

Représentations politiques du monde rural. Éléments d'une analyse textométrique.

Introduction

Cette communication vise à présenter succinctement une branche particulière de la linguistique, que l'on appelle la *textométrie* (ou *lexicométrie*), associée au discours politique concernant le domaine rural. Il s'agit de montrer les intérêts – et les défauts – de cette analyse du discours, et de montrer ce faisant quelles sont les représentations que les hommes et femmes politiques font de la ruralité dans la langue contemporaine. Notre exposé s'articulera en trois temps : tout d'abord, nous présenterons les éléments théoriques dont nous nous sommes servi pour notre étude, ensuite, nous commenterons les unités les plus fréquemment employés dans notre corpus d'étude avant de compléter la chose par la prise en compte de leur contexte d'emploi.

I. Quelques définitions

I.1. Lexicométrie et textométrie

On appelle *lexicométrie* l'étude statistique du vocabulaire – du *lexique* – dans la langue, en général ou dans un discours en particulier. Cette branche assez neuve de la linguistique, puisqu'elle apparut à la fin des années 70 grâce aux avancées de l'informatique, consiste à analyser un texte ou un corpus (soit un ensemble de documents écrits ou oraux réunis par l'intermédiaire d'une ou de plusieurs propriétés communes) du point de vue de l'emploi de ses mots. On relève ainsi, méthodiquement, le nombre de fois que tel terme, ou tel ensemble de mots, apparaît, et on en tire des statistiques qui peuvent nous en apprendre davantage sur les caractéristiques de tel ou tel discours.

L'intérêt de la *lexicométrie* a été, dans un premier temps, de vérifier certaines hypothèses empiriques concernant l'apparition de certains mots en particulier. Si l'on considère, pour être à propos, le mot *informatique*, terme inventé en 1962 sur la base d'*information* et par dérivation suffixale, nous pouvons analyser le journal *Le Monde* et observer à partir de quelle année l'emploi du mot devient signifiant, c'est-à-dire dépasse un certain pourcentage d'emploi dans le corpus que compose ce journal. Cela permet d'estimer à quel moment le terme cesse d'être « spécialisé » pour rentrer pleinement dans le vocabulaire courant, c'est-à-dire le moment où il est approprié par les locuteurs, quels qu'ils soient.

Compte tenu de l'aspect répétitif de la discipline, c'est l'ordinateur qui accélérera la procédure : des logiciels sont très rapidement inventés et ceux-ci relèvent des séquences de lettres qui font mot, et créent automatiquement des données statistiques interprétables. Avec cette avancée informatique, la *lexicométrie* cède progressivement sa place à la *textométrie*, qui cherche à aller au-delà de la simple notion de *mots*. Effectivement, bien qu'il soit souvent utile de rester à cette échelle et de croiser ces données avec les origines socio-politiques des locuteurs¹, peu de choses peuvent être dites de ces résultats. Il est apparu bien plus utile de recontextualiser ces termes dans le discours où nous pouvons les trouver, les *textes* pour ainsi dire, dans lesquels la valeur d'un mot dépend autant de son identité propre – définitions,

¹ La lexicométrie est apparue concomitamment aux grandes découvertes de la sociologie. Nous renvoyons à la bibliographie donnée en fin d'articles pour explorer ces différents domaines de la sociolinguistique.

dénotations et connotations, niveau de la langue... -- que de sa relation avec l'ensemble du discours, et notamment des mots qui les suivent ou les précèdent, soit ceux avec lesquels nous le trouvons le plus fréquemment employé. À nouveau, l'avancée de l'informatique permet d'affiner de plus en plus nos relevés, par exemple en associant un mot à une catégorie grammaticale – le logiciel analysant automatiquement une phrase ce faisant – et en croisant de multiples données qui permettent, tout en restant dans une analyse automatique et itérative des discours, d'obtenir des résultats pertinents pour les études linguistiques.

Le projet PRESTO, auquel nous avons participé², se positionne dans cette lignée : nous avons utilisé pour ce faire une version modifiée du logiciel de lexicométrie « TXM », que nous avons aidé avec mes collègues à développer³, pour mener les analyses suivantes. Une précision avant de rentrer dans le vif du sujet : toutes les données que nous présenterons seront brutes, et seront donc à commenter et à interpréter avec beaucoup de prudence. En tant que telle, toute donnée statistique est *décontextualisée*, c'est-à-dire offerte indépendante des pressions et des phénomènes qui l'ont produite. Pour prendre un exemple connu et issu des études de Pierre Bourdieu, c'est une chose que d'observer statistiquement que les fils de professeurs sont plus souvent professeurs que les enfants des autres catégories socio-professionnelles ; c'est en une autre d'expliquer les phénomènes qui conduisent cette statistique à exister.

Dans cette communication, nous nous consacrerons qu'aux seuls observables statistiques : nous les comparerons entre eux, mais nous ne parlerons pas des contextes de leur production, pour diverses raisons liées à la constitution de notre corpus que nous présentons à présent.

I.2 Présentations et limites du corpus d'étude

Si la notion de *rural* ou de *ruralité* est à prendre ici dans un sens très générique de « ce qui renvoie à la campagne, espace en-dehors des villes », la notion de *discours politique* est en revanche plus complexe à mettre en œuvre et c'est cet aspect-là qui nous empêche, à moins de partir sur une toute autre étude, de comparer les données méthodiquement et ne prendre en compte que des éléments très superficiels.

En linguistique, il est coutume d'organiser les discours, ou les textes, en fonction de leur *réception* et non pas selon leur *production*, ce qui est davantage l'apanage des études littéraires. S'il est très souvent des questions sur l'appartenance générique de tel ou tel objet littéraire, un *roman* est cependant considéré comme tel en fonction d'un certain horizon d'attentes défini selon l'époque de composition du texte. Généralement, c'est l'auteur qui choisira, ou non, de se positionner vis-à-vis de lui, sans que le lecteur n'ait ici son mot à dire. La linguistique, en revanche, considère les discours non pas en fonction de l'intention, réelle ou supposée, du locuteur-source, mais en fonction de sa réception, effective, par un locuteur-cible. Nous parlerons, par exemple, de discours/texte/énoncé *explicatif, narratif, descriptif, argumentatif*... sans nous prononcer sur les intentions véritables de l'auteur, mais plutôt sur l'effet perçu par la communauté des locuteurs.

² <http://presto.ens-lyon.fr/> « PRESTO est un projet franco-allemand financé par l'ANR et la DFG pour trois ans (avril 2013 - avril 2016). Il a pour but l'étude diachronique, adossée à une approche de corpus statistique et distributionnelle, de l'emploi, des valeurs sémantiques et discursives des prépositions françaises du IX^e au XX^e siècle. »

³ <http://textometrie.ens-lyon.fr/> Le logiciel est libre (licence Creative Commons), et un manuel en ligne est disponible à cette adresse : <http://txm.sourceforge.net/doc/manual/manual.xhtml>

Nous pouvons illustrer les choses en reprenant l'exemple du roman : si le texte en tant que tel est perçu comme un ensemble monolithique et stabilisé, avec des bornes clairement établies, il aligne pourtant des extraits visant à décrire un personnage, à expliquer ses motivations, à argumenter ses idées bref, autant de *séquences textuelles* plus ou moins délimitées quant à elle et qui ont une visée locutoire, illocutoire et perlocutoire spécifique. Ce sont ces séquences qui pourront faire l'objet d'une étude linguistique dans la mesure où elles possèdent des caractéristiques discursives (structures syntaxiques, effets de sens, tiroirs verbaux...) quantifiables et analysables. Les séquences sont composées de phrases, les phrases de propositions, les propositions de syntagmes : autant d'éléments analysables par la linguistique mais qui toutes prennent appui sur un principe de réception, et non de production.

Parler alors, comme nous le faisons ici, « d'analyse linguistique de discours politique », est pour le moins malhonnête scientifiquement : l'on se positionne effectivement sous l'angle de la production (*est discours politique un discours prononcé par un acteur de la vie politique*) et non pas de sa réception. Que de différences, pour un président de la république, entre une intervention à l'Assemblée nationale, une ouverture du salon de l'automobile, une interview face à un journaliste ou une remise d'un prix honorifique ! Les effets sont distincts les uns des autres (argumenter, présenter, expliquer, célébrer...) et les mécanismes linguistiques, ce faisant, le sont tout autant.

Il y a cependant une façon de compenser cette difficulté, c'est la voie que nous avons choisie ici : c'est de s'assurer une continuité de signification, ou de thème, et de décontextualiser les propos établis. Quand bien même la visée locutoire des « discours politiques », et plus précisément des « discours d'un homme/d'une femme politique » diffère, nous pouvons considérer qu'un individu possède, sur un intervalle de temps raisonnable (un mandat légal, globalement), le même stock d'objets lexicaux. Bien que notre discours évolue et s'enrichisse, de notre naissance à notre mort, si nous faisons un « instantané » de cette évolution, nous observerons de l'immobilisme, tout comme, mettons, l'on peut analyser l'être humain anatomiquement depuis 2000 ans sans voir de nuances, alors que nous savons que son histoire évolutive, s'échelonnant sur un intervalle de temps bien plus long, est faite de ruptures et de changements notables. Quand, de plus, nous restreignons un sujet de discussion en particulier, la ruralité ici, nous pouvons considérer que les indices linguistiques observables et quantifiables sont relativement stables et comparables de discours en discours, et ce sans prendre en compte la réception des dits discours. Pour résumer, nous pouvons considérer qu'un homme politique, au cours d'un mandat légal, à propos d'un sujet en particulier, s'exprime toujours de la même façon.

Cela est dû également à notre conception de la politique et de ses valeurs : il faut défendre certains intérêts et les conserver pour se distinguer de ses opposants. Un candidat en campagne électorale, de son premier discours d'annonce de participation au scrutin à son discours d'élection en passant par les débats auxquels il participerait, parle de la même façon, précise ou nuance, mais ne fait jamais volte-face : cela serait très mal vue de la part de ses (potentiels) électeurs. Le seul moment, et cela s'est vu, où un discours politique évoluerait serait du fait d'une crise qui inviterait à changer de paradigme : un attentat, une catastrophe naturelle, et ainsi de suite. Mais dans le domaine qui nous occupe, en France et sur les dix dernières années, il n'y a rien eu de tel. La crise s'annonce, nous dit-on et prévoit-on, mais notre corpus s'arrête avant celle-ci.

Pour illustrer ce phénomène, nous avons choisi d'analyser grâce à la textométrie trois corpus de discours politiques traitant de la ruralité. Ces discours ont de commun :

-- Leur origine unique : un seul et même locuteur les produit, sans que nous ne déclarions qu'il en est bien l'auteur. Nous savons effectivement que les hommes politiques ont des rédacteurs, mais nous considérons que les orateurs ont néanmoins un droit de regard sur les mots qu'ils prononcent.

-- Leur date : ces discours ont été prononcés, pour chaque locuteur, dans un intervalle de quatre ans au plus entre le premier et le dernier.

-- Leur sujet : la ruralité, au sens large comme nous le précisons plus haut.

Ces éléments nous permettent de mener une analyse textométrique par locuteur et de les comparer, à nouveau sans les contextualiser : nous présenterons des données brutes et nous nous bornerons à des observations superficielles, qu'il faudra compléter avec un regard plus aiguisé sur les forces politiques en présence. Partant, nous avons analysé les discours de Nicolas Sarkozy, lorsqu'il était président de la République (2007-2011) ; de François Hollande, jusqu'au mois d'août 2016 (2012-2016) ; enfin, de José Bové, en qualité de député européen (2010 – 2014). Nous avons sélectionné ces locuteurs non seulement par souci de pluralisme, mais aussi parce que les corpus étaient accessible librement : les discours de Nicolas Sarkozy ont été compilés par un chercheur de Nantes, ceux de François Hollande sont disponibles sur le site de l'Élysée, et ceux de José Bové sur son site personnel⁴.

II. Unités lexicales les plus fréquemment employées

En analysant les différents discours via le logiciel TXM, nous avons obtenu ce genre de résultats (Figure 1).

word	Fréquence T=12948
,	353
la	266
des	264
d	239
les	235
et	230
l	212
en	158
s	135
le	120
r	118
pour	95
une	86
est	83
qui	75
.	72

Figure 1 – Résultats bruts des trois corpus dans TXM

⁴ Respectivement, les discours analysés se trouvent aux adresses <http://fabienpoulard.info/> (Nicolas Sarkozy), <http://www.elysee.fr/> (François Hollande) et <http://jose-bove.eu/> (José Bové).

Sans surprise aucune, et ce à l'instar de n'importe quel texte, les éléments les plus récurrents sont des « petits mots », des propositions, de la ponctuation, des déterminants... Statistiquement, ces mots dits « grammaticaux », ou « relationnels », sont les plus représentés dans n'importe quel type de production langagière : nous passons plus d'énergie et de temps à déterminer comment s'élaborer nos réseaux de sens qu'à faire proprement sens. Ces éléments du discours n'ont, aux yeux de la sémantique, « aucune signification » : ils se bornent à construire des syntagmes ou à préciser comment les éléments de sens, que sont en général les noms, les verbes, les pronoms, les adjectifs, certains adverbes..., s'articulent les uns avec les autres.

Nous avons dû alors aller un peu plus précisément dans les textes, et nous présentons des statistiques relatives, et non absolues des éléments purement lexicaux employés par nos auteurs : effectivement, nous n'avons pu prendre en considération un élément vital de tout corpus, sa taille. Il est effectivement une chose de relever dix occurrences d'un mot dans un texte de 500 mots, mais cela n'est absolument pas comparable aux dix occurrences du même mot dans un discours de 5000 mots. Une étude plus précise aurait dû faire un « échantillonnage » des corpus, c'est-à-dire considérer des textes de même taille. Ici, et par contrainte de temps surtout, nous n'avons pas été capable de faire les choses comme il le faudrait, ce qui rajoute au problème de l'interprétation de nos résultats. Ces statistiques relatives permettent néanmoins de tracer quelques lignes de force concernant les éléments de sens utilisés par les locuteurs, que nous présentons ici (Figure 2).

	José Bové	François Hollande	Nicolas Sarkozy
Europe	1	5	6
OGM	2		
Agriculture	3	3	3
Nous	4	1	1
Doit (verbe)	5		
Marché	6	11	
Paysans	7		
Produits	8	6	
Alimentaire	9		
Agricole	10	9	
Pays	11		
France		2	2
Européen(ne)		4	
Agriculteurs		7	
Territoire		8	
État		10	
Production		12	
Monde			4
Crise			5
Prix			7
Politique			8
Développement			9

Figure 2 – Mot les plus employés, par ordre de fréquence et par locuteur

Quelques éléments d'observation superficiels tout d'abord :

- Des trois locuteurs, Nicolas Sarkozy a le stock d'unités de sens le plus réduit. La sociolinguistique appelle ces éléments des « concepts opératoires » : ce sont des mots que nous utilisons en force et qui cristallisent notre propos. Plus nous utilisons de concepts opératoires, plus notre propos est complexe et nuancé, du moins, il en a les capacités ; moins nous en usons, plus notre propos se martèle et perd de sa nuance.

- Les trois locuteurs ont des concepts opératoires communs : « Europe », « nous » et « agriculture », mais leur représentation statistique diffère fortement. Il y a nécessairement des effets d'inférence : il est normal que José Bové, en tant que député européen, place l'Europe au centre de ses discours, alors que Nicolas Sarkozy et François Hollande place davantage la France au centre de leurs échanges puisque leurs discours sont avant tout des discours de politique nationale.

- Chaque locuteur utilise des concepts opératoires que n'emploient jamais les autres : *OGM, doit, paysans, alimentaires, pays* pour José Bové, *européen(ne), agriculteurs, territoire, état, production* pour François Hollande, *monde, crise, prix, politique, développement* pour Nicolas Sarkozy. Ces éléments propres à chacun permettent de tracer une carte politique de chaque locuteur, cartes qui s'apparenteraient ici à des idiolectes : nous pouvons considérer que ces termes spécifiques cristallisent les appréciations politiques du thème de la ruralité chez nos locuteurs. Pour les raisons que nous avons présentées cependant, nous ne commenterons point ces éléments en détail.

Il faut enfin veiller aux effets de sens spécifiques : chaque locuteur emploie ces termes dans des sens particuliers, dont il est difficile de rendre compte par cette étude. Quel sens est par exemple donné par un tel au mot *État*, quel sens pour le mot *prix* ou *politique* ? Nous rencontrons là un écueil bien connu des études textométriques, à qui l'on reproche souvent d'aplanir, ou de terrasser, la signification d'un texte qui n'est accessible que par l'intermédiaire d'une étude fine de ses éléments lexicaux. Nous avons cependant à notre disposition un autre outil susceptible de nous aiguiller dans notre analyse : l'étude des cotextes immédiats, ce que l'on appelle encore les pivots lexicaux.

III. Pivots lexicaux

Selon une approche structurelle de la langue, telle que proposée par Saussure et que nous reprenons ici, la valeur d'une unité linguistique dépend tant de ses propriétés spécifiques que de ses relations avec l'ensemble des unités avec lesquelles elle sera en contact. Nous pouvons illustrer cela en reprenant l'image, fameuse, du jeu d'échecs : dans un jeu d'échecs, une pièce comme le cavalier a des propriétés qui en fondent la spécificité vis-à-vis de toutes les autres pièces. On pourrait néanmoins remplacer ce cavalier par un jeton et lui attribuer les mêmes propriétés pour ne pas mettre en péril l'ensemble de la structure : son rôle se saisit alors également par opposition avec les autres éléments de l'échiquier. Pour ainsi dire, une pièce de jeu d'échec, au même titre qu'une unité linguistique, voit sa valeur définie autant par ce qu'elle est, que ce par ce qu'elle n'est pas selon une perspective structurelle.

Par exemple, un mot comme *Europe*, que nous avons relevé précédemment, nous donne une idée de l'orientation du discours d'un auteur : mais cette idée sera enrichie en considérant les éléments avec lesquels nous allons trouver le plus souvent l'unité linguistique. Nous avons observé ce faisant quels éléments étaient les plus fréquemment trouvés à la gauche et à la droite des unités lexicales relevées. Nous avons pris ces mots en position de *pivots* autour desquels viennent se greffer, dans l'environnement cotextuel immédiat (à une distance maximale de deux mots), d'autres éléments lexicaux. Nous les présentons ici par locuteur :

José Bové :

<i>Commission / Commissaire Parlement</i>	européen(ne)	
<i>Une/L'</i>	agriculture	<i>Biologique / de qualité / paysanne</i>
<i>L'UE</i>	doit	
	Marché(s)	<i>Mondial / mondiaux</i>
<i>Les</i>	paysans	
	Produits	<i>Agricole / alimentaire</i>

François Hollande :

	Nous	<i>Allons / avons besoin</i>
<i>La</i>	France	<i>A besoin</i>
<i>Notre</i>	Agriculture	<i>De qualité</i>
<i>Budget / Commission / Conseil / Niveau</i>	Européen(ne)	
	Territoires	<i>Ruraux</i>
<i>Orientation / Politique</i>	Agricole	<i>Commune</i>
	État	<i>doit / et le gouvernement</i>
	Marché	<i>Mondial / financier / unique</i>
	Production	<i>française</i>

Nicolas Sarkozy :

	Nous	<i>Allons</i>
	France	<i>Et en Europe</i>
	Agriculture	<i>Française / de production</i>
	Monde	<i>Rural</i>
	Crise	<i>Du secteur</i>
	Prix	<i>De son travail</i>
	Politique	<i>Agricole commune</i>
	Développement	<i>Durable</i>

Figure 3 – Pivots lexicaux des mots les plus employés, par locuteur

Ces pivots nous permettent d'orienter nos analyses et de préciser nos observations. Nous les répartissons en trois catégories : concernant la ruralité tout d'abord, les acteurs ruraux ensuite, les politiques rurales enfin.

III.1. La ruralité

Il est intéressant de noter que seul Nicolas Sarkozy conçoit la ruralité dans sa dimension nationale et européenne : les autres locuteurs considèrent soit l'un, soit l'autre. Il parle effectivement d'une *agriculture française*, de la *France* et de l'*Europe* et du *monde rural*. Nos autres locuteurs évoquent une agriculture, comme José Bové, politique, de qualité... ce

qui renvoie non pas au domaine de la ruralité à proprement parler, mais plutôt à la production et aux résultats de cette production. Deux visions distinctes de la ruralité semblent alors s'opposer dans les discours :

- D'un côté, Nicolas Sarkozy, qui considère dans son discours la ruralité comme un espace à part entière, déterminé nationalement et par ses propriétés essentielles.
- De l'autre, François Hollande et José Bové, qui considèrent la ruralité comme un espace fonctionnel, déterminé par son emploi et son rôle au sein de la population. Ce serait pour cela que ce qui est *française*, chez Hollande, c'est la production ; ce qui est *biologique*, c'est l'agriculture, pour Bové, soit la conséquence de l'exploitation de la ruralité. À nouveau, nous ne ferons pas d'autres commentaires politiques sur cette question.

III.2. Les acteurs ruraux

La question des *acteurs*, soit des référents animés humains qui font ou subissent des actions, permet d'agencer différemment nos groupes théoriques :

- D'un côté, José Bové, seul à parler des *paysans*.
- De l'autre, Nicolas Sarkozy et François Hollande, qui effacent parfaitement les acteurs ruraux au profit de termes plus génériques, comme la *production* ou l'*Agriculture*. En revanche, nous pouvons nous interroger sur le pronom *nous* qu'emploient ces hommes politiques. Il semblerait qu'il ne s'agisse point ici de *nous* inclusifs, prenant en compte les agriculteurs et les paysans : il s'agit davantage d'un *nous* exclusif, ignorant les principaux concernés et les opposant à la classe politique, c'est-à-dire « l'État et le gouvernement » (Hollande), ou « la France et l'Europe » (Sarkozy).

En revanche, nous remarquons que Nicolas Sarkozy est le seul à parler du *prix de son travail*, sous-entendu *le prix du travail d'un agriculteur*. Il inclut donc certes un référent animé, mais indirectement, par l'intermédiaire du déterminant possessif et dans une structure du type « complément du nom ». Cela repousse par conséquent l'existence de ces référents dans une structure secondaire, induite et non centrale dans l'interprétation.

III.3. Les politiques rurales

Les politiques rurales, à nouveau, nous permettent d'opposer une fois encore Nicolas Sarkozy à nos deux autres acteurs.

- José Bové et François Hollande évoquent, évidemment, la PAC (*Politique Agricole Commune*), de même que les *marchés mondiaux*. Nous sommes bien dans une logique de marché.
- Nicolas Sarkozy en revanche, évoque le *développement rural* et la *crise du secteur*, sans évoquer les autres instances européennes. C'est une conception intéressante et, encore une fois, assez nationaliste du secteur agricole.

Conclusion

Ce tour d'horizon, bref et incomplet, a permis de présenter les avantages d'une étude textométrique pour l'analyse des discours contemporains. C'est là un domaine de la linguistique très prometteur, et qui est fréquemment employé par les chercheurs aujourd'hui pour analyser les discours politiques et médiatiques. Elle permet notamment d'estimer les sous- et sur-représentations de telle idée ou courant politique et idéologique, de comprendre quelles sont les personnalités les plus fréquemment invitées à s'exprimer. Ces données ne sauraient en revanche rien dire du fond de leurs discours : mais nous savons également que plus un discours est répété, plus il a de chances de devenir un discours normatif autour duquel les voix autres apparaîtront dissonantes, voire dérangeantes. Si les études textométriques, à l'instar de toute entreprise scientifique, n'ont pas vocation à devenir politique, elles peuvent en revanche devenir des arguments pour servir une cause politique et ce quelle que soit l'ambition des chercheurs : comme toujours concernant les chiffres, il convient de les manier avec précaution, de les recontextualiser et de les présenter avec prudence et nuance, sous peine de distordre la réalité des faits.

Bibliographie indicative

I. Revue et ouvrage collectif

Mots. Mots, ordinateurs, textes, sociétés : travaux de lexicométrie et de lexicologie politique. Revue fondée en 1980. Lyon : éditions de l'ENS Lyon.

(1966) *Statistique et analyse linguistique : colloque de Strasbourg (20-22 avril 1964).* Paris : Presses Universitaires de France.

II. Ouvrages spécialisés

BAYLON, C. *et al.* (1996). *Sociolinguistique : société, langue et discours.* Paris : Nathan.

LAFON, P. (1984). *Dépouillements et statistiques en lexicométrie.* Paris : Champion.

LEBART, L. *et al.* (1988). *Analyse statistique des données textuelles : questions ouvertes et lexicométrie.* Montreuil : Gauthier-Villars.

Pour une introduction plus large à la linguistique, nous renvoyons à l'ouvrage de Jean PERROT (2010), *La Linguistique*, dans la collection « Que sais-je ? ». Paris : Presses Universitaires de France.