JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

MARIE-LOUISE DUFRÉNOY

Statistique linguistique appliquée aux « lettres persanes ». Distribution des fréquences de mots par phrase

Journal de la société statistique de Paris, tome 107 (1966), p. 130-134

<http://www.numdam.org/item?id=JSFS_1966__107__130_0>

STATISTIQUE LINGUISTIQUE APPLIQUÉE AUX « LETTRES PERSANES »

DISTRIBUTION DES FRÉQUENCES DE MOTS PAR PHRASE

« Que ne vous appliquez-vous plutôt à la recherche de tant de belles vérités qu'un calcul facile nous fait découvrir tous les jours » (*Lettres persanes* CXXVIII).

En écrivant ces lignes, Montesquieu ne pouvait guère prévoir qu'un « calcul facile » permettrait de découvrir l'une de ces belles vérités par l'analyse statistique de ses écrits.

« *Les Pensées et Fragments inédits* (1849) révèlent en Montesquieu un artiste très conscient de son art; de l'art qu'il apporte à la construction de ses phrases, autant que dans la composition même de ses œuvres » (p. XX).

« On lit en effet dans le tome I de ses Pensées : « Bien des gens en France... soutiennent qu'il n'y a point d'harmonie. Je prouve qu'il y en a... »

Comment Montesquieu pensait-il avoir prouvé qu'il y a de l'harmonie? Pouvons-nous trouver une réponse à cette question à la lumière de l'analyse de la construction de ses phrases?

La « Statistique linguistique » connaît un grand intérêt d'actualité : Discutant de « La Linguistique statistique et la Linguistique structurale, A.-J. Greimas (1963) écrit :

« Que savons-nous... de l'économie distributive d'une langue réelle comme le français? A peu près rien... Sur quelles bases pourrait s'appuyer la recherche des normes... sur une théorie des climats à la Montesquieu?... La langue, au sens large du mot, présente un vaste champ de significations... qui... apparaîtrait tout désigné aux méthodes d'approche statistique. »

L'une de ces méthodes d'approche statistique est le recencement du nombre (n) de mots par phrases, et le classement des N phrases d'un texte, par classes de fréquences $f(n)$.

Cette méthode s'applique particulièrement bien aux *Lettres persanes*, constituant une « histoire à tiroirs », c'est-à-dire une séquence où l'on peut distinguer un certain nombre de pastiches, chacun écrit « à la manière de » l'un des auteurs pastichés dont Montesquieu s'est inspiré. Particulièrement instructive à cet égard est la comparaison de l'histoire des Troglodytes (chap. XI, XII, XIII et XIV des *Lettres persannes*, avec la description de la Bétique (livre VII du *Télémaque*).

Le texte concernant la Bétique comporte 105 phrases, dont la plus courte compte 4 mots, la plus longue 62 mots.

Les fréquences $f(n)$ de nombres de mots par phrase ont été transformées en fréquences cumulées, et celles-ci en pourcentages cumulés. Ces pourcentages cumulés ($P(n)$ sont portés en ordonnées sur échelle de probabilité normale, les valeurs de n étant portées en abscisses sur échelle log.

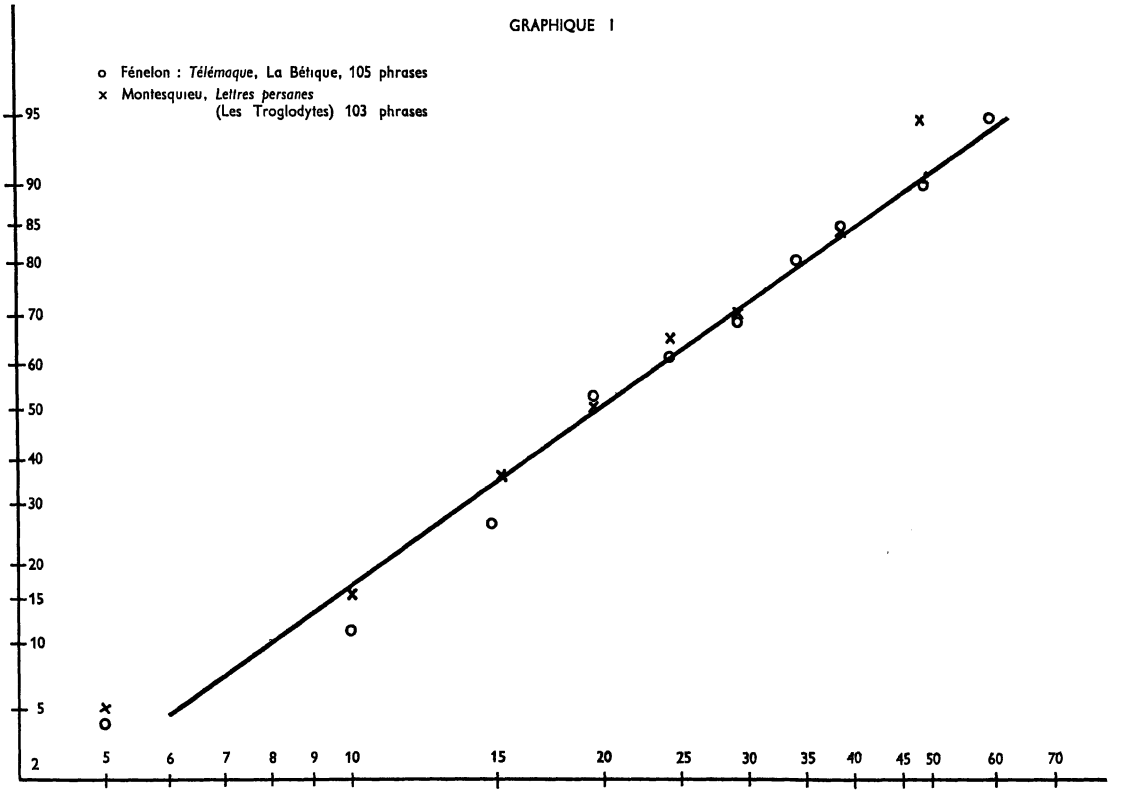
On détermine des points de coordonnées (P_n, log n) qui s'alignent sur une droite, interceptant le niveau 50 % à l'aplomb de l'abscisse correspondant au log de $n = 20$.

Le modèle le plus vraisemblable est celui de distribution log normale, qui, de façon générale, représente les résultats de la « fragmentation » au hasard *d'* une quantité initialement indivise de matériaux fragmentables.

Le texte des Troglodytes comporte 103 phrases, dont la plus courte de 5 mots, la plus longue de 75; à cette exception près d'une phrase anormalement longue, la distribution des $f(n)$ pour les Troglodytes recouvre celle des $f(n)$ pour la Bétique (graphique 1).

Aux Lettres XI à XIV, entièrement consacrées à l'Histoire des Troglodytes, c'est-à-dire à une satire de l'histoire de la Bétique, on peut comparer les « portraits » traités à la manière de La Bruyère : Lettre XLVIII, portraits du « financier » du « prédicateur » et du « vieux guerrier » (9 phrases dont une de 117 mots); Lettre LII, portraits de vieilles coquettes; Lettre LVI, joueurs et joueuses; Lettre LXXII, portrait du « Décisionnaire »

GRAPHIQUE I

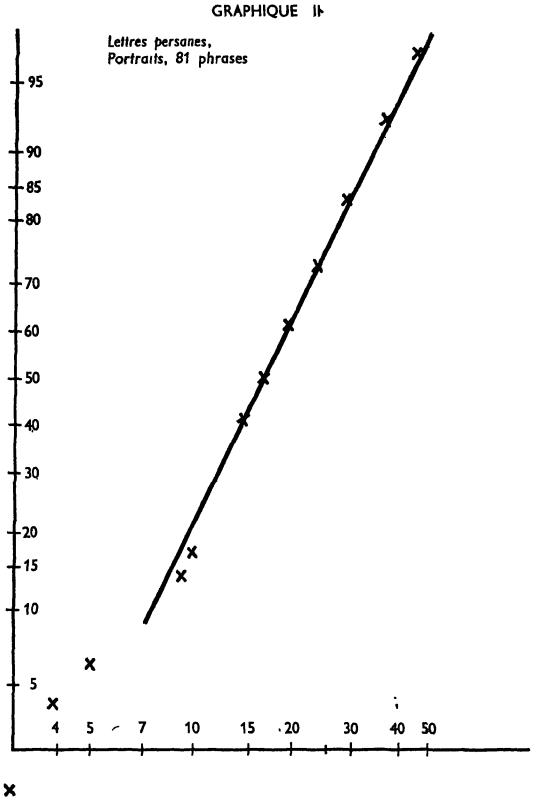


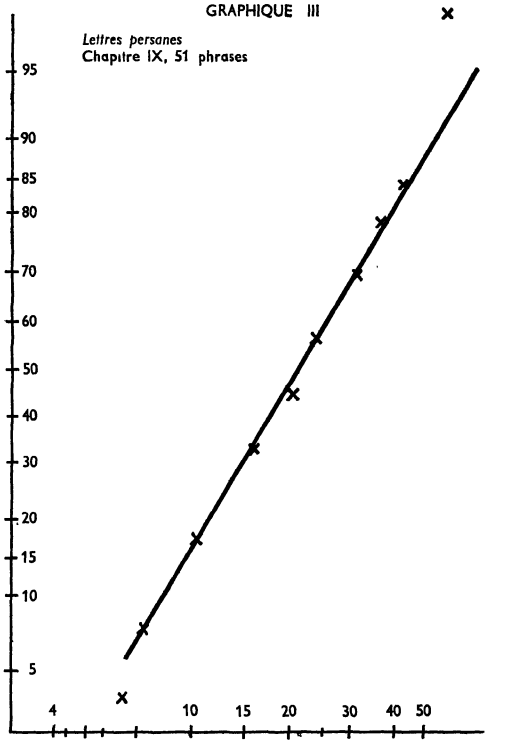
(aussi péremptoire que Arrias de la Bruyère); Lettre CXXXII, portrait du « nouvelliste pessimiste » (où Montesquieu se souvient du Démophile de la Bruyère).

Lettre CXXVIII, portrait du géomètre (qui est aussi un statisticien, s'intéressant à la distribution des nombres de pouces d'eau tombés sur la terre, chaque année, et qui aurait pu suggérer la remarque relative à Zadig qui « ne s'occupait pas à calculer... s'il tombait une ligne cube de pluie dans le mois de la souris plus que dans le mois du mouton » (*Zadig*, III). On pourrait encore citer les portraits « du curieux amateur de la vénérable antiquité » (Lettres CXLII), du « philologue »...

Un échantillon de 81 phrases, tirées de divers portraits, peut être représenté par la distribution log normale du graphique II.

La Lettre XXIV débute par une description des « embarras de Paris » où se trouvent des réminiscences de la Satire VI de Boileau sur les embarras de Paris, suivies de diverses





considérations; cette Lettre, hétérogène, compte 30 phrases dont une de 107 mots; 16 % des phrases comptent moins de 6 mots, 50 %, 31 au plus et 84 % moins de 39.

La Lettre IX, « à la manière du chapitre XIV de la Description du Gouvernement des Persans », de Chardin, compte 51 phrases, dont 16 % de moins de 10 mots, et 16 % de 25 à 53 mots, soit tendance à prédominance de phrases courtes d'une part, de phrases longues d'autre part; malgré cette tendance vers une distribution bimodale, qui se manifeste aussi dans la Lettre XXIV et dans la Lettre XCIX, la distribution observée peut encore être comparée à une distribution log normale avec 50 % de phrases comptant moins de 20 mots (graphique III).

La Lettre CXLI, « Histoire d'Ibrahim et d'Anaïs », est inspirée de l'un des contes persans traduits par Pétis de la Croix (v. M. L. Dufrenoy, *L'Orient Romanesque en France*, t. I, p. 166 et *Lettres persanes*, éd. P. Vernière, p. 302).

Des 104 phrases, 16 % ont au plus 12 mots, 50 % au plus 19, 16 % des phrases ont entre 37 et 104 mots.

Conclusions.

Selon le modèle de distribution log normale.

Les graphiques où les pourcentages cumulés des phrases comptant au moins n mots sont portés en ordonnées sur échelle de probabilité normale, les nombres n de mots par phrase étant portés en abscisses sur échelle log, permettent chacun de tracer une droite de régression définie par un paramètre de position correspondant au « quantile 50 » ou Q_{50} et un paramètre de dispersion que permettent d'estimer les quantiles 16 (Q_{16}) et 84 ou (Q_{84}). Le Tableau ci-dessous indique ces valeurs :

	Q_{16}	Q_{50}	Q_{84}
La Bétique	12	20	37
Les Troglodytes	10	20	40
« Portraits »	10	17	31
« Portrait du Guerrier »	8	15	28
Lettre IX	9	22	40
Lettre CXLI	12	19	37

Les méthodes mathématiques peuvent être appliquées à l'étude du Langage :

1° pour analyser le matériel statistique en vue du choix du modèle mathématique le plus vraisemblable pour rendre compte d'une distribution;

2° pour exprimer en langage mathématique ce qu'exprimait le langage linguistique soumis à l'analyse;

3° pour enrichir le matériau linguistique de formules mathématiques (M. V. Macavariani).

Désormais on peut distinguer, de la linguistique classique, la linguistique moderne ou mathématique, la première représentant la forme inexacte, la seconde, la forme exacte d'une même discipline (G. Herdan).

L'existence de styles divers est rendue possible par l'aptitude du langage à exprimer la même idée de diverses manières; un style peut être caractérisé par l'estimation statistique de fréquences et de probabilités.

E. Souriau, évoquant « la pensée martelant librement le fer du verbe » reconnaît que « les linguistes modernes... se refusent à établir des règles. Ce qu'ils cherchent ce sont des lois... ».

Ils continuent à deux siècles de distance les recherches qu'entreprit Montesquieu avant d'écrire l'*Esprit des Lois*, mais en disposant d'outils statistiques dont Montesquieu ne pouvait même pas soupçonner la réalisation.

Marie-Louise DUFRÉNOY

BIBLIOGRAPHIE

1849 — *Pensées et fragments inédits de Montesquieu*, publiés par le Baron G. de MONTESQUIEU, Bordeaux.

1893 — A. SHERMAN, *Analysis of Literature*, Boston.

1939 — G. U. YULE, On sentence-length as a statistical characteristic of stylein prose. *Biometrika*, 30, III & IV.

1946 — M.-L. DUFRÉNOY, Analyse statistique du Langage, *Journal Soc. Statistique*, 97, 208-218.

1956 — G. HERDAN, *Language as choice and chance*, Groningen.

1957 — WAKE, Sentence-Length distributions, *J. Roy. Statist. Soc.*, 120, 331-346.

1959 — GREXYSTON & G. HERDAN, The authorship of « The pastorals » in the light of statistical linguistics, *The New Testament Studies.*

1960 — G. HERDAN, *Type-Token Mathematics*, The Hague.
P. GUIRAUD, *Problèmes et méthodes de la statistique linguistique*, Paris.

1961 — JACOBSON, Structure of Language and its mathematical aspects. *Am. Math. Soc. Proceed. of Symposia in Appl. Math.*, XIII.

1962 — G. A. LESKIS « Longueur des phrases dans la prose scientifique et d'imagination, en Russie, vers 1960, *Voprosy jazykoznajia*, 2 (cité par M. V. Macavariani).

1963 — A.-J. GREIMAS, La linguistique statistique et la statistique structurale, *Le français moderne*, octobre 1962, 241-254 et janvier 1963, 55-68.
Solomon MARCU, Aspects ale modelärii matematica in linguisticà, *Studii si ceretàri linguistica*, 14, 4, 487-503.

1964 — M. V. MACACAVARIANI, On the relationship between mathematics and linguistics, *Linguistics*, 5 may 1964, 25-34.
G. HERDAN, Analyse de Alvar Ellegàrd : « A statistical Method for determining Authorship » (The Junius Letters). *Linguistics*, 5, 106-115.
G. HERDAN, *Quantitative Linguistics*, London.
F. PAPP, Mathematical Linguistics in the Soviet Union, *Acta Linguistica Acad. Sci. Hungar*, 14, 1,2, 119-137.

1965 — E. SOURIAU, Sur l'esthétique des mots et des langages forgés, *Rev. d'Esthétique*, N. S., 19-48.
D. S. BOOMER, Hesitation and grammatical encoding, *Language and Speech*, 8, 3, 148-158.
G. HERDAN, Lexicality and its statistical reflection *Ibid.*, 190-196.

ADDENDUM

Tableau de concordance établi par une de mes étudiantes, M^lle^ Rebecca Mehl.

Exemples de portraits insérés dans les *Lettres persanes* de Montesquieu (éd. de Paul Vernière, Garnier, Paris, 1963) et dans *Les caractères et les mœurs de ce siècle* de La Bruyère (Club des Libraires de France, 1964).

Sujet	*Lettre*	*Caractères*
Le roi	37	X, 24, 35.
L'obsédé curieux	45	XIII, 2.
Les partisans	48	VI, 13-20.
Le directeur	48	III, 45.
Le poète	48	I, 8.
Le libertin	48	VIII. 18.
Les coquettes	52	III, 8.
L'égoïste	50	XI, 121.
Fausse modestie	50	III, 46. V, 21. IX, 44.
Le décisionnaire	72	V, 9.
Les grands	74	IX (le tout), 50.
Le voyageur	87	VII, 13.
Les magistrats	98	VI, 15, 19, 20, 21.
La mode	99	XIII, 1, 12.
Les femmes	110	III.
Le géomètre	128	pas de vrai modèle sauf. XI, 7
Les nouvellistes	130 & 132	X, 11.
L'homme d'esprit	145	La Bruyère lui-même !