JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

JEAN DUFRÉNOY

La statistique au IXe congrès international de linguistique romane

Journal de la société statistique de Paris, tome 101 (1960), p. 124-130

<http://www.numdam.org/item?id=JSFS_1960__101__124_0>

© Société de statistique de Paris, 1960, tous droits réservés.

L'accès aux archives de la revue « Journal de la société statistique de Paris » (http://publications-sfds.math.cnrs.fr/index.php/J-SFdS) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
http://www.numdam.org/

La statistique au IX^{e} Congrès International de Linguistique romane

(Lisbonne, 31 mars — 4 avril 1959)

Notre Président M. Dumas nous a entretenu du discours prononcé par M. Mac Millan à la Cérémonie du Centenaire de la *Royal Statistical Society :* une telle occurrence pouvait faire apparaître une certaine espérance mathématique d'évocation de la statistique par un homme d'État.

A priori, semblable espérance ne pouvait guère être invoquée pour la Séance d'ouverture du IX^{e} Congrès de Linguistique Romane; cependant, S. Exc. le Ministre de l'Éducation Nationale, le Professeur Leite Pinto, a pris, comme thème de son discours inaugural,

cette notion fondamentale : « Les matériaux sur lesquels portent les études de Linguistique sont des phénomènes de masse, et, comme tels, doivent être soumis à l'analyse statistique. »

Ce discours inaugural a été publié en première page par les grands quotidiens de Lisbonne, avec de gros titres tels que « Le Ministre de l'Éducation, inaugurant le IXe C. I. de L. R. affirme que, comme les Ingénieurs et les Mathématiciens, les philologues peuvent bénéficier de l'utilisation des machines électroniques ».

Telle fut d'ailleurs la conclusion de la communication de Marie-Louise Dufrenoy : sous le titre *De la prodigalité à la parcimonie*, Marie-Louise Dufrenoy a présenté un parallèle et établi un contraste entre Antoine Galland, traducteur des *Mille et une Nuits*, et l' « archiviste électronique » réalisé par l'*I.B.M.* pour lire les publications techniques et en extraire un résumé de quelques mots (1).

L'intérêt que portent certains spécialistes à la résolution mécanique de problèmes linguistiques semble justifier une mise au point, pour exposer les progrès récemment réalisés, depuis l'article publié par le *Journal de la Société de Statistique de Paris* (sept.-oct. 1946).

I. Mots usuels.

Un mot usuel est celui qui se manifeste fréquemment dans un texte : les mots usuels sont peu nombreux; si, dans un texte de 100 000 vocables, on relève les fréquences de manifestation de chaque mot différent, on note que les 9 mots les plus usuels représentent un quart du texte, et 67 mots la moitié du texte : ce que représente le graphique 1.

Pour pouvoir faire figurer sur un graphique la distribution des 732 mots différents qui représentent les 3/4 du texte, il faudrait effectuer l'anamorphose de la courbe du graphique 1 par transformation logarithmique de l'échelle des abscisses et transformation de l'échelle des ordonnées en échelle de probabilité normale.

Dewey a dénombré pour un vocabulaire de 100 000 manifestations de mots, 1 027 mots différents (employés chacun au moins 10 fois); en rangeant ces mots par ordre de fréquence de manifestation décroissante on note que les 9 premiers mots de la liste (soit environ 9 % du total) rendent compte de 25 % des manifestations; les 67 premiers mots (soit environ 6,7 % du total) rendent compte de 50 % des manifestations; tandis que le premier quart (0 à 25 %) de l'échelle verticale ne correspond qu'à moins de 0,9 % de l'échelle horizontale, le second quart (25 à 50 %) correspond à (6,7 — 0,9) soit 5,8 %; et le 3e quart à 71,27 — 6,72 = 64,55 % de l'échelle.

Une analyse statistique de l'évolution de la langue roumaine a révélé à D. Macrea que les mots slaves, hongrois, néo-grec et turcs, représentant des stades dépassés de culture matérielle... ont cédé la place aux mots néo-latins porteurs de la civilisation moderne : de nos jours le vocabulaire roumain s'enrichit en termes empruntés au russe... la plupart de ces mots sont, en russe, d'origine romane, française surtout (2).

L'unité sur laquelle porte l'analyse statistique n'est pas nécessairement un mot. Par exemple, en étudiant les distributions du nombre d'apparitions de chacun des « morphèmes à grande fréquence » (qui sont surtout les désinences affixes et mots-outils) W. Manczack, de l'université de Cracovie, a pu définir le degré d'apparentement de chaque langue romane par rapport à chacune des autres (3).

---

(1) Cf. IXe Congrès International de Linguistique Romane, Lisbonne, 31 mars-4 avril 1959, *Programmes*, p. 49. Marie-Louise Dufrenoy, *De la prodigalité à la parcimonie.*

(2) Cf. IXe Congrès International de Linguistique Romane, Lisbonne, 31 mars-4 avril 1959, *Programmes*, p. 66. D. Macrea, *Les tendances actuelles dans le vocabulaire de la langue roumaine.*

(3) Cf. IXe Congrès International de Linguistique Romane, Lisbonne, 31 mars-4 avril 1959, *Programmes*, p. 68. Witold Manczack, *Le problème de la classification des langues romanes.*

II. Fréquences de manifestation d'un mot.

1. Distribution des fréquenees de 1, 2,... *n* manifestations : la loi harmonique des fréquences relatives.

Dans notre article publié en 1946 nous avions considéré la suite chronologique des *Nuits* (dans la traduction des Contes Arabes de Galland) comme une série d'épreuves, dont chacune permet à un certain mot oriental de se manifester 0, 1, 2, 3, ... fois (fréquences égales à 0, 1, 2, 3,...).

Parmi la centaine de mots représentatifs de l'Orient glanés dans les *Mille et une Nuits*, 15 n'apparaissent qu'une fois dans la suite des *Nuits*, 7 figurent 2 fois, 3 se manifestent 3 fois....

Les distributions de fréquences pour ces mots obéissent à la loi harmonique des fréquences relatives.

Dire d'un langage qu'il obéit à la loi harmonique signifie ceci :

Classons les mots du vocabulaire en plaçant en tête le plus fréquent, et ensuite ceux qui sont de moins en moins fréquemment employés : si le mot le plus fréquent revient tous les 10 mots dans un texte, le second revient tous les 20 mots, le troisième tous les 30 mots, etc...

Appliquée à l'analyse statistique du langage, la loi harmonique permet d'exprimer sous une forme précise le fait que le vocabulaire consiste en une dizaine de milliers de mots dont la plupart n'ont qu'une chance infiniment faible d'apparaître lorsqu'un auteur écrit une ligne; mais si cet auteur écrit des milliers de lignes, il peut donner à chaque mot du vocabulaire la chance de se manifester.

2. Distribution des fréquences de 0, 1, 2,... *n* manifestations : les séries de Poisson.

En 1946, nous avions publié une représentation graphique de distribution de Poisson appliquée à l'analyse statistique du texte des *Mille et une nuits*, en utilisant un abaque préparé par la *Bell Telephone Company.*

Chaque fréquence, transformée en pourcentage cumulatif du total, étant portée sur l'échelle des ordonnées, définit un niveau de probabilité : l'horizontale à chaque niveau intercepte la courbe correspondante de 1, 2, ... *n* manifestations; pour une distribution de Poisson, les points d'intersection définissent une droite verticale, coupant l'échelle des abscisses au point correspondant à la moyenne de la distribution.

Dans la plupart des cas, cependant, les points s'alignent sur une droite qui s'incline vers la droite, et d'autant plus qu'au lieu de correspondre à des tirages indépendants les uns des autres, les pourcentages cumulatifs correspondent à des « tirages contagieux » : c'est ce que nous avions obtenu dans le cas des *Mille et une nuits.*

Rares d'ailleurs sont les textes fournissant des indications sur la fréquence de « zéro manifestation ».

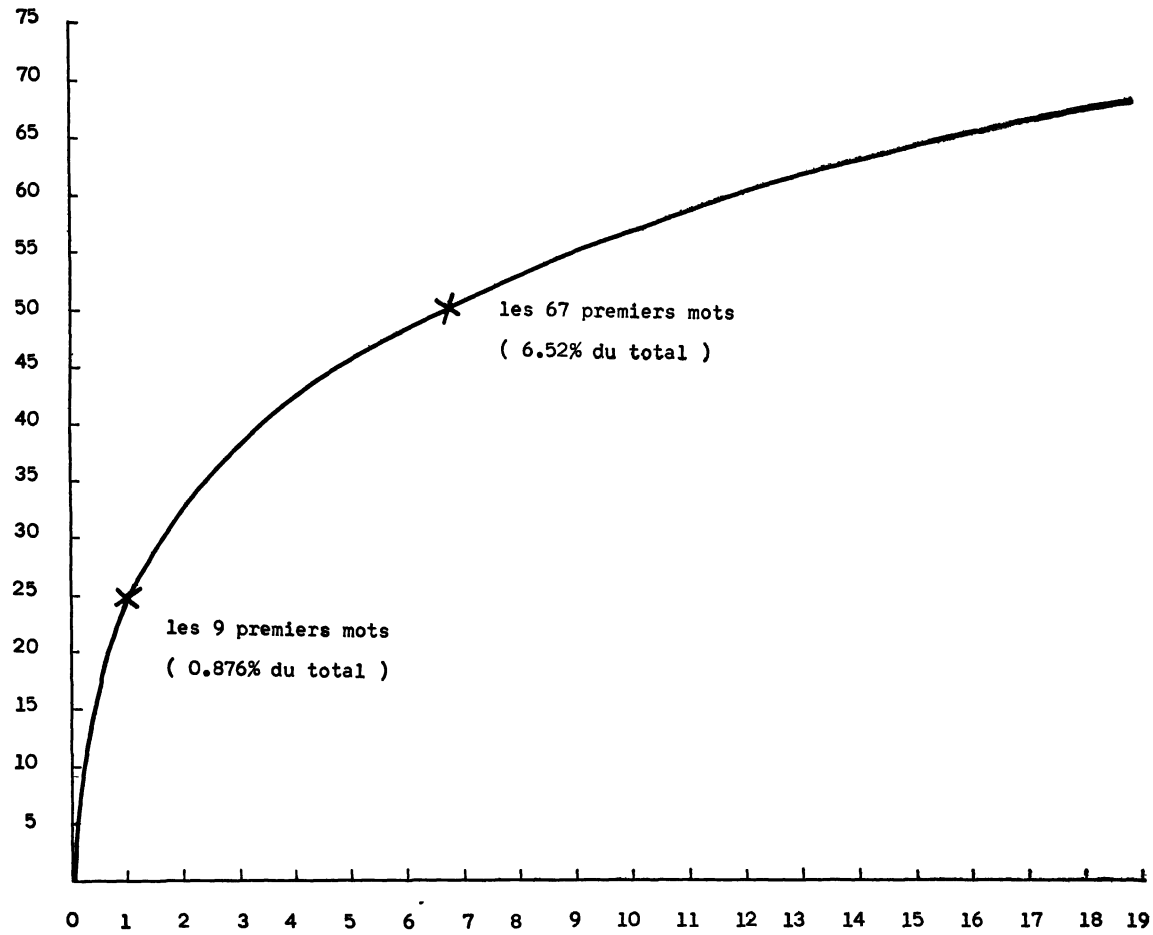
Cependant Herdan (1) a relevé pour deux auteurs, Churchill et Halifax, les distributions de 0, 1, 2, ... 36 utilisations de mots du « Vocabulaire politique ».

Herdan avait étudié une corrélation en dressant une table, indiquant : combien de mots utilisés zéro fois par l'un des membres d'une paire d'auteurs, avaient été utilisés par l'autre 0, 1, 2, ... *n* fois; combien de mots utilisés 1 fois par l'un avaient été utilisés par l'autre 0, 1, 2, *n* fois; combien de mots utilisés par l'un, n fois, avaient été utilisés par l'autre 0, 1, 2, ... *n* fois.

---

(1) HERDAN, C. *Language as choice and Chance*, Groningen, P. Noordhoff, N. V. 1956.

La table de corrélation correspond à une table de distribution de fréquences de deux variates (x) et (y). Au lieu de considérer ces deux variates comme liées à la manifestation d'un même mot chez deux auteurs A et B, on peut considérer l'une des deux variates comme un certain mot dans un texte original et l'autre comme la traduction de ce mot dans une langue différente de celle du texte original : de ce point de vue Herdan propose de représenter statistiquement les « traductions » par des « distributions bi-variates » de symboles choisis



selon un code approprié : les éléments (ou unités) du texte original, apparaissant dans l'ordre $a_1, a_2, a_3 \ldots$, possèdent des propriétés définies par $x_1, x_2, x_3 \ldots$

Les unités équivalentes du texte traduit, $\alpha_1, \alpha_2, \alpha_3$ ont les propriétés $y_1, y_2, y_3 \ldots$

A l'appariage $a_1\,\alpha_1,\ a_2\,\alpha_2 \ldots\ a_i\,\alpha_i$ on peut substituer l'appariage

$$x_1\,y_1,\ x_2\,y_2\ \ldots\ x_i\,y_i$$

et étudier une distribution de fréquence bivariate : la traduction met donc en œuvre une population bivariate de symboles de codes.

Maupertuis a proposé dans ses *Réflexions philosophiques sur l'Origine des Langues* (*Œuvres de* M. de Maupertuis, Dresde, 1752) une mise en code.

Il est regrettable que ce grand mathématicien n'ait donné qu'une ébauche d'un système qui représenterait de nos jours une géniale anticipation.

Maupertuis considérait la capacité de la mémoire humaine comme un facteur limite et ne pouvait imaginer les perspectives pratiquement sans bornes que développeraient les machines électroniques.

Dégagé de ses implications philosophiques, le système de Maupertuis demeure valable. Ce qu'un esprit humain ne peut absorber peut être emmagasiné sous forme graphique et fournir à une machine électronique.

Par un système de signes et d'indices (puissances et exposants), tous les phénomènes de n'importe quelle langue seraient susceptibles d'être mis en code et les codes ainsi constitués seraient la matière la plus adéquate à partir de laquelle pourraient s'opérer les traductions multilingues.

Le travail de mise en code reste à faire dans les différents pays, mais les pionniers de la traduction mécanique se sont mis délibérément à l'œuvre.

Les deux principes fondamentaux sur lesquels reposent les méthodes de traduction mécanique sont d'abord la loi de probabilité de récurrence d'un mot donné et le postulat selon lequel les mots se décomposent en racines et terminaisons.

Les réalisations pratiques ont démontré l'efficacité des méthodes récemment mises au point par plusieurs statisticiens qui ont conjugué leur science à celle des linguistes.

Traduction mécanique (1)

La technique de traduction mécanique est devenue applicable en pratique lorsqu'en 1955 Booth a suggéré une méthode logarithmique d'utilisation du dictionnaire : la position d'un mot dans un dictionnaire de 1 million de mots peut être définie par la machine par une vingtaine d'opérations.

1. Les mots sont classés dans le dictionnaire par ordre de fréquence d'usage : le mot le plus fréquemment employé reçoit le rang 1, le suivant le rang 2, le Γ^{e} le rang Γ. Soit f_{Γ} la fréquence de ce dernier, pour un texte suffisamment long $\Gamma \cdot f_{\Gamma} = K$, c'est-à-dire que le nombre des manifestations du mot de rang Γ est $\frac{K}{\Gamma}$, et le nombre total de mots à explorer, dans un dictionnaire de N mots, est :

$$\int_{\Gamma=1}^{N} \frac{K d\Gamma}{\Gamma} \; K \log._{e} N$$

Chaque découverte du mot de rang Γ exige donc Γ opérations de recherches; le mot apparaissant $K/_{\Gamma}$ fois, la découverte de toutes les répétitions de ce mot dans le texte exige K opérations, soit pour les N mots du dictionnaire, KN opérations; mais puisque selon la méthode logarithmique le nombre total des mots à examiner est K $\log._{e}$ N, le nombre d'opérations à effectuer par mot n'est que N/$\log._{e}$ N.

2). Une économie peut être réalisée par la méthode de « bracketing ».

Le dictionnaire comporte N mots classés par ordre croissant numérique, aux lieux 1, 2 ..., N; N étant une puissance de 2.

Un mot étant offert à la machine, celle-ci soustrait de l'ensemble correspondant à N/2 la grandeur correspondant à ce mot; si le résultat est positif, la grandeur soustraite correspond à un mot situé dans la première moitié de la liste; si le résultat est négatif, la machine doit soustraire la grandeur correspondant à celle qui représente le milieu de la seconde

(1) Mechanical Resolution of Linguistic Problems Andrew D. Booth, D.Sc. L. Brandwood, B.A. J.P. Cleave, Ph. D., B. Sc. London Butterworths Scientific Publications 1958.

moitié, entre N/2 et N/4; si le résultat est encore négatif, la localisation cherchée doit se trouver entre la grandeur correspondant à N/2 et celle correspondant à N/2 + N/4 .... Ces processus de comparaison, répétés jusqu'à repérage final, exigent $\log._2 N$ opérations.

Pour un dictionnaire de 10^4 mots, il suffira de $(4 \log._2 10)$, soit 14.

3). Une économie encore plus grande peut-être réalisée en emmagasinant dans la machine, non plus un seul dictionnaire, mais deux :

1° un « microglossaire » des mots particuliers à la technique faisant l'objet des textes à traduire (et pour lequel suffisent un millier de mots introduits sous la forme de « racine et terminaisons »).

2° une liste de γN mots communs : pour trouver l'un de ces mots la machine emploie $(\log._2 \gamma N)$ opérations; si le mot cherché ne se trouve pas parmi les γN mots communs, la machine doit ensuite explorer le microglossaire contenant

$$(1 - \gamma)\ N \text{ mots, soit en tout } \quad [\log._2 \gamma N + \log._2 (1 - \gamma)\ N]$$

Le nombre des mots à examiner étant $K \log._e N$, la machine doit en examiner dans le deuxième dictionnaire $(k \log._e \gamma N)$ et dans le microglossaire k $(\log._e N - \log._e \gamma N)$.

Concordance

A l'opération la plus simple qui puisse être confiée à une machine, celle du dénombrement des mots, succède cette opération un peu plus complexe qui exige non seulement l'enregistrement des mots du texte et leur classement par ordre alphabétique, mais encore l'enregistrement de la page et de la ligne du texte où le mot se manifeste. Un autre degré de complexité s'introduit lorsqu'on « éduque » la machine à réduire chaque mot à un type de base, de telle manière, par exemple, que toutes les formes du verbe avoir puissent figurer sous le même vocable général.

Les études de concordance, indispensables pour les recherches de linguistique et pour l'analyse stylistique, sont rendues fort aisées par l'emploi de machines qui permettent par exemple d'établir en moins d'un mois une concordance de la Bible.

Mécanisation appliquée aux résumés.

L'effort de mécanisation d'opérations de plus en plus complexes a donné tout récemment des résultats assez surprenants, exposés dans la revue *Chemical and Engineering News* du 8 décembre 1958. Sous le titre *Littérature*, un sous-titre nous apporte cette révélation : *une machine écrit des résumés.* Le système mis au point par la compagnie *IBM* (*International Business Machines*) permet à l'archiviste électronique n° 704 de lire des articles techniques et de fournir un résumé en quelques mots.

Les deux éléments d'information essentiels à considérer pour assurer la validité d'un résumé sont le nombre de manifestations des divers mots dans le texte et la proximité des mots importants et récurrents les uns par rapport aux autres.

Négligeant les articles, prépositions et conjonctions, il faut déceler les « mots-clés » qui se rapportent directement au sujet traité.

La « proximité » acquiert une grande importance; les idées qui présentent entre elles des liens étroits s'exprimeront par des mots intimement associés.

On peut donc concevoir une machine « éduquée » en vue de repérer : 1° les mots

apparaissant le plus fréquemment dans un texte; 2° les phrases où ces mots se manifestent le plus souvent et avec le maximum d'association. Ces phrases seront automatiquement extraites du texte, classées par ordre d'importance selon un code fourni à la machine, laquelle sera ainsi mise en mesure de sélectionner une, deux, trois de ces phrases.

Le texte du résumé sera donc le résultat d'une analyse statistique des propres termes de l'auteur.

Nous pouvons donc conclure avec S. Exc. le Professeur Leite Pinto que les philologues ne doivent pas hésiter plus que les ingénieurs et les mathématiciens à asservir, pour en faire les auxiliaires de leur activité intellectuelle, les machines auxquelles les dernières acquisitions de la Science et de la Technique ont conféré le pouvoir d'affranchir le chercheur des besognes pénibles et fastidieuses de dénombrement et de calcul.

Par l'annexion du domaine de la Linguistique, les Statisticiens sont invités à imaginer de nouveaux problèmes et à inventer des combinaisons originales pour les résoudre.

Marie-Louise DUFRENOY.