La répartition lexicale, problèmes et solutions

par

Charles MULLER

Université de Strasbourg - FRANCE

Dans un temps limité, je ne présenterai qu'un seul exemple : un seul vocable, dans un seul texte. Il va de soi qu'en pratique, on aura toujours à aller au-delà, c'est-à-dire à traiter d'un vocable (ou d'un groupe de vocables) dans plusieurs textes (corpus . . .), ou d'envisager, si l'on se borne à une seule oeuvre, la totalité ou la quasi-totalité de son vocabulaire. J'indique d'autre part que les études de répartition ne sont pas limitées au lexique : comme on étudie la façon dont se répartissent, dans le texte, les occurrences d'une unité lexicale, on peut appliquer les mêmes procédés à tout autre fait de langage : syntaxique, morphologique, rhétorique, etc . . .

En matière de lexique, on parle beaucoup de *fréquence*, relativement peu de *répartition*; et pourtant ces deux notions forment un couple à l'intérieur duquel devrait régner un certain équilibre. Mais s'il est aisé d'exprimer une fréquence, ce qui se fait par un simple nombre, il est moins commode de mesurer et d'exprimer l'une des façons dont *f* occurrences peuvent être localisées parmi N mots (la suite linéaire du texte), en nombre généralement si grand qu'il ne peut s'écrire que sous forme de logarithme.

On a commencé à parler de répartition à l'époque où les dénombrements lexicaux visaient à l'élaboration de vocabulaires fondamentaux; on sentait que la fréquence, si elle servait à un premier classement des vocables par degré d'utilité, devait être corrigée de façon à avantager ceux qui étaient "bien répartis", c'est-à-dire présents dans la plupart des parties du corpus, au détriment de ceux qui n'apparaissaient que dans peu de ces parties; d'où des procédés assez rudimentaires de pondération de F par R; notons sans y insister qu'A. Juilland, dans ses dictionnaires de fréquence, emploie une méthode plus élaborée, fondée sur le calcul des probabilités, et rendue applicable par l'appel à l'ordinateur.

Le but actuel est plutôt de demander à des analyses de répartition, un moyen d'analyser la structure lexicale des textes, de décrire, avec plus de rigueur, la façon dont l'auteur a exploité le lexique et a organisé son oeuvre.

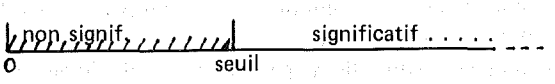
Le choix d'une méthode dépend d'abord des moyens. Au minimum, on aura le texte lui-même; ainsi j'ai sous les yeux une édition (Pléiade) de l'*Emile* de J.-J. Rousseau, un texte de quelque 285.000 mots réparti sur 626 pages : je pourrais y souligner toutes les occurrences d'une forme (je prendrai la forme *esprit*, non sans signaler que ce n'est que l'une de celles que j'ai étudiées; v. les Actes, à paraître prochainement, du colloque de Rome organisé en janvier dernier par Tullio Gregory). Mais on peut aussi disposer d'un index, ce qui est le cas avec celui qui a été établi par Etienne Brunet (éd. Slatkine, 1980, 585 pp.) et où les 265 occurrences de la forme *esprit* sont listées (p. 542-543) avec, pour chacune, le numéro de page où elle se situe (je ne parlerai pas ici de l'usage d'une concordance, qui permettrait de faire des tris sémantiques; nous en resterons au stade de la forme, qu'une vraie recherche aurait évidemment intérêt à dépasser; d'autre part, il conviendrait d'intégrer la forme plurielle *esprits*, car une lemmatisation bien comprise n'a aucune raison de l'écarter avant un tri sémantique; mais nous traitons ici d'une expérience de laboratoire où la simplicité est de rigueur). On peut enfin, si les circonstances sont favorables, disposer du texte enregistré sur support informatique, grâce auquel cet index a été produit, et interroger cet enregistrement pour en obtenir, sous

formes de listages, des données plus complètes. Ainsi, l'obligeance d'E. Brunet, un linguiste qui a parfaitement apprivoisé l'ordinateur, m'a procuré un listage des formes *esprit* dans l'*Emile*, où chacune des 265 occurrences s'inscrit avec le numéro d'ordre qu'elle occupe dans le texte, l'écart ou intervalle entre ce numéro et celui de l'occurrence précédente de la même forme, et la différence entre cet intervalle et l'intervalle précédent. Voici un extrait de ces données (il correspond au début du livre V, de la p. 693 à la p. 720) :

. . . . .	206663	998	-1077
	207892	1229	231
	210631	2739	1510
	210663	32	-2707
	212649	1986	1954
	215281	2632	646
	215849	568	-2064
	215861	12	- 556
	215885	24	12
	218388	2505	2479
	218472	84	-2419
	218513	41	- 43
	219367	854	813
	219672	305	- 549 . . . . .

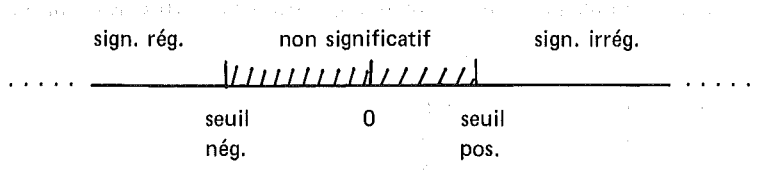
(N.B. : les signes de ponctuation sont comptés comme des "mots").

Muni de données plus ou moins abondantes, on peut obtenir des résultats plus ou moins "riches". On va penser que je cultive le paradoxe si j'avance que les résultats les plus riches ne correspondent pas nécessairement aux moyens les plus puissants (et les plus coûteux à tous points de vue). Je considère comme un résultat "pauvre" celui qui, par un indice numérique, me dit seulement que la forme soumise à l'examen dépasse ou ne dépasse pas un certain seuil. Ainsi, un test comme le X^2 définit un espace linéaire où ce qui sépare l'origine (le zéro) du seuil choisi est traité comme "non significatif", et attribué au hasard, alors que ce qui dépasse ce seuil dénote une répartition "irrégulière", non aléatoire, etc. . .



On peut certes tenir compte de la distance plus ou moins grande au delà du seuil, et même lui donner une appréciation probabiliste, mais en général, on se contente de créer un lot des vocables "irréguliers" et un lot de ceux qui se répartissent aléatoirement dans l'étendue du texte, et de ce fait ne retiennent pas l'attention. Il en est de même avec d'autres tests qui déterminent une zone de l'aléatoire, avec de part et d'autre deux espaces, l'un étant celui de l'irrégulier, l'autre celui des répartitions anormalement

régulières (rarement observées), qui supposent un retour périodique de la forme considérée :



C'est en exploitant les données sur les intervalles entre occurrences, que l'on peut situer une répartition observée dans l'aléatoire ou l'irrégulier. Je renvoie pour cela à deux types de recherches : d'une part celle du groupe de lexicologie politique de Saint-Cloud, et principalement de Pierre Lafon (revue "MOTS"), qui opère sur les intervalles; d'autre part, celle d'E. Brunet (dans sa thèse sur Giraudoux, éd. Slatkine 1978), qui applique les intervalles et leurs écarts à la ponctuation, afin de déterminer si les intervalles, donc les longueurs de phrases ou de membres de phrases, obéissent à des phénomènes de périodicité.

A moins de traiter de textes très courts, où la lecture elle-même suffit souvent pour déceler les répartitions irrégulières, ces méthodes ont l'inconvénient d'exiger absolument l'emploi de l'ordinateur, non seulement pour le recueil des données (mesure des intervalles), mais aussi pour le calcul lui-même; et quand le texte est long, les nombres auxquels fait appel l'application de la loi hypergéométrique dépassent parfois la capacité de l'ordinateur (v. à ce sujet les conclusions d'E. Brunet dans sa communication au colloque de sept. 1980 à Saint-Cloud).

La méthode que je vous présente ici ne nécessite d'autres moyens que l'index dont j'ai parlé, et l'emploi d'une calculatrice de poche.

Je distingue, dans une terminologie que je tente de rendre cohérente, deux sortes de "parties" dans un texte. D'une part, celles qui sont dues à l'auteur, et que je nomme "fragments" : ainsi les 5 livres de l'*Emile*, ou les chapitres d'un roman, d'un traité, les actes d'une pièce de théâtre; voire les rôles de cette pièce; car un "fragment" peut être obtenu en réunissant des morceaux non consécutifs; ce sont des parties organiques de l'oeuvre. D'autre part, des "tranches" obtenues par une division artificielle et arbitraire du texte en séquences d'égale longueur : ainsi les pages d'une édition, si leur typographie et leur organisation permettent d'admettre une égalité satisfaisante; ou bien des tranches de *n* mots ou de *n* lignes, l'idéal est évidemment de prendre le mot comme base du calcul, mais cela n'est généralement possible que si l'on opère sur un listage où les mots sont numérotés, ou si l'ordinateur crée ces tranches.

Dans le cas présent, je me contenterai des chiffres des pages tels que les fournit l'index. Ainsi, pour les 14 premières occurrences du livre V, citées plus haut, je lis ceci :

. . . 693 696 702 702 706 711 713 713 713 718 718 718 720 720 . . .

Il n'est pas difficile de voir que, dans cette suite de 28 pages, il y en a 2 qui comptent 3 occurrences d'*esprit*, 2 qui en comptent 2 et 4 qui n'en ont qu'une; d'où, par soustraction, 20 sans occurrence :

	p.	occ.
0	20	0
1	4	4
2	2	4
3	2	6
	28	14

Le recueil des données, en prenant l'index comme témoin et la page comme unité, est donc d'une grande simplicité.

Mais l'index me dit aussi que les 265 occurrences d'*esprit* sont distribuées entre les 5 livres de la façon suivante : 10, 48, 45, 109, 53; ce qui n'a de sens que si je tiens compte de la longueur relative de chaque livre. Je le ferai soit en me fondant sur le nombre de pages de chacun des livres, soit, si j'en ai les moyens, sur le nombre de mots de chacun (donnée qui est fournie par l'index, p. 564). Avec une précision plus ou moins grande, j'obtiens une répartition calculée (ou "théorique", terme impropre, mais confirmé par l'usage); et un test de X^2 nous permettra de constater que la répartition est très anormale entre ces 5 livres; en même temps, on obtient des écarts réduits qui indiquent que l'écart le plus fort est l'excédent du livre III; puis, presque à égalité, le déficit du livre I, celui du dernier, et l'excédent du livre IV :

livres	effectif calc.	effectif observé	écart	X^2	écart réduit
I	23,5	10	-13,5	7,76	-2,78
II	52,2	48	- 4,2	0,34	-0,58
III	27,4	45	+17,6	11,31	+3,38
IV	86,1	109	+22,9	6,09	+2,47
V	75,8	53	-22,8	6,86	-2,62
	265,0	265	0,0	32,36	

Passons à la répartition entre tranches, c'est-à-dire dans toute l'étendue de l'oeuvre considérée comme une suite continue et homogène (a priori, cela étant une hypothèse nulle, dont une répartition irrégulière exigera le rejet). En prenant la page comme unité, on observe ceci : 1 page avec 8 occurrences (la p. 481); 2 avec 6 (les pp. 486 et 552); 1 avec 5, 1 avec 4, 10 avec 3, 42 avec 2; ce qui donne un total de 143 occurrences; il en reste 122, donc 122 pages à 1 occurrence, cela donne 179 pages; il en reste donc 626 - 179 = 447 sans occurrence.

Ces données n'ont d'intérêt que si on les confronte à un modèle de répartition aléatoire : c'est là que nous passons *des* statistiques à *la* statistique; mais d'avance nous pouvons raisonner ainsi : si la forme étudiée est bien répartie, comme au hasard, la réalité sera proche du modèle; si la forme se signale par une régularité non aléatoire, un retour presque régulier, à intervalles égaux ou presque, il y aura un excédent de pages à 1 occurrence, ou même rien que des pages à 1 ou 0 occurrences; si la répartition est très irrégulière (et le test sur les livres ou fragments nous le fait prévoir), il y aura des excédents sur les pages sans *esprit* et dans celles qui en ont 2 ou plus.

La loi de Poisson nous ayant fourni le modèle (dont l'élaboration sur une simple calculatrice se fait en une ou deux minutes), voici le résultat :

nbre d'occ.	nbre de pages calc.	nbre de pages obs.	écart	éc. réd.
0	409,95	447	+37,05	+1,83
1	173,54	122	-51,54	-3,91
2	36,73	42	+ 5,27	+0,87
3	5,18	10		
4	0,55	1	+ 9,22	+3,84
⩾ 5	0,05	4		
	626,00	626	0,00	

et un X^2 donne un résultat supérieur à 34, ce qui exclut la répartition aléatoire, en confirmant la prévision : très fort déficit dans les pages à fréquence 1, excédents partout ailleurs, et surtout dans les fréquences supérieures à 2; la forme étudiée a donc tendance non seulement à se localiser dans les livres 3 et 4 (test précédent), mais à s'agglomérer dans l'espace d'une même page, c'est-à-dire dans des séquences de moins de 500 mots.

Poussons un peu plus loin en appliquant le même procédé à chaque livre :

	I		II		III		IV		V		T
0	48,8	49	86,1	90	30,8	37	118,7	126	130,2	145	447
1	8,4	8	32,8	26	22,0	18	63,7	53	39,2	17	122
2	0,8	1	6,2	8	7,9	5	17,1	19	5,9	9	42
3			0,9	2	1,9	1	3,1	3	0,6	4	10
⩾ 4					0,4	2	0,4	2	0,1	1	5
	58,0	58	126,0	126	63,0	63	203,0	203	176,0	176	626
X^2	0,00		2,77		2,45		3,10		22,55		
d.d.l.	1		2		2		3		2		

A l'intérieur de chaque livre, les répartitions sont inégalement régulières. La plus normale est observée dans le livre I; la plus irrégulière, et de très loin, dans le dernier; les trois autres livres sont assez proches. Ce qui est intéressant, c'est que ce classement ne correspond pas du tout aux excédents et aux déficits décelés par l'examen des fréquences dans les cinq livres.

Livre	Fréquence	Répartition
I	faible	régulière
II	moyenne	légèrement irrégulière
III	très forte	légèrement irrégulière
IV	forte	légèrement irrégulière
V	faible	très irrégulière

On pourrait ensuite faire varier la longueur des tranches pour serrer de plus près le phénomène de groupement des occurrences dans certaines parties de l'oeuvre; il est facile, par lecture de l'index, de créer des tranches de 2, 3 . . . pages, ou même des unités plus courtes : en effet l'index d'E. Brunet subdivise les pages en 7 zones de 70 mots environ, désignées par les lettres *a, b, . . . g*, qui serviraient de base à une localisation plus précise des formes.

On peut s'interroger sur la validité du test quand on observe que certaines pages sont incomplètes (débuts ou fins de chapitres, présence de notes hors texte . . .), et qu'il n'en a pas été tenu compte ici. J'ai donc tenté une expérience de contrôle, en partant non plus de l'édition ou de l'index, mais de listages où la simple lecture du numéro d'ordre (v. plus haut) permet de situer chaque forme dans des tranches de *n* mots. Pour me rapprocher autant que possible de l'unité "page" (460 mots environ), j'ai pris des tranches de 500 mots; si l'on prend comme exemple l'extrait reproduit plus haut (14 occurrences), il est aisé de déterminer qu'une tranche contient 3 occurrences (entre 215500 et 215999 : 215849, 215861 et 215885), qu'il y en a deux avec 2 occurrences (210631, 210663; - 218388, 218472), et 7 occurrences isolées; de 206500 à 220000, il y a 27 tranches, donc 27 - (1 + 2 + 7) = 17 sans occurrence. On obtient ainsi ce qui suit

	I		II		III		IV		V		T	
0	41,9	41	74,8	76	28,3	33	105,3	109	119,7	128	365,4	387
1	8,2	10	31,5	30	21,3	18	61,0	55	38,4	25	167,5	138
2	0,9	0	6,6	6	8,0	6	17,7	20	6,2	9	38,4	41
3			0,9	1	2,0	1	3,4	3	0,7	2	5,9	8
⩾ 4			0,2	1	0,4	2	0,6	1	0,1	1	0,8	4
	51,0	51	114,0	114	60,0	60	188,0	188	165,1	165	578,0	578
X^2	0,11		0,11		1,06		1,02		8,83		10,85	
d.d.l.	1		2		2		2		2		3	

Les écarts sont moins forts, sans doute en raison d'une meilleure égalité entre les tranches; il serait donc préférable, si l'on prend les pages comme unité, de tenir compte des irrégularités. Mais la tendance générale confirme le résultat précédent, et seule la répartition à l'intérieur du livre V se révèle comme très irrégulière.

La forme *esprit* est donc répartie de façon irrégulière entre les 5 livres de l'*Emile*, et, si je reprends un terme proposé par D. Dugast, elle ne fait pas partie de la "trame" de cette oeuvre; mais elle appartient évidemment au vocabulaire caractéristique des livres III et IV (écarts réduits supérieurs à 2); et les écarts négatifs permettent de l'inscrire dans le vocabulaire caractéristique négatif des livres I et V. A l'intérieur des livres, elle a une légère tendance à des groupements de 2 ou 3 occurrences; à partir du livre III, on trouve des groupements plus nombreux, allant jusqu'à 6 ou même 8 occurrences par page; mais la répartition ne s'écarte significativement d'une répartition aléatoire que dans le livre V.

Ces conclusions ne prendraient leur sens complet que si elles étaient mises en rapport avec les résultats semblables obtenus sur d'autres vocables; tel n'était pas ici notre but.

La méthode illustrée par cet exemple est progressive : elle considère d'abord la façon dont une unité de discours se répartit entre les sous-ensembles organiques de l'oeuvre, ensuite dans la suite continue qui constitue le texte, artificiellement divisée en tranches égales; cela d'abord par rapport à l'ensemble de l'oeuvre, ensuite en prenant chacune des parties comme une unité indépendante des autres. L'avantage est d'abord dans la simplicité des moyens mis en action, ensuite dans le contrôle qui est conservé sur la recherche et ses résultats au cours des opérations. Enfin les conclusions peuvent être plus nuancées que si elles s'appuient sur un indice unique.

5.- omissions :

- 1 or 2 words	10
- 3 to 10 words	14
- 11 to 30 words	20
- more than 30 words	30
omission of complete paragraphs :	
- each of the first 5 paragraphs	20
- each further paragraph	10

6.- words in brackets or on the margin :
- half the weight of a corresponding omission.

This weighting system is only a first attempt to lead the qualitative weighting of conventional methods to a higher degree of comparability and precision. Low weights are associated with readings which can easily be corrected or which are likely to arise independently of one another. High weights are associated with strong indicators of genealogical coherence. For the following stemma constructions, three different ways of weighting have been used :

1.- original weights according to the above list,

2.- cubes of the original weights,

3.- standard weight 5 for each reading.

Cubes enlarge differences of weights, standard weights make them disappear.

Special attention must be drawn to the splitting of weights, as recommended for text words in brackets or on the margin; these may or may not be taken over by a later copyist, so that his model text should be treated as both defective and complete. More general, if a single ms. presents two versions at the same variant place, it is a member of both of the corresponding constellations, but not with the full weights of the versions. Moreover, one version may be more likely to be copied than the other. Let us assume, for example, that at a certain variant place reading A - with weight 9 - is found in the text of mss. 1, 2, 3, that reading B - with weight 3 - is found on the margin of ms. 3, and that only reading B is found in the text of mss. 4 and 5. If in ms. 3 reading A is double as likely to be copied as reading B, the following scheme of constellations and weights seems to be appropriate :

reading	weight	constellation	splitted weight
A	9	{1,2}	3
		{1,2,3}	6
B	3	{3,4,5}	1
		{4,5}	2