

FRÉQUENCES ET PROBABILITÉS

GEORGE TH. GUILBAUD

La notion de *fréquence* est bien installée dans les sciences du langage, tout spécialement dans le domaine lexicologique, et depuis longtemps déjà. On peut se demander: pourquoi? et: pour quoi?

D'abord une acception, la plus ancienne sans doute, mais encore bien vivante: fréquence est une « qualité »; cela semble tout à fait naturel, et même banal: « ... chez Cicéron, *cives nostri* (nos concitoyens) n'est pas rare » (E. BENVENISTE, *Prob. Ling. Gén.* 2, p. 275, 1974); « ... *Criticus* est un mot rare au moyen âge... » (P. MICHAUD-QUANTIN, *Etudes sur le vocab. philos. au M.A.*, Roma, 1970, p. 213). Jules Richepin, en 1880, admirait le style de J.K. Huysmans à cause de « ses substantifs rares, épithètes curieuses, alliances de mots imprévues... » (d'après R. BALDICK, *Huysmans*, 1958, p. 70).

Je laisserai de côté une question difficile: la fréquence (ou la rareté) est-elle consciente chez l'écrivain? est-elle perçue par les lecteurs? Pour la perception, il ne serait pas étonnant d'y trouver quelque lien avec la périodicité des répétitions (entre autres: les fameuses « rafales » tant controversées).

Je me limiterai aux *fréquences* définies seulement par comptage soigné du nombre des occurrences.

Toute fréquence est une construction.

Un index du *Discours de la Méthode* a été établi par P.-A. Cahné et publié par les soins du *Lessico Intellettuale Europeo* (Roma, Ateneo, 1977). Grâce à ce précieux instrument, je peux repérer tous les endroits où Descartes emploie le mot *penser*. Je trouve 83 occurrences. Qu'est-ce à dire? J'ai suivi le conseil souvent donné (et ici même, il n'y a guère): « il convient de *lemmatiser* ». J'ai lemmatisé à fond, au risque d'exagérer: en rassemblant, non seulement toutes les flexions du verbe *penser*, mais aussi le participe devenu substantif: *la pensée*, sans oublier le pluriel: *les pensées*.

Premier point à noter: la détermination numérique d'une fréquence suppose un certain nombre de décisions qu'il faut bien dire arbitraires, ce qui ne signifie pas immotivées, ni fantaisistes.

Toute fréquence est une construction.

Cette élaboration est parfois oubliée; à cause des apparences « rigoureuses » des chiffres. Méfions-nous!

J'aimerais aussi qu'on se débarrasse de l'illusion, persistante, d'une façon de construire l'index des mots qui soit, parmi toutes celles qu'on peut imaginer, la meilleure de toutes. L'avènement des machines a privilégié une définition matérielle du mot indexable (chaîne de caractères isolée par des blancs) et une indexation automatique qui rassemble toutes les formes matériellement identiques. On y ajoute quelques consignes pour distinguer de trop grossières homographies (comme, en français d'aujourd'hui: *car*, *avions*, *tiens*, *été*, *fils*, *savons*, etc. ou en latin: *eam*, *suis*, *est*, ...). Tout cela est très bien. L'évolution de l'art typographique ayant eu l'influence que l'on sait pour unifier l'orthographe, on n'a pas beaucoup à se préoccuper du cas où il faudrait rassembler des formes différentes au lieu de séparer des formes semblables (cependant, en français: *clé*, *clef*). Mais la question se pose différemment quand on dépouille des documents médiévaux (latins ou romans).

Il n'y aura *fréquence* que si l'on a répondu aux deux questions: ce fragment de texte est-il un mot? ce fragment et cet autre sont-ils le « même » mot?

L'éclatement de la fréquence.

Revenons à notre lemme *penser* du *Discours de la Méthode*.
Fréquence = 83, disions-nous. C'est un fait.

Bien entendu ce chiffre est destiné à être introduit, ultérieurement, dans un système de calculs (plus ou moins traditionnels, plus ou moins savants). Ne croyez-vous pas qu'il soit opportun, avant de ce faire, d'examiner, même brièvement, la réalité en quoi consiste ce « fait »?

Il s'agit de 83 morceaux de texte: on les rassemblera, en manière de « concordance », c'est-à-dire non seulement l'occurrence elle-même, mais un petit contexte.

On voit alors apparaître, non seulement les ressemblances (ce sont bien des occurrences du lemme) mais aussi les nuances, et, mieux encore, se dessinent des « constellations » possibles: telle occurrence est « proche » de telle autre et plus « éloignée » d'une troisième.

Une sorte d'*éclatement de la fréquence*, au premier abord; en second lieu se proposent diverses structurations possibles.

De véritables répétitions; je n'en trouve que trois¹:

- le célèbre « je pense donc je suis » (32.19 et 33.17);
- « dont la nature n'est que de penser » (33.05 et 46.18);
- « je pensay qu'il falloit que je taschasse » (22.02 et 30.30).

Des similitudes:

- conduire mes pensées (18.27, 29. 23);
- conduire leurs pensées (15.20);
- conduire nos pensées (02.11).

¹ Références, comme dans l'Index, à l'édition Adam-Tannery.

Des oppositions:

— penser que (5 fois), penser à (2 fois), penser quelque chose (1 fois), penser (absolument, 3 fois).

On pourrait aussi tenter des classements:

— je pense, ma pensée (44 fois), mes pensées (8 fois);

— il pense, sa pensée (2 fois), ils pensent, leurs pensées (8 fois);

— on pense (3 fois);

— penser, la pensée, les pensées (10 fois).

Mais il apparaîtrait bien vite qu'il ne peut y avoir une seule manière de classer.

Ici, derechef, je vais me dérober devant l'obstacle: il s'agit en effet d'une affaire considérable. C'est la *liberté combinatoire* du sujet parlant (ou écrivant) invoquée souvent, à la suite de F. de Saussure, pour fonder l'opposition: Syntagmatique/paradigmatique. On lit, par exemple, dans les *Sources manuscrites du Cours de Linguistique Générale* (pp. 89-90):

(...) il y a ici quelque chose de délicat (...) la frontière de la parole et de la langue est un certain degré de combinaison (...) dans les faits de langue, il y a des syntagmes; mais il y a aussi, probablement, toute une série de phrases qui appartiennent à la langue, que l'individu n'a plus à combiner lui-même (...)

Texte qu'on complètera par le commentaire de la p. 169 et par le C.L.G. lui-même (p. 172).

Lectures concordantielles.

Les suggestions que je viens de faire auraient pu être présentées comme une sorte d'exercice de « lecture concordantielle » (par opposition à la lecture séquentielle, qui est l'ordinaire). Ce n'est pas par le *Discours de la Méthode* que j'avais commencé ces exercices, mais par cet autre ouvrage, chez nous non moins classique, que nous appelons précisément *Pensées de Pascal*.

Qu'est-ce que *pensée* dans les « *Pensées* »? Question amusante. Mais les jeux de mots peuvent être fort sérieux. Quelques fragments pour se faire une idée de l'intérêt de la chose:

Jésus Christ a dit les choses grandes si simplement qu'il semble qu'il ne *les* a pas *pensées*, et si nettement néanmoins qu'on voit bien ce qu'il *en pensait* (309)

la mort est plus aisée à supporter sans *y penser* que la *pensée* de mort sans péril (...) (138)

une seule *pensée* nous occupe, nous ne pouvons *penser* deux choses à la fois (...) (523)

l'homme est visiblement fait pour *penser* (...) tout son devoir est de *penser* comme il faut (...) or, à quoi *pense* le monde? (...) à se faire roi sans *penser* à ce que c'est (...) (620) .

(Les numéros sont ceux de l'édition Lafuma).

Ce qui suffit sans doute à donner l'envie de rassembler tous les fragments où apparaissent les mots: *pensée*, *penser*, sous diverses formes. Aucune des nombreuses éditions ne possède d'index suffisant, il faut utiliser la Concordance de Davidson et Dubé (Cornell Univ. Press) qui me fournit 166 occurrences (sous seize formes). Il ne suffit pas de rassembler, il faut mettre de l'ordre: comme on pouvait prévoir, l'allure générale du système des occurrences est fort différente de ce qu'on a vu chez Descartes.

Ainsi le « je » de la première personne qui domine chez Descartes (52 fois sur 83) est presque absent chez Pascal (8 fois sur 166). Par contre les modalités négatives (« ne pas penser », « sans y penser », « empêcher de penser », etc.) sont, chez Pascal, fort abondantes (38 fois) et nuancées. Si l'on voulait analyser ces contrastes, il faudrait examiner de plus près la combinatoire. Elle n'est pas simple. Bien entendu, on commence par remarquer qu'un verbe, par exemple, peut être pris à la première ou à la troisième personne, au singulier ou au pluriel, au présent, passé, ou futur: cela fait déjà: $2 \times 2 \times 3 = 12$ possibilités. Mais il convient de continuer: négation, affirmation ou interrogation, avec

ou sans complément, quelles prépositions, et ainsi de suite. Alors des incompatibilités peuvent naître, et le calcul devient plus difficile. Ce genre de calcul des possibilités combinatoires de la langue, me fait souvenir de l'étrange obstination que manifestait Jacques Bernoulli dans son *Ars Conjectandi* (Basileae, 1713) pour énumérer les variations d'un hexamètre « *a Bernh. Bauhusio jesuita lovaniensi in laudem Virginis Deiparae constructo: Tot tibi sunt dotes...* ». (*Ars Conj.* pp. 78-81).

Ici nous nous contenterons d'estimations grossières: en prenant les seules déterminations présentes dans le corpus pascalien, j'arrive à presque deux mille constructions possibles; mais un peu plus d'une centaine seulement sont attestées.

Arrêtons ici ce survol. Résumons: pour pouvoir dire une fréquence il faut d'abord rassembler des fragments de texte. Non pas un simple amas d'occurrences, mais une structure, plus ou moins compliquée, liée aussi bien à celles de la langue qu'aux spécificités du discours. Et n'oublions pas que cette collecte ne se fait jamais toute seule: elle requiert des décisions, librement consenties (arbitraires, comme on dit souvent) pour séparer et réunir, pour organiser la lecture concordancielle. Enfin on compte, pour produire du nombre. Mais comment parler chiffres?

Un contraste de fréquence.

Revenons à notre exemple.

Chez Descartes, dans son *Discours*: 83 occurrences;
chez Pascal, dans les *Pensées*: 166.

Vous l'aviez déjà remarqué, je pense: 166 est juste le double de 83! Ce n'est qu'un *hasard*, direz-vous. Qu'est-ce à dire?

— que c'est insignifiant, qu'il n'y a rien à tirer de cette rencontre purement arithmétique, pas de commentaire linguistique ni littéraire.

Et pourtant il y a peut-être quelque chose à dire, non sur la proportion exacte du simple au double, mais sur l'ordre des gran-

deurs: il y a davantage de *pensée* dans les *Pensées* que dans le *Discours*. Mais si l'on s'aventure dans cette voie (on n'hésite guère, me semble-t-il) le « bon sens » s'écrie: Mais l'ouvrage intitulé *Pensées de Pascal* est beaucoup plus gros que le *Discours de la Méthode*!

On voudrait bien savoir mesurer le volume (on dit aussi: la longueur) d'un texte. Faut-il compter les mots, les lignes, les pages, les signes typographiques? La plupart des écrits que j'ai lus en la matière préfèrent compter les mots (j'y reviendrai). Selon l'index de P.-A. Cahné (page X, op. cit.), le texte du *Discours* contient 22 688 mots. La concordance de Davidson ne donne pas, me semble-t-il, le nombre analogue pour les *Pensées*; mais on peut tenter une estimation (à chaque mot correspond une ligne de concordance): l'on trouve entre 110 000 et 120 000: un peu moins de cinq fois plus.

Une autre comparaison: la collection de la Pléiade comporte un volume pour Descartes (1937) et un pour Pascal (1936); la typographie est la même. Le *Discours* occupe 55 pages, et les *Pensées* 273.

Ces estimations, toutes grossières qu'elles soient, nous seront suffisantes; on raisonnera ainsi: les *Pensées* comportent presque *cinq* fois plus de mots que le *Discours*; mais le lemme *penser* ne se rencontre que *deux* fois plus souvent. Personne n'hésitera: ce mot est « moins fréquent » chez Pascal.

Cette manière de parler semble tout à fait naturelle; mais il faut réfléchir un instant à ce qu'un mathématicien appellerait ici « les définitions ». On avait convenu que *fréquence* désignait le nombre des occurrences. De sorte que, pour *penser*, *pensée*, la fréquence (166) dans les *Pensées* de Pascal est *plus* grande que celle (83) dans le *Discours* de Descartes. Mais on ne dira pas cependant: *plus* fréquent. Au contraire, on décide de dire: *moins* fréquent.

Certains préconisent alors de distinguer deux emplois de *fréquent*, *fréquence* en parlant de « fréquences absolues » et de « fréquences relatives ». L'intention est excellente: il faut en effet avouer que la comparaison des fréquences ne peut se faire sur simple examen des chiffres bruts. Il faut une procédure de comparaison. Je préférerais d'ailleurs dire « fréquences comparées » plutôt que « fréquences relatives ». Mais suffit-il de faire des divisions, de calculer des pourcentages?

Les A-peu-près inévitables.

Notre sujet n'étant pas l'étude du Lexique philosophique du XVII^{ème} siècle français, mais seulement les problèmes de méthode, il nous sera permis de faire quelques pas vers la généralité: désormais il ne sera plus question que de deux textes, que j'appellerai PP et DD (ou, plus simplement, le premier, le deuxième) dont les longueurs sont supposées connues. Lorsque, pour fixer les idées, il sera utile de donner les ordres de grandeur, on prendra respectivement cent mille mots et vingt mille mots; sinon les deux nombres de mots seront désignés par T1 et T2 (T comme Texte ou T comme Total).

Une *même* unité lexicale (vocable, lemme ou lexème, etc.) est repérée dans les deux Textes: F1 fois dans l'un et F2 dans l'autre (pour un exemple numérique on pourra prendre, soit les observations précédentes: F1 = 166, F2 = 83, soit d'autres observations, ou même d'autres nombres choisis arbitrairement).

Il n'est pas défendu d'examiner le rapport (au sens de l'arithmétique) entre T1 et T2:

$$T1/T2 = 100\ 000/20\ 000 = 5$$

non plus que:

$$F1/F2 = 166/83 = 2$$

Suffit-il alors de confronter ces deux quotients? et de trancher

(« plus fréquent », « moins fréquent ») selon que l'un de ces rapports est inférieur ou supérieur à l'autre?

Mais peuvent-ils être égaux?

Nous nous sommes étonnés tout à l'heure en constatant que $F1/F2$ était rigoureusement égal à 2, et avons conclu que ce n'était qu'un hasard dépourvu de signification. De même, si l'enquête conduisait à la rigoureuse égalité de nos deux rapports $F1/F2$ et $T1/T2$, nous serions en droit de ne pas nous y appesantir.

Pour la particule « qu' », très fréquente en français, on trouve:

en PP : $F1 = 2\ 099$; en DD : $F2 = 406$,

et nous dirons que le rapport est « à peu près cinq ».

C'est inévitable: dès qu'on introduit, dans la procédure de comparaison, des opérations arithmétiques, il est nécessaire de prévoir des approximations. Mais des approximations contrôlées, sinon la fantaisie de chacun gouvernera le jugement.

Comment contrôler l'à-peu-près?

Fréquence moyenne?

On peut le dire autrement. La logique des proportions, fondement du calcul et de la mesure (depuis les origines, il y a vingt-cinq siècles), nous enseigne que pour comparer $T1/T2$ à $F1/F2$ on peut, aussi bien, comparer $T1/F1$ et $T2/F2$.

Reprenons nos exemples numériques.

Au lieu de dire:

PP est cinq fois plus long que DD,
et $F1$ est le double de $F2$,

on dira:

en PP l'unité lexicale étudiée se présente 166 fois pour environ 100 000 mots, c'est-à-dire (à peu près) 1 fois tous les 600 mots,

en DD: 83 fois pour 20 000 mots
1 fois tous les 240.

On a calculé deux quotients:

$$100\ 000/166 = 602,4\dots$$

$$20\ 000/83 = 240,96\dots$$

et arrondi les résultats.

Mais alors on voit mieux apparaître la modalité de l'approximation: 83 fois pour 20 000 mots, cela fait bien 1 fois pour 240; mais il n'est même pas besoin d'avoir sous les yeux le *Discours* de Descartes pour savoir que *penser*, *pensée* ne s'y trouve pas régulièrement, comme un refrain, tous les deux cent quarante mots.

On prendra bien quelque précaution modale, par exemple en disant:

« tous les 240 mots, *en moyenne* ».

Je constate (non sans regret) que, très souvent, dans la littérature lexicologique, le quotient est fait en sens inverse. Ainsi certains tiennent à ce qu'ils appellent la « fréquence relative »: F/T. Ce qui conduit à des écritures un peu rébarbatives:

$83/22\ 688 = 0,0036583216\dots$ (combien voulez-vous de décimales?)

Qu'on s'y prenne comme on voudra, le problème demeure. Bien sûr, entre:

une fois tous les 600

et: une fois tous les 240,

le contraste pourra être jugé considérable, significatif, etc. Mais, s'il faut établir des règles universelles, il faut savoir quand on commencera à dire que tel ou tel contraste est fort, ou faible, ou trop faible pour être retenu.

Peut-on, en matière de fréquences, instaurer une mesure de l'intensité des contrastes? les experts répondent que ce n'est pas difficile et préconisent, à cet effet, des algorithmes d'usage courant. Les étiquettes les plus répandues sont: « écart réduit », « khi-deux » — qui ne sont d'ailleurs que deux variantes d'une

même attitude théorique. Commençons par le commencement: quels sont les principes?

Echantillonnage.

Pour savoir si tel vocable est moins employé dans PP (cent mille mots) que dans DD (vingt mille mots), on pourrait commencer par comparer le deuxième texte à des fragments de même longueur (vingt mille mots) extraits du premier. Je ne sais si cette procédure a jamais tenté quelque philologue, mais elle ne semble pas déraisonnable².

Si l'on dispose d'une édition lisible en machine, le travail n'est pas trop difficile. Mais il nous suffira ici de l'imaginer.

On s'attend évidemment à ce que divers fragments de vingt mille mots extraits d'un texte de cent mille ne présentent pas tous la même fréquence du vocable soumis à l'enquête; ce qu'on peut vérifier aisément par quelques coups de sonde, même sans machine.

On est alors conduit à une nouvelle façon de comparer: la fréquence F2 (du vocable en DD) confrontée, non plus directement à F1 (fréquence en PP), mais à toute une population de fréquences f , dont chacune correspond à un fragment de même longueur que DD.

Reste le problème, évidemment central, du découpage. Quels fragments choisir? C'est un problème d'échantillonnage: la collection de fragments doit « représenter » le phénomène « fréquence », avec toute sa diversité.

On pourrait ici introduire un débat, long et difficile, sur le découpage des textes. Je ne m'y risquerai pas.

Je dois simplement indiquer la solution brutale, qui a la faveur des statisticiens: prendre *tous* les fragments possibles.

² Cf. cependant le *random partitioning* de G. U. YULE, *The statistical study of literary vocabulary*, Cambridge U. P. 1944 (voir l'Index, s.v.).

Je dis bien: *toutes* les façon de choisir vingt mille mots parmi cent mille. Y compris celles qui ne donnent aucun texte lisible, par exemple prendre un mot sur cinq, ou une ligne sur cinq, ou d'autres façons, encore plus fantaisistes.

Bien entendu, on avoue que c'est un pis-aller: on aimerait mieux, sans doute, ne pas démantibuler la syntaxe et prendre, non les mots, un par un, mais des phrases. Cela pourrait peut-être se faire, mais ce serait très coûteux.

En effet il faut bien comprendre que le nombre des fragments à considérer est extraordinairement élevé.

Une petite remarque: j'ai choisi un exemple numérique ($T_1 = 100\ 000$, $T_2 = 20\ 000$) qui peut induire en illusion: il ne s'agit nullement de découper le second texte en cinq morceaux égaux, mais de le découper d'abord en éléments: des mots, ou bien des phrases, ou des alinéas, ou des pages, etc. Puis de combiner ces éléments de diverses manières, pour atteindre le format désiré (ici vingt mille).

Pour revenir à l'exemple des *Pensées* de Pascal: on sait qu'il s'agit de l'édition posthume de papiers séparés, dont le classement reste encore conjectural. Il y a un millier de fragments: certains n'ont que quelques mots, d'autres plusieurs centaines. Si l'on veut combiner ces fragments pour avoir vingt mille mots, cela peut se faire, pensera-t-on, de bien des manières: vous n'y êtes pas! Si je voulais écrire le nombre de façons *a priori* possibles, il me faudrait écrire un nombre d'environ deux cents chiffres!

Menace des très grands nombres, bien connue de toute combinatoire: il nous faut le secours de la mathématique.

Distributions.

La première vertu du traitement mathématique est celle-ci: on peut dénombrer sans énumérer. Et la seconde: il n'y a pas de « grands » nombres, car la théorie est la même qu'il s'agisse de

sélectionner cent mots dans un texte de cinq mille, ou dix mille en cinq millions, ou cinq parmi vingt.

Les dénombrements combinatoires sont connus depuis longtemps; pour l'époque moderne, les textes fondateurs sont le *Traité du Triangle* (1654) de Pascal, et la *Pars Secunda de l'Arts Coniectandi* (1713) de Bernoulli.

Ainsi commence Pascal:

Combinationis nomen diverse a diversis usurpatur: dicam itaque quo sensu intellegam. Si exponatur multitudo quaevis rerum quarumlibet ex quibus liceat aliquam multitudinem assumere...

Un nombre *quelconque* de choses *quelconques*: on doit en prendre un *certain* nombre...; mais l'exemple qui suit risque de décevoir par sa banalité:

si ex quatuor rebus, per litteras A B C D expressis liceat duas quasvis ad libitum assumere ... experimento igitur patebit duas posse assumi inter quatuor sex modis ...

Il y a 6 manières de prendre 2 parmi 4.

On pourrait suivre cet illustre exemple: prendre ce petit texte déjà cité parmi d'autres:

la mort est plus aisée à supporter sans y *penser* que la *pensée* de la mort sans péril

Il comporte $T = 17$ mots, on y a souligné $F = 2$ occurrences. On décide d'en extraire $t = 14$ mots (pour comparer à cet autre texte cité au même endroit: « une seule *pensée* nous occupe, etc. »).

Il y a 680 manières de faire: certains extraits comportent les deux occurrences, ils sont au nombre de 455 (et représentent donc la majorité des possibles: 67 %), d'autres une seule occur-

rence, il y en a 210; enfin ceux qui restent, au nombre de 15, n'ont aucune occurrence du mot visé.

On voit ce qu'on est capable de calculer: pour un texte de longueur T où l'on a marqué F occurrences, et pour l'ensemble des fragments de longueur t , on décrit la *variété* des valeurs possibles pour f , nombre des occurrences dans chaque fragment.

Pour notre petit exemple, tout est donc résumé par le tableau (appelé *distribution*):

$T = 17$	$F = 2$	$f = 2$	67 %
distribution de f		$f = 1$	31 %
pour les fragments $t = 14$		$f = 0$	2 %

Je me garderai de vous importuner en rappelant ici le détail des calculs nécessaires: ce serait ennui pour ceux qui savent, obscurité pour les autres. Il suffit de dire qu'il s'agit d'opérations toutes simples: multiplications et divisions, en chaîne, dont les termes sont des entiers consécutifs³.

Notons sans pouvoir nous y arrêter qu'on aurait pu poser un autre problème: se donner f et étudier la variabilité de t . En langage ordinaire: on prélèverait des mots un à un jusqu'à obtenir une fréquence donnée (f), le nombre des mots à prélever (noté t) est variable; il ne s'agit pas de réaliser toutes les façons d'ainsi faire, mais d'en imaginer la diversité, et de calculer la distribution des valeurs de t . L'algorithme, comme on peut s'en douter, est fort peu différent.

Une lecture probabilitaire.

Si vous trouvez vraiment trop petit mon exemple numérique, je peux bien vous fournir un cas un peu plus étoffé, sans être

³ Algorithmes que depuis Wallis (1655) on a pris l'habitude d'étiqueter du nom, aujourd'hui suranné: *hypergéométriques*.

gigantesque: $T = 90$, $F = 9$, c'est-à-dire un texte de 90 mots, dont 9 sont marqués; et dressons le catalogue des extraits comportant $t = 25$ mots.

Ne me demandez pas le nombre des possibles⁴; on doit se contenter des proportions, les extraits étant classés selon le nombre (f) d'éléments marqués qu'ils contiennent:

$f = 0$	4,5 %
$f = 1$	17,9
$f = 2$	29,6
$f = 3$	26,9
$f = 4$	14,8
$f = 5$	5,1
$f = 6$	1,1
$f = 7$	0,1
$f = 8$ ou 9	(presque rien)

Comment utiliser pareille distribution? On a pris, depuis longtemps déjà, l'habitude d'employer le langage des « probabilités ». On dira, par exemple: si l'on prélève $t = 25$ mots, il y a « très peu de chances » qu'on y trouve 8 ou 9 mots marqués. Ou bien: il y a 9 chances sur 10 (au vrai: 89,2 pour cent) pour que le nombre des mots marqués soit compris entre 1 et 4. On oublie parfois la condition qui autorise cette façon de parler: estimer que tous les extraits sont également probables. Quand on dit: en prenant 25 mots *au hasard*, on a 9 chances sur 10 d'obtenir entre 1 et 4 éléments marqués — il faut comprendre que sous cette locution *au hasard* se dissimule l'hypothèse: équiprobabilité pour tous les échantillons possibles.

La coutume étant bien implantée, il serait assez vain de vouloir la contrarier, et de s'obstiner à dire *proportion* (de l'ensem-

⁴ C'est 116.10^{20} , « à peu de chose près »!

ble des possibles) là où tout le monde dit *probabilité*. Citons encore Pascal:

j'aurai aussi mes pensée de derrière la tête (...);
 (...) il faut avoir une pensée de derrière et juger de tout par là, en parlant cependant comme le peuple.

Il nous faut revenir maintenant à notre propos: porter jugement sur un contraste de fréquence. Reprenons notre schéma: deux textes, PP et DD (le premier est le plus long). On peut imaginer la construction d'un tableau analogue au précédent (ou toute autre procédure équivalente); on y trouvera décrite avec plus ou moins de précision la collection des extraits du premier texte ayant même étendue que le second. Après quoi il ne reste plus qu'à situer le second texte au sein de cet ensemble. Parfois (et même souvent) on emploiera une variante de ce procédé.

Bien que mon projet ne soit pas ici technique, je dois dire un mot de cette variante. Imaginez que les deux textes à confronter fassent naturellement partie d'un même ensemble. Pour les *Pensées* de Pascal: un peu moins de la moitié des papiers ont été rassemblés en liasses, et, semble-t-il, par l'Auteur lui-même. Mais vous pouvez penser à toute autre œuvre littéraire qu'il vous plaira, divisée, ou divisible, en deux parties.

En pareil cas, on peut décider de comparer l'une des parties à l'autre; mais on préfère souvent comparer une partie à l'ensemble. Dans ce cas, au lieu de dresser le tableau de la distribution pour $T = T_1$, on le fera pour $T = T_1 + T_2$. C'est, le plus souvent, à cette variante (contraste entre la partie et le tout) que se réfèrent les usagers de la procédure dite « test du khi deux ».

En tout cas, il s'agit finalement de situer un texte *donné* au sein d'un ensemble construit (celui de tous les extraits *possibles*): situer le réalisé dans un univers des possibles. Si le réalisé se trouve parmi les plus rares (les plus « improbables »), on dira tout naturellement que le contraste est « significatif ». Sinon —

c'est-à-dire si les caractères du réalisé sont analogues à ceux de la « majorité » des possibles — on conviendra qu'il n'y a pas lieu de chercher plus avant et que le contraste est « négligeable » (du moins, quant aux fréquences).

Tout cela est un peu flou, direz-vous: où finit la rareté? où commence la majorité acceptable? A ces questions, bien naturelles, il faut avoir le courage de répondre: comme il vous plaira. Au risque de décevoir ceux qui imaginent que la statistique mathématique va porter, à leur place, un jugement sans appel.

Il faut le redire: recourir aux méthodes statistiques, dans quelque domaine scientifique que ce soit, n'implique pas qu'on renonce à la liberté.

Probabilités, mais de quoi?

Les distributions: j'en ai, tout à l'heure, donné deux exemples, mais extrêmement réduits, des exemples miniatures. Et j'entends qu'on me dit: il n'y a de vraie statistique que pour les « grands nombres ». Lieu commun, aussi dangereux que la plupart des lieux communs. Qu'est-ce qui change, quand les nombres qu'on traite deviennent « grands »? Pas la structure logique du modèle, ni même les formules, si elles sont exactes.

Imaginons le cas suivant: le texte de référence comporte cent mille mots, on y a marqué, disons, cent soixante occurrences (ce sont à peu près les dimensions du texte PP). On veut constituer l'ensemble des extraits de vingt mille mots, et y étudier la variété des fréquences. Cette fréquence peut varier entre zéro et 160: rien n'empêcherait de calculer, pour chaque valeur possible, de 0 à 160, la proportion (ou, si l'on préfère: la probabilité). Il ne faut pas croire que ces calculs seraient difficiles: les moyens de calcul sont aujourd'hui fort puissants et peu coûteux; il ne s'agit d'ailleurs ici que de longues suites de multiplications et divisions, le seul souci étant la gestion des arrondis.

Mais les informations ponctuelles, ou trop précises, ne sont

pas utiles. Que m'importe de savoir si la fréquence $f = 38$ a une probabilité de $p = 0,0382741\dots$? On préfère évidemment des informations qui soient à la fois sommaires et globales, comme, par exemple:

les grandes valeurs de la fréquence sont très peu probables: à peine une chance sur dix mille pour l'ensemble des fréquences de 53 à 160; du côté des basses fréquences, c'est la même chose, mais de 0 à 14;

ou bien encore, en un autre style, pour d'autres fins:

les fréquences les plus probables sont 31, 32, 33, 34, 35; leurs probabilités sont peu différentes, légèrement supérieures à 7 %; le reste fait environ: 24 % au-dessus de 35 et 39 % au-dessous de 31.

Résumer efficacement une distribution: sur ce thème, on a vu foisonner les règles « pratiques », formules, tables, abaqués, plus ou moins commodes. On ne doit pas trop espérer qu'il y en ait de bonnes à tous usages. Ainsi la routine la plus populaire: *moyenne, variance* et autres *moments*, est fort utile; mais elle ne donne que des approximations, souvent grossières. Il faut savoir, par exemple, que si un écart « réduit » ou un « khi-deux » est très considérable, il y a fort à parier que cela conduira à des estimations tout à fait fantaisistes (et non pas seulement très petites).

Mais n'entrons pas plus avant dans la technique; notons seulement un point: parler le langage des probabilités suppose une certaine habitude de la mesure des probabilités. Je lis ceci: « les statisticiens considèrent en général qu'un écart réduit supérieur à 2 ne s'explique que par une intention particulière de l'auteur » et l'on me précise que « écart réduit supérieur à 2 » signifie « en gros, une fois sur vingt ». J'en conclus que, pour les auteurs de cette règle pratique, une probabilité de « une chance sur vingt » est une « petite » probabilité. Je n'ai jamais compris pourquoi les seuils de 1 % ou de 5 % s'étaient imposés en toutes sortes de

domaines, comme si la sensibilité devait être la même pour tous et partout.

Mais de quelles probabilités parle-t-on? On lit quelquefois: probabilité que ceci (qui a été constaté) soit ou ne soit pas, *effet du hasard*.

Sur un événement, dont on ignore s'il a eu lieu, ou sur une proposition dont on ignore si elle est vraie, on a toujours le droit de placer une probabilité chiffrée, pour représenter, du mieux qu'on peut, *l'incertitude* où l'on se trouve.

Mais dans les circonstances dont nous parlons ici, ce n'est pas tout à fait la même situation: la probabilité que l'on a calculée, c'est celle d'obtenir un extrait de PP qui soit semblable (sous l'aspect de la fréquence) au texte DD. Le « hasard », si hasard il y a, c'est celui que nous avons décidé d'introduire dans le modèle: en considérant que tous les extraits de PP sont, a priori, équiprobables, ou, comme on dit encore, en effectuant (mentalement bien sûr!) un tirage au sort.

Quant à la proposition « ce n'est pas un hasard si Pascal emploie moins souvent *penser* que ne le fait Descartes dans son *Discours* », je ne suis pas capable de lui assigner une probabilité, c'est-à-dire, de traduire par du chiffre, l'incertitude où je serais (d'autant plus que, à donner au mot hasard son sens le plus ordinaire, je suis tout prêt à dire ma certitude: non, ce n'est pas un hasard!).

Certains pensent échapper à cette difficulté: ils parlent de « la probabilité de se tromper » en affirmant que tel ou tel contraste de fréquence est « significatif ». Il s'agit d'un décalque maladroit de la théorie des tests; cette théorie a fait ses preuves dans le contrôle des fabrications industrielles comme dans le montage de certaines procédures d'expérimentation. Mais, pour les études lexicales, il me semble qu'il ne s'agit pas de trancher à la suite d'un calcul « ceci est significatif » ou bien « ceci ne l'est pas ». Heureusement, le plus souvent, ceux qui utilisent le modèle ne

s'en servent en fait que pour trier et non trancher: ranger, en quelque sorte, les contrastes de fréquence dans un ordre d'importance — en évitant les pièges de la proportionnalité.

Quels que soient les débats, toujours d'actualité, sur le bon usage des probabilités — on m'accordera, je pense, que les probabilités dont il a été question jusqu'ici concernent des jugements portés sur les fréquences: « il y a très peu de chances qu'en prenant *au hasard* vingt mille mots de PP je puisse obtenir une fréquence dépassant 60 ».

Mais je ne tiendrais pas la promesse de mon titre si j'omettais de dire quelques mots, pour finir, sur un sujet bien plus délicat, sur une tout autre forme de liaison entre fréquence et probabilité.

Probabilités d'emploi?

Si j'étais poète oulipien, fatrasique ou rhétoriqueur, j'essaierais d'écrire une ode (il faudrait qu'elle soit assez longue) dans laquelle certains mots clefs bien choisis seraient répétés selon un rythme le plus régulier possible: tous les 17, 23 ou 49 mots par exemple. Puéril? non: obsessionnel; c'est l'obsession de la « Règle » que nous appelons « de trois », de la proportionnalité, de la régularité.

Descartes en son *Discours*: 83 fois *penser* en 22688 mots, ce qui fait une fois tous les 273 mots. Mais ce n'est pas régulier. L'auteur nous indique un découpage: « si ce discours semble trop long pour être tout lu en une fois, on le pourra distinguer en six parties... ». Dans la troisième partie, on *pense* tous les 143 mots, dans la quatrième, tous les 479.

Les jeux de hasard ont, comme on l'a observé depuis très longtemps, fourni des modèles de régularité d'un type particulier. Quand on lance un dé plusieurs fois, on peut, d'une part, constater que l'as apparaît à peu près une fois pour six coups, tantôt plus, tantôt moins, — et d'autre part trouver une *explication* de ce phénomène dans l'hypothèse d'une probabilité égale à 1/6. Ce

fut, comme on sait, le grand succès de Jacques Bernoulli d'établir cela sur des bases mathématiques solides.

On a souvent pensé imiter cette réussite, en biologie par exemple: les lois mendéliennes de l'hybridation, le taux de masculinité dans l'espèce humaine, etc.

D'où l'idée de s'acheminer vers un modèle probabiliste en matière de discours. L'objet n'est plus, comme précédemment, des jugements portés sur un texte, comparaisons, contrastes, c'est-à-dire conduites de lecteur: cette fois-ci c'est l'activité discursive elle-même qui est visée, celle de l'écrivain, et non plus celle de son critique.

Tout devient alors beaucoup plus difficile. Peut-on représenter la production d'un texte comme un processus stochastique? les mises en garde ne manquent pas: depuis Cicéron qui nous met au défi de créer un vers latin en jetant au hasard des lettres, jusqu'à Borel installant ses singes devant des machines à écrire.

Un spectre hante le lexique: celui d'une réserve, d'un trésor (*wortschatz*, *treasure-chest*), d'un grand sac où, comme les boules du Loto, l'écrivain puiserait sans l'épuiser: « *rather ridiculous* », dit G.U. Yule, qui s'y connaît.

Il faut rappeler ici une histoire édifiante. Depuis des siècles, les praticiens (et les amateurs) d'écritures secrètes, les cryptologues, comme on dit, ont noté que les lettres de l'alphabet n'ont pas les mêmes fréquences: la lettre E est de beaucoup la plus fréquente en français, cette prédominance étant beaucoup moins accusée en italien, etc. On a ici un magnifique exemple de permanence statistique: fluctuation des fréquences, assortie d'une certaine stabilité. Pendant longtemps on se contente de dire: la fréquence varie un peu; on ajoute: d'autant moins que le texte est plus long; et sans trop oser le dire, on imagine une valeur « vraie », comme au jeu de dés.

Ce n'est qu'en 1913 qu'on voit apparaître les débuts d'une théorie probabiliste digne de ce nom. Le grand probabiliste russe