Exploration sans a priori ou recherche orientée par un modèle : contributions et limites de l'analyse lexicale pour l'étude de corpus documentaires

Jean Moscarola[1], Vassiliki Papatsiba[2], Yves Baulac[3]

[1]IREGE – Université de Savoie (Annecy) – 74000 Annecy le Vieux – France

[2] CREF – EA 1589 Université de PARIS X – Nanterre – France

[3]Sphinx-Développement – 7 rue Blaise Pascal – 74600 Seynod – France

Abstract

The use of textual data analysis methods is steady growing particularly among researchers who are not trained to use such methods. Different approaches can be used which are very different from those involved by traditional mainstreams such as content analysis, semiology or grounded theory. They give to the research a more scientific image although being mostly used only on an exploratory level.

This communication aims at giving an example of these different approaches and discussing them on a methodological point of view. Therefore we present a research led in educational sciences which consists in studying a corpus of 80 students reports about there stay in a foreign university. We present firstly the process of exploratory research led without any a priori from the examination of lexicon and lexical associations. Secondly, following an hypo deductive procedure, we show how to base textual data analysis on an explicit modelling and hypothesis to compute textual measures and test the model. The search for specificities and the measurement of lexical intensities will be used.

Finally, after we discuss, the contributions and limits of each approach. Our conclusion point out the unavoidable implication of the researcher and the necessary flexibility of the tools they have to make us of.

Résumé

L'usage des méthodes d'analyse de données textuelles se développe de plus en plus notamment auprès de chercheurs peu familiarisés avec les méthodes statistiques. Ces méthodes autorisent différentes approches dépendant des techniques et des outils utilisés mais également de la problématique de recherche et de la nature du corpus. Dans l'univers des études d'inspiration littéraire, elles se différencient nettement des courants plus traditionnels du commentaire critique, de la sémiologie ou de l'analyse de contenu. De ce fait elles bénéficient d'une image plus scientifique alors que l'usage le plus souvent exploratoire de ces techniques ne le justifie pas toujours. Le but de cette communication est d'illustrer et de discuter ces approches d'un point de vue méthodologique. On s'appuie pour sur une recherche en sciences de l'éducation portant sur un corpus documentaire constitué de 80 rapports relatant des séjours d'étude à l'étranger.

On examinera successivement les explorations menées sans à priori à partir des lexiques et des associations lexicales pour adopter ensuite une approche hypothético déductive mettant en œuvre l'élaboration préalable d'un modèle. La discussion portera notamment sur la recherche des spécificités lexicales et les mesures d'intensité. A partir de la présentation des résultats de notre cas, on discutera les apports et limites de ces approches pour conclure sur l'inévitable implication du chercheur et la nécessaire flexibilité des outils à mettre en œuvre.

Mots-clés : Corpus documentaire ; analyse de contenu ; statistique lexicale ; méthodologie

1. Les documents analysés

On considère ici un corpus[1] de 80 rapports rédigés par les bénéficiaires d'une bourse régionale ayant effectué un séjour Erasmus[2] de plus 5 mois dans une université étrangère.

Ces rapports ont été rédigés en 1996 et 1997 par des étudiants français, à la demande du conseil régional Rhône Alpes. Ils représentent environ 800 pages sur papier. Quarante-trois rapports sont rédigés par des filles et trente-sept par des garçons. Leur niveau d'études est de bac+3 à bac+5 et les établissements fréquentés sont de type université, grande école et institut universitaire. Des disciplines tant littéraires que scientifiques sont également représentées en passant par des formations professionnelles. Enfin, onze pays de l'Union européenne ont constitué la terre d'accueil pour ce séjour temporaire de cinq à neuf mois des étudiants Erasmus : Allemagne, Autriche, Danemark, Espagne, Grande-Bretagne, Grèce, Italie, Irlande, Pays-Bas, Portugal et Suède.

Dans le contexte d'une recherche en sciences de l'éducation (Papatsiba), on cherche à savoir comment les étudiants appréhendent leur expérience d'études et d'immersion dans un contexte étranger. A l'aide de quelles catégories la construisent-ils et quels registres apparaissent lors de leurs évocations ? Enfin, bien que ces documents soient « de commande[3] », permettent -ils d'avancer dans la réflexion sur le rapport entre le niveau d'implication personnelle et la « valeur » accordée à ces séjours, compris entre une expérience de tourisme culturel, une immersion dans la matérialité d'un quotidien inhabituel et une période de décentrage culturel et de retour sur soi ?

Des informations plus complètes sur cette recherche, les données utilisées et les résultats obtenus sont disponible sur le site cité en bibliographie.

2. L'analyse sans a priori[4] : l'approximation lexicale

Un premier usage de l'analyse des données textuelles consiste à utiliser la statistique pour réduire le corpus à quelques éléments lexicaux sélectionnés sur la base de calculs de fréquence (Muller). Ces procédés sont d'autant plus efficaces qu'ils s'appliquent à des formes lemmatisées du texte (Gavard Perret, Moscarola) et qu'ils concentrent l'analyse sur les formes les plus significatives, substantifs, verbes, adjectifs notamment (Marchand).

Dans une perspective d'analyse de contenu on peut qualifier ces procédés d'approximation lexicale. Il s'agit en effet de prendre connaissance du texte en remplaçant sa lecture par celle d'un substitut lexical (liste de mots, tableaux, cartes factorielles). On gagne ainsi du temps mais c'est aussi un moyen de révéler les structures de l'énonciation difficilement perceptibles par la lecture du texte lui-même.

---

1 Nous tenons à remercier ici le Conseil régional de Rhône Alpes qui a autorisé l'accès aux documents en vue de leur utilisation dans le cadre d'une thèse en Sciences de l'Education.

2 Programme européen pour la mobilité des étudiants

3 Rapport personnel d'une dizaine de pages dactylographiées, rédigées en français, relatant mon vécu à l'étranger.(Eléments pratiques : vie dans l'établissement ou l'entreprise, vie sociale, connaissance du pays, vos conseils pour vos successeurs, etc...)

4 L'expression « sans à priori » est relative au texte étudié. Nous n'évoquons pas là l'hypothèse de travail de toute l'analyse des données textuelles selon laquelle la statistique et notamment l'AFCM permet de révéler les structures linguistiques et les « lieux » de tout texte. L'absence d'à priori s'entend ici par rapport au contenu du texte examiné. Pour mener à bien une analyse lexicale aucun présupposé concernant le contenu du texte n'est en effet nécessaire.

2.1. Découvrir les usages les plus fréquents : les lexiques

La fréquence des termes utilisés dans les rapports Erasmus permet de se faire une idée de leur contenu. Grâce à l'analyse syntaxique on peut isoler dans cet ensemble, les seuls substantifs. Ces référents noyaux pour reprendre la terminologie de l'analyse propositionnelle de discours (Pecheux) pointent sur les objets ou catégories du monde auxquels les auteurs se réfèrent dans leur écrits.

Afin de saisir, en de ça de la variété des expériences individuelles ce qu'il y a de commun à tous ces rapports on se concentre sur les termes les plus fréquemment utilisés et présents dans au moins la moitié des documents analysés.

L'examen du lexique met en évidence les champs lexicaux suivant:
-La vie étudiante et universitaire : *étudiant, cours, université, étude, professeur, matière…*
-Les conditions de vie : *chambre, prix, place, logement, résidence…*
-Les lieux : *ville, pays, région, endroits, lieu, bâtiments, salle…*
-Le temps : *année, heure, semaine, mois, jour…*
-Les contacts : *vie, personne, habitant, contact, ami, échange, occasion…*
-L'expérience, l'épreuve *: problème, expérience, difficulté, chance, différence…*

Sur la base de cette première réduction du corpus les grands thèmes communs à l'ensemble de ces rapports se dessinent. Ils peuvent être enrichis par l'examen d'autre catégories grammaticales. Ainsi dans la liste des adjectifs les plus utilisés par tous trouve-t-on : *différent, bon, étranger, français, possible, difficile seul, international, nouveau, culturel, cher….* Quant aux verbes on peut remarquer : *rester, partir, vivre, comprendre, découvrir, rencontrer, visiter*.

A ce premier stade de l'étude on dispose déjà d'éléments permettant de répondre à la question de recherche.

2.2. Les mots en contexte : les associations lexicales

En sortant les mots de leur contexte, la statistique lexicale soulève autant de questions qu'elle apporte de réponses pour l'interprétation des contenus. La recherche des segments répétés (Salem) est une première manière de répondre à cette critique mais aussi révélatrice qu'elle puisse être, elle ne permet que de repérer les « rigidités » du discours. Demeure la nécessité du retour au texte. Là encore l'analyse lexicale peut faire gagner beaucoup de temps sur la fastidieuse recherche des concordances.

Depuis les travaux de Benzécri, l'analyse factorielle des correspondances multiples est la méthode la plus utilisée pour rendre compte de la manière dont les éléments lexicaux d'un corpus se trouvent associés les uns aux autres. Qu'on y voit un moyen pour révéler les « fondements topiques », lieux du discours (Reinert) ou la trace des modèles cognitifs de l'auteur (Chanal) cette technique appliquée à notre corpus rend compte de la situation des termes du lexique dans leur contexte. La carte présentée en Figure 2 montre comment les mots clés du corpus se trouvent le plus fréquemment associés les uns aux autres. Sa lecture complétée par celle du tableau des contributions peut conduire à l'analyse de contenu suivante :

- En concentrant l'interprétation sur les axes de la carte on peut lire deux oppositions sur l'axe horizontal entre la vie domestique (à droite) et scolaire (à gauche) et sur l'axe vertical entre les motifs du séjour, universitaires (en haut) et de dépaysement (en bas).

- En concentrant l'attention sur les constellations de termes on peut identifier dans la périphérie de la carte les 3 thèmes suivants : La vie pratique et les conditions de vie (en haut à droite), La scolarité (à gauche), La vie sociale (en bas à gauche), enfin on trouve au centre les mots indifféremment associés aux thèmes périphériques, ils renvoient à l'expérience du séjour et des études à l'étranger.

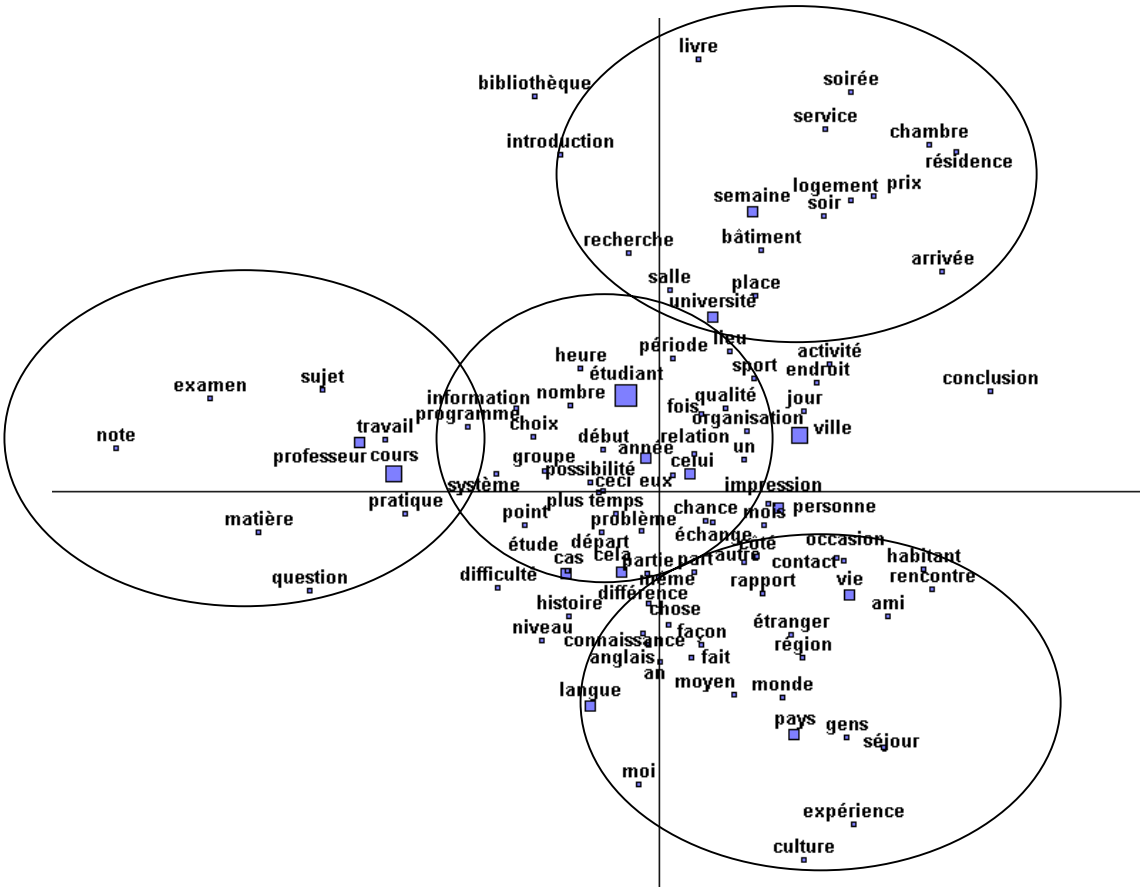

Figure 2 : Associations lexicales établies sur le corpus fragmenté en phrases

Cette analyse peut être confirmée par une classification des fragments du texte selon leur « proximité lexicale » (les éléments d'une même classe ont plus de termes en commun que les éléments de 2 classes différentes). L'examen des phrases les plus significatives de chaque classe est une manière de revenir au texte et de confirmer les interprétations fondées sur les seuls éléments de l'analyse factorielle.

Typologie, LES CONDITIONS DE VIE :	**Typologie, LES ETUDES :**
Le prix des chambres oscille entre 200-300 000 lires pour les chambres doubles et 250-350 000 lires pour les chambres individuelles Chaque université a un "Accommodation Service" qui aide les étudiants à trouver une chambre en résidence universitaire ou un logement indépendant C'est un service bien rodé mais attention, une fois signé le contrat, vous êtes poings et pieds liés et ne comptez pas sur les services de logement " accomodation services " pour être conciliants	La plupart des étudiants qui comptent passer tel examen viennent y voir le type de question ou d'épreuve auxquels ils doivent s'attendre et entourent le candidat, prenant des notes, commentant, écoutant et scrutant tout avec attention Comme les notes de ces examens sont prises en compte pour pouvoir poursuivre en thèse, les examens représentent un gros enjeux pour les étudiants italiens

Figure 3 : Extraits caractéristiques des thèmes mis en évidence par les associations lexicales

Cette manière de faire complète très efficacement le simple examen des éléments les plus fréquents du lexique. Elle présente en outre l'avantage de fournir une vision organisée et

synthétique du contenu à analyser ce qui explique la popularité de ces méthodes[5] : essentiellement fondées sur l'analyse statistique elles donnent l'impression d'un procédé automatique et échappant apparemment à la subjectivité de l'analyste. Elles bénéficient ainsi par rapport aux méthodes classiques de l'analyse de contenu d'une image nettement plus scientifique.

3. Analyses fondées sur un modèle : mesures lexicales

Par opposition aux méthodes exploratoires conduites sans a priori, les recherches fondées sur un modèle utilisent un corps d'hypothèses qu'on cherche à confronter aux données de l'expérience. Ces hypothèses correspondent à la formulation d'une théorie qu'on peut définir comme un ensemble de concepts mis en relations. La confrontation des concepts aux phénomènes nécessite un travail de construction de procédés d'observation et de mesure indispensable pour établir une correspondance entre le domaine de l'empirie et celui de la théorie. Le modèle est à l'articulation de ces 2 mondes.

Ainsi par exemple dans une enquête conduite par questionnaire fermé, le questionnaire est la traduction opérationnelle du modèle et des hypothèses qui en ont été à l'origine. Selon la valeur et les propriétés statistiques des données ainsi recueillies, la théorie sera rejetée ou validée.

3.1. Utiliser le modèle des données

3.1.1.Texte et contexte

Lorsque le document analysé est structuré on dispose d'indications modélisées a priori : identité des personnes interviewées, texte des questions et des réponses, indication relatives aux titres, sous titres et corps du texte, règles éditoriales… Ces informations qualifiées de données de contexte répondent le plus souvent à un modèle (l'age est exprimé en années, les catégories sociales sont l'objet de nomenclatures, un plan s'articule en niveaux…) et peuvent servir de point de départ pour la formulation d'hypothèses répondant à l'idée générale de la détermination du contenu par son contexte.

Une première orientation de la recherche s'impose alors naturellement avec la recherche des ***spécificités lexicales***.

3.1.2. Spécificités lexicales

On appelle spécificité lexicale (Müller, Lebart, Brunet) une indication statistique établissant le fait que la fréquence d'usage de tel vocable dépend du contexte dans lequel il est employé. Il peut s'agir du simple rapport entre la fréquence observée et la fréquence correspondant à l'équi-répartition (l'influence est d'autant plus forte que le rapport est supérieur à 1) ou de la probabilité pour qu'on obtienne la fréquence observée (l'influence est d'autant plus forte que cette probabilité est petite).

Ainsi dans notre exemple peut-on se demander si les différences dans l'usage des termes employés dans les rapports sont explicables par l'identité de l'auteur et la nature du séjour. Cette investigation devrait être préalable à toute autre recherche d'explication d'un autre ordre.

Les résultats obtenus montrent en effet une spécialisation du vocabulaire qui notamment distingue l'Angleterre et l'Irlande par rapport aux autres pays. D'autre-part on constate que les

[5] Elles sont majoritairement utilisées dans les recherches empiriques présentées au cours des dernières Jadt

filles sur utilisent le champ lexical des études et les garçons celui de la vie quotidienne et des contacts.

Caractéristiques selon le sexe

	F	M	Ensemble
Effectif de la catégorie	43	37	80
Effectif ayant répondu	43	37	80
Nombre total de mots	128848	115963	244811
Fréquence de ce mot	10655	10177	20832
Longueur moyenne	2 996,47	3 134,14	3 060,14
Nombre de mots différents	8487	9435	12835
Nombre de mots uniques	2544	3256	5800
Répétition moyenne	15,18	12,29	19,07
Poids des 5087 mots communs	95.9%	94.4%	-
Nombre de mots exclusifs	3400	4348	-
Poids des mots exclusifs	4.1%	5.6%	-

Vocabulaire spécifique selon le sexe

Fille	Nb. (Specif.)	Garçon	Nb. (Specif.)
cours	498 (1,24)	ville	408 (1,25)
professeur	192 (1,28)	logement	115 (1,36)
livre	99 (1,22)	lieu	90 (1,21)
ami	86 (1,22)	étranger	80 (1,43)
bibliothèque	84 (1,32)	partie	78 (1,33)
difficulté	70 (1,28)	échange	71 (1,56)
note	65 (1,24)	monde	69 (1,35)
fait	51 (1,21)	habitant	60 (1,31)
		bâtiment	50 (1,22)
		moyen	48 (1,33)
		qualité	36 (1,25)

Vocabulaire spécifique selon le pays

(analyse factorielle du tableau de contingence pays x mots spécifiques)

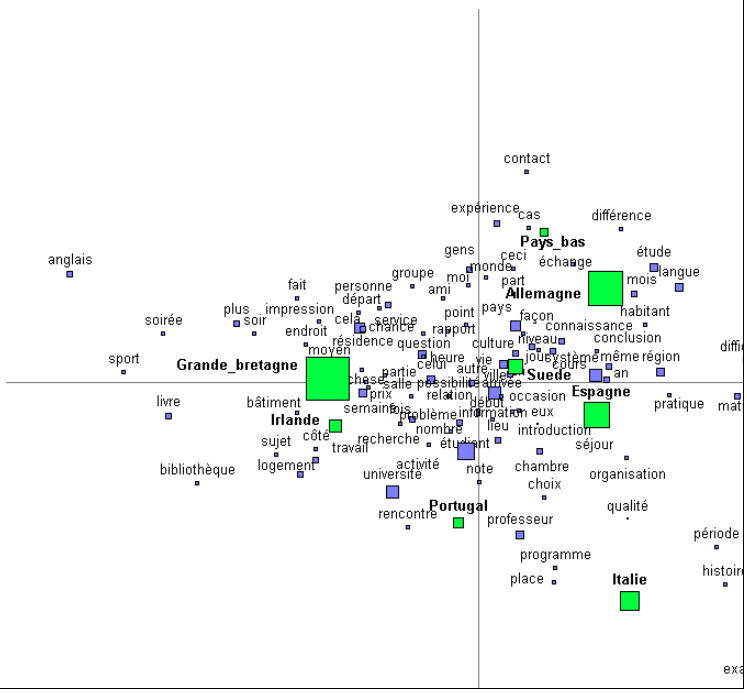

Figure 4 : Les effets du contexte et la recherche des spécificités lexicales

3.2. Elaborer un modèle pour l'analyse du contenu des textes

Pour aller plus loin q'une simple vue du texte comme déterminé par les conditions repérables de sa production il faut engager un travail de modélisation et construire des indicateurs adéquats.

3.2.1. La méthode : dictionnaires et intensités lexicales

Par opposition à la démarche exploratoire la recherche est orientée si on dispose d'une théorie proposant une base d'interprétation du corpus en cours d'analyse. Pour y parvenir il nous faut construire des indicateurs qui permettent :

- La mise en correspondance des concepts de la théorie avec leur manifestation concrète : les vocables choisis dans la production du texte (actes de langage). De la même manière qu'on construit une échelle de mesure on construit pour cela des répertoires appelés aussi ***dictionnaires***, établissant la correspondance entre l'apparition dans le texte d'une forme graphique et la catégorie abstraite à laquelle elle renvoie. Le dictionnaire Dj de la catégorie ou de la dimension j peut être noté comme l'ensemble des éléments $l_{j,i}$ appartenant à la dimension j : Dj, {lj,1…lj,i, lj,n}.
- L'établissement de la fréquence avec laquelle on trouve dans le corpus des formes appartenant à ces catégories. Plus cette fréquence est élevée plus la dimension correspondante est présente dans le texte. On appelle ***intensité lexicale*** de la dimension j cette mesure. Elle s'exprime en pourcentage. Elle est calculée pour chaque

fragment du corpus de la manière suivante : I_j, $\Sigma i\ E(l_{i,j})/n$ avec : $E(l_{i,j})$, nombre d'apparitions de la forme $l_{i,j}$ dans le fragment considéré et n : nombre total de formes présentes dans le fragment. Les mesures d'intensité s'interprètent de la manière suivante : au cours de l'énonciation, l'auteur ou le locuteur décide consciemment ou non d'utiliser tel ou tel mot en puisant dans les différents répertoires correspondant aux dimensions qu'il souhaite exprimer. Les mesures d'intensité donnent ainsi une indication sur la fréquence relative avec laquelle ces dimensions sont évoquées dans le discours.

3.2.2. La modélisation du discours sur l'étranger

Le modèle utilisé ici (Papatsiba) renvoie aux positions discursives des auteurs et à leur articulation avec les différentes approches de l'étranger.

D'une part, les positions adoptées par l'auteur le situent en tant que sujet universel (A. Culioli) ou en tant qu'ego (Benveniste) face à son expérience et expriment par-là de la distanciation ou de l'implication dans son discours.

D'autre part pour rendre compte du discours sur l'étranger on procède à une classification (l'Ecuyer, Glaser et Strauss) des catégories thématiques dont la structuration (Papatsiba) oppose une approche distanciée comportant des données générales sur le pays à une approche qui implique les autres et le sujet. Le récit de l'expérience Erasmus se situant entre le vécu d'une altérité spatiale et matérielle et le vécu d'une altérité interpersonnelle

L'hypothèse principale peut alors être formulée ainsi : A la polarité implication / distanciation dans le discours correspond la thématique extériorité / intériorité culturelle de la narration. La construction des indicateurs suivants permet d'opérationnaliser ce modèle :

- La dimension implication / distanciation : sur la base du corpus lemmatisé on calcule l'intensité de la forme *je.* En moyenne on observe que l'usage du "je" représente 1,18% des actes de langages[6], ce qui nous permet de définir 3 groupes de rapports correspondant respectivement aux 27 rapports dans lesquels l'auteur manifeste une faible implication (intensité je <0,5%), à 26 rapports d'implication moyenne et aux 27 rapports montrant une implication forte (intensité je >1,39).

	Nombre de rapports	Intensité JE(%)	Intensité Je, me mon, mes ma....
je <0,5%) Implication faible	27	0,30	0,60
de 0,50 à 1,39	26	0,97	1,78
je >1,39% Implication forte	27	2,25	3,93
TOTAL	80	1,18	2,11

Test de Fisher :Intensité JE(%) : F, 156,19, 1-p, >99,99%
Implication :, F, 167,58, 1-p, >99,99%

Implication selon le sexe

Classes d'implication/Sexe	F	M	TOTAL
Faible (je <0,5%)	20,9% (9)	48,6% (18)	33,8% (27)
de 0,50 à 1,39	32,6% (14)	32,4% (12)	32,5% (26)
Forte (je >1,39%)	46,5% (20)	18,9% (7)	33,8% (27)
TOTAL	100% (43)	100% (37)	100% (80)

La dépendance est significative. chi2 = 9,01, ddl = 2, 1-p = 98,90%.

Figure 5 : Répartition des rapports selon l'implication estimée par l'intensité lexicale de « je »

- Les dimensions thématiques : 11 thèmes ont été définis, répartis en 5 catégories distribuées selon l'opposition extériorité intériorité : le cadre donné en arrière plan, les repères et fonctionnalités du cadre de vie, la sphère d'action individuelle, la sociabilité et les relations interpersonnelles, les images et empreintes du sujet. Pour chacun des 11 thèmes, des dictionnaires contenant le vocabulaire qui y renvoie ont été établis.

[6] Les pourcentages obtenus sont très faible car les intensités sont calculées par rapport au corpus lemmatisé complet (incluant les mots outils)

Grille thématique

1 Un cadre de vie donné en arrière plan.
1.1 Pays, région, ville
1.2 Patrimoine matériel
1.3 Les particularités culturelles d'un peuple (patrimoine culturel)
2 Repères et fonctionnalité du cadre de vie.
2.1 Le site universitaire
2.2 Le logement.
3. La sphère d'action individuelle.
3.1 L'environnement pratique de la vie quotidienne
3.2 Les études
3.3 Les loisirs
4 Sociabilités et relations interpersonnelles.
4.1 La communication
4.2 Les rencontres
5 Images et empreintes du sujet : l'écriture personnelle de l'expérience
5.1 Le sujet dans l'écriture de l'expérience

Exemple de dictionnaires

Pays région ville

agriculture, alentours, aménagement, armée, artère, capitale, chute, circonscription, conflit, constitution, crise, croissance, démocratie, destruction, développement, domination, économie, élection, empire, époque, est, sud, nord, ouest, frontière, gouvernement, homme_politique, président, reine, parlement, roi, guerre, bataille, bombardement, indépendance, industrie, institution, kilomètre, Land, légende, loi, mer, métropole, milliers, million, millions, nation, pèlerin, population, province, république, réputation, ressource, richesse, révolution, route, royaume, siècle, siège, superficie, symbole, terre, territoire

Site universitaire :

, administration, association, bibliothèque, bureau, cafétéria, centre, département, direction, discipline, équipement, établissement, quartier, faculté, gymnase, institut, Internet, librairie, certificat, mensa_G, organisme, papeterie, règlement, secrétariat, section, service, structure, système, proximité, laboratoire, localisation

Environnement pratique de la vie quotidienne

achat, alimentation, alcool, aliment, boisson, déjeuner, dîner, eau, fruit, nourriture, pain, plat, pluie, poisson, légume, viande, repas, sandwich, argent, budget, dépense, tarif, coût, frais, franc, prix, attente, banque, commerce, boutique, course, marché, magasin, épicerie, supermarché, carte_bancaire, carte_de_crédit, visa, carte, carte, cigarette, tabac, compte, compte_bancaire, compte_en_banque, condition, conditions_de_vie, consommation, délai, démarche, formulaire, formalité, attestation, fermeture, heure, heure, heure, heures, heures, matin, matin, ……..

Figure 6 : Le modèle, de la grille thématique aux dictionnaires…

Les résultats des calculs d'intensité lexicale montrent que les thèmes les plus représentés correspondent à la sphère d'action individuelle. Pour la majorité des étudiants c'est le noyau de leur expérience. En revanche les thèmes relatifs à l'intériorité sont bien plus faiblement représentés.

	Intensité Moyenne	Ecart-type
Cadre de vie	1,95	0,89
Repères et fonctions	1,20	0,63
Action individuelle	4,09	1,13
Sociabilté et relations	0,91	0,44
Image et empreint du sujet	0,70	0,32
Total	1,77	1,44

	C1	C2	C3	C4	C5
C1 : Action individuelle	1,000				
C2 : Cadre de vie	**-0,314**	1,000			
C3 : Repères et fonctions	**0,441**	-0,189	1,000		
C4 : Sociabilté et relations	-0,156	-0,194	-0,115	1,000	
C5 : Image et empreint du sujet	**-0,392**	-0,095	-0,484	**0,612**	1,000

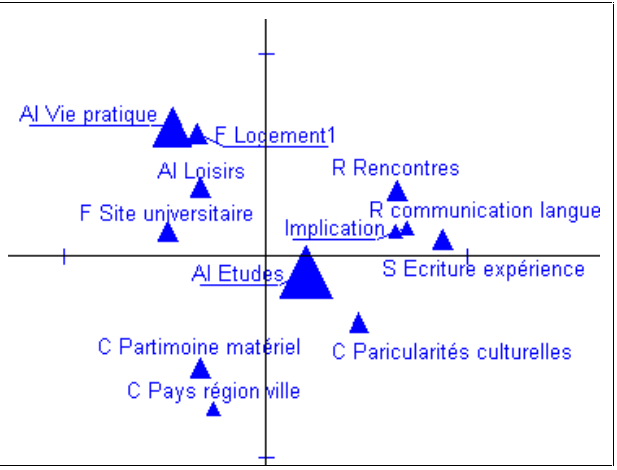

Figure 7 : La confirmation de l'opposition extériorité / intériorité.

D'autre part en examinant, sur l'ensemble des 80 rapports, les corrélations entre les intensités lexicales des différents thèmes on constate que l'opposition extériorité / intériorité est confirmée par l'analyse statistique. L'analyse en composantes principales ci dessus donne un bonne représentation visuelle de cet effet.

3.2.3. Les résultats : postures de l'énonciation et contenu des énoncés

Pour finir de tester notre hypothèse il faut mettre en relation les caractéristiques de l'énonciation relatives à l'implication du sujet avec celles des énoncés. On s'appuie pour cela sur la répartition des rapports en 3 catégories relatives à l'implication de leurs auteurs. Pour chacune de ces catégories on examine l'intensité des dimensions thématiques. Les résultats de cette analyse confirment l'hypothèse.

Tab 1 : Intensité des thèmes selon la catégorie

	Faible (je <0,5%)	de 0,50 à 1,39	Forte (je >1,39%)
Action individuelle	4,31	3,90	4,05
Cadre de vie	2,06	2,15	1,65
Repères et fonctions	1,48	1,14	0,98
Sociabilité et relations	0,77	0,83	1,13
Image et empreint du sujet	0,55	0,65	0,90

Test F : Classes d'implication / Action individuelle p 42,3% ; F 0,88 (NS)
Classes d'implication / Cadre de vie p 8,7% ; F 2,50 (PS)
Classes d'implication / Repères et fonctions p 0,9% ; F 5,03 (TS)
Classes d'implication / Sociabilité et relations p 0,3% ; F 6,19 (TS)
Classes d'implication / Image et empreint du sujet p <0,1% ; F 11,29 (TS)

Tab 2 : Répartition des thèmes selon la catégorie d'implication

	Faible (je <0,5%)	de 0,50 à 1,39	Forte (je >1,39%)	TOTAL
Action individuelle	46,9%	45,3%	46,6%	46,3%
Cadre de vie	23,2%	24,4%	19,5%	22,4%
Repères et fonctions	15,7%	12,5%	10,7%	13,0%
Sociabilité et relations	7,9%	9,9%	12,9%	10,2%
Image et empreint du sujet	6,2%	7,9%	10,3%	8,1%
TOTAL	100%	100%	100%	100%

La dépendance est très significative. chi2 281,24, ddl 8, 1-p >99,99%.

Analyse en composantes principales du tableau 1

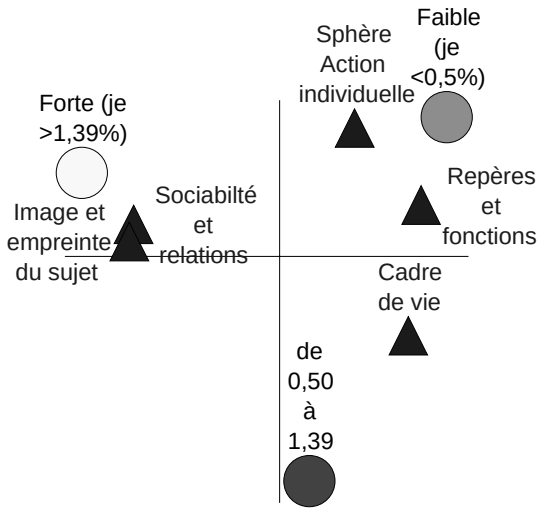

Figure 8 : La relation énonciation énoncé, l'hypothèse est confirmée

Les rapports où l'auteur se trouve fortement impliqué accordent relativement plus d'importance à la sociabilité aux relations et l'image du sujet. Au contraire lorsque l'implication est faible, se trouvent privilégiés les thèmes de la sphère d'action quotidienne, des repères et fonctions du cadre de vie.

On confirme cette analyse en raisonnant non plus sur les intensités mais sur la fréquence absolue (nombre d'appels à un élément d'une thématique).

4. Approche exploratoire ou test d'un modèle

4.1. Les limites de l'analyse sans a priori

Un examen critique de ces méthodes (approche du contenu du texte par l'examen de ses substituts lexicaux - lexiques, cartes…) nous conforte dans l'idée de les considérer comme simples moyens d'approche des corpus. Leur automatisme apparent ne permet pas de faire l'économie d'une implication forte de l'analyste, consciente ou non, selon les outils qu'il utilise. Celle ci se manifeste à différents niveaux :

4.1.1. Le découpage du corpus à analyser

Dès qu'on veut dépasser le simple décompte des termes utilisés il est nécessaire, pour dénombrer la présence simultanée de 2 termes, de découper le corpus en fragments élémentaires ou unités de contexte (Reinert). Faut-il utiliser les découpes naturelles ou marques présentes dans le corpus ou procéder de manière arbitraire en fonction d'un nombre de mots fixés a priori ?

Dans notre exemple, la constitution d'une base de données découpée en rapports peut sembler s'imposer naturellement. Il est satisfaisant tant qu'on se limite à la production du lexique. L'utiliser pour analyser les associations lexicales est plus problématique ; en effet la présence de 2 termes situés à plusieurs pages l'une de l'autre n'a certainement pas la même signification que si ces 2 termes sont présents dans la même phrase ou le même paragraphe. D'autre part selon la technique d'analyse factorielle utilisée les calculs sont effectués sur des

matrices d'occurrence (présence au moins une fois) ou de fréquence (nombre d'apparitions), dans le cas d'un découpage en fragments très longs, raisonner sur les occurrences peut être très pénalisant. Ici on a opté pour un découpage en phrases (ponctuation forte) en limitant toutefois l'analyse des cooccurrences aux seules phrases comportant au moins 4 substantifs. Ce choix a été effectué pour focaliser sur le corps du texte en éliminant les titres de partie.

4.1.2. Du corpus original au choix d'une mise en forme significative

L'analyse du corpus dans sa forme originelle présente bien des inconvénients du fait de la présence de mots vides de contenus (mots-outils), des flexions grammaticales, des polysémies inhérentes à la langue et des formes composées. Là encore l'analyste peut opter pour différentes stratégies vue de réduire l'ambiguïté (lemmatisation, catégorisation selon les formes grammaticales) ou la variété (stemmatisation[7]).

Les analyses précédemment exposées ont été menées sur une restriction du corpus aux seules formes substantives après lemmatisation et identification des locutions usuelles et segments répétés. Cela est pertinent pour une approche du contenu. Ce l'est beaucoup moins pour caractériser l'énonciation comme nous l'avons fait en travaillant sur les intensités des mots outils *je me, ma…*Enfin pour éliminer les particularités propres à chaque auteur et circonstances (pays, type de formation) et pouvoir se concentrer sur ce qui peut être dégagé comme commun à ces diverses expériences, seuls les éléments lexicaux présents dans au moins 40 rapports (1 sur 2) ont été retenus.

Ce choix a été arrêté comme le plus significatif après plusieurs essais portant sur des formes moins restrictives du corpus (corpus complet, lemmatisé, avec et sans mots outils…).

L'analyste a ainsi une première influence par les choix qu'il effectue pour fixer les contours de l'objet auquel sont appliqués les algorithmes de calcul. Ce premier travail est assimilable à un travail de mise au point et de focalisation il exprime une première volonté « d'acuité » (se dégager du flou de la langue) et d'orientation (concentrer l'attention sur ce qui est considéré comme essentiel.

4.1.3. La lecture et l'interprétation des substituts lexicaux

Les biais que nous venons d'évoquer sont indispensables pour éliminer progressivement les bruits de la langue et l'effet des circonstances.

Mais le biais le plus important intervient au moment de la prise de connaissance des résultats de l'analyse statistique. Il s'agit alors de prendre connaissance du texte sur la base de ses substituts lexicaux.

A la différence des méthodes classiques d'analyse de données pour lesquelles les données ont un sens fixé a priori par la définition des variables, l'analyse des données textuelles opère sur des formes graphiques (jusqu'ici abusivement qualifiées de « mots » ou « termes ») dont le sens n'est pas fixé a priori. Par exemple sur les 250 occurrences de la forme région, près d'un 1/3 est associé à Rhône Alpes (le destinataire des rapports) et évoque tout autre chose que la région visitée....

Ceci conduit à réfuter l'opposition souvent faite ente les procédés classiques d'analyse de contenu considérée comme subjective et l'analyse des données textuelles présentée comme beaucoup plus objective. Quand elle est utilisée dans une démarche exploratoire, on ne fait que déplacer l'objet et le moment de la lecture du texte vers son substitut lexical (liste de mots clé, cartes...). C'est ainsi qu'on a pu dégager à la lecture de la liste des principaux substantifs,

[7] procédé automatique consistant à regrouper un ensemble de forme ayant une racine commune

les champs lexicaux évoqués plus haut ou que de la proximité sur la carte des termes *pays, monde, étranger culture expérience..* nous a conduit à identifier le thème de la vie sociale et de l'étranger. Cette synthèse, pour convaincante qu'elle soit, relève beaucoup plus du commentaire que de l'analyse scientifique ! Les procédés de calcul nous révèlent des traits lexicaux peut être cachés, en tout cas difficilement accessibles par une lecture intégrale du corpus, mais seule la capacité d'interprétation de l'analyste peut donner un sens à ces configurations de formes graphiques. A charge pour lui d'en vérifier le bien fondé par un retour au texte.

4.2. Apports et limites de l'approche par un modèle

4.2.1. Exploration sans a priori ou réflexion a priori

Le courant des méthodes que nous venons d'exposer revendique avec raison le statut d'approche exploratoire (Reinert, Chanal). En bénéficiant de la puissance des logiciels d'analyse de données textuelles il permet bien des avancées et peut donner l'illusion de la facilité et de l'objectivité[8]. Nous espérons avoir montré qu'elles n'en nécessitent pas moins une réflexion a priori (découpage et focus) et de réelles capacités d'interprétation plus apparentées peut être à l'art du commentaire qu'à l'analyse de contenu. Cette méthode (Belerson) suppose en effet l'existence d'un modèle a priori, d'une théorie à laquelle on confronte le corpus analysé. La question du sens se trouve alors réglée a priori par l'existence du modèle et sa mise en correspondance avec les significations dégagées par la lecture du texte. La statistique n'intervient que dans un deuxième temps pour analyser la fréquence des éléments de sens ainsi dégagés.

4.2.2. Le modèle informe l'approche du texte…

Le principal apport est évidemment celui de la théorie qui enrichit l'analyse d'une réflexion construite et organisatrice. Formulée indépendamment des données empiriques elle crée les conditions d'une véritable attitude critique : il s'agit de confronter deux univers : celui de « l'expérience du texte » et celui des idées sur le texte. La qualité de cette confrontation critique est à la mesure des méthodes utilisées pour la mener à bien. Par contraste l'approche exploratoire peut apparaître comme aveugle ou tâtonnante. Au contraire le modèle gouverne les procédés de calcul mis en oeuvre. Ainsi la réalisation de mesures lexicales a l'avantage de mettre en œuvre une confrontation avec le texte totalement objective (il ne s'agit que de dénombrer dans les textes des formes graphiques correspondant au dictionnaire de référence).

4.2.3. … mais il faut compter avec la fragilité du « matériau textuel »

La validité de cette approche repose sur la complétude et la qualité des dictionnaires ainsi que sur le postulat de la possible inférence du vocable à la signification. Les dictionnaires peuvent permettre de « lire » le texte comme les échelles des protocoles fermées permettent de lire les opinions du répondant. Les risques de contresens n'en sont pas nécessairement plus élevés et dans tous les cas on peut admettre que leur « bruit » n'affecte pas les tendances mise en évidence par la statistique. Enfin les mesures d'intensité lorsqu'elles s'appliquent à des textes libres ont l'avantage d'être indépendantes de tout biais lié au protocole.

Par souci de précaution et de rigueur on peut également confronter les mesures d'intensité aux résultats d'une analyse de contenu classiquement conduite.

[8] c'est bien sûr un illusion. Il suffit d'avoir été une fois confronté à la complexité des données pour s'en convaincre

4.3. Pour une approche duale

Opposer l'approche sans a priori à l'approche par modèle permet de clarifier les 2 temps de toute recherche. A partir d'une première approximation lexicale, la recherche exploratoire, conduit à orienter la réflexion vers l'élaboration d'un modèle et des dictionnaires thématiques qui lui correspondent. Les mesures d'intensité permettent dans un deuxième temps de tester la validité du modèle comme on le ferait avec toute autre mesure.

C'est ce double mouvement qui a été suivi dans le cas de l'analyse présentée. Il est général à toute analyse de données empiriques. La particularité des données textuelles réside dans le fait que les données traitées n'ont pas de signification a priori et que la quête du sens doit être menée parallèlement à celle des ordres de grandeur et des structures que la statistique permet de mettre en évidence.

D'autre part, que l'on accepte les procédés automatiques de l'approximation et de la mesure lexicale ou que l'on préfère les classiques méthodes de l'analyse de contenu, les outils utilisés doivent permettre la mise en œuvre de l'ensemble de ces démarches complémentaires en permettant aussi bien de partir du texte pour le coder de manière classique que d'y revenir pour vérifier les interprétations déduites des procédés automatiques. Ce besoin conduit à l'intégration des logiciels qui seule peut donner au chercheur la liberté de « manœuvre » nécessaire à l'approche des textes.

Références

Benveniste E. (1996) *Problèmes de linguistique générale, 1*, Editions Gallimard,.

Benzécri J.P. & coll. (1981) *Pratique de l'analyse de données : linguistique et lexicologie*, Paris Dunod

Bolden R. Moscarola J. (2000) Bridging the Quantitative-Qualitative Divide. The Lexical Approach to Textual Data Analysis, *Social Science Computer Review, Vol 19 N°4, Winter 2000 450-460*

Chanal V. Moscarola J (1998) Langage de la théorie et langage de l'action, analyse lexicale d'une recherche action sur l'innovation *Acte des 4ème journées JADT,* 1998.

Culioli, C. Fuchs, M. Pécheux, *Considérations théoriques à propos du traitement formel du langage, tentatives d'application au problème des déterminants*, Paris, Dunod,

Gavart-Perret M.L. Moscarola J. (1998) Enoncé ou énonciation ? deux objets différents de l'analyse lexicale en marketing *Recherche et application en marketing 1998 Vol 13 N°2*

Glaser P. & Strauss A. *(1967) The Discovery of Grounded Theory, Strategies for Qualitative Research, New York* : Aldine Press

L'Ecuyer R. (1987) « L'analyse de contenu : notion et étapes. » *in Les méthodes de la recherche qualitative sous la dir. De J-P. Deslauriers, Presses universitaires du Québec*

Lebart L. , Salem A (1994) *Statistique textuelle*, DUNOD

Marchand P. (1998) *L'analyse de discours assistée par ordinateur*, Paris, Armand Colin

Muller C (1993) *Principes et méthodes de statistique lexicale*, Champion, Genève

Papatsiba V. (2001) *Le séjour d'études à l'étranger : expériences et savoirs. Analyse des rapports d'étudiants français du programme Erasmus*. Sous la dir. Beillerot J., Thèse en cours.

Pecheux M. (1969) *L'analyse automatique du discours*, Paris, Dunod, 1969.

Reinert M**.** (1998) Quel objet pour une analyse statistique du discours ? Quelques réflexions à propos de la réponse Alceste. *Acte des 4ème journées JADT,* 1998.