JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

MARIE-LOUISE DUFRÉNOY

Analyse statistique du langage

Journal de la société statistique de Paris, tome 87 (1946), p. 208-219
<http://www.numdam.org/item?id=JSFS_1946__87__208_0>

V

VARIÉTÉ

Analyse statistique du langage

« Les philosophes se trouvent conduits... à distinguer — leur pensée de toutes conventions : les uns regardent en deçà du langage. Toutefois, ils ne l'ont jamais fait (à ma connaissance) à partir d'une analyse du langage qui le réduise à sa nature statistique. »

P. VALÉRY, *Léonard et les Philosophes.*
(*Variété, III*, p. 193.)

I. Distributions de fréquences :
- A. Séries exponentielles de Poisson;
- B. Courbes de concentration : *Incompatibilité entre « intensité » et « fréquence » de manifestation. Séries harmoniques.*

II. Analyse statistique du vocabulaire Oriental des *Mille et Une Nuits* et des *Mille et Un Jours*.
- A. Loi harmonique des fréquences relatives de manifestation de 100 mots orientaux;
- B. Fréquences de manifestation des mots *calife*, *sultan* et *visir* (tirages contagieux);
- C. Distribution des fréquences d'emploi des mots *eunuque*, *esclave*.....;
- D. Fréquences de manifestation des mots *mille*, *or*.....

Conclusions.

I. — DISTRIBUTIONS DE FRÉQUENCES.

« La manière la plus simple d'appliquer la méthode statistique à la science du langage, consiste à étudier la répartition des mots... On peut dans ce but se placer à deux points de vue très différents selon que l'on considère le matériel utilisable ou le matériel utilisé. Dans le premier cas, c'est le dictionnaire qui constitue le champ d'expérience; dans le second cas, les textes. » (POMARET).

Le mot est l'unité sémantique : le mot peut, au plus haut degré de généralité, être étudié non quant à son sens, mais du point de vue de sa catégorie grammaticale : la valeur (p) de la probabilité pour qu'un mot pris au hasard dans un texte soit un verbe, caractérise ce texte dans une certaine mesure; par exemple, dans 1.000 tranches de 100 mots d'un texte d'Anatole France, les probabilités (p) de manifestation d'un verbe se distribuent à peu près normalement autour d'une valeur moyenne $p = 0{,}1542$, avec une déviation standard $s = \sqrt{0{,}1542 \times 0{,}8452 \times 100} = 3{,}6$ (POMARET).

Un texte peut être considéré : 1° comme une population de mots, groupés par « phrases »; 2° comme une série linéaire de mots, toutes les lignes d'un texte étant supposées être mises bout à bout, sans tenir compte des signes de ponctuation; enfin 3° comme une population de classes fréquences (f_x). Chaque classe (f_x) groupe les mots se manifestant x fois, dans l' « échantillon » étudié statistiquement, qui peut être le texte tout entier, ou chacune des parties, chapitres, paragraphes... dont il est constitué.

Chacun de ces points de vue correspond à un mode différent d'analyse statistique des textes, et, ce qui est plus intéressant encore pour le statisticien, chaque méthode met en œuvre l'une des fonctions fondamentales représentant des distributions de fréquences.

Rappelons qu'il convient de distinguer nettement entre le cas où la distribution des mots est étudiée en conservant leur séquence (soit dans une série linéaire supposée continue d'un bout à l'autre du texte, soit en respectant la ponctuation et en respectant la division en phrases) et le cas contraire, où le texte est considéré comme une population de N mots différents.

Dans un texte considéré comme une série linéaire de mots, nous pouvons étudier la fréquence de récurrence d'un certain mot, parmi les (N-1) autres mots du texte; nous appliquerons cette méthode d'analyse aux *Mille et Une Nuits* traduites par Galland et aux *Mille et Un Jours* traduits par Pétis de la Croix.

Un texte étant composé de phrases, la longueur des phrases peut être prise comme caractère spécifique : dans un texte donné, les fréquences de phrases de 1, 2... mots se distribuent généralement de façon très dissymétrique de part et d'autre de la moyenne, mais la distribution peut être rendue normale en portant en abscisses non plus (x) mais ($\log x$) (C. B. WI-

LIAMS), c'est-à-dire que la distribution est caractérisée par la moyenne géométrique de (x) et par la dispersion de part et d'autre de cette moyenne.

Prenons maintenant comme unité d'échantillonnage, non plus la phrase, mais le paragraphe (ainsi que l'a fait Paul Pelliot pour l'étude de textes chinois) ou le chapitre, ou telle autre division du texte, telle que les « Nuits » des *Mille et Une Nuits* de GALLAND ou les « Jours » des *Mille et Un Jours* de Pétis de la Croix.

A. — *Distributions de fréquences selon les séries exponentielles de Poisson.*

Chaque « Nuit » ou chaque « Jour » peut être considéré comme une épreuve permettant la non-manifestation (0) ou la manifestation (1, 2, ...x fois) d'un certain mot choisi comme test. Nous étudierons, au sujet d'un certain nombre de mots représentatifs du vocabulaire oriental des *Mille et Une Nuits* et des *Mille et Un Jours*, la distribution des fréquences de « Nuits » ou de « Jours » avec 0, 1, 2,... x manifestations, et nous comparerons les fréquences observées à celles que permettent de calculer, pour une même moyenne de fréquence de manifestation, les séries exponentielles de Poisson. Enfin, négligeant ses divisions, nous considérerons le texte dans son ensemble comme une population de classes de fréquences; nous pourrons comparer la distribution des fréquences observées à une série exponentielle de Poisson, si nous faisons figurer dans la « classe de fréquence (0) » tous les mots utilisables, que l'auteur eût pu trouver dans le dictionnaire, mais qu'il n'a pas utilisés.

Le vocabulaire consiste en une dizaine de milliers de mots, dont la plupart n'ont qu'une chance infiniment faible d'être utilisés par un auteur écrivant un mot, mais si cet auteur écrit plusieurs millions de mots, il peut donner à chaque mot du vocabulaire des chances de se manifester. Si chaque mot écrit par un auteur correspond à un tirage représentant pour chaque mot du vocabulaire une opération qui peut ou non le faire sortir, la probabilité de sortie (p) peut être aussi voisine de 0 que l'on veut l'imaginer : si le nombre (n) des opérations devient assez grand, la probabilité moyenne (np) devient mesurable; dans le cas limite où (p) devient infiniment petit et (n) infiniment grand le produit (np) prend la valeur (a), moyenne autour de laquelle se distribuent les fréquences de 0, 1, 2..., x, succès par échantillon, avec des probabilités définies par les séries de Poisson.

Quant à l'espérance mathématique qu'un événement se manifestera au moins 1, 2,..., c fois, elle peut s'estimer par l'intégration des séries exponentielles de Poisson, car si la probabilité n'est pas toujours additive, l'espérance l'est.

Trois conditions doivent être satisfaites pour qu'une distribution observée puisse être comparée à une série d'intégration telle qu'elle vient d'être définie. L'événement étudié peut ou non se manifester, de telle sorte que la fréquence (c) de manifestation soit 0 ou un nombre entier, 1, 2, ...; l'opération, capable de faire apparaître la manifestation, doit pouvoir être indéfiniment répétée dans les mêmes conditions, et, chaque fois, indépendamment des opérations précédentes; l'opération doit faire partie d'un groupe, représentant l'échantillon statistique défini par la moyenne (a) : la valeur expérimentale de (a) est la fréquence moyenne des manifestations dans un groupe d'opérations. Pratiquement, le nombre (n) des opérations est un nombre fini et généralement même assez faible, mais le nombre possible de ces opérations identiques est infini ; il est dès lors possible de définir un type de fonction de probabilité qui puisse être dérivé à la limite, lorsque (n) devient infini alors que (pn) demeure fini (1).

(1) G. A. Campbell a publié les Tables qui permettent de transformer les courbes de probabilité des séries de Poisson en courbes d'intégration. La moyenne (a) s'exprime en fonction de la probabilité (P) d'*au moins* c manifestations .

$$a = c \sum_{n=0}^{\infty} Q_n C^{-\frac{1}{2}n} \qquad (1)$$

les coefficients Q_n étant des fonctions de l'argument (t) correspondant à la probabilité (P) définie par l'équation de la courbe sigmoïde

$$P = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}t^2} dt$$

L'intégration des séries de Poisson met en œuvre les expressions classiques :

$$P = \frac{a^c e^{-a}}{c!} + \frac{a^{c+1} e^{-a}}{(c+1)!} + \frac{a^{c+2} e^{-a}}{(c+2)!} + \dots$$

$$= \sum_{s=c}^{\infty} \frac{a^s e^{-a}}{s!}$$

$$= 1 - \left[1 + \frac{a}{1!} + \frac{a^2}{2!} + \dots + \frac{a^{c-1}}{(c-1)!} \right] e^{-a}$$

$$= \sum_{s=0}^{c-1} \frac{a^s e^{-a}}{s!} = \frac{1}{\Gamma(c)} \int_0^a a^{c-1} e^{-a} \dots$$

Les séries (1) peuvent se déterminer en posant égaux les intégrants de (2) et de (3) :

$$\frac{1}{\Gamma(c)} a^{c-1} e^{-a} da = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2} dt$$

et en résolvant l'équation, pour les valeurs positives de a, sous condition que $t = -\infty$ quand $a = 0$.

De l'idée abstraite de fonction de probabilité, revenons à l'idée concrète de distribution de fréquences, ou mieux encore à celle de distribution des valeurs d'intégration qui définissent la probabilité (p) d'au moins (c) manifestations.

Pour représenter graphiquement l'équation où figurent les trois variables : (a) valeur moyenne, (p) probabilité d'au moins (c) manifestations, et nombre (c) des manifestations on peut donner successivement à l'une des variables (c par exemple), diverses valeurs fixes : chacune des valeurs choisies portées dans l'équation la transforme en équation à deux variables (a) et (p) ; on porte en ordonnées, sur une échelle proportionnelle à l'intégrale de la distribution normale, les valeurs de (p), 10^{-6}, 10^{-5},..... 10^{-1}, 0,25, 0,50, 0,75, 0,90, 0,99, 0,999, 0,99999 ; on porte, en abscisses, les valeurs de (a), de 0 à 15, sur échelle arithmétique, ou, mieux encore, de 0,1 à 100 sur échelle logarithmique (F. Thorndike) ; les axes du quadrillage cotés (a) et (p) se coupent sur la courbe (c). Chaque courbe (c) étant cotée au moyen de sa valeur correspondante dans l'équation, on obtient l'abaque cartésien représentant les courbes de probabilité pour l'intégrale de

$$P = 1 - \left[1 + \frac{a}{1!} + \frac{a^2}{2!} + ... + \frac{a^{c-1}}{(c-1)!}\right] e^{-a} \qquad (3)$$

où (P) est la probabilité d'au moins (c) manifestations dans un groupe d'essais, lorsque le nombre moyen des manifestations est (a).

Sur les graphiques 3 à 6, toute distribution de série de Poisson est représentée par les points d'intersection de la verticale d'abscisse (a) avec les courbes (c) ; (c) représentant le nombre de manifestations de l'événement par échantillon.

Toute distribution observée correspondant à la moyenne (a) est représentée par l'ensemble des points d'intersection de l'horizontale d'ordonnée (P) avec la courbe (c) correspondante. La distance horizontale entre chacun des points ainsi définis et le point correspondant de la verticale d'abscisse (a) mesure la déviation entre fréquence observée et fréquence théorique de la distribution de la série de Poisson ; cependant, du seul fait que le nombre (n) des essais est fini, et que la plus grande valeur observée pour (c) ne peut dépasser (n), la distribution linéaire des points tend à dévier vers la gauche.

De plus, lorsque les échantillons ne sont pas uniformes, la série observée diffère d'une série de Poisson théorique de sorte que la distribution linéaire des points tend à s'incliner vers la droite.

Enfin, si les essais ne sont pas indépendants, mais si les effets de l'un influencent les résultats des autres, la distribution tend à s'incliner : vers la gauche, si la corrélation est négative, vers la droite, si la corrélation est positive ; ce dernier cas, le plus intéressant, est celui des tirages contagieux (1).

B. *Courbes de concentration : courbes hyperbolique et parabolique.*

La fréquence de manifestation des mots dans le vocabulaire d'un auteur peut être étudiée à l'aide des distributions des séries de Poisson si nous considérons le vocabulaire utilisable, et tenons compte des mots non utilisés, c'est à dire des mots qui figurent 0 fois dans l'ensemble du texte étudié. Mais, en fait, ce que nous désirons étudier, c'est plutôt le vocabulaire utilisé que le vocabulaire utilisable, c'est à dire une distribution de fréquence tronquée du fait de l'absence de la classe de fréquence 0.

Au lieu d'une fonction exponentielle $y = ab^x$, nous utiliserons la fonction de la courbe hyperbolique $y = bx^k$ (4) en spécifiant que chaque valeur de (y) représente la fréquence de manifestation de la valeur correspondante de (x) (ici nombre des mots se manifestant chacun (y) fois). L'équation (4) peut s'écrire sous la forme différentielle :

$$\frac{dy}{y} - k\frac{dx}{x} = 0 \quad (5) \quad \text{ou} \quad \frac{dy/y}{dx/x} = k \quad (5^b)$$

qui exprime que les taux de croissance (ou de décroissance) de (x) et de (y) sont respectivement dans un rapport constant (k) et qui met en évidence le caractère purement numérique de (k), rapport entre les deux rapports dy/y et dx/x et, comme tel, indépendant des unités servant à mesurer (x) et (y). Cependant, si la croissance ou la décroissance de l'une des variables (x) ou (y) est mesurée dans un espace à deux dimensions, celle de l'autre étant mesurée dans un espace à trois dimensions, la valeur numérique de (k) se trouvera de l'ordre de 2/3 ou de 3/2. Suivant que (k) est positif ou négatif, la courbe est une parabole ou une hyperbole ; dans l'un ou l'autre cas, elle peut s'écrire sous la forme

$$\log y = k \log x + \log b, \qquad (6)$$

que représente, sur papier logarithmique, une droite, dont la pente (k) exprime les taux relatifs de croissance ou de décroissance de (y) et de (x) ; suivant l'expression de M. d'Ocagne, cette droite, dont le coefficient angulaire est (k) et l'ordonnée à l'origine est $\log b$, est l'« image logarithmique du monome bx^k ».

---

(1) G. A. Campbell : « Probability curves showing Poisson's Exponential Summation ». (*Bell System Techn. J.*, v. 2, pp. 95-113, janv. 1923.)
F. Thorndike : « Applications of Poisson's Probability Summation ». (*Ibid.*, v. 5, pp. 604-624, 1926.)

Dès 1928, E. U. Condon avait exprimé la corrélation inverse qui lie le nombre (n) des différents mots et la fréquence de leur manifestation (f_n) par l'équation :

$$f_n = \frac{k}{n}. \qquad (7)$$

G. K. Zipf (1929, 1932), de son côté, formulait cette loi sémantique : peu de mots sont fréquemment employés; beaucoup de mots sont rarement employés.

En portant, en abscisses, le log. du nombre des mots différents (x) se manifestant chacun avec la fréquence (f_x) et, en ordonnées, log. (f_x), on obtient une ligne droite; à cette loi de la « concentration sémantique » correspond la loi de « concentration urbaine », liant, par une régression rectilinéaire, le log. de la fréquence (f_x) des villes au log. du nombre (x) de leurs habitants.

A. J. Lotka et J. B. Carroll firent bientôt remarquer que de telles distributions, mettant en œuvre une progression géométrique de (x) et une progression géométrique de (y) et centrées sur les moyennes géométriques des (x) et des (y), connaissent de très nombreuses applications; du point de vue physique, elles représentent des applications très diverses de la relation fondamentale exprimée par Boyle et par Mariotte comme « Loi des gaz »; mais trois siècles auparavant, dès 1370, la signification philosophique de ces distributions, mettant en jeu une progression géométrique de (x) et une progression géométrique de (y), avait été explicitée par Nicole Oresme, dans le *Livre de Éthique d'Aristote* (Livre II, ch. 7 et Livre V, ch. 7) : « Or meton donques que la proporcion des personnes quant a dignite ou merite soit comme la proporcion de .xii.a. vi. et que la proporcion de le honeur ou de la pecune distribue soit comme la proporcion de .iiii. a. ii..... Et donques .xii. et .vi, ces .ii. termes con joins ensemble ont tele meisme proporcion a.iiii, et .ii, conjoins ensemble. Et est juste distribucion et est le moien. Et ce que est hors tele proporcionalite est injuste. Car ce que est selon tele proporcionalite est moien et juste. Et les mathematiciens appellent tele proporcionalite geometrique... Car celui qui prent et a du bien plus que selon ceste proporcionalite, il fait injustice..... »

Cinq siècles après Nicole Oresme, les économistes ont en effet constaté que le nombre des « rentiers » dans chaque classe de revenus diminue en progression géométrique à mesure que les revenus croissent en progression géométrique. Niceforo illustre cette loi par la distribution intégrale des 10.000 rentiers de la ville d'Amsterdam en 1912-1913. Si nous désignons par $f(x)$ le nombre des contribuables jouissant d'un revenu d'au moins (X), la « droite de péréquation », tirée de l'équation

$$\log f(x) = 7.7825 - 1.4 \log x$$

apparaît sur le graphique avec une pente de — 1.4, c'est-à dire voisine de la valeur de $-k = \frac{2}{3}$, indiquée plus haut.

L'enrichissement du vocabulaire manifeste une diversification à partir d'un certain nombre de racines sémantiques, et doit obéir à la loi qui régit la diversification des individus (quant à leur aptitude à acquérir ou conserver les richesses) ou celles des êtres animés (sous forme d'espèces morphologiquement différentes au sein d'un genre). Au cours de la diversification des êtres, le nombre des genres croît en proportion géométrique, selon la loi des intérêts composés; le nombre des espèces croît selon la même loi, mais avec un taux plus rapide : la fréquence (f_x) des genres comptant chacun (x) espèces est telle que $x\,(f_x)^n = cte$, c'est à dire que $\log x + n \log f_x - b = 0$. La pente des droites de péréquation ainsi définies peut être comprise entre — 1 (valeur que lui attribue théoriquement Condon) et — 2 (valeur que lui attribue Zipf *a priori*). En fait, nous constaterons qu'elle est voisine de — 2/3, valeur fréquemment indiquée d'ailleurs, comme représentant les phénomènes de « croissance hétérogonique », c'est à dire les phénomènes exprimant les relations entre croissance dans un espace à deux dimensions et croissance dans un espace à trois dimensions.

Transposant cette relation dans l'analyse du vocabulaire, E. U. Condon porte en ordonnées le logarithme de la fréquence (f_x) avec laquelle chaque mot différent se manifeste dans un texte, et en abscisses ($\log x$), c'est à dire le logarithme du nombre des mots dans chaque classe de fréquence (f_x); il obtient une droite de pente voisine de — 1 et peut donc écrire $f_x = k/x$.

Si la loi vaut pour l'ensemble du vocabulaire comprenant (m) mots différents, la constante (k) doit avoir une valeur telle que

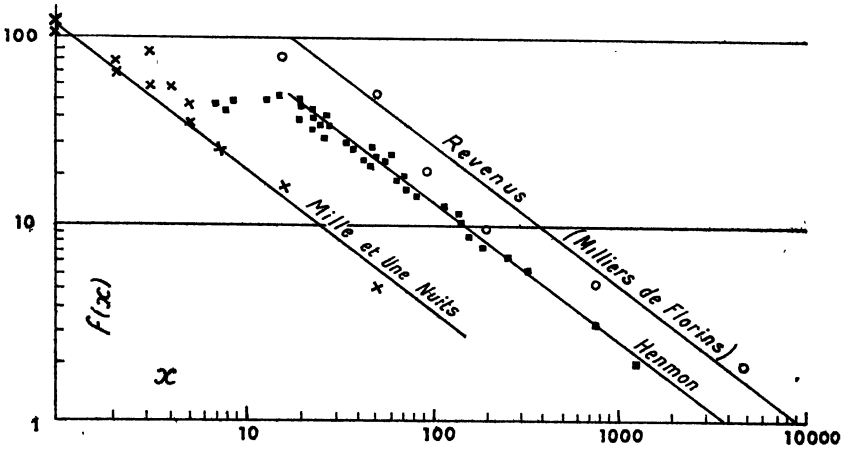
$$k \sum_{n=1}^{m} \frac{1}{n} = 1.$$

D'une analyse de 100.000 mots d'un texte anglais, G. Dewey a conclu que $m = 10.161$ pour cette langue, ce qui donne pour (k) une valeur voisine de 0,1.

La fréquence d'emploi d'un mot mesure son degré d'efficacité dans la transmission des idées; mais l'augmentation de transmissibilité de pensée acquise par un néologisme est d'autant moindre que le nombre (m) des mots différents dans le vocabulaire est déjà plus grand.

La loi de l'enrichissement ou de la diversification du langage doit avoir la même allure que la loi de diversification des formes organiques : toutes deux peuvent s'exprimer par une relation de la forme $f_x = f_1 x^{-r}$. Zipf, considérant que la valeur réelle idéale de r est 2, énonce cette loi sémantique : « Le produit du carré de la fréquence $f(x)$ de manifestation d'un mot par le nombre (x) de mots se manifestant avec cette même fréquence est une constante »; cependant, la valeur de l'exposant (r) dépend du degré d'inflection du langage, c'est à dire du nombre des permutations entre racines, préfixes, suffixes, terminaisons, dans la création des mots du vocabulaire; d'une analyse des fréquences relatives de manifestation de 400.000 mots de la langue française publiée par Henmon, Zipf avait conclu à une valeur de $r = 1{,}39$.

Il se trouve que la droite de péréquation des fréquences de mots dans la langue française (d'après Henmon), celle des fréquences de mots d'origine orientale dans les *Mille et Une Nuits* ou les *Mille et Un Jours* (graphique 1), comme celle des fréquences de classes de revenus étudiés par Niceforo, ou celles des fréquences d'espèce par genres étudiées par Willis, sont toutes parallèles, leur pente étant voisine de — 1/1,4 ou de — 1,4 selon que l'on porte en abscisses les « fréquences d'individus par classes » et en ordonnées les « fréquences de manifestations de classes de fréquences » ou vice versa. Les valeurs portées en abscisses ou en ordonnées, sur échelle logarithmique, étant dans les deux cas des « fréquences », l'une ou l'autre peut être prise comme variable indépendante.



Graphique 1. — Droites de péréquation du nombre des contribuables pour chaque classe de revenus de la ville d'Amsterdam (d'après NICEFORO, *op. cit*, p. 318), du nombre des mots français pour chaque classe de fréquence de manifestation (analyse de Henmon, reproduite par ZIPF, *Psychobiology of Language*, p. 235); du nombre de chacun de 100 mots Orientaux des *Mille et Une Nuits* pour chaque classe de fréquence de manifestation.

Incompatibilité entre « intensité » et « fréquence de manifestation ». — Dans la mesure, où, selon Zipf (*loc. cit.*, p. 27, 1932), le mot est l'unité capable d'évoquer un état de conscience, l'intensité de l'évocation est pour chaque mot du vocabulaire en fonction inverse de la fréquence d'emploi, et, d'une façon générale, si (x) mesure l'intensité de manifestation ou d'évocation, d'un phénomène et si (f_x) est la fréquence de (x), alors $x\ (f_x)^n = cte$ (1).

Séries harmoniques. — E. Baticle (*C. R. Ac. Sc.*, 222, pp. 355 7, 1946), illustre le « Problème des stocks » par cet exemple : si l'on veut remplacer un train ayant à transporter (q) voyageurs par (n) trains, chaque voyageur pouvant emprunter l'un quelconque de ceux-ci, il faut prévoir une capacité totale $q\ (1 + 1/2 + \ldots + 1/n)$ pour qu'en moyenne les (q) voyageurs puissent être transportés sans attente.

Si dans un vocabulaire nous pouvons remplacer un mot ayant à se manifester (q) fois par l'un quelconque de (n) mots, les fréquences de manifestation deviendront proportionnelles à $q\ (1 + 1/2 + \ldots + 1/n)$. Si les mots du vocabulaire sont ordonnés par ordre de fréquence de manifestation, les manifestations consécutives d'un même mot seront séparées par un nombre $(10\ m)$ d'autres mots, tel, en moyenne, que, si le mot le plus fréquent revient tous les 10 mots, le second revient tous les 20 mots, le troisième tous les 30 mots, le n^e tous les $10\ n$ mots. De ce point de vue, Zipf considère qu'un langage est « harmonique » dans son ensemble (anglais) ou harmonique du moins pour les mots usuels.

---

(1) Un exemple très particulier vient d'en être fourni par L F Richardson, qui, entre 1820 et 1929, a dénombré les fréquences suivantes (x) de guerres ayant chacune causé un nombre de victimes (f) :

$\log (f_x)^{\frac{3}{7}} + 0{,}03$	1	3	16	62
	7	6	5	4

Sans doute trouverait-on des relations de même ordre entre « fréquence » des tremblements de terre et « intensité ».

II — Analyse statistique du vocabulaire oriental des *Mille et Une Nuits* et des *Mille et Un Jours*.

A. *Loi harmonique des fréquences relatives.*

Parmi les textes auxquels peut s'appliquer l'analyse statistique du langage, les *Mille et Une Nuits* de Galland et les *Mille et Un Jours* de Pétis de la Croix sont particulièrement intéressants : les traducteurs de ces contes orientaux ont introduit ou acclimaté dans la langue française des mots évocateurs de l'atmosphère orientale; d'autre part (du moins pour la partie des *Mille et Une Nuits* où Galland a respecté la division du récit en « Nuits ») la suite chronologique des « Nuits » ou la série des « Jours » constitue une série d' « épreuves » permettant à chacun des mots du « vocabulaire oriental » de se manifester ou non.

Extrayons des *Mille et Une Nuits* une centaine de mots spécialement représentatifs de l'Orient : 15 de ces mots n'apparaissent chacun qu'une fois dans la suite des « Nuits »; 7 figurent chacun deux fois, 3 apparaissent trois fois, etc... Groupons ces mots par classe de fréquence (F_x), centrons chaque classe autour de sa valeur moyenne F_x; dans chaque classe de fréquence, le nombre (x) des mots est tel que

$$(F_x)^n \cdot x = k,$$

c'est à dire n.log. (F_x) + log. x = log. k.

Cette équation logarithmique est représentée par les droites du graphique (2) dont la pente est — 1/1,5.

En remplaçant n par — 1,5 dans l'équation logarithmique, nous trouvons pour log. k la valeur moyenne 3,10 :

Classes	F_x	x	log k
1- 10	5	52	3,01
11 20	15	15	2,93
21 30	25	7	2,93
31 40	35	5	3,05
41 50	45	5	3,15
51 60	55	3	3,13
61- 70	65	2	3,05
71 80	75	2	3,10
81 90	85	3	3,38
91-100	95	0	—
101 110	105	1	3,05
111-120	115	1	3,10
121-160		0	

La même technique, appliquée à 60 mots spécialement évocateurs de l'Orient où se situent les contes des *Mille et Un Jours*, donne une droite de régression ayant sensiblement la même pente.

La « loi harmonique des fréquences relatives », cependant, ne s'applique pas aux mots les plus usuels : parmi les 100 mots choisis comme tests dans le vocabulaire des *Mille et Une Nuits*, ou les 60 mots choisis dans les *Mille et Un Jours*, ceux qui se manifestent plus de 100 fois se situent fort loin au dessus de la droite de régression (graphiques 1, 2, 3) : tels sont les mots *visir* (268), *calife* (236), *sultan* (231), *esclave* (133), *eunuque* (112), dont la distribution mérite une analyse statistique particulière.

B. *Fréquences de manifestation des mots* calife, sultan et visir *dans les* Mille et Une Nuits *et les Mille et Un Jours.*

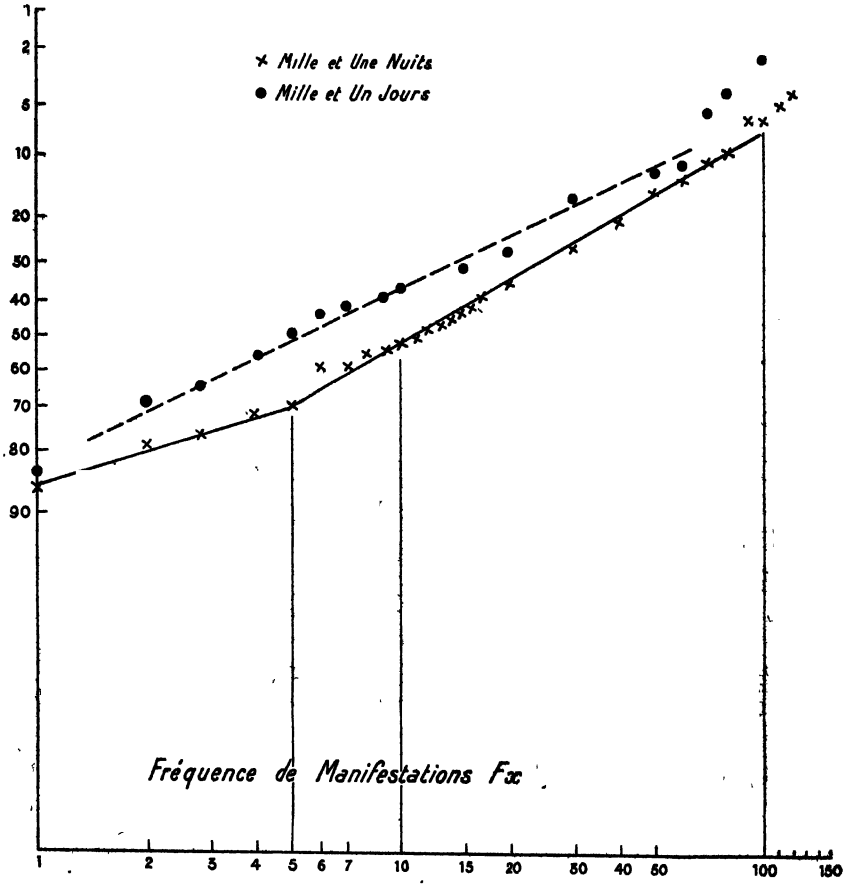
« L'Orient des Mille et Une Nuits... , peuplé de Sultans et de Visirs, d'eunuques et d'esclaves.. .. »
(Paul Hazard, *Revue des Deux Mondes*, III, 228, 15 sept. 1941)

Les « Nuits » de Galland, de la première à la 235e, constituent une série d' « histoires à tiroirs » dont les unes mettent perpétuellement en scènes le sultan, le visir, des esclaves, des eunuques, tandis que d'autres ne les font point apparaître : si chaque « Nuit » est considérée comme une épreuve, permettant à tel ou tel mot oriental de se manifester ou non, la population des 235 « Nuits » est hétérogène; c'est ce que met en évidence par exemple la distribution des fréquences (x_r) de séries de (r) « Nuits » où ne se manifeste pas le mot *visir* ou le mot *esclave*, ou tel autre.

Sur 235 « Nuits », nous en dénombrons 171 où le mot *visir* n'apparaît pas, 169 où le mot *esclave* n'apparaît pas, soit les probabilités de non manifestation :

$$visir,\ p = 171/235 = 0{,}73; \qquad esclave,\ p = 169/235 = 0{,}72.$$

Lisant les 235 « Nuits » à la suite, nous rencontrons une série de (r) « Nuits » où ne figure pas le mot *visir*. En d'autres termes, $(r - 1)$ « Nuits » sans manifestation (considérées chacune comme une épreuve indépendante) suivent chronologiquement la première « Nuit » sans manifestation, tandis que la suivante manifeste le mot *visir* au moins une fois. L'en-



Graphique 2. — En abscisses sur échelle logarithmique, fréquences de manifestation (F_x) d'un mot; en ordonnées, sur échelle de probabilité normale, pourcentage de 100 mots des *Mille et Une Nuits* ou de 60 mots des *Mille et Un Jours* se manifestant au moins 1, 2,... 100 fois.

semble de la distribution des séries de « Nuits sans visir » représente donc la somme de rx_r épreuves indépendantes, dont $(r - 1)x_r$ n'ont pas manifesté ce mot. Pour un certain total de « Nuits sans visir », la valeur de $\hat{p}$, estimée par la méthode de « maximum likelihood » sera d'autant plus grande qu'il y a davantage de longues séries, puisque $\hat{p}$ représente alors une fraction où les $(r - 1)x_r$ figurent au numérateur, les rx_r au dénominateur :

$$visir,\ \hat{p} = \frac{S\,(r-1)\,x_r}{S\,(rx_r)} = \frac{153}{171} = 0{,}891.$$

L'excès $\hat{p} - p = 0{,}891 - 0{,}730 = 0{,}161$, mesure le degré d'association des « Nuits sans visir ».

Les séries x_r observées peuvent d'ailleurs être comparées aux séries géométriques déterminées par les équations :

$$X_1 + r\,X_1 + r^2\,X_1 + \ldots\ldots = S\,(r\,X_r)^{(visir)} = 18 \qquad S\,(r\,X_r)^{(esclave)} = 87$$

TABLEAU I

	Visir				Esclave			
r	X_r	rX_r	$(r-1)X_r$	x_r calc.	X_r	rX_r	$(r-1)X_r$	x_r calc.
1	4	4	0	5,6	63	63	0	47
2	4	8	4	4	9	18	9	22
3	2	6	4	3	4	12	8	11
4				2,2	1	4	3	4,7
5				1,6	5	25	20	2,2
6	1	6	5	1,1	1	6	5	1,0
7				0,8				0,4
8					1	8	7	
9	1	9	8					
10								
11	1	11	10					
12	2	24	22					
13								
15					1	15	14	
21	1	21	20					
32	1	32	31					
50	1	50	49					
	18	171	153		87	169	82	

Ayant opposé aux 171 « Nuits sans manifestation de visir », les 65 « Nuits » où ce mot apparaît au moins une fois, comparons maintenant les fréquences avec lesquelles ce mot se manifeste au moins une fois, au moins deux fois, au moins c fois, avec les séries d'intégration de Poisson; comparons d'ailleurs les fréquences de manifestations des mots *calife* et *sultan* avec celles du mot *visir*, dans les *Mille et Une Nuits* (tableau II).

Les trois distributions ont sensiblement même allure, ce qui était à prévoir, étant donné que le visir accompagne presque toujours le calife ou le sultan, selon le lieu où se situe l'action du conte.

La forte corrélation positive qui lie les manifestations de ces trois mots peut se mesurer par l'angle que fait avec la verticale d'abscisse (a) la distribution des fréquences cumulatives (F) sur le graphique 3.

TABLEAU II

Intégration des séries de Poisson.

c, nombre de manifestations par échantillon (Nuit).
m, nombre d'échantillons avec exactement c manifestations.
f, nombre d'échantillons avec *au moins* c manifestations.
F, fréquence relative d'*au moins* c manifestations par échantillon.

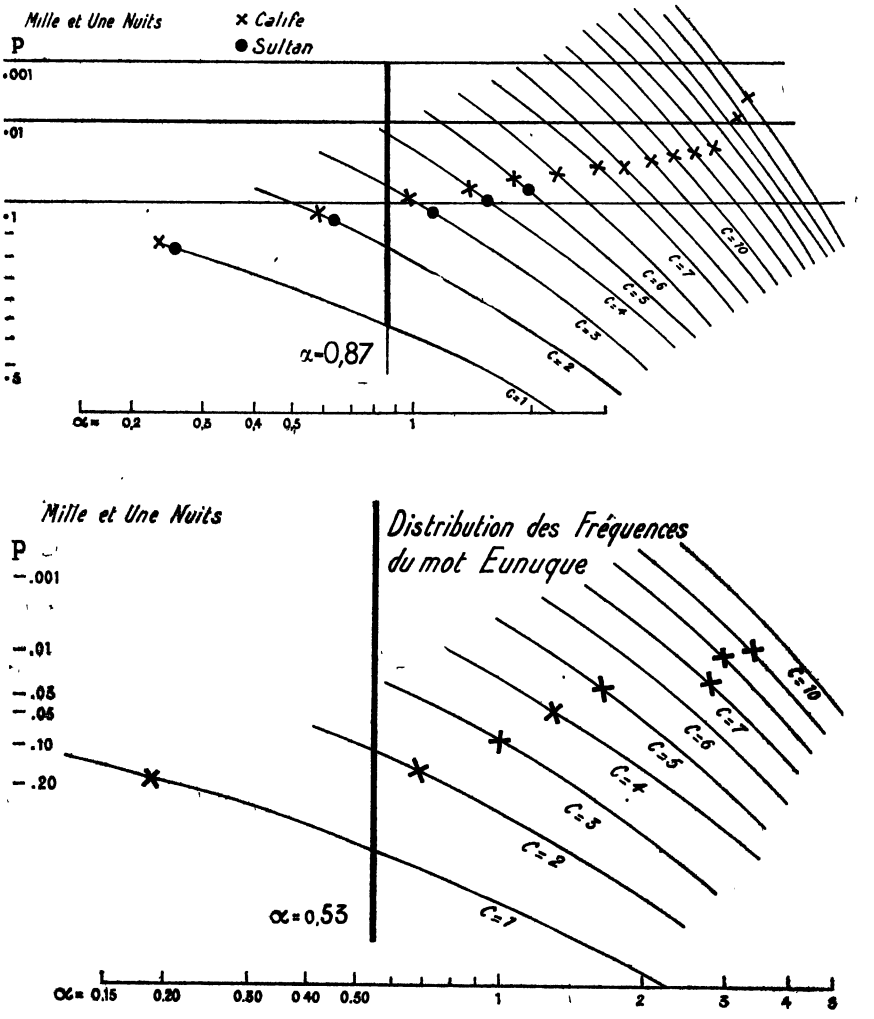
Mille et Une Nuits (Distribution des mots) :

	Calife				Sultan				Visir		
Total	236				231				268		
Moyenne :	0,88				0,87				1,16		
c	m	$c.m.$	f	F	m	$c.m.$	f	F	m	f	F
0	204	0	265	1,000	202	0	265	1,00	171	236	1,00
1	23	23	61	0,235	20	20	63	0,238	18	65	0,281
2	13	26	38	0,114	9	18	43	0,151	11	47	0,195
3	8	24	25	0,098	12	36	34	0,140	6	36	0,160
4	4	16	17	0,066	4	16	22	0,083	9	30	0,130
5	0				2	10	18	0,067	7	21	0,087
6	2	12	13	0,049	6	36	16	0,0425	1	14	0,043
7	0				3	21	10	0,0375	4	13	0,047
8	4	32	11	0,04	2	16	7	0,0262	2	9	0,039
9	0			15	2	18	5	0,0190	1	7	0,031
10	1	10	7	0,02	0				1	6	0,025
11	2	22	6	0,0275	1	11	3	0,011	1	5	0,023
12	2	24	4	0,0125	0				1	4	0,018
13	1	13	2	0,0050	1	13	2	0,0075	1	3	0,013
14	1	14	1	0,0077	1	14	1	0,0037	0		
15				39					0		
16									1	2	0,008
17									0		
18									1	1	0,004
	265	236			265	231			236		

C. — *Distribution des fréquences d'emploi des mots* ennuque, esclave.

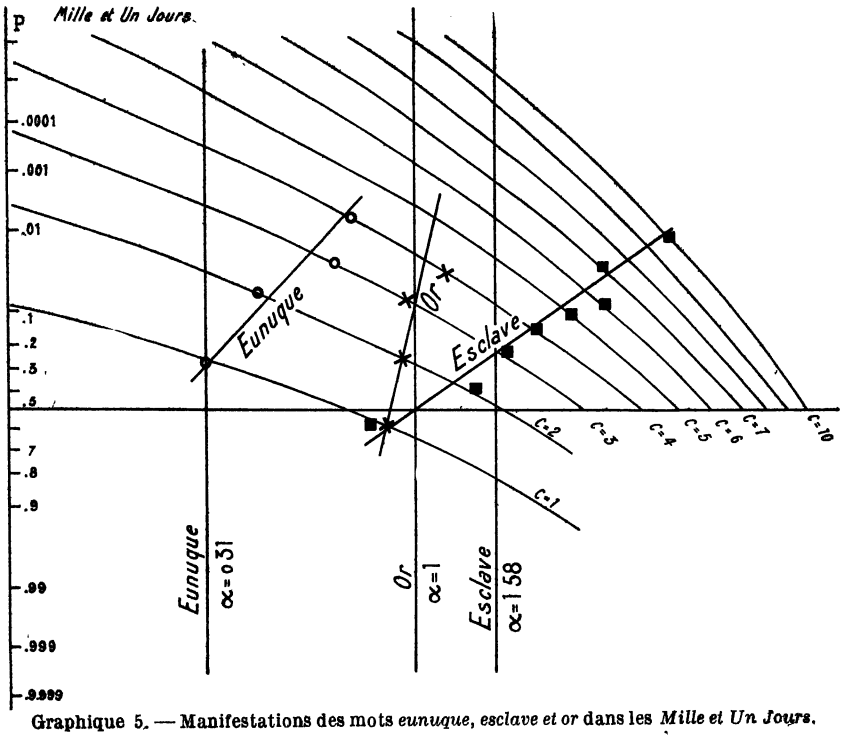
« Mais qui m'expliquera tous ces eunuques? »
(Paul VALÉRY · *Variété II*, p. 72 73, Paris, 1930.)

Le tableau III et les graphiques 4 et 5 permettent de comparer les distributions des fréquences de manifestation du mot *eunuque* dans les *Mille et Une Nuits* et les *Mille et Un Jours*, puis, comparativement les distributions de fréquences, dans les *Mille et Un Jours*, de manifestations du mot *esclave* qu'après les mots *sultan*, *visir*, *eunuque*, P. Hazard considère comme particulièrement évocateurs de l'Orient.

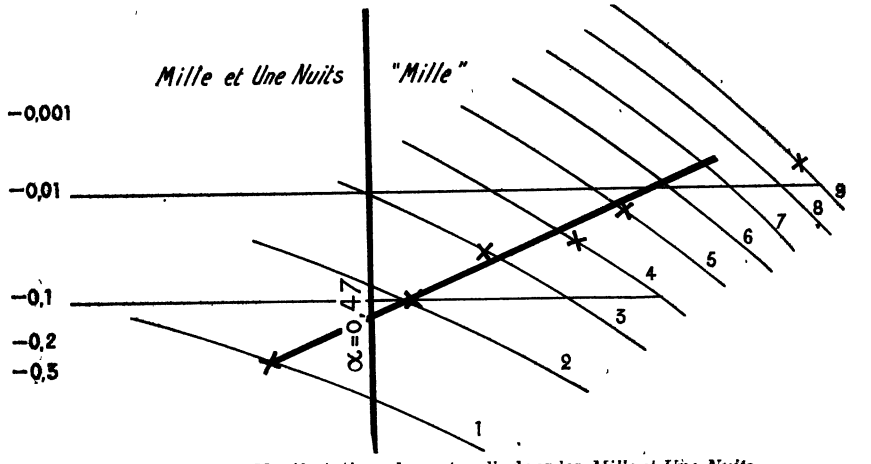


Graphiques 3 et 4. — Manifestations des mots *calife*, *sultan* et *eunuque*, dans les *Mille et Une Nuits*.

Abaques cartésiens des courbes de probabilité donnant l'intégration des séries exponentielles de Poisson : en ordonnées, sur échelle de probabilité, valeurs de la probabilité (P) qu'un événement se manifeste au moins (*c*) fois dans un grand nombre d'essais, lorsque la moyenne des manifestations est (*a*), les valeurs de (*a*) sont portées en abscisses sur échelle logarithmique



Graphique 5. — Manifestations des mots *eunuque, esclave et or* dans les *Mille et Un Jours*.



Graphique 6. — Manifestations du mot *mille* dans les *Mille et Une Nuits*.

TABLEAU III

Mille et Une Nuits. Eunuque Moyenne : 112/227 = 0,53				*Mille et Un Jours.* Eunuque 41/132 = 0,31			Esclave 133/86 = 1,58	
c	*m*	*c.m.*	*f*	F	*m*	*x*	*m*	*x*
0	189	0	227	1,00	105	96,8	33	46
1	13	13	38	0,169	18	29,9	20	27
2	9	18	25	0,110	6	6,6	14	7,7
3	6	18	16	0,071	2	0,5	8	0,86
4	4	16	10	0,044	0	0,05	3	0,17
5	1	5	6	0,026	1	0,001	11	0,02
6	0	0			0	0,0002	5	0,002
7	2	14	5	0,022			2	0,0002
8	1	8	3	0,013			0	
9	0						0	
10	2	20	2	0,009			(1)	
	227	112			132		86	

D. — *Fréquences de manifestation des mots* mille *et* or *dans les* Mille et Un Jours.

Le mot *mille* en tant qu'il figure dans le titre des contes orientaux traduits par Galland et Pétis de la Croix, peut être considéré comme ayant une signification symbolique : sa distribution (figurée sur le graphique 6) est dans les *Mille et Une Nuits* presque superposable à celle du mot *eunuque* (graphique 4).

Le mot *or*, évocateur des richesses fabuleuses de l'Orient, se distribue très sensiblement selon une série de Poisson pour une valeur de $a = 1$ (tableau IV); cette distribution figure (graphique 5) une droite très voisine de la verticale d'abscisse $a = 1$.

De façon générale, les fréquences observées sont d'autant plus voisines des fréquences calculées pour les séries de Poisson qu'elles se rapportent à des mots rares, moins affectés par le phénomène de tirage contagieux.

Cependant, l'apparition d'un certain mot rare, sans provoquer de nouvelles manifestations de ce même mot, peut provoquer la manifestation d'un certain autre mot; ainsi les mots *Égypte* et *Caire* ont sensiblement même distribution; parmi les végétaux, le plus fréquemment cité est le *sandal*, dont la manifestation appelle presque toujours celle d'*aloès* (tabl. IV) :

TABLEAU IV

	or *		Égypte		Caire		Sandal *		Aloès **	
a	m	f	m	f	m	f	m	f	m	f
0	32	27	216	203	216	202,5	82	81,4	225	222
1	24	27	12	28,2	11	28,5	7	8,1	8	12
2	13	13,5	3	2,5	4	2,8	1	0,4	1	0,6
3	2	4,5	3	0,15	1	0,2				
4	1	0,75	1		3				1	
5	3	0,225								
	75	73	235	233,7	235	234	90	89,9	235	234,6

CONCLUSIONS

Les études sémantiques que le professeur Paul Pelliot avait mises au service de l'archéologie avec tant de bonheur peuvent être utilisées avec profit dans la critique littéraire, et la statistique est susceptible de leur conférer un plus grand degré de précision.

Les méthodes statistiques permettraient en particulier de clarifier les données de problèmes particulièrement complexes comme celui des sources des différents contes dans un recueil où se mêlent des éléments imaginatifs, traditionnels et légendaires préservés dans presque tous les foyers de culture du monde oriental, tel le livre des *Mille et Une Nuits*. Peut-être une étude statistique minutieuse du vocabulaire permettrait elle d'identifier ceux dont l'origine est encore controversée, comme la célèbre histoire d'Aladdin, par exemple.

L'étude statistique du langage écrit montre que la récurrence d'un mot est déterminée par cette loi : le vocabulaire est surtout composé de mots dont la probabilité de manifestation est infiniment voisine de zéro; les mots « usuels » sont peu nombreux. Si les mots sont groupés en classes de fréquence de manifestation (b), la population (a) de chaque groupe est définie par l'équation $ab^r = k$: un mot usuel est un vocable qui tend à se manifester d'autant plus qu'il s'est déjà manifesté davantage; le mot usuel est le fait d'un tirage contagieux.

La valeur numérique de l'exposant de b dépend de la langue considérée : elle paraît voisine de 1,4 pour le français.

Pour une même langue, cet exposant varie avec le style : l'auteur peut, abusant de répétitions, exagérer la probabilité de manifestation des mots usuels; il peut au contraire, par la recherche du mot approprié, fût il rare ou fût il un néologisme, diminuer la récurrence.

Entre ces deux tendances, sollicitées par la répétition caractéristique du style oriental et par la multiplicité des idées et images nouvelles suggérant l'emploi de mots jusque là inconnus ou méconnus, les traducteurs semblent s'être bornés à vulgariser les mots clefs, intimement liés au contenu des récits orientaux, faisant ainsi d'un vocable rare un mot usuel, mais ils ont réussi à préserver par ailleurs les caractères statistiques de la langue française.

Si au lieu d'envisager spécifiquement la langue française, nous nous élevons, comme l'avait fait au XVII^e^ siècle le P. Mersenne, à l'idée d'une langue universelle, nous pouvons adopter comme conclusion la lettre que Descartes lui adressait d'Amsterdam le 20 novembre

(*) Les distributions pour *or* ont été calculées d'après 75 des « Jours »; pour *sandal* d'après 90; les mots *sandal* et *aloès* figurent tous deux une fois dans les « Jours » suivants : 1, 31, 47, 74, deux fois au « Jour » 14; *sandal* figure seul aux « Jours » 74, 84, 89; *aloès* seul au « Jour » 59.

(**) La moyenne pour la distribution du mot *aloès* dans les *Mille et Une Nuits* est $a = 14/235 = 0{,}06$; les fréquences calculées (f) indiquent un léger déficit des fréquences d'une manifestation, un léger excès de manifestations multiples : aux « Nuits » 58, 61 et 72, le mot *sandal* est associé au mot *aloès*.

1629 (*Œuvres*, publiées par Ch. ADAM et P. TANNERY, *Correspondance*, I, pp. 76-82, Paris, L. Cerf, 1897 : « On pourrait ajouter à ceci une invention, tant pour composer les mots primitifs de cette langue que pour les caractères : en sorte qu'elle pourrait être enseignée en fort peu de temps et ce, par le moyen de l'ordre, c'est à dire établissant un ordre entre toutes les pensées qui peuvent entrer dans l'esprit humain de même qu'il y en a un natu rellement établi entre les nombres.....

« .....Or je tiens que cette langue est possible et qu'on peut trouver la science de qui elle dépend..... Mais n'espérez pas de la voir jamais en usage : cela présuppose de grands changements en l'ordre des choses et il faudrait que tout le monde ne fût qu'un paradis terrestre, ce qui n'est possible à proposer que dans le pays des romans. »

Marie-Louise DUFRÉNOY.

BIBLIOGRAPHIE

Moyenne géométrique :

Maître Nicole ORESME : *Le livre des Éthiques d'Aristote* (Published from the text of MS. 2902, Bibliothèque Royale de Belgique, with a Critical Introduction and Notes by A. D. Menut, Stechert, New York, 1940).

Mc ALLISTER (D.) : « The law of the geometric mean ». *Proc. Roy. Soc.*, 24, p. 367, 1879.

GALTON (Fr.) : The geometric mean in vital and social statistics (*Natural Inheritance*, London, 1889, p. 138).

Distributions logarithmiques normales :

WILLIAMS (C. B.) : « A note on the statistical analysis of sentencelength as a criterion of literary style », *Biometrika*, 31, pp. 356-61, 1940.

Courbes de concentration :

PARETO (W.) : *Manuel d'Économie politique*, 1909.

RICHARDSON (L.-F.) : « Frequency of occurrences of Wars and other fatal quarrels », *Nature*, 148, p. 598, 1941.

Analyse statistique du langage :

HENMON (V. A. C.) : *A French Wordbook based on a count of 400.000 running words* (Madison, Wisc. Bureau of Educational Research, Bull. nº 3, sept. 1924).

LOTKA (A. J.) : *Elements of physical Biology*, pp. 306 307, 1925.

NICEFORO (A.) : *La méthode statistique et ses applications aux sciences naturelles, aux sciences sociales et à l'art*, Paris, 1925.

CONDON (E. V.) : « Statistics of Vocabulary », *Science*, 67, p. 300, 1928.

VALERY (P.) : Léonard et les Philosophes, *Variété III*, Paris, 1930.

ZIPF (G. K.) : *Selective studies of the principle of relative frequency in language*, Harvard University Press, 1932.

ZIPF (G. K.) : *The psycho biology of language*, Boston, 1935.

POMARET (P.) : « L'architecture du Langage ». *Rev. Gen. Sc.*, 50 , pp. 152 157, 1939.

CARROL (J. B.) : « Zipf's Law of urban concentration », *Science*, 94, p. 610, 26 déc., 1941.

LOTKA (A. J.) : « The law of Urban concentration », *Science*, 94, p. 104, 15 août 1941.

YULE (G. U.) : *The statistical study of literary vocab lary*, Cambridge Univ. Press, 1944.