40 000 pages pour un livre,
Le cas du Débat National sur l'École

Yves Baulac[1], Stéphane Ganassali[2], Jean Moscarola[2]

[1] Le Sphinx-Développement (Chavanod 74650)

[2] Irege Université de Savoie (Annecy le Vieux 74016)

Abstract

This paper presents the methods used to cope with the national consultation on the futur of french schools. The rational , procedures and tools based on textual data analysis that allowede to cope with such a huge amount of information are presented .
Learnings from this expercience are commented with regard to efficiency and legitimity of the methods used.

Résumé

Ce papier présente les méthodes mises en œuvre dans le cadre du débat national pour l'avenir de l'école.
On montre comment les méthodes de l'analyse des données textuelles ont permis de faire face à la masse considérable d'information produite à l'occasion de ce débat.
Les enseignements de cette expérience sont dégagés du point de vue de l'efficacité et de la légitimité des méthodes mise en œuvre.

Mots clé : statistique lexicale, échantillonnage, verbatim, classification, école, système éducatif

1. De l'usage des TIC à la production massive d'information…

1.1. Un cas problématique

Le débat national sur l'avenir de l'école organisé par la commission Thélot, s'est tenu en France entre décembre 2003 et janvier 2004. C'est une innovation dans ce genre de manifestation classique de la vie publique française (Débat public sur l'avenir de l'Europe en 2001, Débat public sur l'environnement 2002, Débat sur la recherche en 2004…).

C'est un classique par ses finalités et son organisation : animé par une commission mise en place par le Président de la République et présidé par Claude Thélot, il s'inscrit dans la préparation d'une loi sur l'école.

« *L'objectif du débat national sur l'avenir de l'École est d'inciter la Nation à s'exprimer sur son École et, par là, aboutir à un diagnostic partagé et une refondation de notre système éducatif*[1] »

Il innove cependant par ses méthodes et les technologies qu'il utilise : la communication sur le débat et son organisation se sont en effet très largement appuyés sur un site Internet[2] dont voici les fonctions :

[1] Site du Débat national : http://www.debatnational.education.fr/index.php?rid=39.
[2] http://www.debatnational.education.fr/.

1/ Mettre à la disposition des participants des éléments documentaires et de réflexions (liste de 22 thèmes proposés au débat) ;

2/ Annoncer les dates et lieu de réunions tenues partout en France ;

3/ Centraliser les synthèses de chacun des débats locaux[3] ;

4/ Afficher la synthèse de chaque débat et en communiquer ainsi le contenu ;

5/ Recueillir des contributions personnelles envoyées par courriel ;

6/ Proposer un forum électronique comme espace de discussion.

Le débat a donné lieu à la publication d'un livre de 649 pages : « Le Miroir du Débat » paru le 6 avril 2005, moins de 4 mois après la tenue de la première réunion.

En exergue le Président de la République annonce : « *un débat, je le souhaite, exemplaire, par sa méthode et son ampleur, qui marquera une étape dans la modernisation de notre vie publique*»[4]

C'est sous cet angle que nous nous intéresserons à ce cas où la technologie de l'information et de la communication joue un rôle essentiel. En effet sans Internet, un débat d'une telle ampleur n'aurait pu donner lieu en si peu de temps à la centralisation et à l'analyse d'autant de contributions. Cependant, dès l'origine, la masse d'information que les TIC permettent de recueillir est présentée de manière polémique par le Canard Enchaîné comme un problème.

« sachant que le Grand Débat sur l'Avenir de l'École organisé par Luc Ferry … aura vu se tenir 15 000 réunions, que chacune de ces réunions aura donné lieu à un compte rendu de 8 pages), qu'il faut y ajouter la petite quarantaine de milliers de messages envoyés sur Internet, que le tout devrait faire dans les 150 000 pages, comment diable les 54 membres de la commission chargés d'en lire l'intégralité, puis d'en rédiger la synthèse finale en deux mois vont-ils procéder (…) ? Il fallait bien un polytechnicien pour venir à bout de ce casse-tête, et c'est Claude Thélot, le président de la commission, qui a trouvé le truc : on va passer tout ça dans l'ordinateur ! (…) Surtout que Thélot a précisé d'emblée l'importance qui sera accordée à la grande synthèse des synthèses : « Si le débat reflète correctement ce que les Français pensent, le gouvernement en tiendra compte » (« Le Monde », 16/11/03). Et comment savoir que ce débat « reflétera correctement » ce qu'ils pensent ? En s'arrangeant pour que la synthèse reflète exactement ce que le gouvernement veut que les Français pensent, évidemment ! »[5].

Nous allons ici reprendre des éléments du rapport présenté à la commission pour montrer comment… « la restitution des milliers de synthèses issues des débats locaux et des contributions électroniques et postales a été organisée avec rigueur. Pour les synthèses, deux démarches complémentaires ont été retenues. Une trentaine de « lecteurs-rédacteurs », choisis pour leur compétence ont pris connaissance de ces textes, soit de manière approfondie, soit de manière cursive, afin d'en dégager le sens. Cette lecture en profondeur a été complétée par une analyse informatisée de l'ensemble du corpus, conduite de manière indépendante et parallèle par une société d'analyse diligentée à cette fin par la Commission. Un conseil scientifique a validé cette seconde phase. »

[3] Le fichier de chaque synthèse rédigée dans Word est « monté » vers le site qui les centralise.
[4] *Jacques Chirac, Introduction générale du rapport « Les Français et leur Ecole : le miroir du débat », Dunod, 2004.*
[5] *Le Canard enchaîné, Mercredi 7 janvier 2004.*

1.2. … par le volume des corpus des informations numérisées

Les informations numérisées rendues très rapidement disponibles grâce au système de collecte porte sur :

- les sujets traités au cours des réunions (choisis parmi une liste de 22 sujets présentés sur le site du débat) ;

- les priorités pour l'école exprimées sous formes de 3 courtes phrases ;

- les synthèses des discussions relatives aux sujets traités :

Ces informations constituent un corpus de plus de 40 000[6] pages analysés par des méthodes statistiques et lexicales, qui concerne 13 000 débats (90% de l'information disponible a pu être recueillie).

13000 débats ont eu lieu rassemblant 1 million de personnes. La représentativité des débats par académie est bonne, l'organisation a conduit à une sous représentation des écoles. Les collèges se trouvent sur représentés.

- la base de données des évènements :

La base de données comporte les données de contexte, les sujets abordés au cours de la réunion et les trois priorités pour l'école. Il s'agit d'une information structurée qui se prête bien à l'analyse statistique. C'est en exploitant ces informations qu'on peut rendre compte de l'ampleur et de l'orientation générale des réunions tenues dans le cadre du débat.

1.2.1. Le corpus des priorités

Il s'agit d'énoncés brefs (une phrase selon les indications données sur le site) donnant les 3 priorités pour l'école retenues au terme du débat. Ces priorités sont déposées sur le site indépendamment de la transmission de la synthèse et viennent utilement compléter l'information relative à la nature des sujets abordés au cours de la réunion. Le corpus des priorités que nous avons analysé porte sur 33 000 courtes propositions (moins de 50 mots) représentant au total 600 000 mots.

1.2.2. Le corpus des synthèses

Il s'agit de comptes rendus rédigés et structurés selon les sujets abordés, à partir de ceux qui sont proposés sur le site. Le corpus d'une synthèse est ainsi décomposable en autant de parties que de sujets abordés. La référence au sujet est explicitée dans le corps du texte par l'auteur de la synthèse et doit être reconnue par le système informatique.

Ces synthèses sont la seule trace que l'on ait du contenu des discussions tenues au cours des réunions. Elles ont été rédigées en connaissance du fait qu'elles seraient toutes consultables sur le site ce qui permet de penser que la publicité ainsi donnée a contribué à des rapports fidèles.

Parmi les 13 000 fichiers de synthèse déposés sur le site, une partie a du être éliminée : les fichiers inutilisables pour des raisons de format, les "doublons" (la synthèse unique correspondant à un ensemble de réunions a été déposée plusieurs fois sur le site du débat), les fichiers quasi-vides ("Le débat n'a pas eu lieu", ...).

Le corpus des 11 000 synthèses utilisables représente un total de 40 000 pages pour plus de 15 millions de mots.

[6] C'est moins que ce que le Canard Enchaîné annonce, c'est toutefois considérable.

2. Coupler analyse quantitative et qualitative

2.1. Le volume et la complexité des données recueillies justifient le recours à une démarche quantitative

Cette approche est naturelle pour produire la statistique des évènements (nombre de réunions, fréquence des sujets abordés…) et la caractériser selon les différents contextes. Pour le reste, priorités et synthèses, l'ampleur du débat est à l'origine d'une masse d'information (plus de 40 000 pages) dont la nature textuelle pose de véritables problèmes d'analyse :

La lecture de l'ensemble est d'une lourdeur dissuasive et l'examen de parties du corpus ne peut permettre qu'une connaissance « impressionniste ».

Le caractère répétitif de ces débats et la structuration somme toute assez forte qui leur est donnée, justifient que l'on porte aussi attention à la répétition, ce qui justifie une démarche statistique et automatique y compris sur la partie textuelle du corpus.

C'est la seule manière de prendre en considération le débat dans son exhaustivité et de garantir l'objectivité de la démarche, notamment pour tous les comptes rendus quantitatifs produits dans de telles circonstances (le sujet x a rencontré le plus d'écho, les écoles sont plus sensibles à tel sujet que les collèges…). Pour contrôler l'approche par le simple comptage et se prémunir contre les risques d'erreur d'interprétation, on a couplé l'analyse quantitative avec la prise de connaissance qualitative qu'elle alimente (voir annexe n°1 pour une illustration).

Le travail quantitatif prend ainsi tout son intérêt en relation avec la lecture attentive des textes. Il permet en effet de :

- préparer le travail qualitatif de lecture en aidant à la sélection de « verbatims », échantillons et extraits des corpus significatifs à proposer aux experts ;

- vérifier les idées a priori ou les impressions dégagées par la lecture et les appuyer sur des données ou arguments chiffrés relatifs à la statistique lexicale de ces textes ;

- stimuler la réflexion et l'examen plus approfondi de ces textes par la mise en évidence de structures lexicales remarquables non repérables dans un travail de lecture classique ;

- aider et accélérer - à la manière de la recherche hypertextuelle - la découverte d'éléments particuliers enfouis dans la masse des informations recueillies.

Ceci conduit à utiliser l'approche quantitative pour produire dans un premier temps des extraits ou « verbatims » des corpus (priorités et synthèses) et pour livrer ensuite une analyse exhaustive du contenu, en mettant en œuvre des méthodes d'analyse lexicale.

2.2. Sélection de verbatim pour préparer l'approche qualitative

En vue de rendre compte du débat sur la base d'une lecture attentive de son contenu, on a procédé à la sélection d'extraits de synthèses et de priorités. Ces extraits étaient destinés aux lecteurs de la commission, à charge pour eux de sélectionner à leur tour les éléments les plus significatifs pour figurer au verbatim destiné à être publié.

La procédure consistant à extraire le verbatim d'une manière aléatoire pour produire un échantillon représentatif des différentes catégories de réunions n'a pas été retenue. Nous avons au contraire choisi de privilégier la restitution de la variété des contenus. Ainsi, pour illustrer la manière dont les sujets ont été traités, chacun a été illustré par le même nombre d'extraits. Le même type d'approche a été utilisé pour le verbatim des priorités. Les

procédures utilisées sont décrites ci-dessous. Elles ont en commun d'utiliser l'analyse lexicale pour produire des extraits visant à restituer les contenus dans toute leur diversité, en mettant l'accent sur les spécificités plutôt que sur les répétitions.

Deux jeux d'extraits ont été définis :

2.2.1. Les extraits les plus significatifs de chaque sujet

On s'appuie pour cela sur la recherche des formes graphiques (corpus lemmatisé) les plus spécifiques de chaque sujet. Les extraits produits sont ceux pour lesquels l'intensité lexicale des formes spécifiques du sujet est maximale.

2.2.2. Les extraits montrant la diversité propre à chaque sujet

Pour compenser le biais introduit par la première approche, on cherche également à rendre compte de la variété à l'intérieur de chaque sujet. On procède à partir d'une classification automatique (par la méthode des nuées dynamiques) appliquée à l'intensité lexicale des 100 premières formes significatives.

2.3. Analyse de données textuelles pour une approche exhaustive de l'ensemble du corpus

Pour analyser le corpus des priorités et celui des synthèses, trois types d'analyses ont été utilisés. Toutes ont été menées sur la base du corpus « lemmatisé » et « catégorisé » (voir annexe n°5).

2.3.1. Analyses sans a priori des structures lexicales du corpus

On cherche à mettre en évidence l'existence de thématiques intrinsèques par l'interprétation des intensités et cooccurrences lexicales. À partir du lexique des formes substantives et verbales les plus utilisées, on construit des cartes factorielles (AFCM ou ACP) qui mettent en évidence les associations lexicales les plus significatives. Les configurations ainsi mises à jour sont prises comme base d'une interprétation de la structuration des contenus.

2.3.2. Analyses par contexte des spécificités lexicales

On cherche à caractériser les éléments de contexte (sujets proposés au débat, type de réunion et zone géographique) par les spécificités lexicales qu'ils induisent. On s'appuie pour cela sur les données de contexte disponibles dans la base des débats : sujets abordés, type de réunion et lieux géographiques. On obtient comme résultat les listes de termes qui rendent le mieux compte des particularités propres à chaque contexte. Ce type d'analyse croisée permet notamment de mettre en évidence les effets propres à l'organisation du système éducatif (type d'établissement notamment) et des contextes socio-économiques.

2.3.3. Analyses de contenu automatiques basées sur une modélisation à priori

Les sujets proposés au débat offrent un cadre d'analyse privilégié. Le contenu des synthèses s'y réfère explicitement, mais pas celui des priorités. Afin de néanmoins pouvoir se rapporter aux 22 sujets, on utilise des dictionnaires contenant les mots-clés (proposés sur le site) et les mots spécifiques des sujets, dégagés à partir de l'examen des lexiques (voir annexe n°2).

De la même manière, on cherche à mesurer la présence dans les synthèses des catégories d'acteurs du système éducatif et des catégories d'actions. On construit pour cela les champs lexicaux caractérisant chaque acteur et par ailleurs, on sélectionne les six verbes enseigner, former, éduquer, instruire, transmettre, apprendre. La présence de ces catégories est évaluée

par le calcul de leur intensité lexicale dans le corpus (rapport de mots appartenant à la catégorie considérée sur le nombre total de mots).

2.4. Les outils et pertinence scientifique

2.4.1. Sphinx Lexica version 5.0

Le Sphinx Lexica est le logiciel standard que nous avons utilisé pour l'analyse quantitative et qualitative. Il a notamment permis la lemmatisation du corpus, la gestion des dictionnaires et des formes composées, la recherche des mots spécifiques, le calcul des intensités lexicales, la mise en œuvre de classification automatique, la production de cartes factorielles : analyse factorielle des correspondances, analyse factorielle multiple, analyse en composantes principales.

2.4.2. Contrôle du conseil scientifique

Ce travail a bénéficié des commentaires et apports du comité scientifique réuni par la commission, auquel les outils et méthodes évoquées plus haut ont été présentés.
Par ailleurs, nous avons pu confronter nos résultats aux analyses menées sur le même corpus par Ludovic Lebart et Romain Vinot. Les résultats qu'ils obtiennent convergent avec les nôtres, et nous avons repris certaines de leurs analyses. Nous les en remercions.

3. Les résultats

3.1. Statistique des débats

Certains débats donnaient lieu à 1 réunion (réunion d'arrondissement par exemple), d'autres (écoles, collèges, lycées) donnaient lieu à 2 réunions. Dans ce document, l'unité statistique considérée est le débat.

900 débats ont eu lieu au mois de Novembre, 11000 au mois de Décembre, 1100 au mois de Janvier.

3.1.1. Répartition par type de lieu

Les débats ont eu lieu dans des établissements scolaires pour la plupart Les tableaux ci-dessous montrent que l'attractivité du débat n'a pas été la même selon le lieu.

Réunion d'arrondissement	**442**	**3.4%**
Réunion de CFA	**6**	**<0.1%**
Réunion de collège privé	**502**	**3.9%**
Réunion de collège public	**5073**	**39.4%**
Réunion de lycée général et technologique privé	**250**	**1.9%**
Réunion de lycée général et technologique public	**1523**	**11.8%**
Réunion de lycée professionnel privé	**162**	**1.3%**
Réunion de lycée professionnel public	**889**	**6.9%**
Réunion de service académique	**180**	**1.4%**
Réunion d'école privée	**237**	**1.8%**
Réunion d'école publique	**3150**	**24.4%**
Réunion d'établissement d'enseignement agricole	**142**	**1.1%**
Autre	**333**	**2.6%**
Total	**12889**	**100.0%**

Ecole	**3387**	**26.3%**
Collège	**5575**	**43.3%**
LGT	**1773**	**13.8%**
LP	**1051**	**8.2%**
Autre	**1103**	**8.6%**
Total	**12889**	**100.0%**

Et. public	**10635**	**82.5%**
Et privé	**1151**	**8.9%**
Autre	**1103**	**8.6%**
Total	**12889**	**100.0%**

Tableau 1 : Les lieux du débat

Les collèges sont sur-représentés, mais la représentativité géographique par académie est bonne.

3.2.2. Les sujets abordés

Les difficultés du système sont au centre des préoccupations des participants. La fréquence relative avec laquelle les sujets proposés aux débats ont été choisis met tout d'abord en évidence les difficultés rencontrées : *motiver les élèves*, la *violence*, l'*adaptation à la diversité*, les *élèves en difficulté*. Il est frappant par ailleurs de constater le peu de succès de sujets comme *préparer l'enseignement supérieur*, la *scolarisation des handicapés* et l'*évaluation*. La préoccupation de faire réussir les élèves est au centre des débats, un sujet au moins de ce groupe a été choisi dans plus de 8 cas sur 10.

3.3.3. Les configurations dans le choix des sujets

Si l'on examine maintenant comment ces sujets ont été associés dans les réunions, les 3 sujets choisis en moyenne font apparaître des configurations très significatives mises en évidence par la carte ci dessous. Les zones de proximité mettent en évidence les sujets fréquemment associés les uns aux autres.

Autour de *motiver,* les sujets les plus souvent choisis sont *élève en difficulté*, *violence*, *adaptation à la diversité*. Un groupe à gauche autour de *parents et partenaires*, *moyens*, *missions*... Un autre associant *collectivités locales*, *autonomie* et *moyens*. En haut à droite les sujets de l'*insertion dans le supérieur*, de la fo*rmation professionnelle* de la *formation continue*…

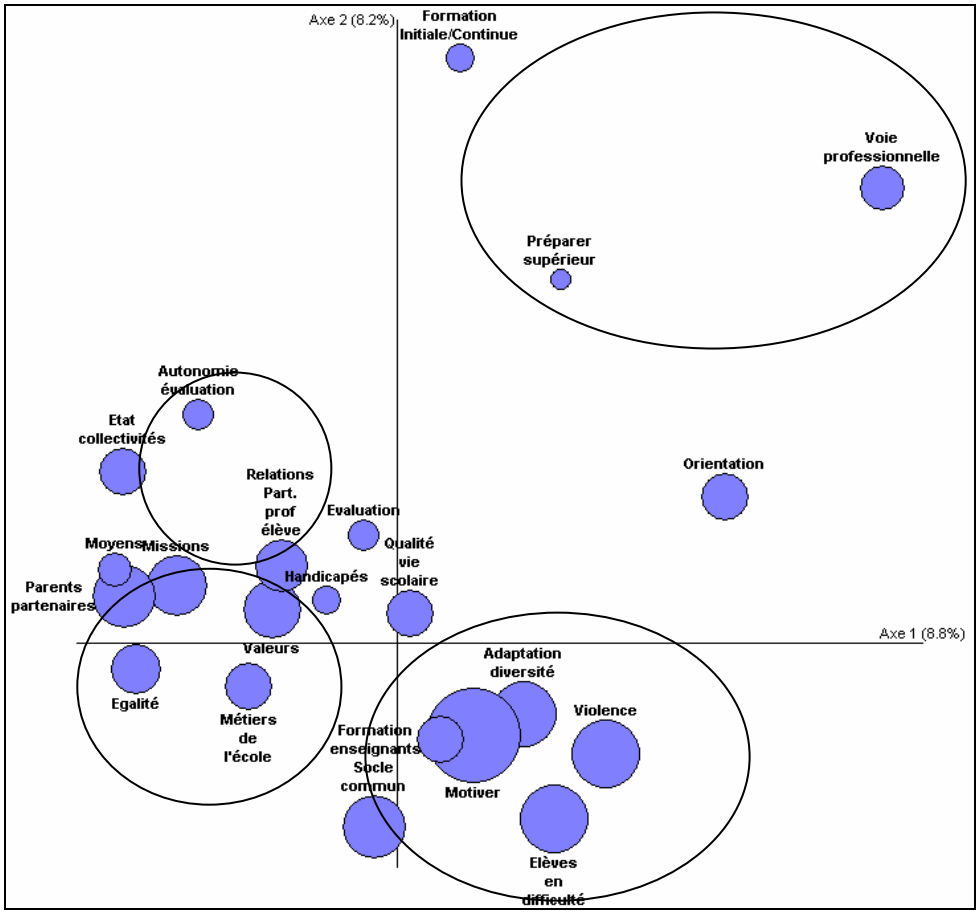

Figure 1 - Configurations des sujets traités au cours des débats

3.4. Les priorités pour l'école : des constats aux projets

3.4.1. Contenu lexical des priorités

L'analyse des 3 priorités énoncée pour chaque réunion sous la forme de courtes phrases complète l'information donnée par le choix des sujets traités.

Une première analyse consiste à analyser les structures lexicales repérables sur les 100 premiers substantifs du texte (qui représentent 47% du corpus). Elle permet de dresser une cartographie des priorités déclarées. Cette carte met en évidence comment le système des priorités s'articule selon 2 systèmes d'opposition : à gauche, les acteurs, à droite les contenus. En bas, l'organisation et les structures ; en haut, les finalités.

L'analyse intrinsèque des structures lexicales pratiquée sur les 100 premiers substantifs du texte (qui représentent 47% du corpus) montre la cartographie des priorités déclarées. Cette carte met en évidence comment le système des priorités s'articule par oppositions.

- On retrouve tout d'abord un premier sujet qui renvoie aux **valeurs de l'école** comme pilier de la République : égalité des chances, confiance, respect, citoyen, valeur, société.
- Les **connaissances générales** mais aussi les compétences apparaissent autour de la notion de "socle", associées au savoir, à la culture et au sens de l'effort.
- Une grande partie des mots se regroupent autour de la notion de **partenariat**, de dialogue entre l'école, l'enfant, la famille et aussi l'État.
- La gestion du personnel et l'animation des **équipes pédagogiques** apparaissent comme des priorités claires avec les mots : encadrement, équipe, concertation.
- Au sein d'une autre thématique, on insiste sur les **moyens** pour faire la classe : les effectifs, les groupes, les heures.
- Par ailleurs, l'organisation même du **système scolaire** est évoquée avec les termes : discipline, matière, filière, passerelle.
- Enfin, un groupe de mots est dédié à **l'insertion professionnelle**, l'entreprise, l'orientation, les langues, les métiers.

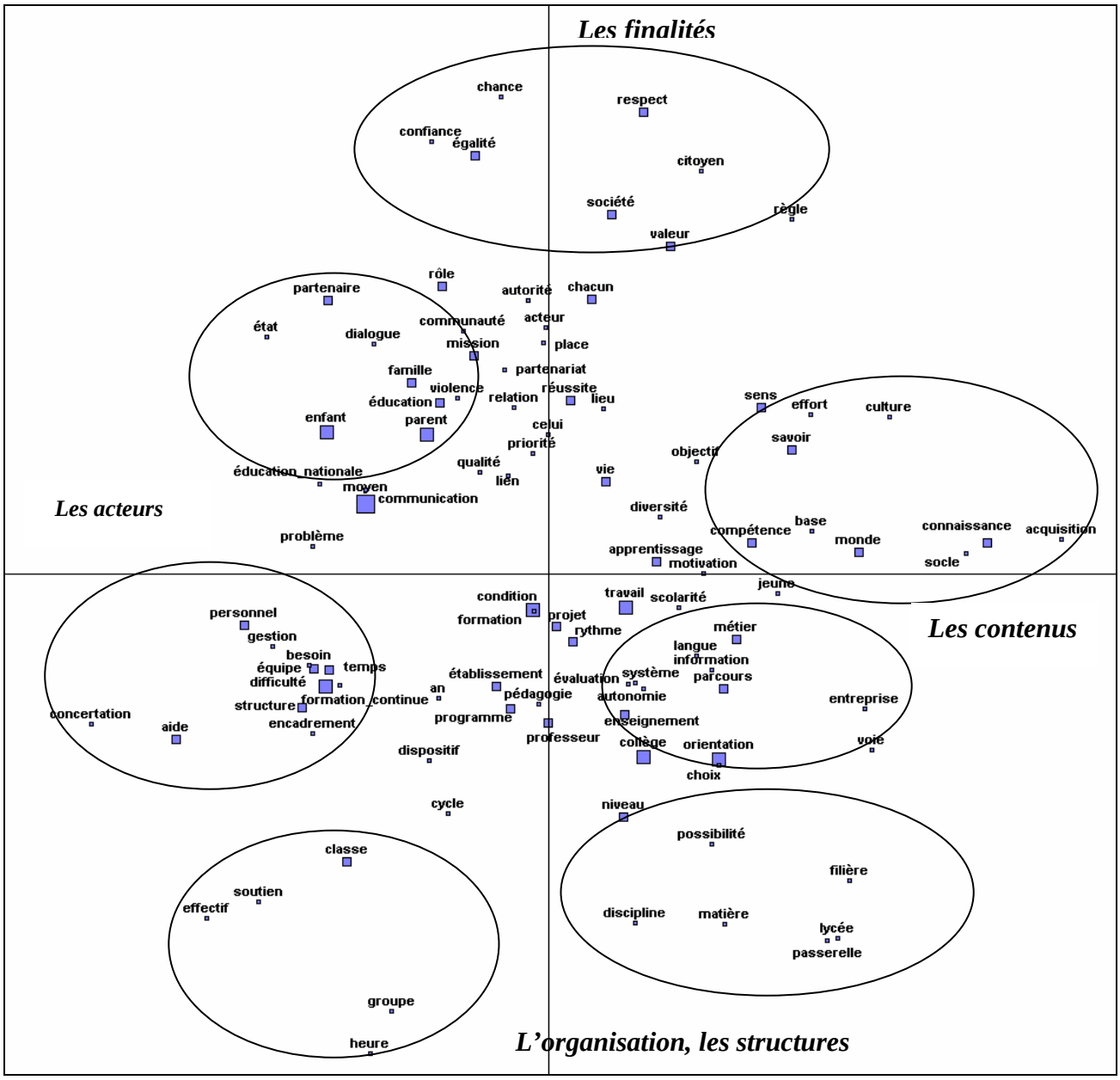

Figure 2 - Les priorités : Associations lexicales des100 premiers substantifs.

3.4.2. Contenu thématique

Nous avons réalisé une analyse en composantes principales sur les 22 variables mesurant l'intensité lexicale de chacun des 22 sujets dans le texte des priorités. Cette analyse fait apparaître plusieurs grandes dimensions parmi les priorités composées par des sujets qui se combinent et qui se corrèlent. Sur ces combinaisons de variables (ou axes factoriels), nous avons positionné des groupes de réponses que nous avons délimités et interprétés de par leur situation respective sur les quatre dimensions identifiées.

Nous avons ensuite fait apparaître les spécificités de chaque groupe sur d'autres variables de l'étude, comme le type d'établissement, l'académie, les mots et les verbes-clés employés. Quatre profils distincts de priorités apparaissent qui se différencient en fonction des solutions qui sont envisagées, des issues qui sont imaginées pour résoudre les problèmes de l'école.

L'examen des contributions aux axes permet d'identifier le groupe de gauche comme correspondant au troisième facteur.

L'équipe pédagogique

Le premier groupe croit aux vertus d'une équipe pédagogique bien formée, évaluée, qui peut agir en autonomie et dont les différents métiers ont été repensés. Comme nous le montre le tableau ci-dessous, ces établissements sont très variés et situés plutôt au Sud de la France.

Le pragmatisme

Le groupe « pragmatique » s'en réfère aux finalités de l'école, à la préparation à l'entrée dans le supérieur ou à des métiers, à une meilleure orientation et à une évaluation mieux organisée des élèves. Il s'agit surtout des lycées professionnels, plutôt dans le Nord de la France.

La tradition

Le troisième groupe en appelle aux valeurs et aux missions traditionnelles de l'école républicaine : il s'agit notamment de garantir et d'assurer la culture pour tous et l'égalité des chances. Dans ce groupe, les lycées généraux sont sur-représentés, et parmi eux de nombreux établissements parisiens.

La co-éducation

Le groupe co-éducation croit au partenariat avec les parents et les personnalités extérieures qu'il faut impliquer et motiver, afin de résoudre les difficultés scolaires mais aussi les problèmes de violence auxquels ces établissements sont visiblement confrontés. On trouve de nombreuses écoles dans ce groupe, des Dom-tom, du Nord et aussi du Sud.

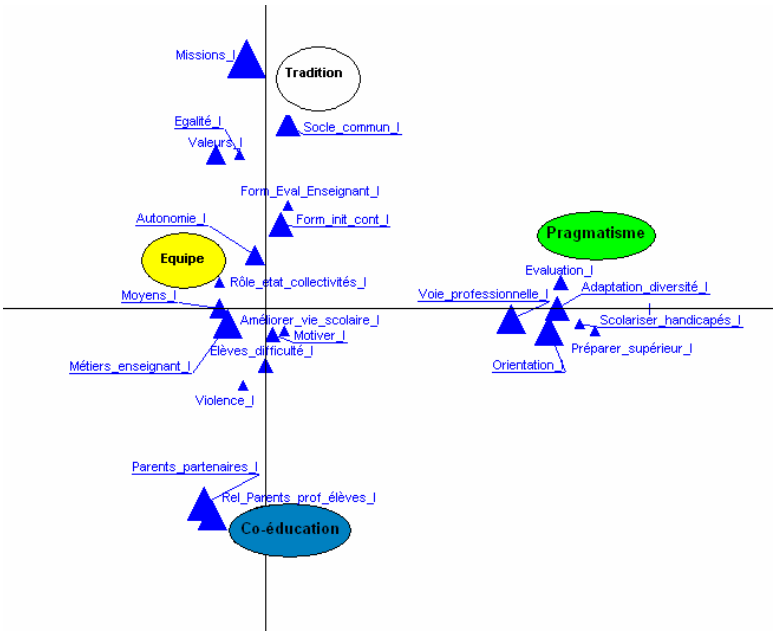

Figure 4 – Typologie thématique des priorités

3.5. Le contenu des synthèses : la dynamique des discours sur l'école

Après les sujets choisis et les priorités retenues, intéressons-nous au contenu des synthèses. Les 9 000 synthèses dont on a pu reconnaître la structuration constituent un corpus de 26 000 paragraphes de 350 mots en moyenne. Chaque paragraphe correspond à un compte rendu d'un des 3 sujets traités en moyenne par réunion.
Ce corpus représente près de 10 millions de mots. Après lemmatisation et suppression des mots outils (voir Annexe 5), on obtient 4 806 064 formes dont 26 438 différentes. Chaque mot se trouve ainsi en moyenne répété 180 fois.

3.5.1. Une variété de débats centrés sur l'école ou sur la société, sur les finalités ou sur les moyens

Pour se faire globalement une idée du contenu des synthèses on va visualiser comment les 100 premiers substantifs et les 60 premiers verbes suivants se trouvent associés dans le corps du texte.

élève	116548	école	79139	parent	50658	enfant	45719
enseignant	44344	collège	30342	classe	29530	difficulté	28471
travail	27430	formation	21988	problème	21915	professeur	21699
moyen	19776	temps	18347	orientation	17352	établissement	17069
enseignement	16597	apprentissage	14456	question	14446	rôle	14049
métier	12874	éducation	12792	société	12546	famille	12498
programme	12177	niveau	12149	projet	10474	débat	10402
vie	10148	savoir	9971	valeur	9934	lycée	9884
connaissance	9698	aide	9673	groupe	9504	évaluation	9473
matière	9294	besoin	9217	entreprise	9024	compétence	8985
système	8843	personnel	8538	monde	8288	réussite	8228
heure	8208	motivation	8018	jeune	7999	filière	7906
an	7397	proposition	7386	sens	7371	année	7290
mission	7281	effort	7228	violence	7202	choix	7120
équipe	7002	structure	6982	cours	6948	rythme	6890
manque	6796	lieu	6684	culture	6669	respect	6614
nombre	6355	relation	6141	activité	6133	égalité	6067
étude	6058	personne	6052	langue	5910	place	5888
règle	5873	stage	5858	base	5791	constat	5778
objectif	5776	notion	5621	possibilité	5598	voie	5559
devoir	5538	situation	5530	discipline	5506	primaire	5453
information	5401	solution	5223	adulte	5203	état	5192
diversité	5104	scolarité	5039	redoublement	4987	cycle	4953
effectif	4945	sanction	4944	échec	4834	méthode	4684
éducation_nationale	4655	pédagogie	4484	cas	4466	autonomie	4457

Tableau 2 - Les 100 premiers substantifs

apprendre	12055	travailler	10685	adapter	8759	aider	7740
favoriser	7019	demander	6804	**former**	6386	créer	6205
mettre_en_place	5835	valoriser	5181	motiver	5037	connaître	4776
enseigner	4757	comprendre	4483	améliorer	4295	assurer	4248
définir	4199	acquérir	3984	renforcer	3930	respecter	3774
intégrer	3771	reconnaître	3688	vivre	3650	réussir	3605
répondre	3568	préparer	3467	organiser	3253	gérer	3175
réduire	3102	impliquer	2979	avoir_besoin	2930	orienter	2718
maintenir	2710	**éduquer**	2695	revaloriser	2668	évaluer	2660
revoir	2642	accepter	2557	imposer	2544	nécessiter	2541
intervenir	2524	**transmettre**	2492	construire	2358	augmenter	2314
maîtriser	2239	accompagner	2114	privilégier	2027	accueillir	2008
redonner	1963	faciliter	1936	redéfinir	1913	instaurer	1799
lutter	1762	partager	1629	conserver	1611	supprimer	1543
réaffirmer	1419	alléger	1410	**instruire**	1309		

Tableau 3 - Les 60 premiers verbes significatifs

La carte ci-dessous est le résultat d'une analyse factorielle des correspondances multiples effectuée sur le corpus découpé en phrases de 10 à 25 mots. On analyse ainsi un tableau de 160 colonnes par 443 031 lignes. L'examen de la projection sur les 2 premiers axes fait ressortir les oppositions du même type que celles mises en évidence sur le contenu des priorités : de gauche à droite s'opposent les champs lexicaux de la société d'une part et de l'école d'autre part. En bas, on trouve le discours des moyens et en haut, celui des finalités. Les configurations plus fines identifiables sur cette carte permettent de préciser cette interprétation. On y retrouve naturellement les contenus des sujets proposés au débat.

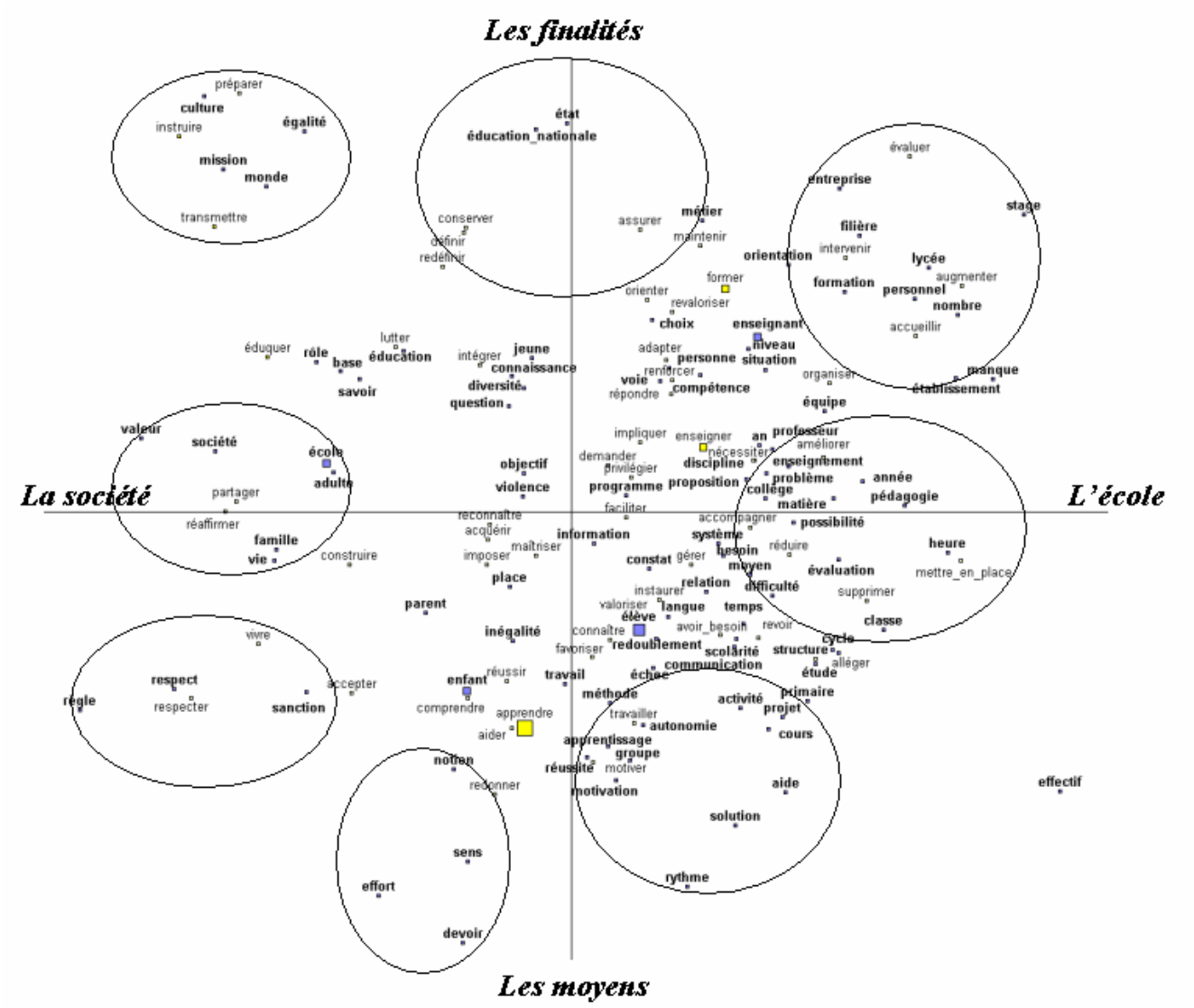

Figure 5 – Corpus des synthèses : Associations lexicales sur les 160 premiers verbes et substantifs

3.5.2. Le type d'établissement a une influence sur les contenus

La méthode ici consiste à rechercher les mots spécifiques, c'est-à-dire ceux qui se trouvent sur-utilisés dans tel ou tel contexte. On utilise pour cela un indice de spécificité.

L'indice de spécificité est égal au rapport entre la fréquence (en %) d'emploi du mot dans le contexte considéré et sa fréquence (en %) dans l'ensemble du corpus.

Le contenu des synthèses des écoles se démarque nettement de celui des autres types d'établissement. Les réunions d'arrondissement et de services académiques se distinguent également. On retrouve là encore le fait que l'organisation du système éducatif détermine les approches du débat.

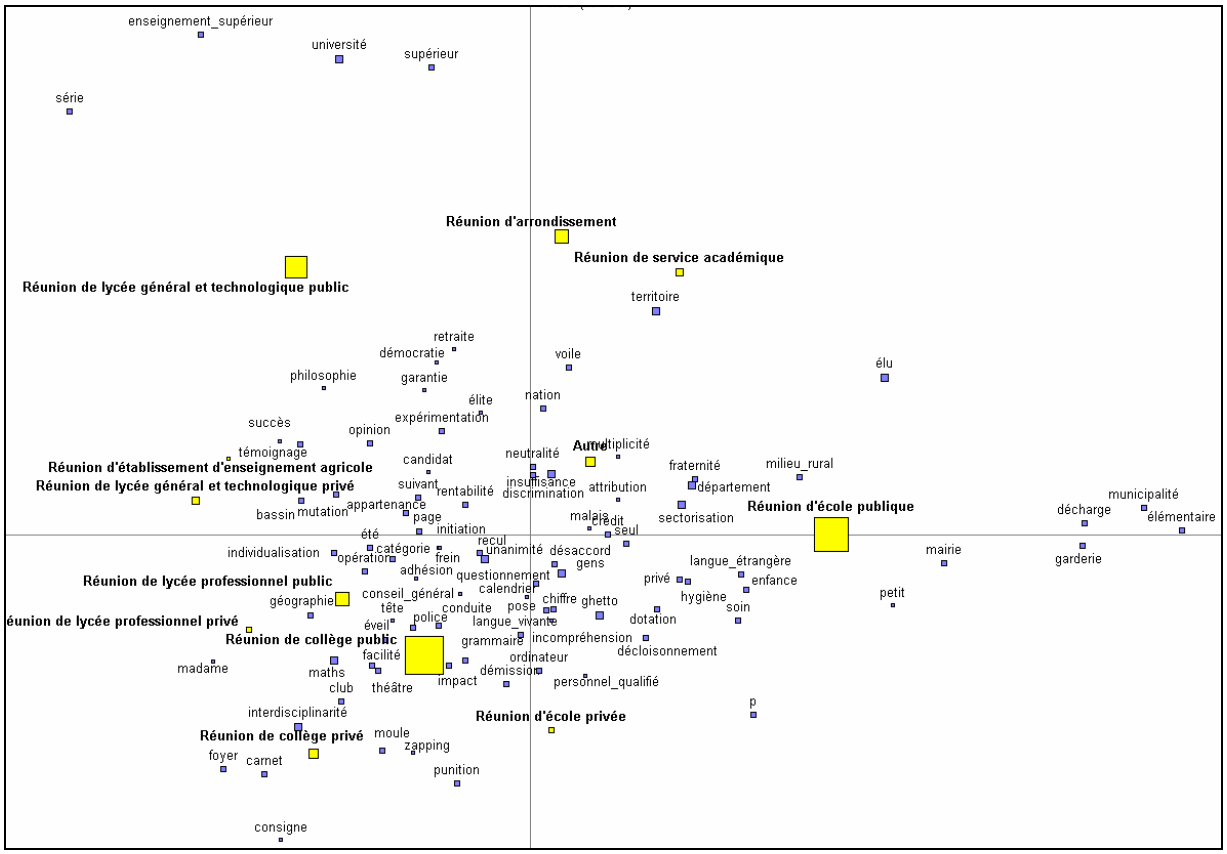

Figure 6 – Synthèse : Spécificités lexicales selon le type d'établissement

(substantifs présents plus de 200 fois, spécificité >1,2) Axe1 51%, Axe 2 27%

3.5.3. Apprendre et former, le face à face élèves enseignants laisse moins de place aux autres acteurs

Les commentaires précédents focalisent naturellement sur les conditions d'organisation du débat. Afin d'examiner d'autres perspectives, nous avons eu recours à d'autres représentations pour rendre compte de la teneur de ce corpus.

Nous avons ainsi recherché à appliquer une analyse en termes d'acteur et d'action en nous référant aux différents acteurs du système éducatif et aux modalités de l'action d'enseigner, repérables par la liste des 6 verbes proposés ci-dessous.

On projette ainsi sur le corpus un modèle a priori pour rechercher dans quelle mesure celui-ci est représenté dans les contenus étudiés. On utilise pour cela des catégories documentées par des champs lexicaux : ensemble de termes illustrant la catégorie. Ainsi la catégorie *Hiérarchie* des acteurs est documentée par les mots : *directeur, proviseur, principal, recteur, IPET*….Les actions sont repérées par l'usage des verbes *apprendre, enseigner, former, éduquer, instruire, transmettre*.

Ces champs lexicaux (ou dictionnaires dans le langage du logiciel) sont utilisés pour calculer l'intensité lexicale des catégories correspondantes. Cet indicateur est égal au rapport du nombre de fois où l'un des éléments du dictionnaire est présent dans le paragraphe étudié par rapport au nombre total de mots du paragraphe. On peut ainsi mesurer pour chaque paragraphe l'intensité avec laquelle chaque catégorie s'y trouve représentée.
Les dictionnaires utilisés sont présentés en Annexe.

Pour rendre compte d'une manière relative de ces intensités, on examine leur co-variation (corrélations entre catégories représentées par une ACP) et on les calcule pour les différentes catégories de contexte : type de débat, géographie, sujets.

3.5.4. Les acteurs et les actions

	Valeur moyenne
A_élève	6,86
A_enseignant	3,30
A_famille	2,68
A_autre_EN	1,25
apprendre	0,85
former	0,54
A_collectivité locale	0,42
A_entreprise	0,38
enseigner	0,36
A_hierarchie	0,22
éduquer	0,22
A_association	0,20
transmettre	0,18
instruire	0,10
Ensemble	**1,25**

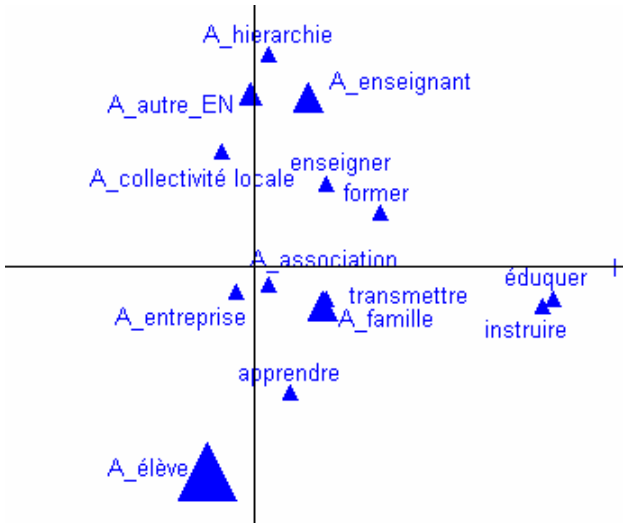

Figure 8 – Acteurs et les modalités de référence à l'action d'enseigner

Les acteurs naturellement privilégiés sont les élèves les enseignants et les familles.

Les enseignants et l'éducation nationale enseignent et forment, les élèves apprennent.

Éduquer, instruire, transmettre dans une moindre mesure varient de manière équivalente et se trouvent plutôt en relation avec la présence dans le texte de l'acteur famille…

3.5.5. L'influence du type d'établissement

Les lycées se différencient des écoles publiques par la place plus forte qu'ils accordent dans leurs synthèses à l'entreprise et aux élèves. Les écoles privilégient la famille et les acteurs de la hiérarchie du système éducatif. Les collectivités locales sont évoquées dans les réunions de service académiques et d'arrondissement. Quand aux verbes, on peut noter le fait que les établissements privés privilégient *transmettre*, *apprendre* et *éduquer* alors que les écoles publiques utilisent plutôt *instruire*.

4. Quelques enseignements

Quels enseignements tirer de ce cas ?

Le fait d'avoir été directement impliqué en tant que prestataires de service dans l'organisation logistique de ce débat et l'expérience acquise dans ce projet nous conduit aux conclusions suivantes :

4.1. La technologie tient ses promesses

Le challenge a été tenu. Le site du débat a certainement contribué à sa publicité mais il a surtout permis, en un temps record de collecter les synthèses rédigées par les organisateurs. Sans la logistique ftp la centralisation de cette information aurait été certainement beaucoup plus coûteuse, moins rapide, mais surtout elle a permis de traiter de manière native la numérisation des données.

Sans les ressources logicielles utilisées les délais n'auraient certainement pas pu être respectés : moins de 3 mois après la tenue de la dernière réunion, l'ouvrage le « Miroir du débat » était publié. Un verbatim représentatif est produit sur la base d'une première sélection statistique dont la pertinence a été contrôlée par une lecture approfondie.

L'exigence d'exhaustivité et d'objectivité a pu être satisfaite grâce aux méthodes d'analyse lexicale évoquées ci-dessus.

4.2. Mais ses procédés restent difficiles à mettre en œuvre et à légitimer…

Dans le cas présent le volume du corpus crée un effet de redondance qui renforce l'assise scientifique des structures lexicales mises à jour. Le procédé statistique permet ainsi de substituer à la lecture intégrale du corpus (entreprise difficilement concevable) l'examen de données (lexiques, cartes, indicateurs…) objectivement produites et communicables.

La mise en œuvre des méthodes utilisées nécessite non seulement une réelle expertise des outils informatiques mais également une bonne connaissance du sujet. Ces 2 conditions sont nécessaires pour parvenir à des résultats dont l'interprétation peut être féconde.

Enfin, pour revenir à l'article du Canard Enchaîné ce qui nous a surtout frappé c'est la difficulté de faire accepter des méthodes inhabituelles quelques aient pu être la qualité des méthodes mise en œuvre et contrôlée par le conseil scientifique créé à cet effet. C'est le véritable obstacle auquel est confrontée la mise en œuvre des nouvelles technologies. Il est à cet égard remarquable de constater qu'alors que ce débat a été lancé sur le mode de l'innovation technologique, la place des nouveaux média dans la suite des travaux de la commission Thélot va décroissant, au point que sur le site du débat n'apparaissent plus maintenant que les 2 ouvrages, le Miroir du Débat et le rapport de la commission.

Cette « étape dans la modernisation de la vie publique » évoquée par le Président de la République, ne résiste donc pas à une structuration des usages dominée beaucoup plus par la prégnance des genres légitimes que par l'influence de l'innovation technologique et la pertinence scientifiques des méthodes utilisées.

Par ailleurs, il est étonnant de constater que l'ouvrage de synthèse « Le Miroir du Débat » ne fait aucune référence à la statistique.

On remarquera également que l'idée force mise en avant dans le projet de loi, le fameux « Socle commun de connaissance » n'est pas le premier choix des participants. Il a fait partie d'une vision « traditionaliste » (voir page 11) de l'éducation, soutenue particulièrement par les lycées généraux, plutôt parisiens.

Références

Bachelet C. Moscarola J. (2005). *La théorie des genres et les modalités de la communication électronique : le cas du débat national sur l'avenir de l'école*. Congrès AIM.

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 : 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*.

Commission Thélot C. (2004). *Les français et leur école*. Le Miroir du Débat, Dunod, http://www.debatnational.education.fr/upload/static/lemiroir/Rapport.pdf

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

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

Moscarola J., Papatsiba V. (2002). Exploration sans a priori ou recherche orientée par un modèle : Contributions et limites de l'analyse lexicale pour l'étude de corpus documentaires. *Actes des JADT 2002*.

Moscarola, J., Baulac, Y. et Ganassali, S. (2004). *Analyse lexicale des données textuelles recueillie lors du Débat National sur l'avenir de l'École*. Le Sphinx – Développement, http://www.debatnational.education.fr/upload/static/lemiroir/pdf/rapport_sphinx.pdf.

Thélot C. (2005). *Débat, démocratie, réformes - Leçons du débat sur la réforme de l'école*. Dunod.

Annexes : Dictionnaires utilisés pour qualifier les acteurs

Élève	Famille	Enseignant	Autres EN	Hiérarchie
élève	parent	professeur	personnel	directeur
enfant	famille	enseignant	médecin	hiérarchie
jeune	père	communauté_éducative	psychologue	proviseur
adolescent	mère	maître	orthophoniste	principal
garçon	frère	équipe	infirmier	adjoint
fille	soeur	pédagogue	spécialiste	recteur
lycéen	parenté	instituteur	adulte	ministre
collégien	tuteur	maître	intervenant	inspecteur
lycéenne	familial	maîtresse	infirmière	IEN
collégienne	maman	assistant	assistante_sociale	IPET
apprenant		surveillant	ATOS	IPR
apprenants		documentaliste	agent	IGEN
apprenti		conseiller_éducation	conseiller	inspectrice
élèves		éducative	spécialiste	chef_de_établissement
écolier		enseignant	Cpe	
stagiaire		corps_professoral	surveillant	
		formateur	médiateur	
		formatrice		
		prof		

Collectivités locales	Associations	Entreprises
mairie	association	entreprise
commune	club	employeur
département	associatif	acteur_economique
ville	bénévole	partenaire_extérieur
municipalité	syndicat	monde_travail
municipal	acteur_social	acteur_économique
région	milieu_associatif	milieu_professionnel
territorial	partenaire_associatif	partenaire_économique
collectivité_locale	partenaire_extérieur	pme
service_social	milieu_social	pmi
partenaire_local		
collectivité		
acteur_local		
partenaire_extérieur		

Utilisation des dictionnaires pour le calcul d'intensités lexicales.

Pour rendre compte d'une manière relative de ces intensités, on examine leur co-variation (corrélations entre catégories représentées par une ACP) et on les calcule pour les différentes catégories de contexte : type de débat, géographie, sujets. On s'intéresse ainsi à la seule variation d'une mesure dont l'étalonnage en valeur absolue (liste des éléments constituant un dictionnaire) est bien sûr discutable.