son comportement, dans ce texte, est tout autre que celui que nous venons d'étudier; mais nous n'aurons pas le temps de traiter cette question, ni de rechercher pourquoi *raison* a une répartition tout à fait régulière dans les livres I, III et V, et non dans II et IV. J'ai également fait des tests sur *esprit*. Vous trouverez, en bas et à droite, deux résultats pour les trois vocables: sous 1, les χ^2 obtenus par la répartition entre les cinq livres: *coeur* est le plus irrégulier, et nous savons pourquoi (le thème des passions apparaissant au livre IV); *raison* s'écarte le moins de la répartition idéale; sous 2, résultat obtenu

sur des tranches artificielles (pages): c'est exactement inverse: c'est *raison* qui apparaît 3, 4 ou 5 fois dans la même page, trois fois plus souvent que le calcul ne le faisait prévoir (19 contre 6,4), alors que pour *coeur* ce phénomène ne se produisait que 17 fois pour 9,6, et sans dépasser 4 occurrences par page.

Voilà donc une première façon de regarder comment un vocable, une unité lexicale se distribue soit à l'intérieur des parties organiques du texte (chapitres, livres...), soit à travers des tranches que nous découpons de manière artificielle, mécanique, dans le texte: des pages, si nous travaillons sur un index bien fait, ou des séries de *n* mots, si un listage nous donne des numéros d'ordre pour les occurrences.

Cela dit, on a proposé d'autres moyens, peut-être plus puissants. L'un d'eux se fonde sur la loi hypergéométrique et a été exposé une première fois dans une petite publication ronéotypée que Guilbaud connaît bien. Son auteur, que je ne connais que de nom, Thomas Réginal (un de vos élèves, sans doute), présentait en 1975 une expérience sur un texte très court, 600 mots environ, dans lequel il étudiait la répartition d'un vocable à très haute fréquence: la préposition *de* sous ses différentes formes (*de*, *d'*, *du*, *des*), bien entendu sans limitation sémantique. Il proposait tout simplement de considérer les intervalles qui séparent les retours de cet unité, et comme il y a une cinquantaine d'occurrences, il a entre deux d'entre elles un intervalle X qui est en moyenne de douze mots environ: c'est la longueur du texte en mots, représentée par T, moins les F occurrences du vocable *de*, divisée par le nombre d'intervalles. Il y a là une petite question accessoire: faut-il compter l'intervalle entre le début du texte et la première occurrence, et celui qui sépare la dernière de la fin du texte? Si oui, on aura F + 1 intervalles, sinon il y en aura F — 1; on peut aussi les réunir en un seul, traitant le texte comme une boucle, et alors le nombre est égal à F; ce sont des détails qui, du reste, ne jouent qu'un rôle négligeable quand les

nombres sont grands, et que je ne mentionne que pour mémoire.

Le sujet a été repris et développé par Pierre Lafon dans une communication récente à Saint-Cloud. Il a donné des formules plus détaillées que j'ai reproduites (en bas à gauche du tableau), calculant la moyenne et l'écart type de la variable. J'ai essayé d'appliquer cette méthode aux cas de quelques vocables de l'*Emile*; je ne fournis ici que les résultats pour *coeur*, *esprit* et *raison* (à droite, sous 4). Mais cela pose un petit problème de méthode. Les deux auteurs que je viens de citer disent tous deux qu'on peut bien calculer un intervalle de part et d'autre de la moyenne, mais qu'on ne peut accorder une valeur probabiliste à ce qui est contenu dans cet intervalle. Moi, cela ne m'intéresse pas beaucoup, car si l'on dit simplement: « dans tout ce qui est compris entre telle et telle valeurs (« seuils »), la répartition n'est pas significative, en dehors de cet intervalle on dira qu'elle est irrégulière », il m'intéresse surtout de savoir si c'est tout près du bord de l'intervalle (du seuil), ou très loin. Donc je préfère quand même calculer un écart réduit, quitte à faire protester les mathématiciens, et c'est sous cette forme que j'ai donné mes résultats. Et vous voyez qu'ils sont énormes: 13 pour *raison*, 27 pour *coeur*, 30 pour *esprit*. Evidemment on ne peut accorder aucune valeur de probabilité à un écart réduit de 30: cela ne signifie à peu près rien; cela permet seulement de comparer les trois vocables, et de constater que le classement obtenu par ce moyen, en donnant à *raison* la répartition la moins anormale et à *esprit* la plus anormale, ne s'accorde pas avec celui de ma précédente enquête.

L'autre méthode est celle qui a été employée par Etienne Brunet dans un article qui est à la fin de son index; il calcule un indice η dont je reproduis les formules; cet indice part non point des intervalles entre deux occurrences, mais de la différence entre un intervalle et celui qui le précède; je m'explique rapidement et de façon peut-être rudimentaire: si les intervalles entre

les occurrences varient beaucoup, cela peut se produire de plusieurs façons: par exemple les intervalles courts alternent avec les intervalles longs; ou au contraire la variation peut être cyclique, si dans toute une partie les intervalles sont longs, comme pour *coeur* dans les trois premiers livres, pour devenir ensuite plus courts quand, à partir du livre IV, le vocable devient plus fréquent.

Evidemment les deux phénomènes ne sont pas comparables, ce qui conduit à penser qu'une vue globale et statique sur les intervalles peut ne pas nous suffire; il vaudrait mieux voir dans quel ordre ils se présentent, et c'est ce que fait l'indice η de Brunet. Par ce procédé (sous 3), c'est *esprit* qui s'écarte le moins de la régularité, *coeur* beaucoup plus, et *raison* encore un peu plus. Ce que je retiens pour l'instant, c'est que les quatre procédés donnent des classements différents. Chacun d'eux a sans doute sa valeur et sa signification. La difficulté est de les traduire dans un langage « sémantique », et c'est là que je constate que, dans les limites de mon information, nous ne disposons à peu près de rien. Car j'ai employé, tout au long de ces expériences, des termes que je n'ai pas définis: normal, anormal, irrégulier, bien ou mal réparti, etc. Comment exprimer nos résultats?

On parle volontiers de « répartition aléatoire », c'est-à-dire d'une répartition qui serait facilement obtenue par le hasard. Si les 289 000 mots de l'*Emile* étaient mélangés, en les classant en deux catégories, d'une part les 320 occurrences de *coeur*, d'autre part tous les autres; si, après ce mélange, on les tirait au sort en notant l'ordre de sortie des 320 *coeur* parmi les autres mots, on obtiendrait une des répartitions possibles, qui sont au nombre de:

$$\binom{T-1}{f-1}$$ ou, avec la notation française: $$\mathbf{C}_{T-1}^{f-1}$$

On peut calculer alors un intervalle à l'intérieur duquel se situeraient les types de répartition les plus nombreux et les plus probables, que l'on déclare aléatoires; à droite et à gauche de cet intervalle, limité par deux seuils, des répartitions peuvent s'écarter de l'aléatoire, soit par un excès de régularité, soit à l'inverse par un excès d'irrégularité; d'un côté, à droite par exemple, le vocable sera déclaré anormalement régulier, c'est-à-dire qu'il revient, peut-être par une volonté consciente de l'auteur, de façon quasi régulière; à l'opposé, à gauche si l'on veut, on aura le petit nombre de distributions dans lesquelles le vocable se concentre dans certaines zones du texte, alors qu'il est absent ou rare dans les autres; et c'est ce qui se produira le plus souvent. Si j'anticipe sur ce que dira M. Guilbaud, puis-je avancer qu'alors la probabilité d'emploi du vocable varie suivant les lieux du texte?

On dispose donc de ce vocabulaire un peu rudimentaire qui oppose « régulier » à « irrégulier », mais où souvent « régulier », faute d'une bonne définition, devient synonyme d'« aléatoire », alors qu'il ne devrait s'appliquer qu'à ce qui est significativement plus régulier que l'aléatoire:

seuil		seuil
irrégulier (significatif)	aléatoire (non significatif)	régulier (significatif)

Alors on a proposé des terminologies plus expressives, peut-être plus précises du fait qu'on les accompagne d'une définition.

Chez P. Lafon et ailleurs, j'ai trouvé le terme de « rafale », avec création d'un adjectif « rafaleux » et d'un substantif « rafalité »: la « rafalité d'un texte ». Je serais favorable à l'adoption d'un terme qui n'aurait pas de précédents dans le domaine philologique, qui aurait au moins l'avantage d'être neuf, vierge, et à qui on pourrait donner la définition qu'on veut; toute convention

terminologique est possible, alors que « régulier » ou « aléatoire » sont déjà contaminés, ont déjà été utilisés et peut-être de façon imprécise. Mais « rafale » me paraît très mal choisi, je le dis franchement; il a un avantage, c'est de bien exprimer que les choses se passent dans un univers à une dimension, linéaire, alors que quand on parle de « densité » on se place plutôt dans l'espace à trois dimensions. Mais ce mot de « rafale » fait penser à des intervalles très courts et surtout très réguliers, à des occurrences très rapprochées et à intervalles fixes. Alors, pour le mot *coeur*, qui serait déclaré « rafaleux » par P. Lafon, eh bien! non, ce n'est pas cela du tout. La preuve en est que quand je prends des tranches très courtes, de 100 mots, on devrait les voir, les rafales: on aurait de nombreuses tranches avec 2 occurrences, ou même 3 ou 4; or il n'y en a aucune avec 3 occurrences ou plus, et seulement 18 avec 2, alors que le calcul en annonce 16... C'est pourquoi ce choix ne me paraît pas heureux. Mais il serait bon qu'il y ait un terme pour exprimer que le vocable considéré, dans l'oeuvre étudiée, a tendance à venir à certains endroits très fréquemment, alors qu'ailleurs il est réparti à peu près aléatoirement. Je ne vois pas actuellement de mot qui exprime cela commodément.

Il y a une autre proposition, celle d'un chercheur qu'E. Brunet connaît aussi bien que moi, M.D. Dugast; il a un peu tendance à la métaphore, mais peut-être n'est-il pas mauvais d'avoir des métaphores expressives dans la terminologie. Il propose de distinguer les vocables qui constituent la « trame » du texte et ceux qui constituent des « motifs »; terminologie empruntée aux arts picturaux, si l'on peut dire. La trame, ce serait ce qui apparaît (aléatoirement) un peu partout dans le texte, mots outils ou non; les motifs, ce sont alors les vocables qui sont chargés par l'auteur, à un certain moment du texte, et pendant une partie assez brève, de jouer un rôle thématique.

Evidemment une terminologie utile de la répartition aurait

un double rôle: d'une part une tâche purement descriptive, pour nous dire, avec peu de mots: ce vocable est réparti au hasard, je ne lui trouve rien de significatif, il n'y a ni creux ni bosses dans la manière dont il apparaît dans mon texte; ou, au contraire, que ce n'est pas le cas, et sous quelle forme il s'écarte du non significatif; car il peut être régulier et fréquent dans certaines parties, et régulier et peu fréquent dans d'autres (c'était le cas de *coeur*, sur lequel je suis tombé par hasard); ou bien les irrégularités se distribuent un peu partout, avec des concentrations. Il y a une infinité de situations possibles, entre lesquelles il faudrait tracer des limites. Le deuxième rôle que devrait jouer cette terminologie, ce serait un rôle non plus descriptif, mais interprétatif; c'est de dire: si ce vocable est fréquent ici, c'est pour des raisons thématiques (le vocabulaire des passions à partir du livre IV); et il sera bon de regrouper tout ce qui contribue à un même thème. Ou bien il s'agira de raisons stylistiques... Là je vous citerai un petit exemple anecdotique. Quand j'ai demandé à Etienne Brunet des listages pour un certain nombre de vocables, dont je n'ai cité ici que trois substantifs, j'y avais inclus quelques mots dont je me disais: pour ceux-là, cela doit être plat; ils doivent se répartir d'une façon absolument banale, étant inertes aux thèmes comme au style. J'y avais pris le mot *fois* (pas la *foi* religieuse, mais les *fois* des multiplications); et je m'étais dit qu'à moins que le texte se mette à faire de l'arithmétique, ce qui n'est pas le cas, on ne voyait aucune raison pour des irrégularités. En fait, j'ai trouvé pour ses 182 occurrences une distribution très proche de celle de la théorie; mais il y avait quand même un certain excédent dans les tranches ayant plus de 3 occurrences. Alors je suis allé voir les passages en question, et qu'est-ce qui s'était passé? Tout simplement que plusieurs phrases de suite commençaient par « Combien de fois... »; autrement dit que l'accident lexical était dû à un phénomène rhétorique, auquel le mot *fois* n'est pas plus voué que beaucoup d'autres, mais qui,

dans ce cas, lui faisait tenir sa place dans un fait de style[1].

Pour conclure, je dirai ceci: à partir du moment où nous quittons le recueil des données numériques (fréquences, intervalles, etc.), puis la fabrication des tests ou des indices comme ceux dont je vous ai proposé des échantillons, à partir du moment où on aborde l'interprétation on quitte un terrain solide; on quitte les chiffres pour le langage des mots, on est sur un terrain où la subjectivité, l'intuition et l'approximation règnent en maîtres, et je crois que c'est un peu inquiétant. C'est cette inquiétude que j'ai voulu vous exprimer et que je livre à votre discussion. Je vous remercie de votre patience.

[1] "Heureux amans, dignes époux!... Combien de *fois*, contemplant en eux mon ouvrage, je me sens saisi d'un ravissement qui fait palpiter mon coeur! Combien de *fois* je joins leurs mains dans les miennes en bénissant la providence et en poussant d'ardens soupirs!..." (p. 867). "Montagne dit qu'il demandoit un jour un Seigneur de Langey combien de *fois* dans ses négociations d'Allemagne il s'étoit enivré pour le service du Roi. Je demanderois volontiers au Gouverneur de certain jeune homme combien de *fois* il est entré dans un mauvais lieu pour le service de son élève. Combien de *fois*? Je me trompe..." (p. 664).

	COEUR									RAISON				
I	0,0888	28,4	10	— 18,4	11,92					24,6	14	—10,6	4,57
II	0,1970	63,0	23	— 40,0	25,40					54,6	57	+ 2,4	0,11
III	0,1034	33,1	11	— 22,1	14,76					28,6	18	—10,6	3,93
IV	0,3250	104,0	158	+ 54,0	28,04					90,0	119	+29,0	9,34
V	0,2858	91,5	118	+ 26,5	7,67					79,2	69	—10,7	1,31
	1,0000	320,0	320	0,0	87,79					277,0	277	0,0	19,26
0	375,5	398	+22,5		1,35		95,5	124	+28,5	8,51				
1	191,9	158	—33,9		5,99		105,7	80	—25,7	6,24				
						I-						402,2	425	+22,8	1,29
												178,0	145	—33,0	6,12
2 (T)	49,0	53	+ 4,0		0,33	III	58,5	38	—20,5	7,24	(T)	39,4	37	— 2,4	0,15
3	8,4	12	+ 3,6	}	5,70		21,6	30	+ 8,4	3,27		5,8	10	+ 4,2	}
4	1,2	5	+ 3,8	}			6,0	11	+ 5,0	}	11,23	0,6	7	+ 6,4	} 24,80
5							1,7	6	+ 4,3	}		0,0	2	+ 2,0	}
	626,0	626	0,0		13,37		289,0	289	0,0	36,49		626,0	626	0,0	32,36
	206,7	206	— 0,7		0,00		75,6	74	— 1,6	0,03		45,6	48	+ 2,4	0,13
I-	36,8	38	+ 1,2		0,04		29,7	33	+ 3,3	0,37	(I)	11,0	8	= 3,0	
III	3,3	3	— 0,3	}	0,07		5,8	4	— 1,8	}	0,43	1,3	1	— 0,3	} 0,46
	0,2	0	— 0,2	}			0,9	1	+ 0,1	}		0,1	0	— 0,1	}
	247,0	247	0,0		0,11		112,0	112	0,0	0,83		58,0	58	0,0	0,59
	93,2	95	+ 1,8		0,03		18,0	23	+ 5,0	1,39		80,2	92	—11,8	1,74
	72,5	72	— 0,5		0,00		29,9	26	— 3,9	0,51		36,2	20	—16,2	7,25
IV	28,2	25	— 3,2		0,36		24,9	19	— 5,9	1,40	II	8,2	8	— 0,2	
	7,3	8	+ 0,7				13,8	17	+ 3,2	0,74		1,2	4	+ 2,8	}
	1,4	3	+ 1,6	}	0,40		5,7	7	+ 1,4	}	0,30	0,1	1	+ 0,9	} 2,01
	0,4	0	— 0,4	}			2,7	3	+ 0,3	}		0,0	1	+ 1,0	}
	203,0	203	0,0		0,79		95,0	95	0,0	4,34		126,0	126	0,0	11,00
	90,0	97	+ 7,0		0,54		19,4	27	+ 7,6	2,98		47,4	45	— 2,4	0,13
	60,4	48	—12,4		2,55		28,0	21	— 7,0	1,75		13,5	16	+ 2,5	}
V	20,2	25	+ 4,8		1,14		20,1	15	— 5,1	1,29	III	1,9	2	+ 0,1	} 0,33
	4,5	4	— 0,5				9,7	12	+ 2,3	0,55		0,2	0	— 0,2	}

	4,5	4	— 0,5			9,7	12	+ 2,3	0,55
	0,8	2	+ 1,2	0,07		3,3	4	+ 0,5	1,00
	0,1	0	— 0,1			1,3	3	+ 1,7	
	176,0	176	0,0	4,30		82,0	82	0,0	7,57
	331,4	352	+20,6	1,38		1079,9	1079	— 0,9	0,00
	183,8	150	—33,8	6,22		42,3	44	+ 1,7	
T	51,0	61	+10,0	1,96	I-				
	9,4	8	— 1,4	0,95	III	0,8	0	— 0,8	
	1,4	6	+ 4,6			1123,0	1123	0,0	0,02
	577,0	577	0,0	10,41					
	2582,1	2583	+ 0,9	0,00		1507,5	1505	— 2,5	0,00
	286,4	284	— 2,4	0,02		236,0	240	+ 4,0	0,07
T	15,9	18	+ 2,1		IV	18,5	18	— 0,5	
				0,38	-V	1,0	0	— 1,0	
	0,6	0	— 0,6						
	2885,0	2885	0,0	0,40		1763,0	1763	0,0	0,19

	0,2	0	— 0,2	
	63,0	63	0,0	0,46
	113,0	118	+ 5,0	0,22
	66,2	59	— 7,2	0,78
	19,4	16	— 3,4	0,60
IV	3,8	5	+ 1,2	
	0,6	4	+ 3,4	6,34
	0,2	1	+ 0,8	
	203,0	203	0,0	7,94
	118,9	122	+ 3,1	0,08
	46,6	42	— 4,6	0,45
	9,1	10	+ 0,9	
V	1,2	1	— 0,2	0,21
	0,2	1	+ 0,8	
	176,0	176	0,0	0,74

$$\bar{x} = \frac{T\text{-}f}{f+1} \qquad \text{Var.} = \frac{\sum x_i^2}{f+1} - \bar{x}^2$$

$$\mu_2 = \frac{(T\text{-}f)\ (T\text{-}f\text{-}1)}{f\ (f+1)}$$

$$\sigma = \sqrt{\frac{\mu_2 T(T+1)\ (f-1)}{f\ (f+1)\ (f+2)\ (f+3)}}$$

$$\eta = \frac{\delta^2}{d^2} = \frac{\Sigma(x_{i+1} - x_i)^2}{\Sigma(x_i - \bar{x}^2)}$$

$$\sigma_\eta = 2\sqrt{\frac{f-2}{f\ (f-1)\ (f+1)}}$$

	(1)	(2)	(3)	(4)
COEUR	87,8	13,4	3,69	27,3
ESPRIT	32,4	26,9 *	2,06	30,5
RAISON	19,3	32,4	3,80	13,8

* par livre: I: 0,00 II: 2,77 III: 1,22 IV: 3,10 V: 22,55

En regard des tableaux numériques, T signifie « Texte » (en totalité); les chiffres romains renvoient aux livres de l'*Emile*.