

CAD

Mémoire reçu : Richesse et mode d'investissement d'un vocabulaire

Les cahiers de l'analyse des données, tome 11, n° 4 (1986),
p. 491-496

http://www.numdam.org/item?id=CAD_1986__11_4_491_0

© Les cahiers de l'analyse des données, Dunod, 1986, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » 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/>

MÉMOIRE REÇU : RICHESSE ET MODE D'INVESTISSEMENT D'UN VOCABULAIRE

par E. Evrard

Mémoire reçu : Richesse et mode d'investissement d'un vocabulaire
par E. Evrard.

"Un texte est constitué d'une succession de mots ou *occurrences*; chaque occurrence relève d'un *vocabulaire* susceptible de répétitions... La fréquence des répétitions de vocabulaires dépend à la fois de la structure de la langue et des tendances stylistiques du texte... De toute manière... à mesure qu'on avance dans le texte, le nombre de vocabulaires déjà employés a tendance à croître, ce qui multiplie les occasions de répétitions; par ailleurs en raison même des exigences du thème traité, la difficulté de trouver de nouveaux vocabulaires qui soient adaptés aux circonstances se fait de plus en plus sentir ...".

Partant de semblables considérations, l'auteur a été conduit à définir un indice que nous appellerons en bref "entropie d'un texte"; il en a étudié la variation pour des textes constitués par des segments de plus en plus longs d'une même oeuvre. Il a découvert que cette variation, pouvait être bien décrite par une fonction quadratique du logarithme de la longueur du texte.

Dans le présent compte rendu nous reprendrons d'abord, avec des notations mathématiques adaptées à notre propos, les calculs de E. Evrard. Puis nous justifierons la formule logarithmique dans le cadre du modèle d'urne et de la loi de Mandelbrot-Zipf. Chemin faisant, nous examinerons dans quelle mesure cette justification vaut pour un texte réel, qui ne s'accorde qu'en partie avec le modèle d'urne.

1 Entropie d'une langue et entropie d'un texte : En vue d'utiliser un modèle d'urne (cf. § 2), c'est-à-dire un modèle qui engendre les mots d'un texte par des tirages aléatoires successifs indépendants, nous supposons que le vocabulaire d'une langue est un ensemble V dont chaque élément v , ou vocabulaire a une probabilité déterminée p_v , de telle sorte que la somme des probabilités de tous les mots soit 1.

$$1 = \sum \{p_v | v \in V\}.$$

Dans la réalité, cette probabilité p_v n'existe que si on se place dans un genre de discours stable, à la fois quant au thème et quant au style; ce qu'il n'est pas facile de définir avec précision. D'autre part pour la simplicité des calculs du § 3, il convient de supposer que V est infini et dénombrable: hypothèse qui n'entre pas en conflit avec la réalité si, par exemple, on considère des textes où ne sont attestés que 1000 à 2000 vocabulaires différents; alors que la langue en compte plusieurs dizaines de milliers.

Ceci posé, on définit l'entropie de la langue par la formule :

$$H = - \sum \{pv \log pv | v \in V\} ;$$

(où on a introduit le signe -, afin que H soit positif : en sorte que le terme exact serait "néguentropie" ou information...).

Soit maintenant un texte t , ou suite d'occurrences de vocables de V . On notera xt la longueur du texte t , ou nombre total des occurrences qu'il comprend ; et $x(v,t)$ le nombre des occurrences du vocable v dans t . Il est clair que xt est la somme des nombres d'occ. de tous les mots, i.e. :

$$xt = \sum \{x(v,t) | v \in V\}.$$

En fait, seuls sont attestés dans le texte t , certains vocables, ceux pour lesquels $x(v,t) \neq 0$. On désignera par V_t , le sous-ensemble de ces vocables ; V_t peut être appelé le vocabulaire du texte t ; le nombre des vocables qu'il comprend sera noté mt . On distingue dans V_t plusieurs segments : le segment V_{1t} des hapax, ou mots dits une fois, dont le nombre est noté mlt ;... ; le segment V_{7t} des mots dits 7 fois ou $\xi\pi\tau\acute{\alpha}\kappa\iota\varsigma$ λεγομενα ;... etc. . Soit en formules :

$$V_t = \{v | v \in V ; x(v,t) \neq 0\} ; mt = \text{Card } V_t ;$$

$$V_{1t} = \{v | v \in V ; x(v,t) = 1\} ; mlt = \text{Card } V_{1t} ; \dots$$

$$V_{it} = \{v | v \in V ; x(v,t) = i\} ; mit = \text{Card } V_{it} ; \dots$$

Pour le texte t , on définit une entropie H_t , comme on a défini H ; mais en se bornant au vocabulaire V_t , et en substituant à la probabilité pv (inconnue) la fréquence fv observée dans le texte t :

$$fv = f(v,t) = x(v,t)/xt ; (\text{nombre d'occ. de } v \text{ dans } t/\text{long. de } t)$$

$$H_t = - \sum \{fv \log fv | v \in V_t\}.$$

On peut calculer H_t en considérant les vocables v de V_t dans l'ordre des fréquences croissantes : d'abord l'ensemble V_{1t} des mlt hapax ; puis l'ensemble V_{2t} des m_{2t} mots dits deux fois ; etc. ; on obtient alors pour H_t la formule suivante, où ne figurent que la longueur xt du texte t et les nombres successifs mlt, m_{2t}, \dots :

$$H_t = - \sum \{mit(i/xt) \log(i/xt) \quad i = 1, 2, \dots\} ;$$

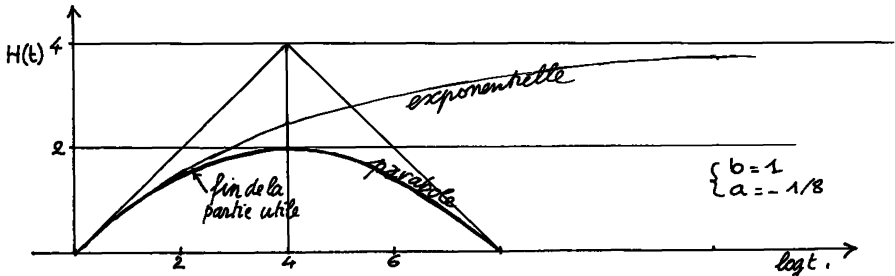
en somme, cette formule tient compte de l'ensemble des caractéristiques de la richesse du vocabulaire du texte t .

Pour étudier l'enrichissement du vocabulaire, quand on avance dans une oeuvre, E. Evrard considère la suite des textes t formés chacun des t premiers mots de cette oeuvre. Ainsi H_t devient une fonction de $H(t)$ de l'entier t c'est ce qu'on appellera, la fonction d'enrichissement du vocabulaire. On a nécessairement $H(1) = 0$, l'entropie d'un texte formé d'une seule occurrence d'un seul vocable étant nulle ; $H(t)$ varie ensuite, en connaissant de nombreuses fluctuations, mais avec une nette tendance à la croissance, qui s'affaiblit pour les fortes valeurs de t . En étudiant 13 oeuvres de la littérature latine, E. Evrard a trouvé dans chaque cas une fonction $H(t)$ bien approchée par un polynôme de degré 2 en $\log t$:

$$H(t) \approx b \log t + a(\log t)^2.$$

Il va sans dire que dans cette formule les coefficients a et b dépendent de l'oeuvre considérée : par exemple, pour le Jugurtha de Salluste, $b = 1,21$; et $a = -0,12$; et les valeurs trouvées pour les autres oeuvres oscillent pour b de 1,01 à 1,27 ; et pour a de -0,08 à

-0,16. En bref, initialement $H(t)$ croît linéairement en fonction de $\log t$; puis l'effet du terme quadratique, affecté d'un coefficient a négatif, se manifeste par ce que l'auteur appelle "un essoufflement" dans l'enrichissement du vocabulaire. S'il était permis de prendre à la lettre la formule, la fonction $H(t)$, après avoir atteint un maximum, décroîtrait, s'annulerait et deviendrait même négative ; ce qui est strictement impossible de par la définition même de l'entropie d'un texte. En fait la partie utile de la parabole représentative de $H(t)$ s'arrête avant le maximum ; à la valeur de t qui correspond à la longueur de l'oeuvre dont on prend les segments initiaux successifs.



Il est clair que, dans sa partie utile, la fonction $b \log t + a(\log t)^2$ peut être remplacée par une grande diversité de fonctions. Pour des raisons qui apparaîtront au § 3, nous proposons la forme :

$$H(t) \approx H(1 - \exp(-B \log t)).$$

Pour passer des paramètres (a, b) aux paramètres (H, B) , le plus simple est de demander que la parabole et l'exponentielle soient osculatrices à l'origine ; (encore que pour l'ajustement il soit plus précis d'appliquer à l'exponentielle la méthode des moindres carrés). Il vient (en notant $\lambda = \log t$) :

$$b \lambda + a \lambda^2 \approx H(1 - \exp(-B \lambda)) \approx H(B \lambda - (B^2/2) \lambda^2 \dots)$$

$$b = HB ; -a = (HB^2/2) ;$$

$$B = -2a/b ; H = -b^2/(2a).$$

Comme dans ces calculs, E. Evrard prend pour λ le log décimal, l'exponentielle $\exp(-B \lambda)$ peut s'écrire (en notant Log le ℓ . népérien) :

$$\exp(-B \lambda) = \exp(-B \text{Log } t (\log_{10} e)) = \exp(-0,43 B \text{Log } t).$$

D'où finalement pour $H(t)$ la formule :

$$H(t) \approx H(1 - t^{-\beta}) ; \text{ où } \beta = 0,43 B.$$

Il nous reste maintenant à voir dans quelle mesure, cette formule équivalente quant à l'approximation de la loi observée par E. E., peut se justifier dans le cadre du modèle d'urne et de la loi de Mandelbrot-Zipf.

2 Modèle d'urne et décomposition de la fonction d'enrichissement du vocabulaire ; $H(t)$

Supposons qu'une oeuvre de longueur indéfinie, soit engendrée par tirages successifs indépendants de vocables v du vocabulaire V ; chaque mot v ayant la probabilité p_v de sortir à chaque tirage. La fonction $x(v, t)$ donnant le nombre d'occurrences de v dans les t premiers tirages est une fonction aléatoire ; dont la valeur à chaque temps entier t , est distribuée, comme il est bien connu, suivant une loi binomiale.

$$\text{Prob}\{x(v, t) = x\} = p v^x (1-pv)^{t-x} C_t^x .$$

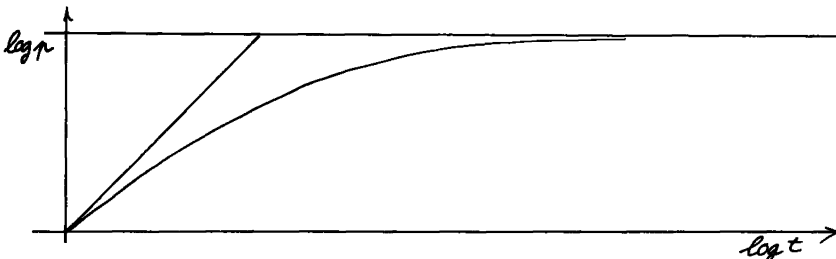
De même la fréquence $f(v, t) = x(v, t)/t$, a une loi connue ; ainsi que le produit $f(v, t) \cdot \log f(v, t)$. Enfin $H(t)$ apparaît comme une fonction aléatoire, somme de termes dont chacun suit une loi connue. L'étude détaillée de la fonction aléatoire $H(t)$ semble difficile ; et il n'y a pas à le regretter du point de vue du linguiste ; dans la mesure où le modèle d'urne ne s'accorde qu'en partie avec la réalité ! Il nous suffira d'étudier la fonction $EH(t) = \text{Esp}(H(t))$ donnant pour chaque valeur de t , l'espérance mathématique de $H(t)$. A des fluctuations près qui échappent à notre étude, la fonction d'enrichissement étudiée empiriquement au § 1, doit s'accorder avec la fonction moyenne $EH(t)$. Plus précisément notre but est de montrer que la formule d'approximation proposée à la fin du § 1, vaut pour $EH(t)$, moyennant des hypothèses sur le vocabulaire ; hypothèses que l'on considère au § 3. Dans le présent § nous considérerons seulement la variation en fonction de t de l'espérance mathématique du terme $f(v, t) \cdot \log f(v, t)$, de $H(t)$, afférent à un seul mot, pour simplifier les notations, la probabilité de ce mot sera notée p (et non pv). On a :

$$\begin{aligned} \text{Esp}\{f(v, t) \cdot \log (f(v, t))\} = \\ - \sum \{(x/t) \log(x, t) p^x (1-p)^{t-x} C_t^x \mid x = 1, \dots, t\} \end{aligned}$$

Excepté quand p est grand (e.g. sup. à 0,05) on peut utiliser au lieu de la loi binomiale la loi de Poisson. Le cas où p est grand ne peut concerner que peu de vocables (puisque $\sum pv = 1$) : vaut alors l'approximation par la loi normale ; mais peu nous importe car, pour ces vocables la fréquence (x, t) atteint rapidement la limite que lui assigne la loi des grands nombres ; et ces vocables n'apportent pas de contribution importante à la croissance de $EH(t)$ en fonction de t (De façon précise, $\text{Esp}\{f(t) \log f(t)\}$ rejoint $p \log p$ en croissant ; la différence étant de l'ordre de la variance de $f(t)$, c'est-à-dire de $(1/t)$. On se bornera donc à :

$$\text{Esp} = - \sum \{(x/t) \log(x/t) \exp(-pt) pt^x / (x!) \mid x = 1, 2, \dots\}$$

Quand le produit pt est petit, seul compte dans cette somme le terme $x = 1$; $\text{Esp} \approx p \log t$. Quand devient grand le produit pt (qui n'est autre que l'espérance mathématique du nombre d'occurrences de p dans le texte formé par les n premiers tirages), la fréquence f se concentre vers sa valeur limite p et on a $\text{Esp} \approx p \log p$. Ceci



permet de dessiner la courbe représentative de la fonction (Esp/p) . Il est possible dans le cadre du modèle de préciser la branche asymptotique ($\log t \rightarrow \infty$) ; mais l'allure précise de cette branche dépend en réalité de la tendance des mots à survenir en grappes, et n'est donc pas fournie par le modèle ; et de plus seule nous intéresse la contribution apportée par la fonction $\text{Esp}(t)$ à la croissance de $EH(t)$. De ce point de vue seuls comptent les vocables pour lesquels pt est faible ; inférieur à 1 ou de l'ordre de 1 ; en sorte que la dérivée de $\text{Esp}(t)$ reste ≈ 1 , comme pour $pt \approx 0$; tandis que quand pt dépasse 5,

la loi des grands nombres s'applique, et $\text{Esp}(t)$ est donc proche de sa valeur limite. Ces considérations dépendent peu du modèle d'urne ; dans la mesure où si les vocables surviennent, e.g., par grappes de 2, il en résulte simplement que la pente de la tangente à l'origine à $\text{Esp}(t)$ est divisée par 2.

En résumé, nous retiendrons que la fonction $(1/p) \text{Esp}(f(v,t), \log(v,t))$ croît comme $\log t$ tant que $pt \leq 1$; et nous négligerons sa croissance pour $pt > 5$.

3 Hypothèses sur la structure du vocabulaire, et croissance de la fonction d'enrichissement : Pour appliquer à la fonction $\text{EH}(t)$ (fonction moyenne d'enrichissement) les conclusions du § 2, nous adopterons le schéma suivant :

$$d \text{EH}(t) / d \log t \approx \Sigma \{pv | v \in V ; pv.t \leq 1\} ;$$

les hypothèses faites nous paraissent suffisamment robustes pour permettre de justifier par cette voie une formule de $\text{EH}(t)$ telle que celle introduite au § 1. Le problème est donc ramené à estimer le membre de droite de l'équivalence ci-dessus : cela ne peut se faire que moyennant de nouvelles hypothèses relatives aux probabilités p_v : on est dans le domaine classique, découvert par Zipf et profondément étudié par Mandelbrot, de la relation entre fréquences p_v et rang des vocables (ceux-ci étant rangés dans l'ordre de p_v décroissants). Il est commode ici de formuler des hypothèses sur la densité limite de la distribution des vocables en fonction de la variable q , inverse de la fréquence p . On note :

$$\text{densité} = \zeta(q) \approx q^{-\beta} ; \text{ où } 0 < \beta < 1$$

ce qui signifie que pour q grand, le nombre des mots dont la fréquence est comprise entre $(1/q)$ et $(1/(q + \Delta q))$ est de l'ordre de $\Delta q \cdot q^{-\beta}$. Cette formule est satisfaisante en ce que d'une part elle assure l'existence d'un vocabulaire infini :

$$\int^Q dq q^{-\beta} \text{ diverge comme } Q^{1-\beta} \text{ pour } Q \rightarrow \infty$$

et d'autre part permet de normaliser à 1 la somme des probabilités :

$$\int^Q dq (1/q) q^{-\beta} \text{ converge comme } Q^{-\beta} \text{ pour } Q \rightarrow \infty$$

ici, la condition que β soit entre 0 et 1, est essentielle. De ce point de vue la formule approchée initiale de Zipf qui correspond à $\beta = 0$, pêche en ce que la somme des probabilités est assimilée à une intégrale divergente comme $\log Q$. Sous l'hypothèse faite, la somme des probabilités des mots ayant chacun une probabilité inférieure à $(1/t)$ est estimée par l'intégrale :

$$\int_t^\infty (1/q) q^{-\beta} dq \approx t^{-\beta} ;$$

d'où pour la fonction d'enrichissement $\text{EH}(t)$:

$$d \text{EH}(t) / d \log t \approx t^{-\beta}$$

et en notant $\log t = \lambda$; et tenant compte de ce que $\text{EH} = 0$ si $\lambda = 0$:

$$d \text{EH}(\lambda) / d \lambda \approx \exp(-\beta \lambda)$$

$$\text{EH} = H(1 - \exp(-\beta \lambda)) = H(1 - t^{-\beta}) ;$$

c'est la formule du § 1.

Pour comparer avec la relation entre rang et fréquence, on notera que le nombre des mots de fréquence supérieure à $(1/Q)$ est fourni par l'intégrale $\int_Q^\infty dq q^{-\beta}$, laquelle diverge comme $Q^{-1-\beta}$; en sorte que le log du rang, $\log R$ apparaît lié au log de la fréquence $\log P = -\log Q$, par la relation :

$$\log R = K + (1 - \beta) \log Q = K - (1 - \beta) \log P$$

$$\log P = C - (1 - \beta)^{-1} \log R.$$

C'est bien ce que donne la loi de Mandelbrot pour les valeurs élevées de R ; avec un coefficient $(1 - \beta)^{-1}$ qui dépasse d'assez peu 1; ce qui correspond à un β faible.

Il importe de noter qu'il est possible de tracer pour le vocabulaire de chacune des oeuvres latines étudiées par E. Evrard, le diagramme en $(\log p, \log R)$ (log de la fréquence dans l'oeuvre; log du rang) de mesurer empiriquement la pente limite de ce diagramme pour les grandes valeurs de R (faibles valeurs de P); et de comparer avec ce que donnent les valeurs de β déjà calculées d'après la fonction d'enrichissement. Ce sera le moyen de vérifier si la loi d'Evrard, avec ses deux constantes H et β (telles que nous les avons notées) s'accordent suivant notre modèle (plusieurs fois périlleux...) avec la loi de Mandelbrot; laquelle fournit β et permet de calculer H par une simple intégrale.