Indicateurs de diversité et exploitation statistique d'une question ouverte

Michèle Moine

LABSAD – Laboratoire de Statistique et Analyse des Données, Département STID-IUT 2 – Université Pierre Mendès France – BSHM-BP47 – 38 040 Grenoble Cedex 09 – France – Michele.Moine@iut2.upmf-grenoble.fr

Abstract

The interest and the limits of indicators concerning diversity, inequality, concentration, poverty, and also graphics associated, are illustrated in the case of the study of an open-ended question.

Résumé

L'intérêt et les limites d'indicateurs de diversité, d'inégalité, de concentration, de pauvreté ainsi que des représentations graphiques associées, sont illustrées dans le cadre de l'exploitation d'une question ouverte.

Mots-clés : indicateurs de diversité, indicateurs d'inégalité, indicateurs de concentration, indicateurs de pauvreté.

1. Introduction

L'étude de la diversité d'un texte est complexe car sa description est de nature multidimensionnelle. Il est souvent nécessaire et intéressant de résumer cette information par des indicateurs.

Dans ce contexte, le choix de résumés statistiques est délicat. Ce choix renvoie à des problèmes très divers, du choix de l'unité d'information à analyser (vocables, thèmes,..), aux qualités exigées des indicateurs retenus (vérification d'axiomes, mais aussi sensibilité aux fréquences élevées ou faibles, aptitude à mettre en évidence des phénomènes de spécialisation de corpus,….).

Les travaux concernant les indicateurs de diversité, d'inégalité, de concentration, de pauvreté foisonnent dans les domaines de la statistique textuelle, de l'écologie et de l'économie. L'intention de cet article n'est pas d'effectuer un recensement de ces résultats, elle est de montrer l'intérêt de certains d'entre eux pour la comparaison de textes produits par des enquêtés répondant à une question ouverte.

2. Indices de diversité, d'inégalité

2.1. Statistique lexicale

Des **indicateurs de diversité** ont été proposés par Yule (1944), puis Guiraud (1954) et Herdan (1957). Leurs travaux visaient entre autres à étudier la stabilité d'indices par rapport à la taille des corpus. Dans le cadre d'une modélisation binomiale, Muller (1993) propose une

méthode de comparaison de la richesse de textes de tailles différentes (technique de réduction de textes). De nombreux autres indicateurs ont été proposés par les linguistes. Une revue des principaux résultats se trouve dans les ouvrages suivants : Brunet, 1978 ; Thoiron, Labbé, Serant, 1988; Cossette, 1997 ; Baayen, Tweedie, 1998 (1), Baayen, Tweedie, 1998 (2).

R. Baayen et F. J. Tweedie (1998) proposent une classification des indices usuels en trois catégories selon qu'ils tiennent compte de la taille du vocabulaire et de la taille du texte seuls (indices de répétition moyenne, de Guiraud, de Herdan, de Brunet, de Dugast), de la distribution de la « gamme des fréquences » du texte (indice de Simpson, Yule, famille d'indices de Good, entropie de Shannon, de Yule, de Simpson), ou encore sont déduits de paramètres de modèles (modèle de Zipf, modèle de Sichel, modèle LNRE).

Un grand nombre de travaux a pour préoccupation la dépendance de ces indicateurs par rapport à la taille des textes étudiés. Cette dépendance pose problème dans le cadre d'une comparaison de textes de longueurs différentes.

Les **représentations graphiques** des fonctions de répartitions empiriques associées à la gamme des fréquences (diagrammes de Pareto, par exemple) permettent d'analyser et de comparer les richesses de textes (Lebart & Salem, 1994).

2.2. Statistique écologique

Deux propriétés sont habituellement imposées aux indicateurs de diversité :

- Pour une nombre d'espèces donné, une communauté est d'autant plus diversifiée que la distribution d'abondance est proche de l'équidistribution,
- Pour des communautés de vecteur d'abondance équiréparti, la diversité est d'autant plus forte que le nombre d'espèces est élevé.

Une communauté notée C est constituée de N spécimens regroupés en s espèces : 1, 2,…i…s. La fréquence de l'espèce i est notée π_i.

π est le vecteur d'abondance des espèces : $(\pi_1, \pi_2, \ldots \pi_i, \ldots \pi_s)$.

Une communauté C est caractérisée par s et π.

Le vecteur π^* est le vecteur d'abondance dont les composantes ont été rangées par ordre décroissant.

Définition de familles d'indices

A chaque espèce i, on associe un indicateur de rareté i : $R(i; \pi)$**.** La mesure de la diversité de la communauté C est la moyenne de la rareté des espèces.

$$\Delta(C) = \sum_{i=1}^{s} \pi_i \, R(i;\pi)$$

A partir de cette définition, plusieurs familles d'indices sont définis : les indices dichotomiques, les indices de rang, les nombres équivalents.

Un indicateur est de type « **dichotomique** » si la rareté de l'espèce i ne dépend que de la valeur de π_i : $R(i; \pi) = R(\pi_i)$.

Un **indicateur est de type « rang »** si la rareté d'une espèce ne dépend que de son rang dans le vecteur $\pi*$: R(i; π)=R**(i)**.

Des indices tels le nombre d'espèces (richesse taxonomique), l'indice de Shannon, l'indice de Simpson sont des cas particuliers de ces mesures de diversité de type dichotomique.

- Richesse taxonomique : s – 1
- Indice de Shannon : $-\sum_{i=1}^{s}\pi_i \log(\pi_i)$
- Indice de Simpson : $1-\sum_{i=1}^{s}\pi_i^2$

Dans certains travaux, les fréquences sont remplacées par des surfaces de recouvrement (Vanpeene Bruhier, 1998).

Une classe particulière d'indices de type dichotomique appelés ***indices de diversité de degré β*** est définie :

- $\Delta\beta = \sum_{i=1}^{s}\pi_i \, R(i;\pi) = \sum_{i=1}^{s}\pi_i \frac{\left(1-\pi_i^{\beta}\right)}{\beta} = \sum_{i=1}^{s}\frac{\left(1-\pi_i^{\beta+1}\right)}{\beta}$, si $\beta \neq 0$
- $\Delta_\beta = \sum_{i=1}^{s}\pi_i \, R(i;\pi) = -\sum_{i=1}^{s}\pi_i \, Log(\pi_i)$, si $\beta = 0$

Le nombre d'espèces, l'indice de Shannon, l'indice de Simpson appartiennent à cette famille d'indices.

Les nombres équivalents

Les nombres équivalents sont des indices de richesse taxonomique construits à partir d'indices de diversité. Un indice de diversité est calculé pour une communauté donnée C, le nombre d'espèces d'une communauté ayant une distribution d'abondance uniforme pour laquelle l'indice de diversité correspond à l'observation est appelé nombre de diversité.

Les **indices de Hill** sont les nombres équivalents de la famille d'indices de diversité de degré β :

$$S_\beta = \left(\frac{1}{\sum_i \pi_i^{\beta+1}}\right)^{1/\beta}$$

On peut en déduire un autre indicateur de diversité : log(S_β) qui est l'**entropie généralisée de Rényi.**

$$H_\beta = \log(S\beta) = -\frac{\log(\sum_i \pi_i^{\beta+1})}{\beta}$$

Propriétés concernant les indices dichotomiques (Patil et Taillie, 1982)

Les propriétés décrites au début du paragraphe 2.2 mènent aux conditions suivantes :

(1) R est une fonction décroissante de π, définie sur [0,1], R(1) = 0 .

(Une espèce est d'autant plus rare que son poids π_i est faible),

(2) $\Delta(C) \leq \Delta(C')$ lorsque C' est obtenu à partir de C après introduction d'une espèce[1] ou par transfert d'abondance [2].

- Si $R(\pi)$ est une fonction décroissante de π (condition (1)), l'introduction d'une nouvelle espèce accroît la mesure de la diversité de la communauté. Par contre la relation (1) n'assure pas que la diversité croît avec une tendance à l'uniformité de la distribution.
- La condition (2) entraîne une condition sur la fonction $V(\pi) = \pi\ R(\pi)$. En particulier, si la fonction V est concave, (2) est vérifiée.
- Tout indice de type Δ_β satisfait la propriété (1). Δ_β satisfait la condition (2) si et seulement si $\beta \geq -1$. De ce fait, le nombre d'espèces, l'indice de Shannon et l'indice de Simpson vérifient ces conditions.

Classements de communautés selon différents indices

Deux indices peuvent donner un classement différent de deux communautés. Il est important de pouvoir caractériser des familles d'indices classant de la même façon des communautés. C'est dans ce but que la définition suivante est introduite :

Une communauté ***C' est intrinsèquement plus diversifiée*** que C si C' est obtenue à partir de C par un nombre fini de séquences d'opérations de type suivant : introduction de nouvelles espèces, transfert d'abondance, permutation des composantes du vecteur d'abondance.

La communauté C est caractérisée par les paramètres s et π, C' par s' et π'

□On démontre que **C' est intrinsèquement plus diversifiée** que C si et seulement si :

---

[1] Définition de l'**inclusion de deux communautés par introduction d'une espèce**: C est inclus dans C' par introduction d'une espèce si : s'=s+1 et si $\exists$ i et j > 0 tels que :

$$\pi'_k = \begin{cases} \pi_k & , \text{si } k \neq i, j \\ \pi_i - h, & \text{si } k = i \\ h & , \text{si } k = j \end{cases}$$

avec $0 < h < \pi_i$ (Hypothèse de partage des ressources de l'espèce i avec la nouvelle espèce introduite).

[2] Définition du **transfert d'abondance** de C vers C' : les deux communauté C et C' ont le même nombre d'espèces (s=s'). De plus, $\exists$ i, j / $\pi_i > \pi_j > 0$ tels que :

$$\pi'_k = \begin{cases} \pi_k & , \text{si } k \neq i, j \\ \pi_i - h, & \text{si } k = i \\ \pi_j + h, & \text{si } k = j \end{cases}$$

lorsque $0 < h < \pi_i - \pi_j$. (π' est plus proche d'une distribution uniforme que π).

$$\sum_{i>k} \pi'^{*}_{i} \geq \sum_{i>k} \pi^{*}_{i} \qquad , k = 1, 2\ldots$$

ou de façon équivalente :

$$\sum_{i\leq k} \pi^{*}_{i} \geq \sum_{i\leq k} \pi'^{*}_{i} \qquad , k = 1,2\ \ldots$$

Une famille d'indices est complète si une communauté C est intrinsèquement plus diversifiée que C' lorsque $\forall$ l'indice Δ de cette famille, $\Delta(C) > \Delta(C')$: les indices de cette famille classent de la même façon deux communautés quelconques.

La famille d'indices $\{\Delta_\beta , \beta \geq -1\}$ n'est pas complète. La comparaison des valeurs de Δ_β de deux communautés s'avère intéressante car l'indicateur est sensible aux fréquences faibles (espèces rares) lorsque β est de valeur faible et l'est moins lorsque β est élevé.

Indices basés sur le rang d'espèces d'une communauté

Dans le cas d'indices basés sur les rangs d'espèces, une condition nécessaire et suffisante pour que les conditions (1) et (2) soient vérifiées conduisent à « R est une fonction croissante de i ». Une classe spécifique d'indices vérifie cette propriété :

$$\Delta_\rho^{(rang)} = \sum_{i\geq 1} \left(\sum_{j>i} \pi^{*}_{j} \right) \rho^{i-1}, \ \rho \geq 0$$

$\Delta_\rho^{(rang)}$ est un indicateur d'autant plus sensible aux fréquences faibles (espèces rares) de la distribution considérée que ρ est élevé.

Si R est définie de la façon suivante : $R(i) = \begin{cases} 1 \text{ si } > j \\ 0 \text{ si } i \leq j \end{cases}$

La rareté moyenne des espèces sont alors égale à T_j :

$$T_j = \sum_{i=j+1}^{s} \pi^{*}_{i} \ , j=1,\ldots.s-1$$

Les **graphes** des points (j, T_j) sont appelés **profils de diversité**. La comparaison des profils de plusieurs communautés permet de repérer l'existence de relations d'ordre intrinsèque (Patil & Rao, 1994).

- Certains indicateurs sont déduits des indicateurs précédents par une transformation de normalisation (comparaison à la valeur maximale ou minimale de l'indice, pour un nombre d'espèces observé).
- Les diagrammes rang-fréquence sont les diagrammes obtenus en reportant en abscisse les espèces rangées par ordre décroissant d'abondance et en ordonnée les fréquences associées. Ils visent à décrire la diversité. Ces diagrammes ont été utilisés en particulier pour décrire des écosystèmes à différents stades d'évolution (D. Mouillot, 1999 ; R. Dajoz, 2000). Plusieurs modèles ont été proposés : modèles de Zipf-Mandelbrot, de Preston, de

Mac Arthur,… Des indicateurs ont été proposés à partir de ces modélisations (Mouillot, 1999).

Les diagrammes de ce type sont également utilisés en linguistique.

La comparaison de communautés s'avère inadéquate lorsque les aires d'observation sont de mesures différentes. Des méthodes dites de raréfaction ou l'utilisation de relations entre le type indicateur et la taille du piège (relations aire-espèces) sont parfois utilisées.

2.3. Mesures de concentration industrielle, d'inégalité de répartition des revenus, de pauvreté en économie

Mesures de concentration

Des entreprises se partagent un marché. La part du marché de l'entreprise i est π_i. Les indicateurs de Shannon et Simpson sont souvent utilisés dans les travaux d'économistes.

La fonction décroissante de l'index de Simpson : $\sum_i \pi_i^2$ est également utilisée.

Mesures d'inégalité

Chaque individu ω_i d'une population est caractérisé par la valeur prise x_i d'une variable X (le revenu, par exemple). Les modalités prises par X sont notées a_1, a_i,…a_p et sont classées par ordre croissant (de revenu individuel).

ν_i est la proportion du revenu global correspondant aux unités statistiques pour lesquelles X = a_i, π_i la fréquence de a_i. $\nu_i = a_i \pi_i \Big/ \sum_i a_i \pi_i$.

La courbe de Lorenz (Theil, 1967; Kendall et Stuart, 1958)

La courbe de Lorenz est la courbe linéaire par morceaux qui joint les points de coordonnées P_k, k=1,…s+1.

$$P_1=(0,0)$$

$$P_k = \left(\sum_{i\le k} \pi_i, \sum_{i\le k} \nu_i\right)$$

Une population C caractérisée par le couple (π, ν) est d'autant plus inégalitaire que la courbe de Lorenz s'éloigne de la première bissectrice de [0,1] x [0,1].

Indices d'inégalité

L'indice de Gini est égal à deux fois le valeur de la surface située entre la courbe de Lorenz et et la 1ère bissectrice de la courbe de Lorenz. Cet indice varie entre 0 et 1 et est d'autant plus élevé que l'inégalité de répartition des revenus est importante.

La comparaison d'indices de Gini de deux populations a un intérêt si l'une est intrinsèquement plus inégalitaire que l'autre.

3. Analyse d'une question ouverte

Dans le paragraphe précédent, des indicateurs et des représentations graphiques permettant de caractériser la diversité, la concentration ou la pauvreté d'une population ont été décrits. Selon le cas, la population étudiée est constituée des occurrences d'un texte, des spécimens d'un écosystème, d'agents économiques. Le caractère est alors soit le vocabulaire, soit l'espèce soit encore le revenu.

Le principe de construction des indicateurs et des graphes décrits a été appliqué à l'étude des textes de réponses à une question ouverte. Pour ce qui concerne les indicateurs de diversités des écologistes, les thèmes d'un texte (ou les lemmes) se substituent aux espèces, la distribution d'abondance devient alors la distribution des thèmes (ou des lemmes) de ce texte.

Les indices de concentration, d'inégalité des revenus, de pauvreté sont transposés à l'étude de textes de la façon suivante : le caractère X dont on étudie la concentration ou l'inégalité de répartition est la fréquence des thèmes (ou lemmes) évoqués dans le texte. On s'intéresse alors à l'inégalité de répartition de la fréquence des thèmes, à l'analyse de la concentration des fréquences des thèmes (lemmes) … Selon les paramètres choisis, il sera possible de discerner des phénomènes relatifs aux thèmes rares (ou lemmes rares), ou plus fréquents.

Les graphes des fonctions de répartition empiriques de la fréquence des thèmes (diagramme de Pareto), les courbes de Lorenz décrivant la contribution au corpus des thèmes selon leur fréquence, permettent de classer des corpus selon leur diversité (mesure de la contribution de thèmes rares), de repérer des phénomènes de spécialisation de corpus (contribution de thèmes peu fréquents dans une sous-partie du corpus),…

Présentation de l'enquête

Une enquête administrée par téléphone a été réalisée dans la cadre de l'enseignement du DESS « PROGIS Etudes d'opinion et de marchés » durant l'année universitaire 2000-2001. L'objet de cette enquête était de mieux cerner les « usages et représentations sociales du Centre Ville de Grenoble » (Denni B., Caillot P., 2001). Le corpus analysé est constitué des réponses à l'une des questions ouvertes du questionnaire :

« Si vous deviez imaginer les deux choses les plus importantes qui devraient être améliorées d'ici 20 ans dans le Centre Ville, quelles seraient-elles ? »

Une étude réalisée en juin 2001, a mis en évidence les liaisons entre les thèmes des réponses et les variables socio-démographiques caractérisant les personnes interrogées, indépendamment d'une effet enquêteur (Caillot P. Moine M., 2001). Le nombre de thèmes d'une réponse donnée varie en fonction du sexe, de l'âge, du statut économique et social.

A chaque personne interrogée est donc associée une séquence de thèmes. La liste des thèmes retenus a été construite à partir des résultats d'une classification automatique des réponses recueillies (méthode « Alceste »). De ce fait, le corpus étudié est constitué de l'ensemble des formes graphiques associées aux thèmes évoqués par les enquêtés.

Le but de l'étude est d'analyser la richesse de textes produits par des enquêtés selon leurs caractéristiques socio-démographiques. Les strates de personnes comparées ont été ramenées à des groupes de tailles égales par échantillonnage (tirage à PESR).

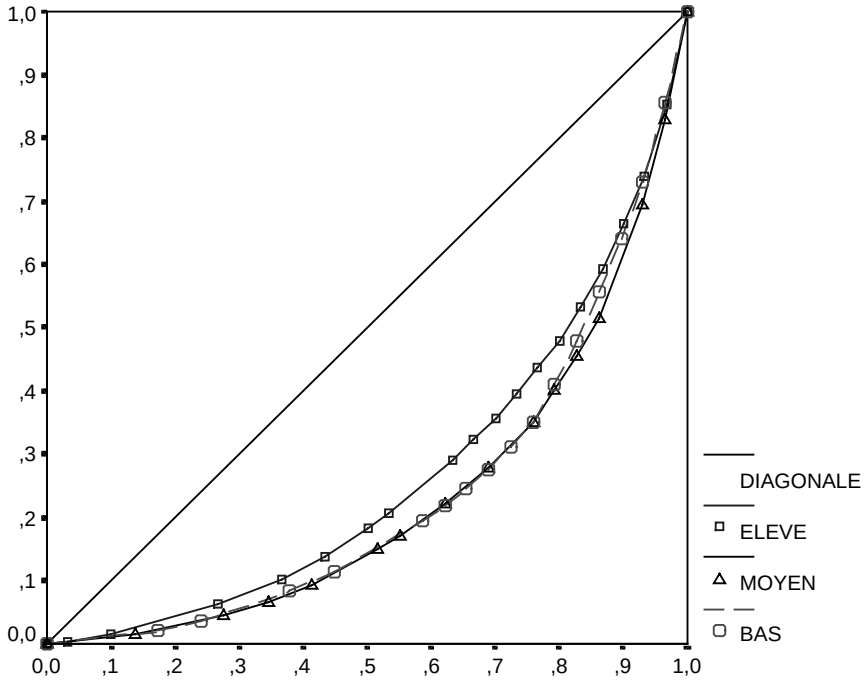
Les résultats

Chaque thème du lexique t_i est associé une fréquence x_i. Le tri à plat de la fréquence des thèmes (gamme des fréquences) conduit à la construction d'une courbe de Lorenz. A chaque valeur x de la fréquence observée, on associe un point dont l'abscisse est le pourcentage de thèmes de fréquence inférieure ou égale à x et dont l'ordonnée est la contribution en pourcentage de ces mêmes thèmes au corpus.

Les courbes de Lorenz permettent de comparer les textes produits par différentes catégories de locuteurs.

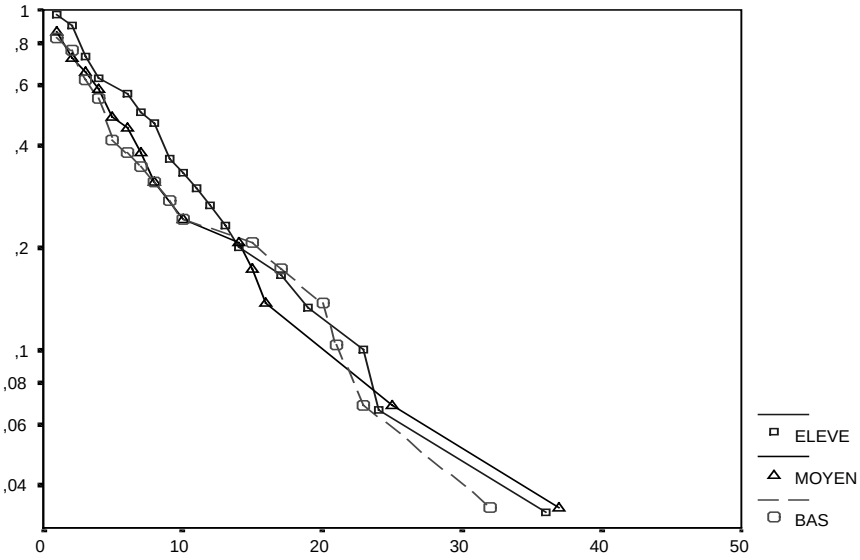
Les écarts entre les courbes associées à une partition du corpus peuvent être plus ou moins importantes, les relations plus ou moins complexes comme on peut le constater dans les exemples qui suivent (partition selon le statut économique et social, selon l'âge).

Dans le cas des corpus associés au statut économique et social, la richesse (nombre de thèmes différents) et l'indice de Shannon indiquent que la diversité des réponses est plus importante pour un niveau socioéconomique élevé. L'indice de Gini indique l'existence d'une répartition des fréquences plus inégalitaire pour le niveau moyen. Le contribution de thèmes de fréquences très élevées est plus important dans le corpus des enquêtés qui ont un statut économique et social moyen (la pente de la courbe de Lorenz est plus forte pour les fréquences élevées dans le cas de cette strate d'enquêtés). Les indices de Foster et le diagramme de Pareto des distributions des fréquences illustrent ces phénomènes.

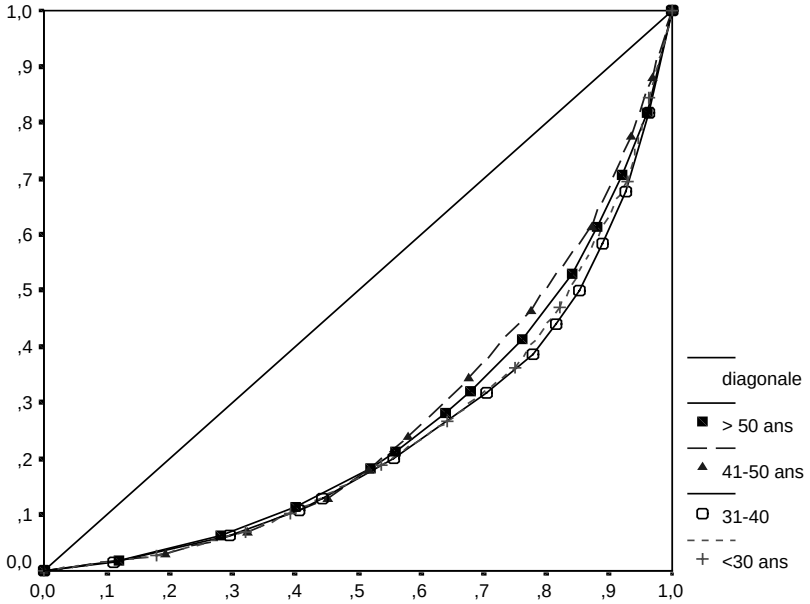


Indicateurs	AS	MOYEN	ELEVE
Richesse (nombre de thèmes différents)	29	29	30
Indice de Shannon	2.89	2.86	3.03
Indice de Gini	0.53	0.54	0.47

Seuil	AS	MOYEN	ELEVE
5	0.45	0.41	0.37
10	0.72	0.69	0.63
15	0.76	0.79	0.80
20	0.83	0.86	0.87

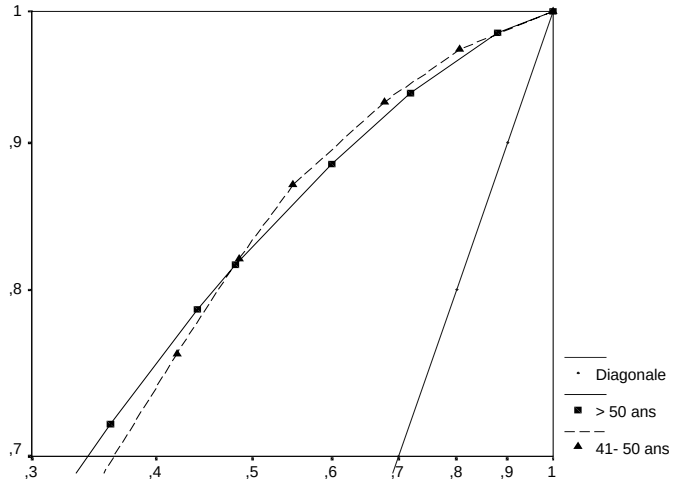


Ci-dessous, comparaison de la diversité des corpus associés aux classes d'âge :

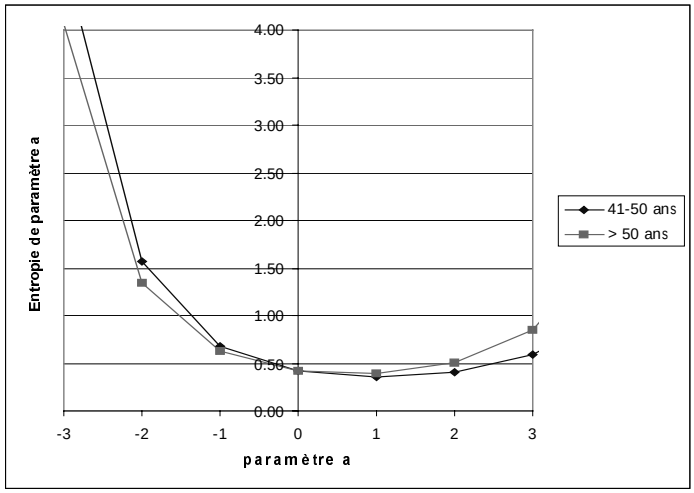


Pour une meilleure lisibilité, une courbe de Lorenz transformée est obtenue en joignant les points de coordonnées P'_k, k=1,…s+1, selon une échelle log-log. (Pour des lois de Pareto, une relation linéaire lie le logarithme de l'abscisse et le logarithme des ordonnées des points P'_k).

$$P'_k = (1 - \sum_{i \le k} \pi_i ,\ \ 1 - \sum_{i \le k} \nu_i)$$



Le graphique suivant (entropie généralisée en fonction du paramètre a), comme le précédent, met en évidence la contribution plus importante des thèmes de fréquences moyennes et plutôt faibles pour les 41-50 ans.



4. Conclusion

L'exploitation dans le domaine de la statistique textuelle d'indicateurs de diversité écologique, de concentration, d'inégalité de répartition des revenus en économie permet d'approcher selon différents critères la richesse et la diversité d'un corpus de données textuelles.

De façon plus précise, l'influence de thèmes évaluée sur la base de leurs fréquences peut s'effectuer par des représentations graphiques appropriées, des indicateurs adaptés. Cet aspect est d'une importance particulière pour le praticien des enquêtes qui s'intéresse souvent à des événements de fréquence de réalisation faible ou moyenne.

Le travail effectué relève de la statistique descriptive. L'étude de la précision des résumés serait à développer.

Références

Baayen R.H., Tweedie F.J. (1998) (1) Sample-Size Invariance of LNRE Model Parameters : Problems and Opportunities. , Vol. (5), No 3. Pp. 145-154.

Baayen R.H., Tweedie F.J. (1998) (2) How may a constant be ? Measures of Lexical Richness in Perspective , vol (32). Pp. 324-332.

Benzecri J.-P. et al. (1981 , t. 3. Dunod, Paris.

Brunet E. (1978) . Genève-Paris : Slatkine-Champion.

Bernet C. Brainerd B., Brunet E., Dubrocard M., Holmes D.I. Hubert P., Labbé D., Serant D., Cossette A. (1994). . Genève-Paris : Slatkine-Champion.

Caillot P. Moine M. (2001)

intervention de la Journée d'Etude "Traitement des questions ouvertes dans les enquêtes et les sondages", 8 juin 2001 à Grenoble.

Cossette A.. (1994) . Genève-Paris : Slatkine-Champion.

Cowell FA. (2000) , in AB Atkinson, F. Bourguignon (Eds), Handbook of Income Distribution, Amsterdam.

Dajoz R. (2000) . Dunod, Paris

Denni B. Caillot P.

, Rapport IEP Grenoble, DESS "PROGIS Etudes d'opinion et de marché".

Lallich S. (1996) , intervention séminaire LABSAD.

Lebart L., Salem A. (1994). . Dunod, Paris.

Legendre P & Legendre L. (1984) . Elsevier, Amsterdam.

Mouillot D. (1999)

. Thèse Université de Corse

Muller C. (1993) , Genève-Paris : Slatkine-Champion.

Patil G.P., Taillie C. (1982) Diversity as a concept and its measurement Vol 77. Pp. 548-561.

Patil G.P., Rao C.R. (1994), , eds, Handbook of Statistics, Vol.12

Thoiron P., Labbé D., Serant D. (1988). . Genève-Paris : Slatkine-Champion.

Vanpeene Bruhier S. (1998), Thèse de Docteur de l'ENGREF, "Sciences de l'environnement", Grenoble.

Zeileis A. Add-on package de R - A language for Data Analysis and Graphics : INEQ (Inequality, concentration and poverty measures and Lorenz curves).