

Dominique LABBE  
**LEXICOMETRIE  
ET  
ANALYSE DU DISCOURS POLITIQUE**

Publications et résumé  
Classement par ordre chronologique  
(novembre 2002)

**I - OUVRAGES**

- *Le discours communiste*, Paris, Presses de la F.N.S.P., 1977, 204 p.  
Louis Althusser, Georges Marchais, le paysan rouge du Huelgoat et 400.000 autres français ont plus en commun que le simple fait d'avoir en poche la carte du PCF et de voter régulièrement pour les candidats qu'il présente. Ce "plus" réside en particulier dans un même discours sur le monde. Suivant la culture propre à l'émetteur, la nature du destinataire et le lieu où il est produit, ce discours sera plus ou moins riche, nuancé et savant, mais il s'y trouvera toujours à l'œuvre une seule thématique que ce livre décrit dans sa structure fondamentale, à l'aide de la linguistique, ainsi que sous l'angle de ses capacités d'adaptation et de transformation. Cette étude conduit naturellement à s'interroger sur la nature du discours communiste : idéologie dominante ou idées révolutionnaires ?
  
- *François Mitterrand : essai sur le discours*, Grenoble, La Pensée Sauvage, 1983, 191 p.  
Peu de personnalités sont aussi difficiles à cerner que celle de François Mitterrand. Il est saisi ici à travers ses oeuvres et ses déclarations publiques auxquelles ont été appliquées les méthodes et les instruments de l'analyse du discours. La description de son vocabulaire, de ses métaphores et de son style démontre progressivement la manière dont François Mitterrand use du pouvoir des mots pour accomplir la destinée peu ordinaire qu'il s'est choisie et révèle la façon dont il se voit lui-même et dont il envisage le gouvernement de la France.
  
- En collaboration avec : SERANT (Daniel), THOIRON (Philippe), *Etudes sur la richesse et la structure lexicales*, Paris-Genève, Champion-Slatkine, avril 1988, 172 p.  
La richesse lexicale d'un texte est un concept souvent utilisé mais pour lequel il n'existe pas encore d'indice de mesure unique. Toute avancée dans les domaines de la richesse et de la structure du lexique ne peut être que le fruit d'une collaboration étroite entre la linguistique et les mathématiques. Ce volume réunit des travaux, en anglais et en français, de spécialistes des études lexico-statistiques, les uns mathématiciens, les autres linguistes. Par le truchement de simulations et d'applications portant sur des textes anglais, danois ou français, sont affinées les méthodes classiques d'investigation et sont élaborées de nouvelles procédures de description et de modélisation du vocabulaire et du lexique.
  
- *Le vocabulaire de François Mitterrand*, Paris, Presses de la Fondation Nationale des Sciences Politiques, mars 1990, 326 p.  
L'analyse porte sur les 305.124 mots prononcés lors des 68 interventions radio-télévisées du premier septennat de François Mitterrand. Le président se singularise essentiellement par un excédent de substantifs et de verbes exprimant la volonté. Son lexique s'organise autour d'un pivot: "je suis le président" auquel s'associent "Nous, la France" puis "Moi et les Français". Une place restreinte est accordée aux autres acteurs politiques et aux problèmes économiques et sociaux. La mesure de la richesse du vocabulaire montre que le président choisit avec soin ses mots (il dépasse ses rivaux sur ce point). En revanche, il spécialise peu ses propos et montre une inclination pour les généralités. Des tests statistiques permettent de

découper quatre périodes dont on analyse le vocabulaire caractéristique : "L'ère des réformes" (mai 1981-mai 1983), "L'effort pour la modernisation" (mai 1983-janvier 1985), "Majorité contre opposition" (avril 1985-octobre 1986), "Le président et le premier ministre" (octobre 1986-mars 1988). Enfin l'étude du style du président révèle ses figures rhétoriques favorites, et une construction de phrase très particulière. A la fin de l'ouvrage, les mots sont classés dans un index où sont mentionnées leur fréquence, leur première occurrence et leur répartition dans le discours du septennat.

- *Normes de saisie et de dépouillement des textes politiques, Cahier du CERAT n° 7, Grenoble : CERAT-I.E.P, avril 1990, 135 p.*

Présentation des procédures de saisie et de dépouillement des textes politiques en vue de leur analyse lexicométrique. La première partie présente la norme de Saint-Cloud qui unifie la saisie et le traitement des "types" ou "formes graphiques". Elle édicte également les règles de reconnaissance des mots composés et des locutions figées. La seconde partie présente la "norme Muller" régissant la lemmatisation. Cette opération consiste à ajouter à chaque occurrence du texte une forme canonique et un code grammatical à la manière d'une entrée de dictionnaire. Le rapport présente les règles de résolution des homographies du groupe verbal, du nom et des mots invariables. Dans ces différentes opérations, l'opérateur est assisté par l'ordinateur grâce à une série de programmes informatiques. En annexes, le rapport comporte des tables récapitulatives (mots composés et locutions, désinences du verbe, homographies des participes, des autres formes verbales, des substantifs...); une présentation de la méthode de calcul de l'indice de répartition utilisé pour mesurer la régularité d'apparition d'un mot dans un corpus ; un index des principales homographies traitées dans l'ouvrage.

- En collaboration avec BRUGIDOU (Mathieu), *Le discours syndical français contemporain (CFDT, CGT, FO en 1996-98)*, Paris, EDF - Division Recherche et Développement, 2000, 150 p.

Présentation des résultats d'une analyse de discours réalisée sur un corpus d'éditoriaux de la presse syndicale des confédérations françaises (CGT, CFDT et FO) et des deux principales fédérations de l'énergie (CGT et CFDT) en 1996 et 1998. Deux types d'approches ont été utilisées : la statistique lexicale, telle qu'elle a été développée par C. Muller et ses disciples, et l'analyse des données textuelles. On cherche expérimentalement, sur un corpus de textes, à dégager les convergences dans les résultats et à préciser les spécificités de chaque approche à l'aide de deux logiciels : *Alceste* de M. Reinert et *Lexicométrie* de D. Labbé). L'analyse des données textuelles propose une approche essentiellement exploratoire en mettant en lumière la structure des données. La statistique lexicale permet d'approfondir et d'enrichir les hypothèses issues de la première analyse et de les vérifier empiriquement.

- En collaboration avec MONIERE (Denis), *Le vocabulaire gouvernemental. Canada, Québec, France (1945-2000)*, à paraître chez Champion en 2003.

Les déclarations des chefs de gouvernement, depuis plus d'un demi-siècle, dans trois grandes démocraties de langue française (Canada, France, Québec) ont été passées au crible de l'analyse lexicométrique. En Amérique du nord, surtout au Québec, les idéologies apparaissent assez nettement malgré le moule de la tradition parlementaire. A partir de la fin des années 1970, l'ancienne opposition entre libéraux et conservateurs s'efface alors qu'apparaît un nouveau clivage entre fédéralistes et indépendantistes. En France, il n'y a pas d'écart significatif entre les discours de gauche et de droite, ni entre la IVe et la Ve République. La seule différence tient à la solidité du gouvernement : en position de force, les chefs présentent des programmes ambitieux ; ils tiennent des discours plus modestes quand ils sont en difficulté. Enfin, la comparaison entre les trois pays révèle que les discours gouvernementaux tendent à se ressembler de plus en plus, gommant ainsi les diversités nationales et institutionnelles.

## II - ARTICLES et COMMUNICATIONS depuis 1981

- "Moi et l'autre. Le débat Giscard d'Estaing—Mitterrand", *Revue Française de science politique*, XXXI-5-6, Octobre-décembre 1981, p. 951-981.  
La comparaison des interventions de MM. Giscard d'Estaing et Mitterrand, lors du débat qui les opposa le 5 mai 1981, révèle deux «énonciations» très différentes. L'analyse porte sur les pronoms, la structure actantielle, le temps et la «modalisation» du discours. Elle fait apparaître les présuppositions des deux adversaires et permet de comprendre quelles furent leurs stratégies de persuasion respectives.
- "Le discours de la CGT", *Que faire aujourd'hui*, n° 19, mai 1982, p. 8-11.  
Compte-rendu d'une analyse lexicographique menée sur des textes de la CGT (1979-1981). A partir du vocabulaire de la Confédération se dessine une certaine vision du monde, des rapports sociaux et du rôle qu'y joue l'organisation.
- "Structure de l'idéologie communiste : le cas du parti communiste français", *ECPR (European Consortium of political research)*, Freiburg, mars 1983, 20 p.  
Modèle de description de l'idéologie à partir du discours : le cas du PCF de 1961 à 1982. Contenu du code idéologique ; mécanismes d'actualisation. Thématique et lexique communistes. Comment l'idéologie change et s'adapte. Facteurs d'expansion ou de sclérose.
- "Nous les communistes...", *Mots*, n°10, mars 1985, p. 133-146.  
Le dépouillement des résolutions politiques des congrès du PCF depuis 60 ans (1961-1979) dévoile le code se trouvant à la source du discours communiste : "Nous, les communistes, sommes le parti de la classe ouvrière". Suivant que l'accent sera mis sur l'un ou l'autre de ces termes, le visage du discours change : avec "parti", il est institutionnel et unitaire ; fusionnel et sectaire avec "nous" et "classe ouvrière". La thématique et la lexique des textes communistes sont également décrits dans leurs grandes lignes.
- "La France et moi, Jacques Chirac (Analyse de la déclaration de politique générale du 9 avril 1986)", *Rapport de recherche*, Grenoble, CERAT-I.E.P., avril 1986, 17 p.  
Le rapport présente les principaux résultats de traitements lexicographiques réalisés sur le discours prononcé par Jacques Chirac devant l'Assemblée nationale : richesse de vocabulaire, mots employés, thèmes abordés, acteurs en présence, rôle des verbes et des adverbes, qualité de la déclaration suivant les passages.
- "La France et moi", *Esprit*, n°7, juillet 1986, p. 101-103.  
Présentation résumée d'une analyse lexicale et stylistique de la déclaration de politique générale prononcée le 9 avril 1986 par Jacques Chirac devant l'Assemblée nationale.
- "Le Barre comme il se parle", *Libération*, n° 1753, janvier 1987, p. 9.  
Extraits d'une étude lexicologique et thématique menée sur certains discours de R. Barre (prononcés entre 1984 et 1986). L'article donne quelques aperçus du lexique et de la rhétorique de l'ancien premier ministre. On montre comment R. Barre glisse, dans un propos à tonalité pédagogique, de nombreux traits polémiques contre ses rivaux de droite. Par certains côtés, le discours de R. Barre est gaullien, par d'autres il se rattache à la tradition parlementaire et au radical socialisme.
- "Une mesure de la richesse du vocabulaire : l'indice de Gini", *Mots*, n° 15, octobre 1987, p.171-184.  
La lexicométrie recherche une expression quantitative simple et uniformisée de la "richesse lexicale". La plupart des indices utilisés reposent sur la mesure de la fréquence moyenne des formes - ou des vocables - employés dans un texte. On montre ici que cette mesure peut se révéler trompeuse car elle est fortement influencée par la forte dispersion des fréquences d'emploi. L'objet de cet article est de rappeler l'intérêt que présentent, pour l'étude et la comparaison des structures lexicales, les instruments proposés par C. Gini et M. Lorenz (en particulier la fonction de concentration de Gini). Une étude empirique en est faite ici par application au débat entre V. Giscard d'Estaing et F. Mitterrand (mai 1981). Elle révèle que

les indicateurs classiques surestiment probablement la "richesse de vocabulaire" de Mitterrand par rapport à celle de Giscard d'Estaing.

- En collaboration avec HUBERT (Pierre) "Un modèle de partition du vocabulaire", in LABBE (Dominique), SERANT (Daniel), THOIRON (Philippe), *Etudes sur la richesse et la structure lexicales*, Paris-Genève : Champion-Slatkine, avril 1988, p. 93-114.

On propose ici un modèle de description du vocabulaire employé dans un corpus ; il est partagé en deux groupes : un vocabulaire général et des vocabulaires locaux (ou "spécialisés") dont chacun est mobilisé dans une partie seulement du corpus. Les vocables généraux peuvent apparaître en n'importe quel point du texte et leur accroissement, en fonction de la taille du corpus, peut être estimé grâce à la formule de Muller. Dans le modèle, un paramètre de partition estime le poids relatif des deux vocabulaires : la valeur de ce paramètre donne donc une estimation de la spécialisation lexicale à l'oeuvre dans le corpus. Des applications de ce modèle sont conduites sur l'oeuvre de Racine et sur des débats télévisés (Giscard-Mitterrand et Chirac-Fabius). Le modèle de partition peut être également utilisé pour calculer l'accroissement du vocabulaire dans un corpus, pour y localiser des variations stylistiques ou pour comparer plusieurs textes du point de vue de leur "richesse de vocabulaire".

- En collaboration avec HUBERT (Pierre), "Note sur l'approximation de la loi hypergéométrique par la formule de Muller", in LABBE (Dominique), SERANT (Daniel), THOIRON (Philippe), *Etudes sur la richesse et la structure lexicales*, Paris-Genève : Champion-Slatkine, avril 1988, p. 77-91.

Le raisonnement part de l'estimation de la probabilité d'absence d'un vocable dans un échantillon exhaustif prélevé dans un corpus, connaissant la distribution des fréquences des vocables qui constituent ce corpus. C'est la formule qui a été proposée, il y a plus de vingt ans, par Charles Muller et qui est ici comparée avec la loi hypergéométrique. Deux applications sont examinées : le calcul de l'accroissement du vocabulaire dans des corpus et le prélèvement aléatoire d'un grand nombre d'échantillons exhaustifs sur ces corpus. On démontre ainsi, théoriquement et empiriquement, que la formule de Muller, représente une bonne approximation de la loi hypergéométrique. On montre également la nécessité d'associer aux valeurs calculées un écart type qui permettra d'estimer l'intervalle de confiance attachée aux valeurs obtenues grâce à cette formule de Muller.

- En collaboration avec HUBERT (Pierre), "A model of Vocabulary Partition", *Literary and Linguistic Computing*, Vol. 3, n° 4, 1988, p. 223-225.

Présentation en anglais du modèle de partition du vocabulaire qui permet de mesurer, dans un corpus, la part du vocabulaire général, utilisé quel que soit le thème, et celle du vocabulaire spécialisé (qui apparaît seulement dans une partie). Le paramètre de partition mesure le poids relatif de ces deux vocabulaires et donne une estimation de la spécialisation lexicale. Le modèle permet de décrire le style d'un auteur et de localiser les ruptures thématiques dans un corpus.

- "Compte rendu de PECHANSKI (Denis), *Et pourtant ils tournent. Vocabulaire et stratégie du PCF*", *Communisme*, 22-23, 1989-III, p 195-197.

Denis Pechanski applique les méthodes de la statistique textuelle aux éditoriaux de l'Humanité entre janvier 1934 et décembre 1936. Ce livre comporte des descriptions intéressantes des techniques d'analyse du discours et du vocabulaire communiste. Les interprétations historiques sont plus contestables.

- "Des réformes à la cohabitation. Les quatre périodes du premier septennat Mitterrand", *Mots*, n° 22, mars 1990, p. 62-78.

Au cours de son premier septennat F. Mitterrand est intervenu 68 fois à la radio ou à la télévision. Dans ce corpus, trois ruptures majeures apparaissent, localisées grâce aux fluctuations dans l'apparition des mots nouveaux. Quatre périodes sont délimitées. Le

septennat s'ouvre avec "L'ère des réformes" (1981-1985). Puis vient "Le temps de l'effort et de la modernisation" (1983-1985) auquel succède la lutte de "La majorité contre l'opposition" (1985-1986). Enfin, le septennat se clôt sur la rivalité entre "Le président et le premier ministre" (1986-1988).

- "Compte rendu de PAPADOPOULOS (Ionnis), *Dynamique du discours politique et conquête du pouvoir. Le cas du PASOK : 1974-1981*", *Mots*, n° 22, mars 1990, p. 122-124.  
La thèse est nourrie par une très bonne connaissance de la vie politique grecque. L'ouvrage montre clairement que le principal ressort de l'idéologie du PASOK réside dans un "nationalisme défensif". L'analyse du discours est parfois moins convaincante.
- "Bibliographie des études en langue française sur le «discours socialiste»", *Mots*, n° 22, mars 1990, p. 105-106.
- En collaboration avec HUBERT (Pierre), "La répartition des mots dans le vocabulaire présidentiel", *Mots*, n° 22, mars 1990, p. 80-88.  
Les mots employés par F. Mitterrand lors de son premier septennat sont analysés sous l'angle de leur répartition, c'est-à-dire de leur localisation dans les interventions du président. L'indice de répartition met en évidence deux vocabulaires caractéristiques. Le vocabulaire habituel, employé quelles que soient les circonstances, éclaire le rôle du président selon F. Mitterrand. Les mots qui appartiennent au vocabulaire circonstanciel sont localisés en certains points du corpus. Ils révèlent l'impact de certaines crises ou de préoccupations importantes mais passagères.
- "Un modèle d'analyse du vocabulaire", *Communication aux secondes journées d'analyse de données textuelles*, Montpellier, 21-22 octobre 1993, 12 p.  
Présentation du "modèle de partition du vocabulaire" et exemple d'application pour la recherche des ruptures thématique dans les discours du général de Gaulle (1958-1969). Le modèle permet ainsi un découpage non-arbitraire des corpus en parties.
- "Compte-rendu de MONIERE (Denis), *Le combat des chefs*", *Mots*, n° 37, décembre 1993, p 111-115.  
Le livre analyse les débats télévisés opposant les trois principaux leaders lors des élections au Québec en 1962 et pour l'assemblée fédérale en 1968, 1979, 1984 et 1988 : étude du vocabulaire, des thèmes, de l'énonciation, de la gestuelle des participants. Ainsi se dévoilent les stratégies de persuasion et les personnalités des orateurs.
- En collaboration avec HUBERT (Pierre), "La richesse du vocabulaire", *Communication au Colloque de l'ALLC-ACH*, Paris, 19-23 avril 1994.  
La "richesse du vocabulaire" est analysée grâce à trois indicateurs : la diversité, la spécialisation et l'originalité. Le "modèle de partition du vocabulaire" permet de mesurer les deux premières dimensions. La communication présente une application de ces calculs aux discours du général de Gaulle (1958-1969). Le modèle est aisé à programmer et apporte des dimensions nouvelles à la statistique lexicale.
- "Déportation : les difficultés du témoignage", *Communication au Congrès international des femmes déportées*, Turin, 20-21 octobre 1994. Publié dans MONACO Lucio (dir), *La deportazione femminile nei lager nazisti*, Turin, FrancoAngeli, 1995, p 47-61.  
Analyse thématique de récits de femmes déportées : Buber-Neuman, Delbo, Heftler, Maurel, Millu, Tillion, Toulouse-Lautrec... Seuls les témoignages dotés de qualités littéraires évidentes remplissent leur fonction et pourront lutter efficacement contre l'oubli.
- En collaboration avec LABBE (Cyril), *Que mesure la spécificité du vocabulaire ?*, Grenoble, CERAT, décembre 1994 et juin 1997. Publié dans *Lexicométrica*, 3, 2001.

Cette note comporte une présentation de la formule de P.Lafon qui est utilisée habituellement pour l'étude des spécificités du vocabulaire d'un corpus découpé en sous-ensembles, puis elle analyse le comportement des résultats en fonction de la fréquence et de la taille des parties. Enfin, la formule est appliquée à un corpus de textes politiques contemporains. Il apparaît que les résultats sont influencés par la fréquence des mots et la taille des parties. En conclusion, on préconise certaines précautions dans l'utilisation de la formule et dans la présentation des résultats.

- "Les métaphores du général de Gaulle", *Mots*, 43, juin 1995.

Un relevé systématique des métaphores employées par le général de Gaulle révèle que ses images appartiennent à deux registres. L'image de l'élévation qui le place au sommet de l'Etat ; la mer et la navigation qui suggèrent une vision pessimiste de l'histoire et une conception très personnelle de l'autorité politique.

- En Collaboration avec HUBERT (Pierre), "La structure du vocabulaire du général de Gaulle" (communication aux 3e journées internationales d'analyse des données textuelles, Rome : 11-13 décembre 1995), in BOLASCO (Sergio), LEBART (Ludovic), SALEM (André), *III Giornate internazionali di Analisi Statistica dei Dati Testuali*, Rome, Centro d'Informazione e stampa Universitaria, 1995, tome II, p 165-176.

Description de la structure du vocabulaire d'un corpus à partir de ses principaux "univers lexicaux". Les liens entre les mots sont calculés grâce à la loi hypergéométrique, ce qui permet de déterminer si leurs cooccurrences sont statistiquement significatives. Le calcul a été appliqué aux discours télévisés et aux conférences de presse du général de Gaulle entre mai 1958 et avril 1969.

- En collaboration avec PIBAROT (André) et PICARD (Jacques), "Un outil de statistique textuelle : le lemmatiseur", *Travaux scientifique du Service de Santé des Armées*, XVI, 1995, p 305-307.

Le lemmatiseur est une suite de programmes permettant la mise à la norme des textes et le codage grammatical des mots. Son portage sur plate-forme DOS/Windows autorise la recherche de thèmes dans des corpus importants. Une application sur des questions ouvertes concernant une enquête psychosociologique sur le moral des armées est en cours.

- "Le «nous» du général de Gaulle", Communication au colloque *La comunicazione politica : aspetti socio-linguistici e pragmatici*, Rome, Université La Sapienza, 9-10 mai 1997, 16 p. Publié dans *Quaderni di studi linguistici*, 4/5, 1998, p 331-354.

On commence par rappeler le statut des pronoms personnels dans la langue et les sens possibles de "nous" d'après la théorie standard. Puis la notion d'«univers lexical» est présentée et appliquée aux allocutions radiotélévisées du général de Gaulle entre mai 1958 et avril 1969. L'univers du "nous" recouvre essentiellement les questions économiques et sociales ainsi qu'une partie des relations internationales. En revanche, le première personne du pluriel est exclue du jeu politique qui est le domaine des pronoms "je" et "vous". En définitive, de Gaulle s'adressait aux Français alternativement sur le mode de l'interpellation et sur celui de l'inclusion. En annexe, tableaux présentant les univers lexicaux des pronoms : "je", "nous", "vous" et de "France".

- En collaboration avec HUBERT (Pierre), "Vocabulary Richness", *Lexicometrica*, numero O, hiver 1997-98.

A model for analysis of the vocabulary of a corpus. This vocabulary is divided into two groups. First, the author uses the same general words whatever the circumstances. Second, several specialised vocabularies are used in only one part of the corpus. General words may appear everywhere in the text : their increase with the corpus' size can be estimated with Muller's formula. On the contrary, specialised vocabularies grow proportionally according to the corpus' size. We calculate the relative importance of the two vocabularies. This calculus gives an estimation of the lexical 'specialisation' in the text.

- En collaboration avec HUBERT (Pierre), "La connexion des vocabulaires" in MELLET (Sylvie), *IVe journées internationales d'analyse statistique des données textuelles*, Université de Nice-Sophia Antipolis, Nice, février 1998, p 361-369.

La connexion lexicale mesure la proximité ou l'écart existant entre les vocabulaires de plusieurs textes. On calcule d'abord le nombre théorique de mots que ces textes devraient avoir en commun et en propre s'ils appartenaient à la même oeuvre. Puis l'on compte les mots propres à chaque texte. L'indice de connexion des textes est le rapport entre le nombre théorique et les effectifs réellement observés. Appliqué aux tragédies de Corneille et Racine, le calcul montre que — sauf pour les deux dernières pièces de Racine — le vocabulaire des deux auteurs est très proche. Le décompte des "mots absents" — sans tenir compte de leur fréquence — n'est probablement pas une technique fiable pour l'attribution des textes dont l'auteur est inconnu ou douteux.

- En collaboration avec PIBAROT (André) et PICARD (Jacques), "Les syntagmes répétés dans l'analyse des commentaires libres" in MELLET (Sylvie), *IVe journées internationales d'analyse statistique des données textuelles*, Université de Nice-Sophia Antipolis, Nice, février 1998, p 507-515.

Méthode d'exploitation des réponses aux questions ouvertes dans les enquêtes sociologiques. Chaque mot reçoit un lemme et une catégorie grammaticale (lemmatisation), puis un programme extrait les groupes nominaux et verbaux significatifs en neutralisant les variations dans les adverbes, articles, prépositions, pronoms et conjonctions. Cette technique permet d'extraire les principaux thèmes développés par les enquêtés. Deux enquêtes sont présentées portant sur la santé au travail et la restauration d'entreprise.

- En collaboration avec FABRE (Cécile), HABERT (Benoît), "La polysémie dans la langue générale et les discours spécialisés", *Sémiotiques*, 13, décembre 1997, p 15-30.

Analyse des contextes d'emploi des substantifs et des adjectifs dans deux corpus. Le vocabulaire spécialisé est étudié dans un recueil de textes médicaux portant sur les maladies coronariennes (Menelas) ; la langue générale à travers les interventions radio-télévisées du premier septennat de François Mitterrand. L'univocité conceptuelle du langage spécialisé s'oppose à la polysémie massive de la langue générale.

- "La France chez de Gaulle et Mitterrand", in FIALA (Pierre), LAFON (Pierre) (dir), *Des mots en liberté. Mélanges Maurice Tournier*, Fontenay-aux-Roses, ENS Editions, 1998, p 183-193.

"France" est le substantif le plus employé dans les discours présidentiels chez de Gaulle comme chez Mitterrand. Un test statistique permet de comparer les contextes dans lesquels ce mot est employé. Les deux hommes sont d'accord pour réserver l'essentiel des emplois de France à la politique étrangère mais, pour de Gaulle, il s'agit d'aide, de coopération, d'amitié alors que chez Mitterrand, la diplomatie, la défense nucléaire et les questions militaires dominent le discours.

- "La richesse du vocabulaire politique : de Gaulle et Mitterrand", in MELLET (Sylvie), VUILLAUME (Marcel), *Mots chiffrés et déchiffrés. Mélanges offerts à Etienne Brunet*, Paris, Champion, 1998, p 173-186

La mesure de la "richesse du vocabulaire" chez E Brunet. Application aux allocutions radio-télévisées de C. de Gaulle et F. Mitterrand. On propose de scinder la notion en deux dimensions : la diversité du vocabulaire et sa spécialisation. Les mesures confirment les valeurs obtenues avec l'indice de Brunet tout en les affinant. On peut alors isoler l'oral et l'écrit et opposer la préparation soignée et l'unité thématique chez de Gaulle au style oral et l'adaptation à l'événement chez Mitterrand.

- "La recherche de l'information dans les textes", *Séminaire de l'Institut Catalan de Statistique*, Barcelone, 13 juin 1999, 18 p.

Les méthodes statistiques employées pour l'étude des "données textuelles" sont extraordinairement diverses dans leur origine et leurs objectifs. Mais elles partagent une préoccupation commune qui intéresse toutes les sciences sociales : le contenu des discours, la recherche du sens. En effet, les mots forment le principal matériel sur lequel travaillent les sociologues ou les politistes : transcriptions d'entretiens, de discours, articles, livres, groupes de textes... Nous avons choisi de montrer l'intérêt de ces recherches à l'aide d'un exemple : la comparaison du discours des deux hommes qui ont marqué l'histoire politique de la France au cours de ce dernier demi-siècle : le général de Gaulle et F. Mitterrand.

- "Compte rendu de VILLONE (Massimo) et ZULIANI (Alberto) (dir), *L'attività dei governi della Repubblica italiana (1948-1994)*", *Mots*, 62, mars 2000, p 117-119.  
Remarquable exemple de coopération interdisciplinaire pour la constitution d'une base de données sur l'activité gouvernementale en Italie depuis la fondation de la République : composition du parlement et des gouvernements, programmes des partis, discours d'investiture, lois de finances, délibérations des conseils des ministres. L'analyse des textes met en valeur le passage d'un vocabulaire sobre et courant à un lexique de plus en plus technique voire bureaucratique.
- "Compte rendu de MONIERE (Denis), *Démocratie médiatique et représentation politique*", *Mots*, 62, mars 2000, p 121-122.  
Analyse de contenu sur les bulletins d'information de quatre chaînes de télévision francophones (Belgique, Canada, France, Suisse) pendant 14 semaines. Il s'agit de l'unique analyse empirique, de qualité et de vaste ampleur, sur l'information télévisée.
- "*Analyse des données textuelles et Statistique lexicale (Textual Data Analysis and Lexical Statistics)*", conférence introductive aux 5<sup>e</sup> journées internationales d'analyse des données textuelles, Lausanne, Ecole polytechnique fédérale, 2000, reproduite dans *Lexicometrica*, 4, 2002.  
Cette conférence plaide pour des données textuelles de qualité, normalisées et étiquetées. Elle illustre leur utilité à l'aide d'un exemple : le sens du mot "amour" dans l'oeuvre de Corneille. La technique de l'étiquetage est présentée. Enfin, on évoque la nécessaire coopération entre les chercheurs pour la réalisation des outils de normalisation et d'étiquetage et pour la constitution de corpus de référence.
- En collaboration avec BRUGIDOU (Mathieu), « Le vocabulaire syndical français à la lumière de l'analyse des données textuelles et de la statistique lexicale. », RAJMAN (Martin) et CHAPPELIER (Jean-Cédric) (eds), *Actes des 5<sup>e</sup> journées internationales d'analyse des données textuelles*, Lausanne, Ecole polytechnique fédérale, 2000, vol 1, p 85-94.  
Analyse de discours réalisée sur un corpus d'éditoriaux de la presse syndicale confédérale des trois principales centrales françaises (CGT, CFDT et FO) en 1996 et 1998. Deux approches ont été privilégiées : la statistique lexicale telle qu'elle a été développée par C. Muller et ses disciples et l'analyse des données textuelles. On cherche expérimentalement sur un corpus de textes à dégager les convergences dans les résultats produits et à préciser les spécificités de chaque approche. Ces analyses sont réalisées grâce à différents logiciels (*Alceste* de M. Reinert et *Lexicométrie* de D. Labbé). On observe des convergences réelles entre les deux types de méthodes. L'analyse des données textuelles propose une approche essentiellement exploratoire en mettant en lumière la structure des données. La statistique lexicale permet de d'approfondir et d'enrichir les hypothèses interprétatives issues de la première analyse et de mieux les vérifier empiriquement.

- En collaboration avec MONIERE (Denis), « La connexion intertextuelle. Application au discours gouvernemental québécois », RAJMAN (Martin) et CHAPPELIER (Jean-Cédric) (eds), *Actes des 5<sup>e</sup> journées internationales d'analyse des données textuelles*, Lausanne, Ecole polytechnique fédérale, 2000, vol 1, p 85-94.

La connexion intertextuelle mesure la distance entre les vocabulaires de plusieurs textes. Pour chacun des mots, on calcule la différence entre une fréquence théorique et la fréquence observée. L'indice est insensible aux différences de longueur entre les textes. Il est appliqué aux discours prononcés par les Premiers ministres québécois pour ouvrir les sessions parlementaires depuis 1945. Appliquée à ces données, la classification automatique met en valeur quelques grands épisodes dans la vie politique de la province et souligne la singularité des deux passages au pouvoir du parti québécois (1977-84 et 1996-).

- En collaboration avec BERGERON (Jean-Guy), "L'évaluation de la négociation raisonnée par les acteurs. Une analyse lexicométrique", *Communication au XVI<sup>e</sup> Congrès international de l'Association internationale des sociologues de langue française*, Québec, juillet 2000, 12 p.

La négociation collective raisonnée entre employeurs et syndicats a rencontré une audience importante au Québec au cours des années 1990. Comment les acteurs ont-ils utilisé la méthode et comment évaluent-ils son utilité ? Cette communication présente une méthode originale pour répondre à ces questions : la statistique lexicale appliquée aux entretiens réalisés auprès d'un pannel de parties à cinq négociations (représentants des employeurs, syndicalistes et conciliateurs). Trois outils sont présentés : le calcul de la distance intertextuelle, la classification automatique et le calcul des vocabulaires spécifiques.

- En collaboration avec LABBE (Cyril), "Discrimination et classement au sein d'un groupe d'entretiens. Le cas du confort électrique", , *Journées d'études du CIDSP*, 9 mars 2001, 28 p.

Présentation du calcul de la distance intertextuelle et de deux méthodes de classification (classification hiérarchique ascendante, analyse arborée). Caractérisation du vocabulaire spécifique des différentes classes. Application à un groupe d'entretiens sur le confort électrique.

- "Normalisation et lemmatisation d'une question ouverte. Les femmes face au changement familial", *Traitement des questions ouvertes dans les enquêtes et sondages*, Journées d'études de la Société Française de Statistique, Grenoble, 8 juin 2001, 19 p. Reproduit dans *Journal de la Société Française de Statistique*, 142-4, décembre 2001.

La normalisation consiste à réduire les majuscules des noms communs, à uniformiser les orthographes multiples des noms propres, des dates et des chiffres ou de certains mots communs, à déployer les abréviations, etc. La lemmatisation associe à ces graphies normalisées un lemme correspondant à l'entrée du dictionnaire et une catégorie grammaticale. Ces tâches sont confiées à un automate dont l'efficacité est testée sur les réponses à une question ouverte dans une enquête sur les causes de divorce. Par rapport aux formes graphiques brutes, les données lemmatisées réduisent le nombre de mots différents et permettent de retrouver les principaux thèmes. Elles mettent également à jour certaines déformations produites par la manière dont les enquêteurs retranscrivent les réponses.

- En collaboration avec LABBE (Cyril), "Inter-Textual Distance and Authorship Attribution Corneille and Molière", *Journal of Quantitative Linguistics*, 8-3, December 2001, p 213-231.

The calculation proposed in this paper, measures neighborhood between several texts. It leads to a normalized metric and a distance scale which can be used for authorship attribution. An experiment is presented on one of the famous cases in French literature : Corneille and Molière. The calculation clearly makes the difference between the two works but it also demonstrates that Corneille contributed to many of Molière's masterpieces.

- En collaboration avec HUBERT (Pierre) et LABBE (Cyril), "Segmentation automatique des corpus. Voyages de l'autre côté de J.-M. Le Clézio" in MORIN (Annie) et SEBILLOT (Pascale) (eds), *VIe Journées Internationales d'Analyse des Données Textuelles (Saint-Malo 13-15 mars 2002)*, IRISA-INRIA, 2002, tome I, p 359-369.  
Méthode originale pour segmenter un corpus en sous-parties homogènes. On calcule l'accroissement du vocabulaire et les variations de sa diversité. Un algorithme de segmentation associé à un test de validité donne le découpage optimal des deux séries. Application à un roman de Jean-Marie Le Clezio : Voyages de l'autre côté.
  
- En collaboration avec LESELBAUM (Jean), "Lexicographie assistée par ordinateur. Signification de "Banque" dans le vocabulaire économique" in MORIN (Annie) et SEBILLOT (Pascale) (eds), *VIe Journées Internationales d'Analyse des Données Textuelles (Saint-Malo 13-15 mars 2002)*, IRISA-INRIA, 2002, tome II, p 447-456.  
Méthode pour définir les sens précis d'un mot dans un corpus ou chez un auteur. On recherche le vocabulaire associé à ce mot (univers lexical) puis tous les synonymes potentiels : vocables de même catégorie grammaticale et employés dans des contextes semblables. La méthode est illustrée avec le mot "banque" dans le vocabulaire économique et social français contemporain. Ce mot possède deux sens principaux : opérateur sur les marchés financiers et groupe financier.
  
- En collaboration avec MONIERE (Denis), "Essai de stylistique quantitative. Duplessis, Bourassa et Lévesque" in MORIN (Annie) et SEBILLOT (Pascale) (eds), *VIe Journées Internationales d'Analyse des Données Textuelles (Saint-Malo 13-15 mars 2002)*, IRISA-INRIA, 2002, tome II, p 561-569.  
Comparaison des discours inauguraux prononcés par les trois Premiers ministres qui ont le plus marqué l'histoire moderne du Québec: Maurice Duplessis, Robert Bourassa et René Lévesque. Afin d'analyser de façon systématique et comparative leur style discursif respectif, nous utilisons une série d'indices comme la richesse et la spécialisation du vocabulaire, la longueur et la structure des phrases, la densité des catégories grammaticales, ce qui permet de dégager les caractéristiques de chacun. Nous montrons qu'en dépit des fortes contraintes institutionnelles imposées par le cadre de ces discours, chaque Premier ministre laisse la marque de son style personnel.
  
- "La lemmatisation des grandes bases de textes. Un exemple : Corneille, Molière et Racine", Communication au colloque *L'édition électronique en littérature et dictionnaire, évaluation et bilan*, Rouen, 17-21 juin 2002, 19 p.  
Avec l'exemple des pièces de Corneille, Molière et Racine, on montre quelques-uns des nombreux usages possibles des bases de données textuelles normalisées et lemmatisées. Elles sont d'une consultation aisée. Elles fournissent de nombreux renseignements sur le vocabulaire, le style, le sens des mots... Pour cela, il faut réduire les graphies multiples et rattacher chaque mot à son entrée de dictionnaire.
  
- "Le général de Gaulle en campagne", Communication aux IIIe Journées de l'ERLA, *Aspects linguistiques du texte de propagande*, Brest, 15-16 novembre 2002, 16 p.  
Alors que le général de Gaulle préparait toujours avec soin ses allocutions radio-télévisées, il a réalisé trois entretiens sans préparation, entre les deux tours de l'élection présidentielle de 1965. Comparés à ses autres interventions, ces trois émissions font apparaître des caractéristiques lexicales et stylistiques remarquables inconnues par ailleurs dans les discours du Général : vocabulaire restreint où les verbes usuels sont privilégiés ; forte personnalisation et tension importante ; nombre anormal de phrases courtes avec de nombreuses interpellations, interrogations rhétoriques et dénégations. Il s'agit des principales caractéristiques du discours propagandiste.