MATHÉMATIQUES ET SCIENCES HUMAINES

M. PETRUSZEWYCZ

L'histoire de la loi d'Estoup-Zipf : documents

Mathématiques et sciences humaines, tome 44 (1973), p. 41-56

<http://www.numdam.org/item?id=MSH_1973__44__41_0>

Math. Sci. hum. (11e année, n° 44, 1973, p. 41-56).

L'HISTOIRE DE LA LOI D'ESTOUP-ZIPF : DOCUMENTS

par

M. PETRUSZEWYCZ

RÉSUMÉ

En 1896, *V. Pareto découvrait une loi empirique qui porte depuis son nom ; il a montré que le logarithme du nombre cumulé d'individus percevant un revenu* r *supérieur ou égal à* x *est une fonction du* log *de* x *:* $\log N_{r \geqslant x} = -\alpha \log x$. *On trouva ensuite de nombreuses applications de la loi de Pareto par exemple en statistique lexicale. Dans ce domaine plusieurs auteurs contribuèrent à la mise en forme de la loi d'Estoup-Zipf : cet article présente quelques points de repère* (1912-1928-1935).

SUMMARY

In 1896 *V. Pareto discovered the empirical law which is named after him ; he demonstrated that the logarithm of the accumulated number of individuals collecting an income* r *greater than or equal to x is a function of log x, namely log* $N_{r \geqslant x} = -\alpha \log x$. *Afterwards one found numerous applications of Pareto's law, for example in lexical statistics. In this field, many authors have contributed in giving form to Estoup-Zipf's law. This article presents some benchmarks :* 1912-1928-1935.

Lorsqu'on aborde l'étude de la distribution des revenus, on rencontre tout de suite la loi de Pareto. Si l'on veut remonter aux textes originaux il suffit de lire le *Cours d'économie politique,* paru en 1896-1897, de V. Pareto (chapitre 1er : « La courbe des revenus » du Livre III : *La répartition et la consommation*). En effet aussi bien la nature de la liaison entre l'effectif cumulé des individus ayant un revenu supérieur ou égal à *r,* que sa représentation par une droite d'équation $\log N_r = -\alpha \log x$ en coordonnées logarithmiques, dans ce seul chapitre tout est mis en place. Seul manque un modèle — on en discute encore — l'auteur présentant sa découverte comme une loi empirique. Depuis cette découverte on a trouvé à la loi de Pareto de très nombreux domaines d'application entre autres celui des statistiques lexicales. En effet, une loi souvent nommée loi d'Estoup-Zipf, après B. Mandelbrot [1], ou seulement loi de Zipf, établit une relation entre le rang des mots d'un

texte ordonnés par ordre décroissant de fréquences d'apparition et cette fréquence. Cette double dénomination éveille la curiosité ; retracer la genèse de la loi d'Estoup-Zipf appartiendrait à un historien des sciences mais, dans l'attente d'une recherche plus systématique, il a paru intéressant de présenter ici quelques jalons de cette histoire, certains des textes reproduits, commentés ou utilisés étant d'un accès difficile. Il est par contre assez facile de rencontrer de bons exemples d'application aux statistiques lexicales de la loi d'Estoup-Zipf ; citons par exemple « L'élaboration du français fondamental (1er degré) » [2 : p. 125-127-134]. La fréquence est pour le linguiste ce qu'est le revenu pour l'économiste, c'est la variable dont on étudie la distribution entre les unités de la population statistique considérée.

C'est par la notion de fréquence, ou plutôt celle de tables de fréquence que sténographie et cryptographie sont à l'origine de la statistique lexicale ainsi que le rappelle B. Mandelbrot [1]. Pour le cryptologue, les « tableaux montrant, pour une langue donnée, les fréquences relatives des lettres, bigrammes, trigrammes, syllabes ou mots dans un texte normal... » [3] sont les outils nécessaires et suffisants pour son travail. Les premiers tableaux que nous ayons conservés parurent dans le *Liber Zifrorum* de Cicco Simonetta, conseiller des Sforza entre 1375 et 1383 : ils concernaient les bigrammes et les trigrammes. Beaucoup plus tard, en France, dans les années 1890, le Commandant Bazeries considéra les 587 groupes différents qu'on distinguait dans quelques messages jusque-là non déchiffrés, mais dont on savait qu'ils étaient écrits dans le Grand Chiffre de Louis XIV, composé à l'époque du siège de La Rochelle par le célèbre Antoine Rossignol et dont la clé était perdue depuis des siècles. Il dressa les tables de fréquence des syllabes de la langue française et de proche en proche reconstitua tout le Grand Chiffre. C'est tout ce que nous dirons de la cryptographie à laquelle il ne semble pas que G.-K. Zipf se soit intéressé mais en revanche il a plusieurs fois utilisé les recherches des sténographes de son époque et c'est lui qui nous a transmis le nom de J.-B. Estoup.

I. L'APPORT DE J.-B. ESTOUP (1868-1950)

G.-K. Zipf cite, à notre connaissance, trois fois le nom de J.-B. Estoup. La première dans une publication qu'il donne lui-même ultérieurement comme sa première prise de position publique [4] : « Relative frequency as a determinent of phonetic change ». Cet article, paru dans le volume XL des Harvard Studies in classical philology en 1929, contient une note pp. 2-3, où l'auteur rend hommage aux personnes lui ayant recueilli l'énorme documentation phonétique qu'il a utilisée. La liste, assez longue, débute par le nom d'un professeur de Dresde, sténographe, et se termine par celui de « M. J.-B. Estoup, Paris ». Il ne semble pas que les descendants de M. Estoup aient gardé le souvenir de cette collaboration dont on ne sait sous quelle forme elle a été établie. Par ailleurs il est certain, puisqu'il le cite, que Zipf connaît G. Dewey [5] auteur de décomptes phonétiques et autres qui, pour présenter son livre se donne le titre de « author of personnel shorthand, demotic shorthand, Dewey Shorthand, etc. ». Zipf donc eut l'idée de s'adresser à des sténographes pour obtenir les données nécessaires pour vérifier son « Principe de fréquence : l'accent ou degré de difficulté de tout mot, syllabe ou son, est inversement proportionnel à la fréquence relative de ce mot, cette syllabe ou ce son, parmi les autres mots, syllabes ou sons dans le discours. Plus l'usage est fréquent, moins la forme est accentuée c'est-à-dire plus facile à prononcer, et vice versa. »

Cependant ce ne sont pas les travaux d'Estoup que Zipf retient pour représenter le français dans cet article mais ceux de Dujardin, repris d'ailleurs dans Karl Faulmann [1]. De toutes

1. K. Faulmann, *Historische Grammatik der Stenographie*, Vienne, Bermann und Altmann, 1887 ; F. Dujardin, *Journal des Connaissances usuelles*, 1834. L'auteur de cet article n'a pas pu consulter ces ouvrages.

façons il ne s'agit que de statistiques de lettres ou de sons : consonnes, voyelles nasalisées, voyelles et diphtongues, dans le français de l'époque, dans le but de fonder un système de sténographie. G.-K. Zipf assure avoir confronté ces données avec celles d'un autre auteur de système sténographique : J.-J. Thierry-Mieg [7]. Ces deux auteurs sont d'ailleurs cités par P. Guiraud [8, p. 10], comme les précurseurs des travaux phonologiques français en tant que sténographes.

De nombreux auteurs avaient inventé d'aussi nombreux systèmes sténographiques depuis la renaissance au XVII[e] siècle en Angleterre de cette technique qui remonte à l'Antiquité classique et qui permet de « noter la parole aussi vite qu'elle est émise ». Bien qu'il n'ait pas laissé son nom à la méthode qu'il avait apprise et qu'il a si fortement contribué à améliorer sous le nom de Métagraphie Duployé, l'œuvre de J.-B. Estoup est bien différente de celle des nombreux auteurs de système sténographique. Esprit curieux de toutes les acquisitions de la recherche contemporaine dans la mesure où cela lui permettait d'étendre la diffusion de la sténographie, pédagogue né, épris de précision, il instaura un enseignement de la sténographie qui permet d'accéder aux vitesses de sténographie de discours d'une manière rationnelle. Tous les sténographes se forgeaient des sténogrammes abrégés pour les mots qui revenaient *souvent* : J.-B. Estoup, le premier, va donner un sens précis à cet adverbe et fonder la progression de son enseignement et le vocabulaire de ses gammes sur les fréquences d'apparition des mots usuels. D'ailleurs P. Guiraud [8, p. 50] ne s'y trompe pas qui classe J.-B. Estoup sous la rubrique : « *Examen critique* des listes de fréquence, leur valeur pédagogique, leur établissement et leur emploi ».

J.-B. Estoup a édité une Méthode et deux ouvrages de Gammes : Gammes de cinquante à cent mots par minute, Gammes de cent vingt-cinq à cent quarante mots ; ces différents livres ayant connu plusieurs éditions modifiées. L'ouvrage « Gammes sténographiques, méthode et exercices pour l'acquisition de la vitesse », dont la première édition est de 1907 si l'on en croit les « Éphémérides sténographiques de 1898 à 1908 » recueillis par l'« Histoire générale de la sténographie et de l'écriture à travers les âges » [9], pose quelques problèmes. D'abord le sous-titre a pu varier mais surtout il est difficile de préciser son contenu, les exemplaires que l'on peut consulter à la Bibliothèque Nationale de Paris étant tous incomplets car c'est un ouvrage en fascicules. Il semble que, sinon à l'origine, du moins dans la 4[e] édition de 1916 existait un fascicule précédant ceux des gammes proprement dites, intitulé *Exposé théorique.* C'est cette 4[e] édition que Zipf aurait eu en mains, en la consultant peut-être à la Bibliothèque Municipale de New York où B. Mandelbrot l'a consultée lui-même, et qui l'amenait à dire en 1946, deuxième citation, dans son article « The psychology of language » [10] : « En étudiant le problème sténographique, J.-B. Estoup (*g.st.*, 4[e] éd., 1916) a observé la relation en gros hyperbolique entre le nombre des mots nouveaux et différents dans des tranches consécutives de 1 000 mots de français d'une part, et l'effectif cumulé des mots d'autre part ». P. Guiraud reprendra ce commentaire en l'atténuant car il supprime hyperbolique et parle de relations dans sa référence 5 C. 2 : Estoup, *Gammes sténographiques* [8, p. 50].

Pour notre part, il nous a été possible de noter :

• que les publicités parues au verso des couvertures de diverses publications sténographiques détaillent le contenu de la 5[e] édition et font état d'une « Statistique des mots usuels » ;

• que la 3[e] édition 1912, intitulée elle : « *Recueil de textes choisis pour l'acquisition méthodique de la vitesse* », ne contient pas les textes que nous reproduisons plus loin mais affirme, dans l'Introduction : « Le nombre des mots employés par l'orateur et surtout par l'improvisateur est bien loin d'être illimité. On a évalué à un maximum de 3 000 le nombre de mots employés par une personne cultivée. L'orateur en a, à sa disposition, une quantité bien moindre. » On verra que cette affirmation impliquait une statistique et son interprétation.

L'obligeance des descendants de J.-B. Estoup nous a permis d'avoir en main la 7e édition, s.d., du fascicule *Exposé théorique de la méthode pour l'acquisition de la vitesse.* Nous présentons, en hommage à la mémoire de J.-B. Estoup, la reproduction *in-extenso* du paragraphe : Le nombre et la fréquence des mots usuels (pp. 21 à 23) et l'Annexe II : Les mots usuels, leur nombre et leur fréquence (p. 51 à 57).

LE NOMBRE ET LA FRÉQUENCE DES MOTS USUELS

[...] Sans doute, le nombre total des mots d'une langue est très grand. Au dire de certains philologues, on n'en compterait pas moins de 90 000 dans la langue française. Mais il s'en faut que tous soient d'un usage courant. Il en est un petit nombre qui reviennent à tout instant et qui forment comme une petite troupe active, toujours en avant, toujours prête à servir, tandis que les autres constituent des réserves et même d'immenses territoriales rarement dérangées.

L'orateur, surtout celui qui improvise, le seul à qui le sténographe ait réellement à faire, ne dispose que d'un nombre de mots assez restreint.

Il nous a paru intéressant d'établir, avec quelques précisions, la liste et le nombre de ces mots d'usage courant et de les classer par ordre de fréquence. A cet effet, des statistiques ont été dressées, portant sur des textes assez longs. Nous avons compté : 1) combien de fois, sur un texte d'une longueur totale de 20 000 mots, sont répétés les vocables les plus fréquents ; 2) combien un texte de longueur totale de 30 000 mots comprend d'expressions différentes.

Voici, résumés, les résultats généraux de ces décomptes :

1) *Combien de fois sont répétés les mots les plus fréquents ?*

Sur un texte d'une longueur totale de 20 000 mots, les articles *le, la, les* (1 950 fréquences), *du, de, des* (1 700 fréquences), les particules conjonctives *que, à, et* (750 fréquences), quelques autres expressions très générales (*un, il, ne, en*), en tout une douzaine d'expressions, présentent ensemble, en chiffres ronds, 8 000 fréquences, soit 40 % du nombre total des mots.

16 particules (auxiliaires *être* et *avoir, dans, par, faire,* etc.) présentent chacune de 100 à 200 fréquences, c'est-à-dire se rencontrent en moyenne une fois ou deux tous les 200 mots.

22 autres expressions présentent de 40 à 100 fréquences, c'est-à-dire se rencontrent une fois ou deux tous les 500 mots.

55 présentent de 20 à 40 fréquences, c'est-à-dire se rencontrent une fois ou deux tous les 1 000 mots.

150 présentent de 10 à 20 fréquences, c'est-à-dire se rencontrent une fois ou deux tous les 2 000 mots.

280 présentent de 5 à 10 fréquences, c'est-à-dire se rencontrent une fois ou deux tous les 4 000 mots.

Ces 540 vocables, que l'on peut qualifier de généraux, forment par leurs fréquences les 85 % du texte entier.

2) *Combien le discours comprend-il de mots différents ?*

Le texte, d'une longueur totale de 30 000 mots, a été préalablement divisé par tranches de 1 000 mots.

Nous avons compté, successivement, dans chaque tranche, le nombre de mots différents et non rencontrés déjà dans les tranches antérieures. Les chiffres ainsi établis montrent combien est rapide la décroissance du nombre des mots différents et nouveaux.

Dès le deuxième mille, leur proportion par rapport à ceux déjà vus, est seulement de 23 % ; au 10e mille cette proportion tombe à 9 % ; au 30e mille, à 4 %.

En totalisant les nombres des vocables différents, rencontrés dans chaque tranche, on constate que 2 870 ont suffi pour former le texte entier de 30 000 mots. Si la statistique était poussée plus loin et poursuivie, par exemple, jusqu'au 60e mille, la proportion des mots nouveaux décroissant toujours et tombant à la fin à 1 % et peut-être plus bas encore, ce seraient à peine 650 mots nouveaux à ajouter aux 2 780 précédents, soit, pour un texte de 60 000 mots, 3 500 expressions d'où il convient de défalquer, pour le calcul qui nous occupe, 700 ou 800 mots qui ne sont pas d'un usage courant. Or, parvenus à ce point, nous pouvons être sûrs d'avoir épuisé le vocabulaire entier qui constitue la trame de tout discours.

On peut donc affirmer, sans crainte d'erreur, que le nombre des mots couramment employés par un orateur n'est pas supérieur à 3 000.

Tel est également le nombre des signes sténographiques à apprendre jusqu'au point de les rendre automatiques.

ANNEXE II

LES MOTS USUELS

Leur nombre et leur fréquence

Le nombre des mots d'une langue comme la nôtre est certainement très élevé. Il est difficile de donner un chiffre même approximatif. Il y a des dictionnaires de 80 000, de 100 000 mots et plus. Mais il s'en faut que tous soient d'un usage courant. Max Muller évalue à 3 000 ou 4 000 le nombre des mots employés habituellement par un anglais cultivé.

Il peut être intéressant à divers points de vue, pour les sténographes, d'avoir des précisions à cet égard, de connaître non seulement combien de mots sont d'un usage courant, forment la trame ordinaire du discours, mais encore quels sont ces mots, et quel est leur ordre de fréquence.

J'ai fait, dans ce but, une statistique [1] qui a porté sur un texte d'une longueur totale de 30 000 mots. Afin que l'expérimentation fut très probante, j'ai eu soin d'opérer non sur un texte suivi d'un même auteur traitant un même sujet, mais sur des extraits de discours et d'articles (75 extraits) [2] portant sur les sujets les plus variés, tout en restant cependant dans le style oratoire ou épistolaire qui seul nous intéresse.

J'ai d'abord découpé le texte entier en tranches de 1 000 mots. J'ai recherché, en considérant successivement chaque tranche :

1) Combien chaque tranche contient de mots différents et ne figurant pas dans les tranches précédentes.

2) Combien de fois revient chaque mot répété.

En procédant ainsi, j'ai compté dans la première tranche 336 mots différents.

Dans la deuxième, j'ai compté 233 mots différents et ne figurant pas dans la première.

J'ai opéré de même successivement pour les tranches suivantes, comptant toujours combien y sont contenus de mots différents et nouveaux, c'est-à-dire ne figurant pas dans les précédentes.

Voici le résultat de ces décomptes (voir p. 46) :

1. Une autre statistique a été faite dans le même ordre d'idée par un de nos dévoués et excellents professeurs, M. Touzeau. Son travail corrobore en tous points les résultats que nous publions ici.

2. Gammes à 110, 115, 120, 125 mots.

Tranches	*Nombres de mots différents et nouveaux*	*%*	*Moyennes sur 4 tranches %*
1re	336	33,6	
2e	233	23,3	
3e	201	20,1	
4e	152	15,2	
5e	134	13,4	
6e	121	12,1	
7e	105	10,5	
8e	92	9,2	9,30
9e	87	8,7	
10e	88	8,8	
11e	89	8,9	
12e	87	8,7	8,10
13e	66	6,6	
14e	82	8,2	
16e	66	6,6	
15e	50	5,0	7,20
17e	105	10,5	
18e	67	6,7	
19e	58	5,8	
20e	59	5,9	
21e	81	8,1	5,90
22e	43	4,3	
23e	44	4,4	
24e	56	5,6	
25e	39	3,9	5,20
26e	69	6,9	
27e	51	5,1	
28e	30	3,0	
29e	56	5,6	4,32
30e	36	3,6	
Total	2 172		

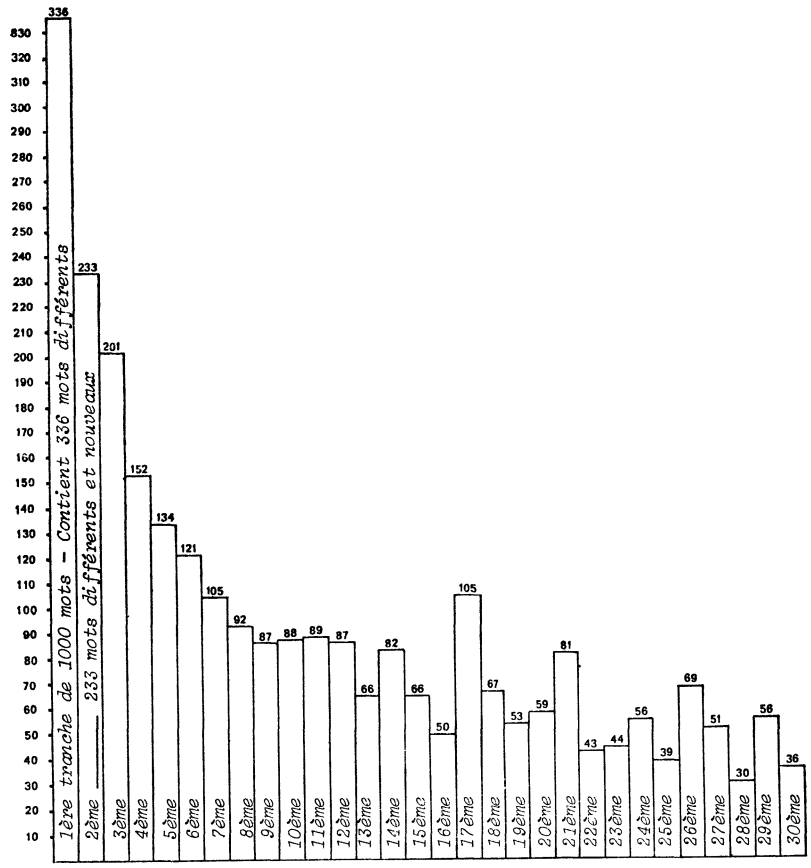


Tableau graphique montrant la courbe décroissante du nombre des mots différents et nouveaux rencontrés successivement dans chaque tranche de 1 000

Ce graphique montre clairement que le pourcentage des mots différents et nouveaux trouvés dans chaque tranche décroît très rapidement au fur et à mesure que l'on avance dans le décompte. Dans le 2e mille la proportion des mots nouveaux est de 23 %, dans le 10e mille elle est d'environ 9 %, dans le 20e mille de 6 %, dans le 30e mille de 4 %.

Il n'est pas nécessaire d'affirmer que si l'on poursuivait le décompte jusqu'à un 50e ou à un 60e mille, la proportion descendrait à 1 % ou peut-être à moins, c'est-à-dire qu'à ce point on aurait épuisé, à peu près, le vocabulaire entier qui constitue la trame habituelle de tout discours [4].

Pour former le texte entier d'une longueur de 30 000 mots, 2 780 vocables ont suffi.

Si le décompte était poursuivi jusqu'au 60e mille, nous pouvons admettre que l'augmentation par tranche de 10 000 mots serait la suivante :

Tranche du 31e au 40e mille	3 %,	soit 300 mots
Tranche du 41e au 50e mille	2 %,	soit 200 mots
Tranche du 51e au 60e mille	1,50 %,	soit 150 mots
Total		650 mots

4. Les textes des Gammes, première et deuxième parties, comprenant des séries d'exercices depuis la vitesse de 50 mots jusqu'à celle de 140 mots inclus, contiennent environ 65 000 mots.

Ainsi, ces 30 nouvelles tranches de mille ne doivent pas compter plus de 650 mots différents et nouveaux. Additionnés aux 2 780 précédents, nous obtenons un chiffre de 3 430, disons en nombre rond 3 500 mots différents, pour un texte d'une longueur totale de 60 000 mots.

Examinons maintenant l'ordre de fréquence de ces mots.

Ils se rangent ainsi : (le décompte ne porte plus ici que sur 20 000 mots).

	Nombre des répétitions		Nombre des répétitions
Le, la, les	1 949	bien	53
de, du, des	1 712	industrie	50
que, qui	766	français	48
à, au	743	certain	44
et, est, si	741	grand	44
un, une	413	avec	41
ce, se	393	sans	40
il	282	venir (et temps)	39
ne	279	quel	39
en	258	commerce	38
ces, ses, cette	232	nouveau	37
être (et temps)	200	très	37
dans	181	vouloir (et temps)	37
pas	179	droit	36
nous	167	autre	36
plus	162	produit	36
pour	157	environ	35
par	155	falloir (et temps)	35
on	155	voir (et temps)	35
son, sont	141	donner (et temps)	34
je	137	savoir (et temps)	33
faire (et temps)	136	nation	32
tout, toute	119	développer	31
avoir (et temps)	118	économique	31
nos	115	messieurs	30
vous	108	sa	30
ou, où	100	France	29
leur	94	quelque	29
si	93	trouver (et temps)	29
elle	92	intérêt	29
même	84	aller (et temps)	28
sur	82	lui	27
mais	82	nombre	27
y	70	point	26
pouvoir (et temps)	66	public	26
celui, celle, ceux	62	année	25
dont, donc	62	moins	25
devoir (et temps)	60	demander	25
travailler (et temps)	60	tant, temps	24
peut, peu	58	chose	23
dire	57	État	23
comme	56	me	23

Sont répétés 22 *fois :*

depuis, deux, enseigner, moyen, ni, toujours.

Sont répétés 21 *fois :*

considérable, entre, petit, premier.

Sont répétés 20 *fois :*

bon, consommer, général, homme, non, obliger, possible.

Sont répétés 19 *fois :*

actuel, fabriquer, jour, jusque, loi, mieux, question.

Sont répétés 18 *fois :*

cela, connaître, industriel, qualité, richesse, seul.

Sont répétés 17 *fois :*

arriver, besoin, croire, effort, école, esprit, importation, mesure, parler, vin, vos.

Sont répétés 16 *fois :*

aujourd'hui, cause, chaque, effet, mon, ouvrier, parce que, politique, surtout, vie.

Sont répétés 15 *fois :*

ainsi, alors, budget, chambre, devenir, élever, goût, jeunesse, marché, ministre, personne, porter, prix, prime, société.

Sont répétés 14 *fois :*

car, commission, emploi, lorsque, orient, partie, passer, permettre, prendre, progrès, situation, spécial, vraiment.

Sont répétés 13 *fois :*

cas, différent, étranger, exister, formuler, important, mettre, nécessaire, perdre, rendre, concurrence, eux, facile, heure, lieu, monnaie, particulier, tel.

Sont répétés 12 *fois :*

aucun, aussi, compter, compagnie, condition, diminuer, énergie, époque, étude, exemple, force, fournir, million, penser, préparer, priver, prospérité, quand, résultat, rester, souvent, tenir, technique, usine.

Sont répétés 11 *fois :*

agriculture, après, appel, apporter, atteindre, augmenter, beaucoup, colonie, crise, dernier, fortement, Gouvernement, jamais, largeur, longtemps, manquer, marchand, moment, monde, nécessité, place, pratique, présent, principal, représenter, République, servir, simple, sous, sort, trop.

Sont répétés 10 *fois :*

d'ailleurs, alimenter, assurer, charger, chez, client, culture, démontrer, défendre, direct, douane, enfin, essayer, façon, grave, idée, immédiat, matériel, moral, repos, œuvre, presque, profession, rapport, service, suivre, utile, ville, vue.

Sont répétés 9 *fois :*

acheter, article, actif, autant, avantage, beau, chiffre, comprendre, continuer, côté, démocratie, devant, dépense, dix, entreprise, exact, fer, dinance, guère, haut, loin, méthode, multiplier, raison, rien, siècle, somme, suivant, vers, voyage.

Sont répétés 8 *fois :*

bénéfice, but, commencer, constater, contre, cours, créer, disparaître, égal, entrer, entendre, épuiser, exporter, extrême, famille, finir, frapper, groupe, grâce, huile, ici, indigène, intelligence, marché, or, partout, Paris, parmi, peuple, protéger, puisque, rapide, recherche, reconnaître, revenu, révolution, suite, toucher, trois, valeur.

Sont répétés 7 *fois :*

action, adresse, admettre, affaire, agir, Allemagne, Allemand, Amérique, apparaître, apprenti, assez, atelier, avenir, cependant, chemin, civiliser, colon, combien, constituer, déjà, dessin, divers, doute, échange, entier, établir, extérieur, fois, fromage, garder, houille, impôt, instruction, instituer, justice, liberté, long, lutte, mai, manger, maison, naturel, opération, organiser, patron, pendant, pièce, point de vue, près, preuve, procéder, producteur, profond, propre, ressource, réserver, retard, rival, rôle, social, supposer, véritable, volonté, wagon.

Puis viennent :

62 mots répétés 6 fois [5] ;
82 mots répétés 5 fois ;
131 mots répétés 4 fois ;
194 mots répétés 3 fois ;
329 mots répétés 2 fois ;

Enfin 922 mots ne se présentent qu'une fois.

---

5. La liste n'a plus d'intérêt.

Remarques

D'abord en ce qui concerne la terminologie, *mots, expressions, vocables* sont tour à tour utilisés pour désigner, soit les occurrences, soit les lexèmes. En fait si l'on se reporte aux listes précises des unités de texte ayant la même fréquence d'apparition on s'aperçoit que J.-B. Estoup présente ses relevés après avoir effectué une lemmatisation, au moins partielle : il compte sous la forme de l'infinitif toutes les formes fléchies d'un verbe ; mais aussi en sténographe, il groupe par exemple, *ces ses, cette* et *dont, donc.* Par ailleurs il donne bien les relevés par fréquence décroissante mais n'affecte pas un rang c'est-à-dire un effectif cumulé si l'on tient compte des *ex-æquo,* aux unités ou classes d'unités de même fréquence.

Le graphique hyperbolique traduit ce que les linguistes appelleraient actuellement l'apport lexical. J.-B. Estoup paraît donc être le premier à avoir mis en évidence cette notion très intéressante du point de vue stylistique mais qu'il interprète en sténographe. Elle lui permet de poser sa thèse : la sténographie doit être un système de signes unique à un seul degré. En cela il s'opposait aux tenants d'un système à deux degrés comportant des signes conventionnels de tracé plus rapide pour les mots fréquents. Cette unicité est le pivot de sa méthode. Reçu au concours de sténographe de la Chambre en 1896, il s'était entraîné sur les textes des journaux officiels, pratiquant donc les vocabulaires des orateurs de la Chambre. Il évoque dans une préface de 1912 « quinze années de pratique de l'enseignement » ce qui correspond bien à la période écoulée depuis 1896 et on a vu que si le principe même n'est pas formulé dans cette édition, les connaissances qui fondent la méthode y sont présentes. Au cours de ces quinze années il avait sélectionné, classé, comparé tous les discours publiés par le Journal officiel, et donc du domaine public. Ces textes, plus ou moins modifiés par ses soins pour avoir un contenu fermé, une longueur adéquate, ont été constitués par lui en gammes. S'il n'y avait dans les deux premières éditions aucune formulation même incomplète de la relation, nous avons la preuve que les calculs étaient faits lors de la 3^{e} mais il faut attendre la 4^{e} pour que les calculs et graphiques soient présentés. Il ne semble pas qu'ils aient été modifiés lors des éditions ultérieures. Ce n'est qu'au moment de la codification de la méthode Duployé en Métagraphie Duployé qu'avec l'aide de son fils J.-H. Estoup seront vérifiées les fréquences et qu'il y aura quelques ajouts. M. J.-H. Estoup, dans le cadre des recherches qui le conduiront à l'invention du télétype, fit des études au niveau des lettres et bigrammes, trigrammes, le télétype comportant à son clavier des touches de deux ou trois lettres.

Enfin Zipf citera une troisième fois Estoup dans son œuvre la plus connue [11], publiée en 1949. Dans la note 4 au chapitre II [11, p. 546], il affirme : « La première personne (à ma connaissance) à avoir remarqué la nature hyperbolique de la fréquence d'usage des mots fut le sténographe français. J.-B. Estoup qui effectua des études statistiques sur le français, *cf.* ses gammes sténographiques, Paris, 4^{e} éd. 1916 (je n'ai pas vu les éditions antérieures) ».

2. LES ÉTAPES DE G.-K. ZIPF (1902-1950)

Les publications de G.-K. Zipf sont nombreuses. Ne citant que deux articles de cet auteur et quelques chapitres de ses deux ouvrages [10, 11], on ne prétendra pas présenter ici un compte rendu exhaustif de sa pensée, même dans le domaine étroit de la statistique lexicale, et ce qu'on en dira ne permet pas de mesurer l'étendue de ses recherches ni de connaître les nombreux résultats qu'il a présentés.

C'est dans une approche phonétique des langues que Zipf prend conscience de l'existence d'une « puissante loi du langage », mais on a vu que les données ont été relevées par des sténographes essentiellement. Et ceux-ci, ainsi que le note P. Guiraud [8, p. 6] et comme nous l'avons vu pour J.-B. Estoup, ne différencient pas très nettement les divers niveaux de segmentation possibles d'un texte.

Dans son premier article [4] Zipf énonce cette loi, ou *Principe de fréquence relative,* cité textuellement p. 42, de la façon la plus générale mais ne l'appuie que de données phonétiques. Sans changement le Principe est aussi exprimé dans le deuxième article [12], mais là Zipf étend sa démonstration à la segmentation en « mots-formes ». La différence essentielle entre les deux étapes n'étant d'ailleurs pas là mais dans la formalisation de la distribution et sa représentation graphique.

En [4], l'auteur exprime la distribution par la formule :

$Y = \frac{k}{X}$ où Y est le degré de difficulté ou complexité,

X est le nombre d'occurrences ou fréquence d'apparition,

k une constante.

Zipf rattache, sans entrer dans le détail, cette distribution à la loi de Weber-Fechner dont il rappelle qu'elle « agit logarithmiquement ». Mais c'est en coordonnées arithmétiques qu'il présente la distribution, sous forme d'hyperbole ; il n'explicite pas d'ailleurs, et il lui en sera fait reproche, comment il gradue l'axe des Y sur lequel figure un ordre et non une mesure. Rappelons que les données françaises sont celles de Dujardin.

Toute référence à la loi de Weber-Fechner a disparu de [12] alors qu'apparaissent des graphiques bi-logarithmiques. La droite d'ajustement est alors traduite par la formule :

$a\ b^2 = k$ où a est l'effectif des classes de fréquence (mots de même fréquence),

b est le nombre des occurrences ou fréquence d'apparition,

k une constante.

Les ordonnées sont ici les occurrences, c'est-à-dire la fréquence, car répondant aux critiques faites à sa première présentation l'auteur déclare que malgré les progrès de la phonétique il ne sera sans doute jamais possible de mesurer la complexité attachée à tel son, même relativement aux autres. Mais, dit-il, si le Principe de fréquence est vrai, «on peut s'attendre à observer un plus grand nombre d'occlusives non voisées que d'occlusives voisées », par exemple, et ses données le vérifient. Il adopte donc les fréquences pour graduation de l'axe des Y et met en abcisses les effectifs correspondants. Parmi les exemples traités deux sont particulièrement intéressants. D'abord le chinois de Pékin, car cette langue est étudiée à deux niveaux différents de segmentation : les syllabes et les mots. L'idée de confronter les deux niveaux est intéressante mais le fait qu'il s'agisse du chinois rend l'exemple même peu accessible. Mais il y a aussi les mots-formes de Plaute dans quatre de ses comédies, pour lesquels l'auteur présentera ultérieurement [13] un véritable ajustement parétien.

Que peut-on dire en général des ajustements qu'il présente ? Qu'ils sont satisfaisants mais pour la partie de la distribution qu'ils représentent. En effet, l'auteur tronque ses distributions pour vérifier la formule. Par exemple pour les formes de Plaute il ne représente que les fréquences 1 à 47, les hautes fréquences : 48 à 514 (89 formes) étant exclues pour une raison dont on reparlera plus loin car elle touche à un caractère propre aux distributions lexicales par rapport aux distributions parétiennes en général : le caractère discret des graduations des axes. Zipf ne l'aborde

ici que dans un sens. Prendre en compte ces hautes fréquences l'obligerait, d'après sa formule, à admettre l'existence de « mots fractionnaires ». En effet, k ayant une certaine valeur pour un texte déterminé, si l'on considère un mot très fréquent, un article par exemple comme Zipf lui-même donne l'exemple de *the* dans les données d'Eldrige, on a : $a\,(4\,290)^2 = 4\,200$, a, effectif du mot apparaissant 4 290 fois représente 0,000025 d'un mot, « notion tout à fait absurde quelle que soit la définition du mot que l'on donne » [13, p. 43].

Les graphiques paretiens, caractérisés par la présence d'effectifs cumulés sur l'un des axes se trouvent pour la première fois chez Zipf dans son livre de 1935 : *The psychobiology of language* [13]. Alors que Pareto avait présenté d'emblée les effectifs cumulés par rapport à un niveau donné de revenu distribué, dans le domaine linguistique il faut intercaler une procédure que nombre des premiers lecteurs a considéré comme un subterfuge [1] : l'affectation d'un rang aux unités rangée par ordre décroissant des fréquences. Considérer le rang et non l'effectif c'est se ramener à étudier la relation « [...] a une fréquence $f \geqslant x$ ». En effet, si on considère le mot le plus fréquent et qu'on lui affecte le premier rang, on peut écrire :

Rang	Effectifs de classe de fréquence	Fréquence	$f_i n_i$	Effectifs cumulés
1	$n_1 = 1$	f_{MAX}	$f_{\text{MAX}} \times n_1 = N_1$	$n_1 = 1$ a une fréquence $\geqslant f_{MAX}$
2	$n_2 = 1$	f_2	$f_2 \times n_2 = N_2$	$n_1 + n_2 = 2$ ont une fréquence $\geqslant f_2$
i	$n_i = p$	f_i	$f_i \times n_i = N_i$	$\sum_{s=1}^{i} n_s = i$ ont une fréquence $\geqslant f_i$
j	n_j	$f_j = 2$	$2 \times n_j$	$\sum_{s=1}^{j} n_s$ ont une fréquence $\geqslant 2$
$R = V$	n_R	$f_{min} = 1$	$1 \times n_R$	$\sum_{s=1}^{R} n_s$ ont une fréquence $\geqslant 1$

V = nombre des unités de texte différentes, c'est le vocabulaire

n_R = nombre des apax

$\sum_{i=1}^{R} N_i = T$ c'est le nombre total des unités constituant le texte $p \neq i$, $p,i \in [1, 2, 3, \ldots$

Cette procédure lui a été, dit-il [13, p. 44] suggérée par un ami. Il nous paraît possible de supposer que celui-ci puisse être Alan N. Holden, cité comme critique [4, p. 2], ingénieur à la Bell Telephone Company. Cette célèbre entreprise est le lien que nous voyons entre Zipf [2] et

1. Pour les lecteurs qu'inquiéterait le problème, voir par exemple B. Mandelbrot [1].

2. Zipf citera d'ailleurs d'autres travaux exécutés par d'autres ingénieurs, de la même compagnie, dans ce domaine de l'étude des fréquences relatives de sons en anglais parlé [14, p. 316].

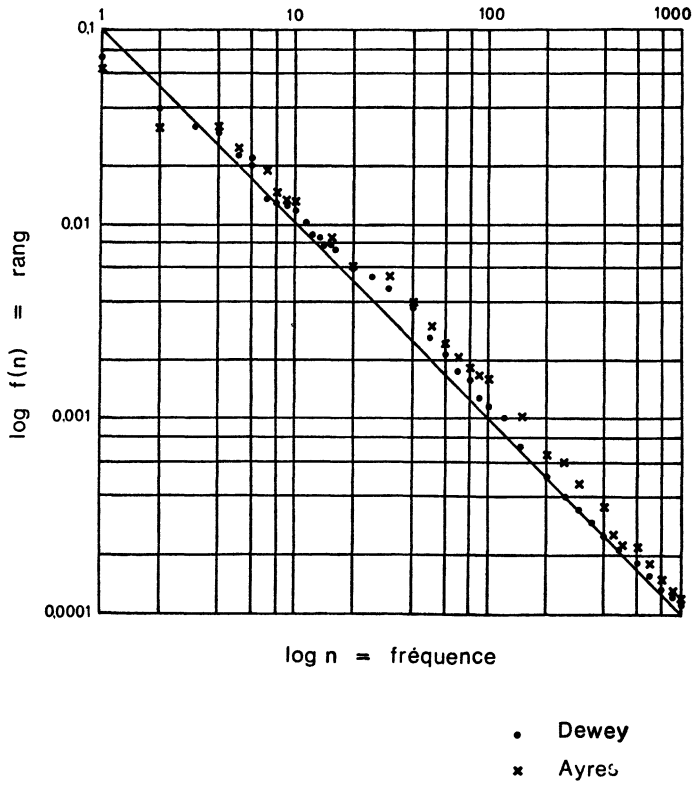
E.-V. Condon, autre ingénieur de la même com pagnie. Du moins l'était-il en 1928 quand il adressa une lettre à la revue *Science* [14], organe officiel de l'American Society for the Advancement of Sciences. Dans cette lettre, E.-V. Condon annonçait qu'il publiait sans avoir le goût ni le temps de l'approfondir, une relation fonctionnelle qu'il avait découverte au cours d'une étude sur les fréquences relatives des sons de la langue anglaise, relation susceptible d'intéresser les linguistes. Rangeant les mots dans l'ordre des fréquences absolues décroissantes et leur affectant de ce fait un rang, il note que pour tout rang n, la fréquence $f(n)$ vérifie la relation :

$$f(n) = \frac{k}{n}$$

la valeur de la constante k se déduisant de la relation :

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

Et l'auteur présente deux ajustements en graphiques bi-logarithmiques (fréquences en ordonnées, effectifs cumulés c'est-à-dire rangs correspondants en abscisses) ; l'un sur les données de Dewey [5] et l'autre sur celles de Ayres : « A measuring scale for ability in spelling », Russel Sage Foundation, 1915. Les deux décomptes sont d'environ 100 000 mots. Ce sont, semble-t-il, les premiers graphiques parétiens de statistiques lexicales. Les deux ajustements sont très bons malgré des écarts systématiques. L'auteur souligne la valeur — 1 de la pente de la droite d'ajustement et pour terminer il se demande s'il ne faut pas voir sous cette relation l'analogue linguistique de la loi de Weber-Fechner ou de la loi de l'utilité décroissante chère aux économistes. Il est cité par P. Guiraud [8, p. 42].



L'hypothèse formulée ici n'a pas été retenue par B. Mandelbrot qui pense que les travaux de Condon et ceux de Zipf sont tout à fait indépendants. Nous revenons donc à celui-ci.

C'est sous le nom de « standard curve of english » et sur les données d'Eldridge [15] que Zipf présente une distribution lexicale sous forme parétienne [14, p. 44] ; sur le même graphique, mais sans rappel des données dans le texte, figurent les mots-formes de Plaute. Parallèlement, l'auteur dégage la notion de « longueur d'onde » qu'on appellerait plutôt maintenant intervalle moyen entre répétitions, dénomination qu'il utilisera ultérieurement [11]. C'est, pour le mot le plus fréquent par exemple, le nombre moyen de mots entre deux occurrences de ce mot le plus fréquent. Zipf ne formalise pas et ne donne que le résultat des calculs. On pourrait obtenir ces valeurs de deux façons pour un mot donné : soit en divisant le nombre total des unités de texte T par la fréquence du mot, soit en faisant la moyenne des effectifs des séquences successives de mots séparant les occurrences consécutives du mot. Il est évident que la deuxième procédure permettrait une utilisation statistique plus raffinée par l'étude des écarts à la moyenne par exemple qui fait de cette notion une notion stylistique. Zipf montre que les valeurs successives ainsi attachées aux mots ordonnés sont les multiples successifs de 10 et forment donc une « série harmonique » ou progression géométrique. Pour le mot le plus fréquent l'intervalle moyen est de 10 mots environ, pour le mot venant au second rang il est à peu près de 20 mots, etc. On a donc :

$\frac{T}{f_1} = 10 \times 1, \frac{T}{f_2} = 10 \times 2, \ldots \frac{T}{f_n} = 10 \times n$ où T : nombre total des occurrences,

f_l : nombre d'occurrences du mot de rang l ou sa fréquence,

n : rang.

Mais $\frac{T}{a} = f_n \times a$ ou *Constante* = *fréquence* × *rang*, qui est la formulation la plus usuelle de la loi de Zipf, ne sera présentée que dans le dernier livre de l'auteur [11, pp. 23-24]. Cette présentation de la distribution lexicale a d'une part l'avantage d'être quasi-indépendante vis-à-vis de la longueur de texte dépouillé par rapport à la formalisation de $a\ b^2 = k$. Mais surtout elle permet à la série harmonique de servir de loi sous-jacente à laquelle confronter les données réelles des langages les plus divers. C'est en ce sens que Zipf la qualifie de « standard », mais elle ne permet pas non plus de rendre compte des hautes fréquences.

L'auteur aborde de façon explicite pour la première fois le problème des classes de fréquence, c'est-à-dire des fréquences représentées par plusieurs mots. Il présente deux calculs d'intervalles moyens, l'un basé sur le rang du premier mot, l'autre sur le rang du mot médian sans donner le critère de rangement des mots dans la classe.

Enfin, il ne reprend pas sur des données, dans le texte, la formulation présentée dans la Préface, exprimant les effectifs des classes de fréquence en fonction de l'effectif des apax ; si cet effectif est x, alors les classes successives ont pour effectifs :

$$x, \frac{1}{2^2} x, \frac{1}{3^2} x, \ldots \frac{1}{n^2} x$$

C'est dans son dernier ouvrage publié [11] que Zipf a donné la formulation habituellement utilisée de la distribution lexicale :

$$rang \times fréquence = \text{Constante} = \frac{1}{10} T.$$

La table et le graphique sont établis [11, p. 24-25] sur les données de l'Index de Hanley pour *Ulysses* de J. Joyce. Sur le même graphique figurent les données d'Eldridge. La distribution rang-fréquence est, dit-il, une hyperbole équilatère puis qu'elle est figurée en coordonnées logarithmiques par une droite de pente négative — 1/2. Il met les fréquences en ordonnées et les rangs en abscisses — la tradition économique, suivant Pareto, fait le contraire. Les pentes obtenues, toujours négatives, sont alors en valeurs absolues l'inverse l'une de l'autre $\frac{dy}{dx} = \frac{1}{\frac{dx}{dy}}$. Il faut y faire attention quand on compare deux distributions.

Quelques étapes dans la découverte de la loi d'Estoup-Zipf ont été évoquées ici ; un article qui paraîtra ultérieurement dans cette revue présentera des exemples d'ajustements parétiens de statistiques lexicales [1].

1. Des documents ont été gracieusement prêtés ou offerts à l'auteur par M^lle^ Barascud, M. Chouvet, M. Coste, M. Germanet et par les descendants de J.-B. Estoup. Mais ici, outre le prêt de documents, l'auteur doit les plus vifs remerciements à Madame H. Estoup et à Monsieur J. Estoup pour l'extrême bienveillance avec laquelle ils l'ont tous deux longuement reçu, lui fournissant de nombreux renseignements sur les activités si diverses de J.-B. Estoup et J.-H. Estoup. Madame Estoup qui est la fille de R. Havette a prêté aussi quelques-uns des livres de la célèbre bibliothèque de son père.

De plus, la revue Mathématiques et Sciences humaines et ses éditeurs se joignent à l'auteur pour remercier les descendants de J.-B. Estoup d'avoir autorisé la reproduction de l'unique exemplaire en leur possession du livret *Exposé théorique des gammes sténographiques.*

BIBLIOGRAPHIE

[1] MANDELBROT, B., « Les constantes chiffrées du discours », *Encyclopédie de la Pléiade Le Langage*, vol. publié sous la direction d'A. Martinet, Paris, Gallimard, 1966, pp. 46-56.

[1 bis] MANDELBROT, B., « Word frequencies and Markovian models of discourse », *Structure of language and its mathematical aspects*, Proceedings of symposia in applied math., vol. XII, *Ann. math. statist.*, Providence, USA, 1961.

[2] GOUGENHEIM, C., MICHEA, R., RIVENC, P., SAUVAGEOT, A., *L'élaboration du français fondamental* (1er degré), *Étude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris, Didier, 1964.

[3] PRATT, F., *Histoire de la cryptographie : Les écritures secrètes depuis l'Antiquité jusqu'à nos jours*, Trad. du capitaine E. Arnaud, Paris, Payot, 1940.

[4] ZIPF, G. K., *Relative frequency as a determinant of phonetic change*, Harvard Studies in classical Philology, vol. 40, Cambridge, Mass., Harvard University Press, 1929.

[5] DEWEY, G., *Relative frequency of English speech sounds*, Harvard Studies in Education, Cambridge Mass., Harvard University Press, 1923.
Très curieux livre écrit dans une orthographe simplifiée prônée par le Simplified Spelling Bord. De nombreuses recherches sur les fréquences de lettres, bigrammes [...] mots ont été à cette époque effectuées dans ce cadre aux USA.

[6] GERMANET, F., « La sténographie, ses origines et son histoire, ses principes et son avenir, Paris, Aix-en-Provence, 1899.

[7] THIERRY-MIEG, J.-J., *Phonographie à pente unique, nouveau système d'écriture abrégée*, Paris, Firmin Didot, 1853.

[8] GUIRAUD, P., *Bibliographie critique de la statistique linguistique*, Utrecht, Spectrum, 1954.

[9] NAVARRE, A., *Histoire générale de la sténographie et de l'écriture à travers les âges*, Paris, Delagrave, s.d.
Ouvrage introuvable que m'a aimablement communiqué Mlle Barrascud, Directrice de l'Institut Sténographique de France.

[10] ZIPF, G. K., « The psychology of language », *Encyclopedia of psychology*, edited by P. L. Harriman, New York, Philosophical Library, 1946.

[11] ZIPF, G. K., *Human Behavior and the principle of least effort : An introduction to human ecology*, New York, Hafner, 1949.

[12] ZIPF, G. K., « Selected studies of the Principle of Relative Frequency », *Language*, Cambridge, Mass., 1932.

[13] ZIPF, G. K., *The psychobiology of language : An introduction to dynamic philology*, Boston, Mass., Houghton-Mifflin, 1935.

[14] CONDON, E. U., « Statistics of vocabulary », *Science*, vol. 57, march 16, 1928, n° 1733, p. 300.

[15] ELDRIDGE, R. C., *Six thousand common English words*, privately printed at Niagara Falls, NY.

[16] PETRUSZEWYCZ, M., « Loi de Pareto et processus markovien », *Math. Sci. hum.*, n° 3, avril 1963, pp. 21-29.