Comparaison de corpus de tailles différentes : l'apport de techniques d'échantillonnage

Yvette Cottavoz

GREIF, Groupe de Recherche en Evaluation, et Ingénierie de la Formation,
Université Pierre Mendès France, BSHM- BP47- 38 040 Grenoble Cedex 09

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

Abstract

The stability of results of lexicometric analysis in the case of text comparison is examined in a situation where the text lengths are unequal. The results were obtained for educational evaluation. The texts analysed are responses to open-ended questions.

Résumé

L'objet de cette étude est la stabilité de résultats d'analyses lexicométriques dans le cas de comparaison de textes de longueurs inégales. L'étude concerne l'exploitation de données d'enquête dans le cadre d'une démarche évaluative. Les textes analysés sont ceux d'une question ouverte.

Mots clés : Analyse des Données, mesures de la richesse du vocabulaire, courbe de Lorenz, méthodes de rééchantillonnage.

1. Introduction

Le contexte

A la suite d' un appel d' offre du Ministère de l' Agriculture et de la Forêt, les chercheurs du GREIF, sous la direction de G. Figari, ont évalué les Brevets de Techniciens Supérieurs Agricoles (BTSA). Pour cela, ils ont mis en place et exploité les résultats d' une investigation sur le terrain (entretiens, enquêtes auprès des acteurs) et ont ainsi cherché à "mesurer les effets d' un dispositif sur l' utilisation d' un curriculum par les acteurs " (Greif, 1994). Ils ont mis en évidence **l' existence d' un effet recrutement sur des options aux ambitions inégales** et des **"parcours aux stratégies cachées"**. En effet les étudiants du monde agricole titulaires d' un BTA (Brevet de Technicien Agricole) sont les plus nombreux dans les options de production (ACSE : Analyse et Conduite des Systèmes d' Exploitation, TV. Techno-Végétale). Dans les options de commerce (TC : Techno-Commercial), d' Aménagement (GME : Gestion et Maîtrise de l' Eau) et de Transformation (IAA : Industrie Agro-Alimentaire) l' origine sociale est plus diversifiée et les titulaires d' un Bac de l' Education Nationale sont plus nombreux. Les deux premières options sont en "régression", les trois dernières en "développement". D' autre part, le projet individuel des étudiants, seul objet de notre étude, présente des orientations différentes selon les options : "rejoindre le monde du travail" (59,4% des étudiants de l' option ACSE), "poursuivre des études" (61,8% pour les étudiants de l' option IAA ; 50,9% pour ceux de

l' option TV ; 41,7% pour ceux de l' option GME). Enfin, certains étudiants sont "des recrutés contraints" dans la mesure où 62% des étudiants provenant d' un BTA suivent une formation qui correspond à leur premier choix, contre seulement 25% des étudiants provenant d' un BAC de l' Education Nationale et ont peut-être élaboré des stratégies.

Analyse de questions ouvertes concernant la réussite

Afin de mieux saisir l' étudiant dans son parcours, les réponses aux questions ouvertes[1] de l' enquête du GREIF ont été étudiées. L' étude visait à montrer que la perception de la réussite, l' expression de doutes et de motivations différaient selon les options. L' utilisation de techniques statistiques appliquées à des données textuelles pose de multiples problèmes (existence de phénomènes rares, variation des caractéristiques de la distribution du vocabulaire en fonction de la taille du corpus...). Nous avons cherché à observer dans quelle mesure les résultats des comparaisons des distributions du vocabulaire des réponses des étudiants des différentes options étaient sensibles à la taille des corpus comparés. Dans ce texte, nous n' exposerons que les conclusions apportées par la mise en œuvre de techniques d' échantillonnage (voir Y. Cottavoz, M. Moine, 1999, pour plus détails).

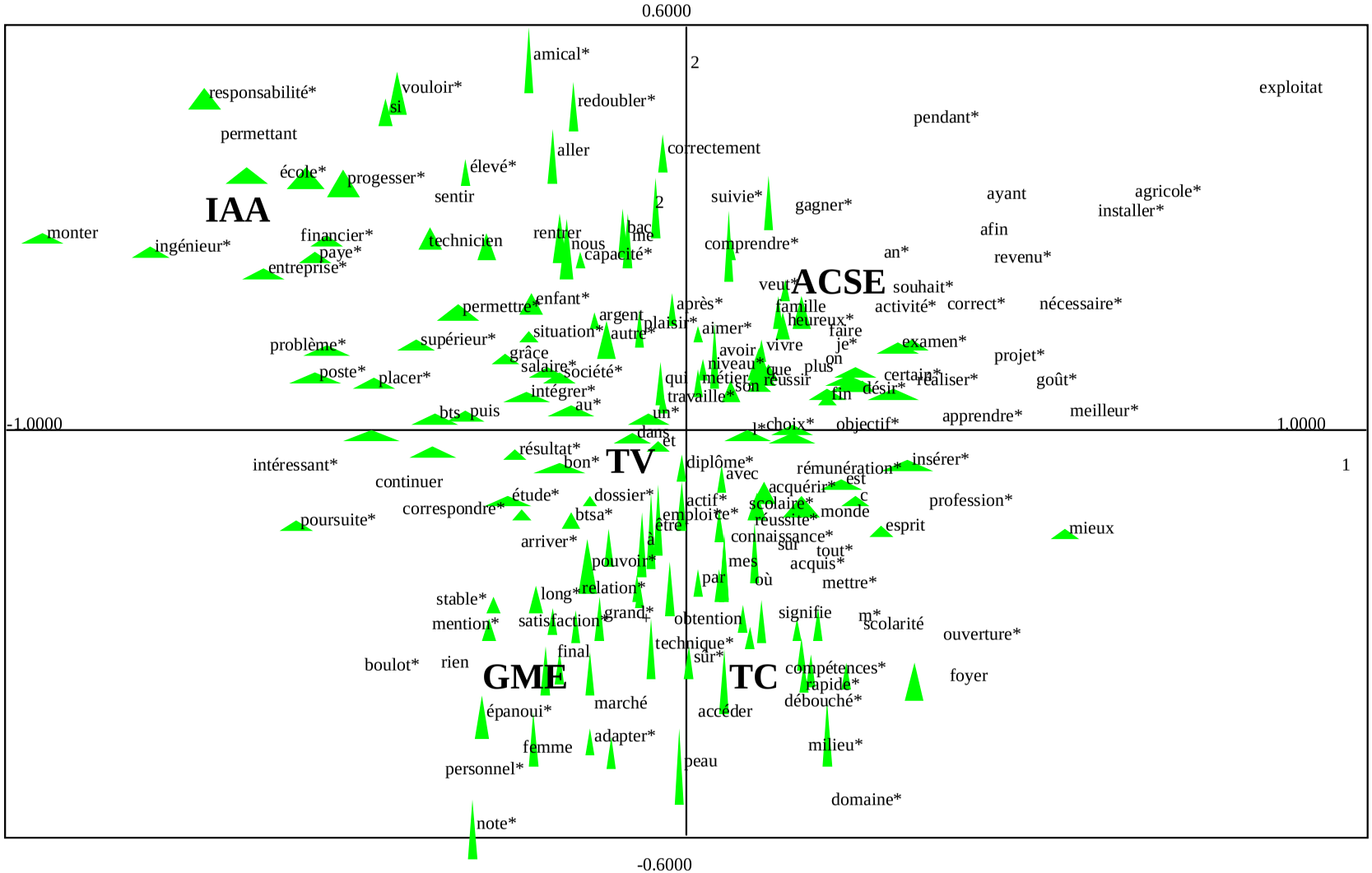
2. La statistique lexicale et l'analyse de contenu : des approches exploratoires

Les tableaux lexicaux étudiés sont les tableaux de contingence dont les lignes correspondent aux mots du corpus, les colonnes à des parties de ce corpus. Ces parties sont soit les énoncés de chaque étudiant-locuteur, soit la réunion d' énoncés de catégories de locuteurs (les énoncés des étudiants regroupés par option). Ainsi, à l' intersection de la $i^{ème}$ ligne et de la $j^{ème}$ colonne de ces tableaux correspond le nombre d' occurrences du mot i dans la $j^{ème}$ partie du texte. Les analyses effectuées comportent des calculs de termes spécifiques à chaque option, le détermination des réponses caractéristiques des options, des Analyses Factorielles des Correspondances et des Classifications Hiérarchiques des tableaux précédemment décrits. L' exploitation statistique a été bien entendu accompagnée de l' analyse du contenu des textes originaux recueillis.

Le premier plan principal de l' Analyse Factorielle des Correspondances du tableau de données décrivant la distribution du lexique de chaque option a été l' un des graphes analysés. Pour ce graphe, les mots retenus sont de fréquence supérieure à 15, et, pour des raisons de lisibilité, seuls ceux de corrélation avec le 1^{er} plan supérieur à un seuil ont été représentés :

---

[1] Les questions ouvertes dont les énoncés des réponses ont été analysés étaient libellées ainsi :
Question 32 : Que signifie pour vous réussir au terme d'un BTSA ?".....…….
Question 33 : Comment définissez-vous en deux phrases chacun de ces trois types de réussite ?
-Réussite scolaire : .....
-Réussite professionnelle:.....
-Réussite sociale:.....
Au total 1144 étudiants ont été interrogés, le nombre de réponses aux questions ouvertes est de 1092 (ACSE : 436; TV : 272; TC : 147; GME : 118; IAA : 119).


0.6000
2
amical*
responsabilité*
vouloir*
si
redoubler*
exploitat
pendant*
permettant
aller
correctement
école*
progesser*
élevé*
sentir
IAA
2
suivie*
gagner*
ayant
agricole*
installer*
monter
financier*
technicien
rentrer
nous
bac
me
comprendre*
afin
ingénieur*
paye*
capacité*
an*
revenu*
entreprise*
ACSE
veut*
souhait*
permettre*
enfant*
après*
famille
activité*
correct*
nécessaire*
argent
heureux*
plaisir*
aimer*
faire
problème*
supérieur*
situation*
autre*
avoir
vivre
je*
examen*
grâce
on
projet*
salaire*
niveau*
que
plus
poste*
placer*
société*
qui
métier
réussir
certain*
réaliser*
goût*
intégrer*
travaille*
son
fin
désir*
au*
bts
puis
un*
choix*
objectif*
apprendre*
meilleur*
-1.0000
1.0000
dans
et
résultat*
TV
diplôme*
rémunération*
insérer*
intéressant*
bon*
avec
1
continuer
acquérir*
est
étude*
dossier*
actif*
scolaire*
c
profession*
correspondre*
btsa*
emploi*e*
monde
poursuite*
être
réussite*
esprit
mieux
connaissance*
arriver*
a
sur
tout*
pouvoir*
mes
acquis*
par
où
mettre*
stable*
long*
relation*
signifie
m*
mention*
satisfaction*
grand*
obtention
scolarité
ouverture*
technique*
final
sur*
boulot*
rien
GME
TC
compétences*
foyer
rapide*
débouché*
épanoui*
marché
accéder
adapter*
milieu*
femme
peau
personnel*
domaine*
note*
-0.6000

En ordonnant la lecture des résultats par regroupement des termes noyau du vocabulaire, nous saisissons l'étudiant dans son parcours mais aussi face à son environnement.

Les thèmes de la **scolarité** et de l'**avenir professionnel** sont présents dans les corpus de chaque option. Les mots "diplôme" et "examen" , "Travailler", "trouver "(du travail), "profession", "métier" sont parmi ceux qui sont les plus utilisés dans le corpus. Le contenu de la formation semble être une préoccupation générale. L'op tion ACSE (secteur production en régression) est la plus sensible à la réussite à l' examen. L' étudiant s' exprime comme un gestionnaire pour l'option ACSE, comme un demandeur d'emploi inquiet pour l'option TV, comme un commerçant pour l'option TC, comme un élément soucieux de son insertion sur le "marché" pour l'option GME, comme un futur titulaire de poste à responsabilité soucieux de sa carrière pour l'option IAA.

La **rémunération** est un paramètre évoqué par tous. Seuls les étudiants de l'option GME minimisent l'aspect financier. Pour les étudiants de l'option ACSE, le vocabulaire traduit le souci d'une ascension sociale rendue visible par des revenus corrects.

Le désir de **reconnaissance sociale et professionnelle** de ces étudiants est fort. ("intégrer", "insérer", "social", "décrocher", "reconnaissance"…). Pour certains, l'ambition s'exprime clairement (options IAA, TC).

Les thèmes du **plaisir, des désirs** mais aussi des possibilités de choix sont évoqués dans toutes les options ("plaisir", "intéressant", "souhait", "désir", "heureux", "satisfaction", "choix"...). Les énoncés des étudiants de l'option GME se distinguent par une connotation un peu plus marquée à ces valeurs. Les termes exprimant un doute sur les possibilités de correspondance entre leur formation et le travail qui les attend ("exercer un métier qui correspond" ou "qui répond, ... à mes désirs"..."(si) possible", "permettre", "possibilité", "pouvoir", "être capable"...) sont moins présents pour l' option TC que pour les autres options.

Des traces concernant **la construction de la réponse, le niveau de langage, la complexité des phrases et plus généralement des traces d'énonciation peuvent être observées** ("et", "mais", "afin", "si",... personnalisation du discours : "je", "nous", "on",....).

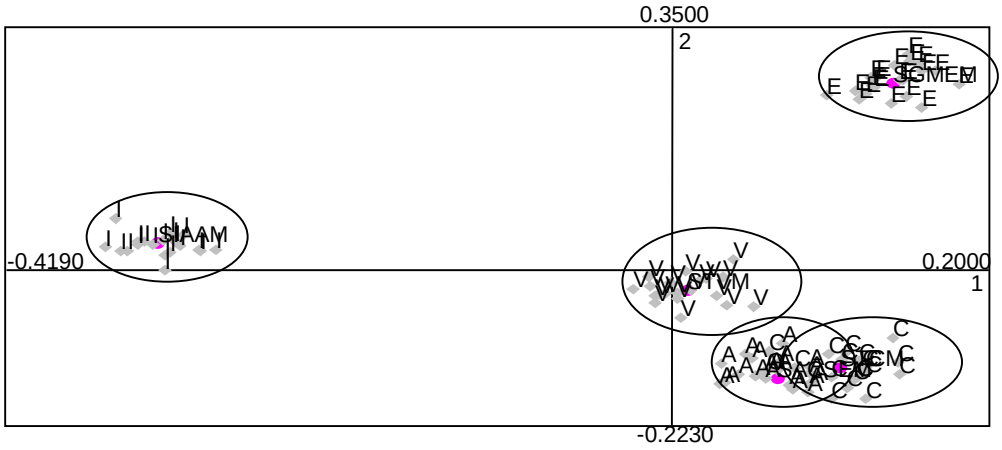
En conclusion : des sous-cultures en fonction des options

L'hétérogénéité du lexique reflète bien la spécificité des options et a été interprétée comme l'expression de sous-cultures en relation avec leur futur métier et l'origine sociale. Des doutes sont exprimés par les étudiants relativement à leur réussite aussi bien sociale que professionnelle et à la qualité de leur formation. Un nombre relativement important d'étudiants des options IAA et GME ont une ambition de poursuite d'études. Les étudiants de l'option TC expriment moins de doute que ceux des options ACSE et TV pour qui le travail est un moyen de vivre et d'insertion sociale. Néanmoins, le désir de reconnaissance sociale qui passe par un "travail intéressant" est exprimé par tous. Ils placent la réussite professionnelle en correspondance avec leur choix et même leur plaisir.

3. Etude de la stabilité des Analyses des Correspondances des Tableaux lexicaux

Des analyses utilisant des **techniques d'échantillonnage** ont été réalisées dans le but de tester si les distributions conditionnelles du vocabulaire différaient de façon significative selon l'option. Nous notons les textes produits par les étudiants des cinq options : T_k, k=1,…5, et leurs tailles N_k, k=1,…5. Selon le modèle binomial développé par C. Muller (1993), nous

considérons des textes T'_k , k=1,…5 de tailles N'_k , k=1,…5 extraits de T_k, k=1,…5 . Chaque texte T'_k est obtenu par tirage de N'_k mots du texte T_k , chaque occurrence du texte T_k ayant une probabilité égale d'être choisie. Afin de neutraliser l'effet de la taille des textes produits, nous prenons $N'_1 = \ldots = N'_5 = N'$. Sous cette hypothèse, la distribution du vocabulaire pour le texte T'_k est une observation d'une loi multinomiale de paramètre N' et de probabilités de tirage les fréquences observées des modalités du vocabulaire dans le texte T_k. Pour chaque option, 20 réplications ont été simulées. Nous avons ensuite calculé la moyenne des 20 réplications simulées. Une AFC du tableau de contingence dont les lignes correspondent à des mots et les colonnes aux options a été effectuée. A l' intersection d' une ligne i et d' une colonne j est associé l' effectif moyen du mot i calculé sur les 20 échantillons simulés pour l' option j. Les colonnes correspondant aux réplications sont projetées en tant que modalités supplémentaires sur les différents axes factoriels (Lebart L. et al., 1995).



Premier plan factoriel de l'AFC des données obtenues par rééchantillonnage (seuil de fréquence : 8)

Note : Les lettres A, V, C, E, I désignent les profils des réplications des options ACSE, TV, TC, GME, et IAA. Les profils moyens sont identifiés par SACSEM, STVM, STCM, SGMEM, SIAAM.

Des résultats confirmés

Les résultats obtenus à partir du tableau de données brutes ne sont pas fondamentalement remis en cause :

- Les options ont des "profils" (distribution du vocabulaire des étudiants des options) significativement différents si l' on tient compte des 4 axes des analyses.
- L'oppo sition entre les options IAA, GME, TC apparaît sur le plan 1-2 des deux types d' analyses.

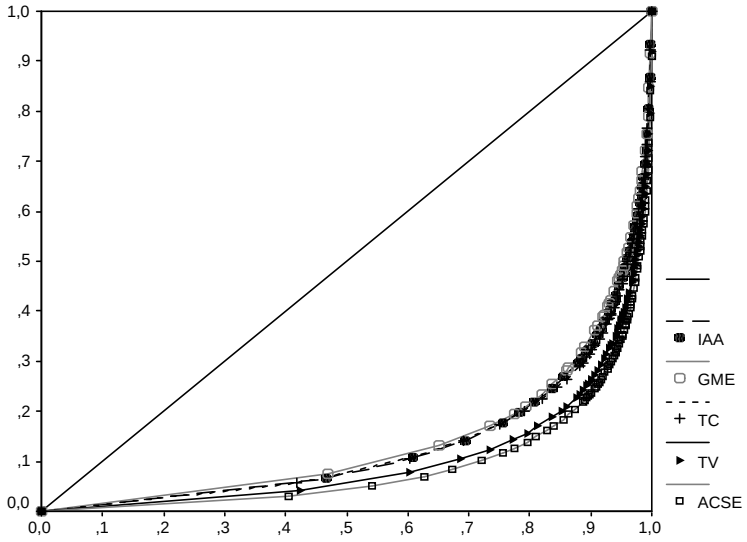
Modulation des résultats obtenus sur les données brutes

- Pour les AFC effectuées à partir des données brutes, l' option ACSE contribue fortement au 1er axe factoriel et s' oppose de façon nette à l' option IAA, ce qui n' est pas le cas des résultats d' analyses sur les données échantillonnées. Ainsi, la singularité de l' option ACSE serait due en grande partie à l' effet de la taille de l' échantillon de personnes interrogées. (la taille du corpus de cette option est beaucoup plus importante que celle des autres options, ce qui a une influence sur la structure de son vocabulaire).
- Dans le plan engendré par les axes 2 et 3, on peut observer que les options ACSE et TV ont un vocabulaire comportant des similitudes en relation avec les thèmes de l' insertion sociale et professionnelle, de la réussite scolaire, de la satisfaction personnelle, des doutes. Ce phénomène n'était pas aussi clairement mis en valeur dans les analyses obtenues à partir des données brutes.

4. Etude de la stabilité de courbes de concentration

Les courbes de concentration (ou de Lorenz) sont utilisées en économie. Dans le contexte de cette étude, elles permettent d' analyser la contribution des termes (classés selon leur fréquence dans le corpus) à la « masse » du corpus. Elles sont construites de la façon suivante : à chaque valeur f de la fréquence de termes observée, on associe un point dont l'abscisse est le pourcentage de termes de fréquence inférieure ou égale à f, et dont l'ordonnée est la contribution en pourcentage de ces mêmes vocables à la « masse » du corpus[2].

Le graphe ci-dessous met en évidence la contribution importante des vocables de fréquence élevée pour toutes les options, ce phénomène est plus marqué pour les options ACSE et TV (vocabulaire moins riche).



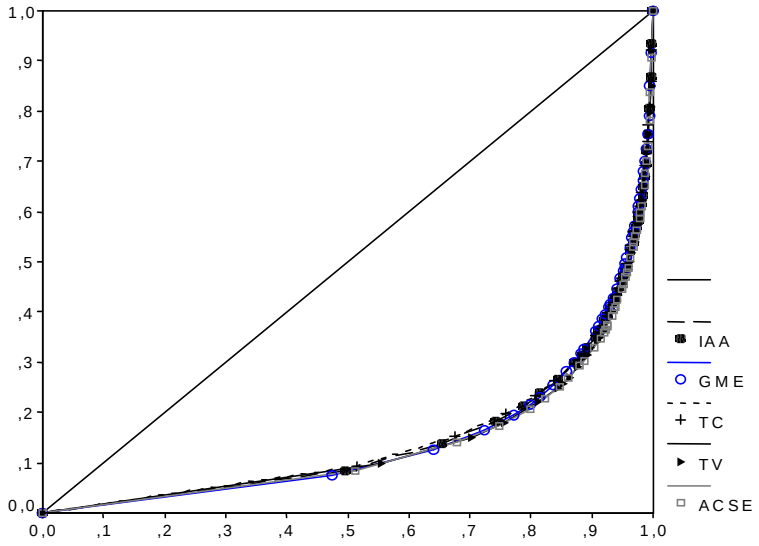
Courbes de Lorenz sur les données brutes

Afin de vérifier dans quelle mesure les différences observées sont dues à des effets de taille des corpus, trois démarches ont été entreprises.

[2] Ces graphes sont précieux lorsque l'on effectue des analyses de données pour lesquelles les mots les moins fréquents du corpus sont éliminés : ils permettent de mesurer la perte d'information qui résulte de cette suppression.

Confrontation des courbes de concentration des données brutes et des courbes de concentration des données obtenues par réplications des profils de chaque option

Les graphes des courbes de Lorenz construites à partir des moyennes des simulations de lois multinomiales décrites au paragraphe 3 révèlent que ces courbes sont peu différenciées du fait que l'on raisonne à taille de corpus égal. Un examen plus précis des courbes (zoom) révèle que le corpus des étudiants de l' option TC se caractérise par plus de richesse du vocabulaire si on considère les termes de fréquence faible et par plus de pauvreté si on considère les mots utilisés plus fréquemment. Les termes liés à l' effet curriculum (lexique de type commercial pour lequel la communication est primordiale) sont plus particulièrement présents dans les gammes de fréquence faible (fréquences comprises entre 2 et 5).



Courbes de Lorenz sur les données obtenues par rééchantillonnage

Comparaison des courbes de concentration construites à partir des données obtenues par réplications de chaque option et des résultats théoriques spécifiques au modèle binomial

Dans le contexte d'un modèle binomial, C. Muller (1993), remarque que « partant d'un texte connu T de taille N, la distribution théorique » de la fréquence du vocabulaire « d'un texte plus court de taille N' extrait de celui-ci peut être calculée». A toute fréquence observée f dans T, on associe f+1 sous-fréquences (notées f') de probabilité :

$$\mathbf{P[f'=k]} = \mathbf{C_f^k\, p^k (1-p)^{n-k}}, \quad \text{où } \mathbf{p} = \frac{\mathbf{N'}}{\mathbf{N}}$$

dont on déduit les effectifs théoriques des mots de fréquence f' pour le texte réduit. Nous avons confronté les graphes des courbes de Lorenz obtenues à l'aide de tables de sous-fréquences théoriques et les courbes correspondant aux moyennes des simulations de lois multinomiales. Les résultats obtenus par les deux procédés concordent et confirment donc les conclusions du paragraphe précédent.

Analyse d'échantillons de réponses d'étudiants

Un échantillon de réponses d'étudiants a été constitué pour chaque option (les échantillons extraits, de tailles égales, sont obtenus selon un procédé de tirage à PESR). Les courbes de concentration construites à partir des textes de ces réponses échantillonnées de chaque option

ont confirmé les résultats précédents : les courbes de Lorenz associées aux options sont proches, il apparaît une spécialisation du vocabulaire pour les étudiants de l'option TC. Néanmoins, des écarts entre les tailles des corpus associés aux options existent dans la mesure où la longueur moyenne des réponses varie selon les groupes d'étudiants concernés. Les réponses des étudiants de l'option GME sont nettement plus courtes que celles des étudiants des autres options et ceci explique vraisemblablement que les vocables de fréquence faible aient une contribution plus importante à la masse totale de leur corpus.

5. En conclusion : des sous-cultures différenciées selon les options, l'effet curriculum a un impact sur le vocabulaire

La mise en œuvre de techniques d' échantillonnage a permis de montrer que les profils des options étaient significativement différents. Ceci confirme donc l' analyse descriptive réalisée sur les données brutes.

La prise en compte de simulations a permis de relativiser l' écart entre le vocabulaire des étudiants de l' option ACSE et celui des autres options, et de mettre en évidence des préoccupations communes à celles des étudiants des options TV et ACSE.

La construction de courbes de concentration de la fréquence du vocabulaire et leur comparaison indiquent que les corpus ont une diversité de vocabulaire proche lorsqu'on compare des corpus de tailles peu différentes. Un phénomène de spécialisation est plus marquée pour l'une des options.

Références

Authier-Revuz, J (1982). *Hétérogénéité montrée et hétérogénéité constitutive : éléments pour une approche de l'autre dans le discours*, D.R.L.A.V n°26.

Benzecri J.-P. et al.. (1981). *Pratique de l'analyse des données*, t. 3. Dunod, Paris.

Bernet Ch. (1983). *Le vocabulaire des tragédies de Racine*. Genève-Paris : Slatkine-Champion

Bernet C., Brainerd B., Brunet E., Dubrocard M., Holmes D.I., Hubert P., Labbé D., Serant D., Thoiron P. (1988). *Etudes sur la richesse et la structure lexicales*. Genève-Paris : Slatkine-Champion

Cossette A. (1994). *La richesse lexicale et sa mesure*, Genève-Paris : Slatkine-Champion

Cottavoz Y., Moine M. (1999*) l'évaluation d'un curriculum de BTS de l'enseignement agricole : de l'évaluation par les acteurs à l'évaluation par les sujets*, Grenoble. Rapport, Laboratoires GREIF et LABSAD, Université Pierre Mendès France, France.

Figari G. (1994). *Evaluer : quel référentiel ?* Bruxelles, Ed. De Boeck Université.

GREIF (1994). *Etude sur les BTS de l'enseignement agricole Rapport 3.* Laboratoires GREIF, Université Pierre Mendès France, France.

Lebart L., Salem A. (1994). *Statistique Textuelle*. Dunod, Paris.

Lebart L., Morineau A., Piron M. (1995). *Statistique exploratoire multidimensionnelle*. Dunod

Mainguenaud D. (1991). *L'analyse du discours*, Hachette, Paris.

Muller Ch. *Principes et méthodes de statistique lexicale.* Genève-Paris : Slatkine-Champion

Patil G.P., Taillie C. (1982). Diversity as a concept of measurement, *JASA*, September 1982, Vol. 77, N° 379.

Pottier B. (1992). *Théorie et analyse en linguistique*, Paris, Hachette.