Analyse discriminante et vocabulaire politique Essai méthodologique

par Jacqueline CHAPELLE,

Chef de travaux à l'Université de Mons-Hainaut

Pierre COUVREUR

Maître de conférence à l'Université de Mons-Hainaut

et Giuseppe PAGANO,

Assistant à l'Université de Mons-Hainaut

I. Introduction

La disparition du bloc communiste et l'ampleur des problèmes économiques, sociaux, écologiques,... expliquent sans doute que nous vivons une époque de dépolitisation où le pragmatique a pris le pas sur l'idéologie. Ce phénomène se traduit en Europe notamment par l'étonnante convergence des politiques économiques pourtant mises en oeuvre par des gouvernements divers qu'ils soient mono ou pluri-partis, socialistes, centristes ou libéraux.

De nombreux auteurs soutiennent que cette *désidéologisation* ne toucherait pas seulement l'action mais aurait également atteint le vocabulaire politique. Les formules auxquelles les hommes politiques ont recours ne seraient plus libérales ou marxistes mais délibérément apolitiques et culturelles [1].

Cette affirmation est difficile à vérifier. Si l'uniformisation des politiques économiques apparaît assez clairement à travers le choix des instruments et la fixation des objectifs, la banalisation du discours ne peut être attestée qu'à travers une analyse plus longue.

L'objet de cet article est double. Il s'agit d'abord, précisément, d'exposer une *méthode* relativement simple qui, si l'on dispose d'un échantillon suffisant, permettrait de vérifier le bien-fondé de cette assertion. Il s'agit ensuite d'exposer les principaux *résultats* d'une première application prospective sur un échantillon limité.

L'étude se base sur la mise en évidence de *formes lexicales* dont la présence ou l'absence dans le vocabulaire majeur des hommes politiques est spécifique de la droite ou de la gauche, du discours libéral ou socialiste. Le mot spécifique doit être pris au sens *statistique* [2]. L'approche des occurrences est mathématique et non linguistique. Elle n'englobe ni les relations syntaxiques, ni les processus énonciatifs.

(1) D. LABBE, *Le vocabulaire de François Mitterand.* Paris, 1990, p. 25 ; G. THOVERON, *La communication politique aujourd'hui.* Bruxelles, 1990, p. 216.

(2) A. BERGOUNIAUX, M. LAUNAY, R. MOURIAUX, J.P. SUEUR, M. TOURNIER, *La parole syndicale.* Paris, 1982, p. 97.

Dans une première approche nous nous limitons à une situation minimaliste en choisissant l'échantillon représentatif de ce vocabulaire dans la strate du sommet, à savoir les discours des présidents du Parti Réformateur Libéral (P.R.L.), Daniel Ducarme et Antoine Duquesne et du président du Parti Socialiste (P.S.), Guy Spitaels.

Notre recherche vise à mettre en évidence les mots que ces hommes politiques préfèrent ou évitent en postulant que les fréquences d'emploi des mots signalent des goûts ou des répugnances.

Selon O. REBOUL [3], certains mots se chargent de lourds parfums idéologiques et traduisent la position de celui qui les utilise. Il s'agit alors de déceler ces mots à appartenance politique, les désignants socio-politiques, ou, si l'on veut, les mots-étendards, les mots-valeurs, les mots-pôles, à partir du dénombrement de tout le vocabulaire des discours.

La question à laquelle il sera tenté de répondre est la suivante. Est-il possible de distinguer des textes écrits par des personnalités d'obédience politique différente et bien connue, sur la base de mots typiques ?

La démarche comporte deux étapes. Il s'agit d'abord de l'identification des mots discriminants à partir de leurs fréquences. Si, par exemple, il y a deux classes à discriminer (le discours socialiste d'une part, le discours libéral de l'autre), seront retenus les mots ou groupes de mots présentant des fréquences d'apparition très différentes dans l'une et l'autre de ces classes.

En second lieu, on utilisera *l'analyse factorielle discriminante* pour tester la capacité des mots ou groupes de mots retenus à discriminer entre les origines politiques des discours.

II. La méthode

A. Principe

L'analyse discriminante est présentée dans de nombreux ouvrages traitant de l'analyse des données multidimensionnelles [4], nous nous contenterons donc d'en expliquer brièvement le principe général et son application à la discrimination lexicale.

1. Principe général

Soit un ensemble de p variables quantitatives $X^1,...,X^p$ et une variable qualitative Y ayant q valeurs possibles. Ces variables prennent leurs valeurs sur un ensemble de n éléments ou individus. En pratique, on disposera donc d'un tableau de données du type suivant (figure I).

(3) O. REBOUL, *Langage et idéologie.* Paris, 1980.

(4) Voir, par exemple : J.M. ROMEDER, *Méthodes et programmes d'analyse discriminante.* Paris, 1973 ; E. DIDAY, J. LEMAIRE, J. POUGET, F. TESTU, *Eléments d'analyse de données.* Paris, 1982 ; G. SAPORTA, *Probabilités, analyse des données et statistique.* Technip, 1990 ; ou M. TATSUOKA, *Multivariate Analysis.* New York, 1971.

FIGURE I

Le principe de l'analyse discriminante

INDIVIDUS ↓	VARIABLES →	QUANTITATIVES					QUALITATIVE
		X_1	....	X^j	....	X^p	Y
1		$X^1{}_1$		$X^j{}_1$		$X^p{}_1$	Y_1
i		$X^1{}_i$		$X^j{}_i$		$X^p{}_i$	Y_i
n		$X^1{}_n$		$X^j{}_n$		$X^p{}_n$	Y_n

La variable Y définit une partition de l'ensemble des individus en q classes distinctes. Dans notre exemple d'analyse lexicale, la variable qualitative (Y) est *l'appartenance politique ;* elle peut prendre deux "valeurs": socialiste ou libéral. Les variables quantitatives (X_i) sont *les fréquences d'apparition* de certains mots ou groupes de mots que nous appellerons les *vocables discriminants.*

Classiquement, on distingue deux formes d'analyse discriminante : l'analyse discriminante *à but descriptif* et l'analyse discriminante *à but décisionnel.*

La première permet de voir si les p variables quantitatives sont aptes à différencier les q classes définies par la variable qualitative. Dans notre exemple, les vocables discriminants sont-ils capables de différencier les discours − ou parties de discours − libéral et socialiste ?

L'idée de base est de construire de nouvelles variables quantitatives, combinaisons linéaires des variables initiales, qui prennent des valeurs "proches" pour des individus d'une même classe et des valeurs "éloignées" pour des individus de classes distinctes.

Ces nouvelles variables sont appelées *facteurs discriminants* et on parle ainsi d'analyse factorielle discriminante. Concrètement, on procède comme suit. Le premier facteur discriminant est celui qui a le report variance interclasses / variance intra-classes maximum ; le deuxième facteur discriminant est indépendant du premier et maximise encore le précédent rapport, et ainsi de suite. La théorie montre qu'il y a q-1 facteurs discriminants et que les pouvoirs discriminants successifs sont décroissants.

Le problème dans le cas de l'analyse discriminante à but décisionnel consiste à "décider" à quelle classe appartient un individu pour lequel connaît les valeurs des X_j ($j = 1,...,p$) mais pas celle de Y qui détermine sa classe.

En utilisant les facteurs discriminants présentés ci-devant et en se fixant une règle d'affectation, on conçoit qu'il sera possible d'identifier la classe d'un individu "anonyme". Son application à l'analyse lexicale consisterait à classer un discours dont on ignore l'origine, en socialiste ou libéral en fonction des mots qui le composent.

L'analyse discriminante à but décisionnel est donc une suite de l'analyse discriminante à but descriptif. Elle comporte essentiellement trois phases.

Il s'agit, d'abord, de l'élaboration du modèle discriminant à partir d'un échantillon "de base". Celui-ci et l'échantillon "test" introduit ci-après sont constitués d'individus dont la classe est connue. La deuxième phase correspond à l'estimation des performances du modèle (pourcentages d'individus classés correctement) à l'intervention d'un échantillon "test", l'échantillon "de base" pouvant surestimer ces performances. Enfin, la troisième phase consiste en l'application du modèle à des individus "anonymes".

2. Application à la discrimination lexicale

L'analogie entre l'analyse discriminante et la séparation lexicale peut être résumée comme suit (figure II).

FIGURE II

Analyse discriminante et séparation lexicale

	ANALYSE DISCRIMINANTE	SEPARATION LEXICALE
Individu à classer	i	Discours ou partie de discours (groupes de propositions)
Nombre total d'individus	n	169 groupes de propositions
Variable quantitative	X_i avec i = 1,..., p	Fréquence d'apparition d'un mot ou groupe de mots appelé *vocable discriminant*
Variable qualitative	Y (pouvant prendre q valeurs)	Appartenance politique : pouvant prendre 2 valeurs : libéral ou socialiste

L'objectif est ici essentiellement de faire de l'analyse discriminante à but descriptif puisque la question est de savoir si représenter un discours politique sur la seule base des fréquences lexicales peut rendre compte de son appartenance politique.

Les deux étapes retenues actuellement sont l'élaboration d'un modèle discriminant à partir d'un échantillon comprenant trois textes et l'estimation des performances du modèle (pourcentage de parties du discours classées correctement).

B. Les étapes de l'analyse

1. L'échantillon

L'échantillon que nous avons retenu pour cette analyse comprend trois discours regroupés en deux sous-échantillons.

Le premier sous-échantillon correspond au discours prononcé par M. Guy Spitaels, président du Parti Socialiste, à Ottignies – Louvain-la-Neuve, à l'occasion de la Convention – Forum des 24 et 25 novembre 1990. Il comporte 9.257 mots que nous avons regroupés en 707 propositions – la proposition est une unité

indépendante énonçable – puis en 66 *groupes de propositions*. En effet, c'est le groupe de propositions et non la proposition ou la phrase qui est choisi comme unité de segmentation du texte.

Le second sous-échantillon a été obtenu par addition de deux discours prononcés respectivement par M. Antoine Duquesne, président du Parti Réformateur Libéral, à l'occasion du Comité permanent de Gembloux, le 23 février 1991 ; et par M. Daniel Ducarme, vice-président de ce parti, à l'occasion du Congrès doctrinal de Liège, le 23 décembre 1989. Ensemble, ces deux discours comptent 13.413 mots regroupés en 1.091 propositions et en 103 groupes de propositions.

Comme on peut le constater les échantillons utilisés présentent un caractère limité. Rien ne nous permet d'affirmer, en effet, qu'ils soient représentatifs, respectivement, du discours libéral et du discours socialiste.

D'abord parce que l'évolution dans le temps du vocabulaire politique limite leur représentativité éventuelle à la période actuelle.

Ensuite, le nombre restreint d'auteurs ne nous permet pas d'exclure a priori, le caractère atypique du vocabulaire utilisé par rapport au discours actuel de l'ensemble du Parti Socialiste et de l'ensemble du Parti Réformateur Libéral.

Notre analyse ne permet donc aucune généralisation ; mais, comme nous l'avons signalé dans l'introduction, la méthode utilisée est encore à un stade *exploratoire* et *notre objectif se limite ici à un premier test.*

2. Le relevé du vocabulaire politique

En toute première analyse, il nous a semblé logique d'étudier les mots utilisés dans chaque discours et leur fréquence d'apparition. Nous avons retenu tous les mots ou vocables *significatifs :* les substantifs, les adjectifs qualificatifs, verbes et adverbes, en n'envisageant pas leur univers sémantique. Après avoir regroupé toutes les formes se rapportant visiblement à un même vocable (masculin, féminin, singulier, pluriel, formes conjuguées, pronoms remplaçant le vocable) nous avons retenu les vocables le plus fréquents, c'est-à-dire *tous ceux apparaissant au moins deux fois dans l'un ou l'autre des sous-échantillons.*

De cette façon, on a retenu 441 vocables dont les plus représentatifs sont repris au tableau I ci-dessous avec leurs fréquences d'apparition dans chacun des sous-échantillons.

3. Choix de l'unité statistique

Ces vocables les plus fréquents permettent-ils de reclasser correctement les parties des deux sous-échantillons, le sous-échantillon socialiste et le sous-échantillon libéral ? Pour répondre à cette question, il convenait d'abord de préciser ce que l'on entend par *partie* des sous-échantillons.

L'unité de base était la *proposition* telle que nous l'avons définie ci-dessus. Cependant, il restait possible de regrouper plusieurs propositions en blocs. Pour cette première analyse, nous avons retenu des *blocs (ou groupes) de dix propositions* environ. De cette façon sont apparus 66 blocs dans le discours socialiste et 103 blocs pour le sous-échantillon libéral, soit un total général de 169 blocs. Ce sont ces blocs que l'analyse discriminante devrait reclasser correctement sur la base des vocables discriminants.

III. Les résultats

Les résultats doivent être analysés à deux égards.

D'une part, d'un point de vue technique, il s'agit de *tester* une méthode d'analyse. On examinera donc d'abord les performances du modèle. Celles-ci sont appréciées à travers les taux de reconnaissance, c'est-à-dire le pourcentage de groupes de propositions qui seront reclassées correctement.

D'autre part, si la méthode est efficace, les vocables discriminants deviennent représentatifs d'un type de discours au moins sur l'échantillon considéré. Ils peuvent alors être analysés d'un point de vue socio-politique.

A. Les vocables discriminants

1. Sélection

La première question qui se pose pour l'application d'une telle méthode est celle de la *sélection des éléments discriminants.* Quels vocables sont, plus que d'autres, susceptibles de caractériser le sous-échantillon socialiste par rapport au libéral? Quel critères utiliser pour les sélectionner? Combien en retenir?

Dans un premier temps, nous avons rentenu *tous les vocables qui apparaissaient au moins deux fois dans l'un ou l'autre des sous-échantillons.* De cette façon, 441 vocables ont été repris. Il est clair, cependant, que le pouvoir discriminant de ces 441 vocables est variable, de sorte que la deuxième étape consistait à les ranger en fonction de leur pouvoir discriminant [5].

A cette fin, nous avons mis en oeuvre un test X^2 de dépendance entre la fréquence du vocable et l'appartenance politique du discours. Ce test s'est avéré significatif au seuil de 1% pour 34 vocables, à un seuil compris entre 1 et 5% pour 47 vocables et à un seuil compris entre 5 et 10% pour 41 vocables.

On trouvera au tableau I ci-dessous la liste de ces 122 vocables discriminants ainsi que le seuil de signification du test.

TABLEAU I

Vocables discriminants et seuils de signification

SEUIL		
$0.05 < \alpha \leqslant 0.1$	$0.01 < \alpha \leqslant 0.05$	$\alpha \leqslant 0.01$
1) Militant	Groupe	Collectif
2) Revendication	Déséquilibre	Démographie
3) Capitalisme	Politique	Action
4) Dépenses	Aide	Mouvement
5) Socialisme	Crise	Contrôle
6) Parti	Projet	Formation
7) Pouvoirs	Ressources	Investissement

(5) La procédure discriminante du logiciel SPSS permet, en principe, de sélectionner les éléments les plus discriminants. Cependant, le logiciel ne traite qu'un maximum de 200 variables. Il convenait donc de procéder à une première sélection.

8)	Chômage	Vie	Structure
9)	Région	Gouvernement	Libéral
10)	Wallon	Homme	Amis
11)	Choix	Société	Ecologie
12)	Consommateur	Gauche	Protection
13)	Masculin	Développement	Société
14)	Peuple	Croissance	Marché
15)	Familial	Individu	Economie
16)	Féminin	Système	Tiers-monde
17)	Inégalité	Facilité	Citoyen
18)	Marginal	Faveur	Industrialisé
19)	Pacifisme	Harmoniser	Production
20)	Accroître	Infrastructure	Public
21)	Assumer	Monde	Revenus
22)	Central	Nécessaire	Femme
23)	Conseil	Privatisation	Secteur
24)	Croire	Etudiant	Améliorer
25)	Demande	Alliance	Convient
26)	Fermeté	Autonomie	Durée
27)	Fin	Cadre	Entente
28)	Intégration	Certain	Renforcer
29)	Intérêt	Clair	Renouveler
30)	Matière	Dire	Terme
31)	Moment	Décentralisation	Toutefois
32)	Norme	Essentiel	Elément
33)	Oeuvre	Expérience	Evidence
34)	Orientation	Fondateurs	Egalité
35)	Permanent	Garde	
36)	Permettre	Habitude	
37)	Pratique	Imagination	
38)	Préférence	Menace	
39)	Réalisation	Mesure	
40)	Réunion	Occidental	
41)	Souhait	Priorité	
42)		Prétendre	
43)		Rapide	
44)		Rapport	
45)		Rester	
46)		Stabilité	
47)		Liberté	

2. Valeur discriminante

En reprenant les 81 vocables dont le seuil de signification au test X^2 est inférieur ou égal à 5%, le logiciel SPSS permet de classer les vocables les plus discriminants sur la base des coefficients de corrélation entre les variables discriminantes et la fonction discriminante. Le tableau II fournit la liste des 30 vocables les plus discriminants ainsi que leur coefficient de corrélation.

TABLEAU II

Corrélation entre variables discriminantes et fonction discriminante

Vocables discriminants	Occurrences Disc. Soc.	Disc. Lib.	Coefficient de corrélation
1) Marché	40	2	.20449
2) Action	41	13	.19593
3) Social	46	13	.19555
4) Collectif	19	1	.17149
5) Amis	0	22	-.17090
6) Convient	11	1	.16907
7) Renforcement	12	2	.16449
8) Structure	12	2	.16436
9) Tiers-monde	19	2	.16309
10) Amélioration	8	0	.16072
11) Libéral	1	46	-.15588
12) Public	29	9	.15413
13) Mouvement	16	2	.15352
14) Industries	9	0	.15336
15) Femme	26	3	.15149
16) Economie	39	21	.15149
17) Développement	30	10	.14873
18) Protection	25	4	.14147
19) Egalité	11	0	.14142
20) Durée	8	1	.13842
21) Société	24	13	.13806
22) Investissement	6	0	.13685
23) Démographie	6	0	.13685
24) Terme	9	1	.13573
25) Liberté	1	26	-.13121
26) Revenu	7	0	.12976
27) Masse	13	3	.12770
28) Renouveler	16	8	.12582
29) Projet	14	2	.12582
30) Toutefois	7	1	.12582

B. Les taux de reconnaissance

Ces vocables permettent-ils de classer correctement un bref extrait de discours, ce que nous avons appelé un groupe de propositions ? Le tableau III ci-dessous indique que le taux de reconnaissance moyen est particulièrement élevé. Il montre en outre qu'un nombre limité de vocables suffit à obtenir des taux de reconnaissance déjà largement significatifs.

TABLEAU III

Les taux de reconnaissance

	Disc. Soc.	Disc. Lib.	Total
N. de groupes de propositions	66	103	169
	10 vocables		
Groupes cl. Corr.	49	101	150
Tx. Recon. (%)	74.2	98.1	88.8
	20 vocables		
Groupes cl. Corr.	56	101	157
Tx. Recon. (%)	84.8	98.1	92.9
	34 vocables		
Groupes cl. Corr.	62	103	165
Tx. Recon. (%)	93.9	100	97.6
	81 vocables		
Groupes cl. Corr.	66	103	199
Tx. Recon. (%)	100.0	100.0	100.0

Si l'on retient les 10 vocables les plus significatifs selon le test X^2, on s'aperçoit que 88% des groupes de propositions sont reclassés correctement.

Ce pourcentage est plus élevé pour les discours libéraux (98.1%) que pour le discours socialiste (74.2%). Les 20 vocables les plus significatifs permettent de reclasser correctement 92.9% des groupes de propositions. L'amélioration est due exclusivement au discours socialiste dont 84.8% des groupes de propositions sont désormais identifiés correctement. Les 34 vocables pour lesquels le test X^2 est significatif au seuil de 0.01, permettent de reclasser correctement 97.63% des groupes de propositions dont tous les groupes de propositions contenus dans les deux discours libéraux et 93.9% du discours socialiste. Enfin, les 81 vocables pour lesquels le test est significatif au seuil de 0.05, permettent de reclasser correctement tous les groupes de propositions.

Les taux de reconnaissance plus élevés des groupes de propositions libérales tiennent sans doute à l'importance des échantillons. Les deux textes libéraux étant ensemble plus longs que le texte socialiste, la méthode a pu exercer ses capacités discriminantes sur une matière plus abondante, ce qui en accroît l'efficacité.

En dehors de cette particularité, il ressort clairement de cet exercice que l'analyse factorielle discriminante appliquée à partir de critères simples, se prête bien à l'étude du discours politique.

IV. Conclusions et perspectives

A. La méthode

A ce stade, deux conclusions peuvent être tirées.

D'une part, l'analyse discriminante semble se prêter utilement à l'étude du discours politique. Sur la base de technique simples et au moyen d'un logiciel statistique bien connu, elle permet une différenciation efficace des groupes de propositions repris dans un discours socialiste et dans deux discours libéraux. Les taux de reconnaissance obtenus à partir d'un nombre restreint de vocables sont certainement remarquables.

Il faut, d'autre part, limiter la portée de ce résultat compte tenu de l'étroitesse des échantillons utilisés. A ce stade, toute généralisation serait prématurée. On ne peut notamment affirmer que les vocables discriminants retenus pourraient également reclasser correctement des propositions extraites d'autres discours socialistes ou libéraux.

L'influence des échantillons est d'ailleurs perceptible lorsqu'on observe les vocables les plus discriminants repris au tableau I. Deux d'entre eux, *toutefois* et *renouveler* ne sont sans doute pas représentatifs d'un discours politique mais plutôt d'une tournure propre à l'auteur. De tels phénomènes seraient vraisemblablement atténués et sans doute même disparaîtraient si les échantillons étaient plus larges et comprenaient un nombre d'auteurs sensiblement plus élevé.

Enfin, les développements futurs de la méthode peuvent être envisagés de deux façons.

La première consiste à élargir les échantillons qui ont servi à l'identification des vocables discriminants. Le nombre de textes devrait être accru d'abord pour élargir l'échantillon à d'autres partis et ensuite pour multiplier les auteurs au sein d'un même parti.

La seconde consiste à utiliser la batterie de vocables discriminants identifiés (que ce soit ceux repris ci-dessus ou ceux obtenus à partir d'échantillons plus larges) pour tester leur capacité à reclasser correctement des propositions tirées d'autres discours socialistes ou libéraux : des discours écrits par d'autres personnalités ou à d'autres époques.

B. Le discours politique

Les résultats de cette première étude infirment au moins partiellement l'idée de l'*uniformisation* du discours politique.

Comment, en effet, expliquer alors que quelques mots suffisent à identifier presque à coup sûr l'origine politique d'un groupe de propositions ? D'autant que les vocables retenus comme les plus discriminants ne sont pas des termes passe-partout mais restent chargés d'idéologie. Des termes tels que *marché, social, collectif, structure, tiers-monde, libéral, public, protection, liberté* ou *égalité* relèvent incontestablement du débat idéologique.

On observera, cependant, que la différenciation idéologique n'est pas nettement marquée même si elle est statistiquement perceptible. D'une part, les deux sous-échantillons reprennent un certain nombre de *vocables non discriminants,* c'est-à-dire présents dans les deux sous-échantillons avec des fréquences largement comparables. Ces mots – Europe, écologie, fédéralisme, fiscalité, traduisent

des préoccupations communes et constituent des thèmes stables du débat politique. D'autre part, certains termes nettement chargés d'idéologie sont *totalement absents :* prolétaire, syndicat, patronat, camarade, marxisme,...

Ainsi, si les discours socialiste et libéraux de notre échantillon contenaient assez de spécificités pour qu'une technique relativement simple les identifie presque à coup sûr, ils semblent néanmoins dénoter un rapprochement du vocabulaire des uns et des autres et un recul de l'idéologie du moins dans sa forme la plus marquée.

Summary : **Discriminant analysis and the study of political vocabulary**

This paper aims at testing the hypothesis of growing ideological uniformity of political speeches. If political speeches lack ideological differences, it should be difficult to re-classify them only by analyzing the presence or absence of lexical items. We first worked out a method to classify political speeches and then carried a test on two speeches by leading Belgian French-speaking politicians.

The method is based on discriminant analysis. It utilizes the words most encountered in one speech and not in the other as discriminant factors. Statistical softwares then assess a discriminant function used to re-classify short parts of each speech called blocks. The most discriminating 10 factors re-classify correctly 89% of the blocks. The percentage increases to 93% with 20 factors and to 98% with 30 factors.

However the results should be taken with caution because of the limited sample, the test tends to question the growing uniformity of political speeches. The sampled ones had enough specific features for allowing a rather simple method to re-classify most parts of them correctly, even if some typically ideological items are not to be found.