Dominique LABBE

Dominique.labbe@umrpacte.fr
https://www.pacte-grenoble.fr/membres/dominique-labbe

LEXICOMETRIE
ET
ANALYSE DU DISCOURS

Publications et résumés

Classement par ordre chronologique inverse
(2019-1981)

La plupart de ces publications sont consultables aux Archives ouvertes du CNRS
(http://hal.archives-ouvertes.fr/)
Ou sur ResearchGate (https://www.researchgate.net/)

Depuis 1969 : statistique appliquée au langage, conception de logiciels et constitution de corpus.

Mise au point d'algorithmes de balisage des textes, de standardisation des graphies et d'étiquetage des mots (lemmatisation).

A l'aide de ces outils, constitution d'une bibliothèque électronique du français moderne (XVIIe-XXIe siècle) : actuellement 62 millions de mots étiquetés et indexés (discours politique, littérature, presse, français oral).

Principales recherches (et logiciels) utilisant la bibliothèque électronique :

- Le vocabulaire français (lexicographie assistée par ordinateur) ;
- La répartition des mots dans les grandes collections de texte. Au-delà de la fréquence d'emploi, la répartition offre une seconde dimension pour caractériser l'utilisation d'un vocable dans une oeuvre, un groupe humain, une langue.
- Vocabulaire caractéristique d'une œuvre, d'un auteur, d'un groupe social, d'une époque.

- Le sens des mots (univers lexicaux) chez un auteur, un groupe, une époque et dans la langue,
- Les combinaisons de mots les plus fréquentes avec une attention particulière à la modalité verbale,
- Localisation des ruptures thématiques et stylistiques dans un corpus,
- Distance intertextuelle, classification, automatique,
- Les principaux thèmes d'une œuvre, d'un auteur, d'un groupe social…
- Stylométrie (richesse et spécialisation du vocabulaire, figements) et étude de la phrase en fonction de sa longueur et de sa construction,
- Les genres littéraires. Qu'est-ce qu'un genre ? (lexique, syntaxe, stylistique).
- L'attribution d'auteur (authorship attribution) : détermination de l'auteur d'un texte d'origine douteuse ou inconnue.
- La détection des fraudes à la publication scientifique

Participation à un réseau informel de recherche, comprenant notamment : E. Arnold (Département de Français de Trinity College Dublin), M. Brugidou (EDF et laboratoire PACTE), P. Hubert (Université de Paris VI), C. Labbé (Laboratoire d'Informatique de Grenoble) - D. Monière (Université de Montréal), J. Savoy (Département d'Informatique de l'Université de Neuchatel).

I – OUVRAGES – RAPPORTS DE RECHERCHE

Une expérience d'attribution d'auteur. Le corpus Saint-Jean. Grenoble : Pacte, octobre 2017.

Avec la collaboration de J. Savoy, dans le but de tester les méthodes d'attribution d'auteur, il a été constitué un corpus de 200 extraits tirés de 68 romans par 31 auteurs différents. Les différences de vocabulaire entre les textes sont mesurées grâce à la distance intertextuelle. Tous les extraits sont correctement attribués avec la technique du plus proche voisin mais cette méthode exige que les auteurs aient au moins deux textes dans le corpus. En l'absence de cette condition, on utilise les plus petites distances, en définissant un intervalle de confiance. Cette méthode permet d'attribuer, sans erreur, 8 extraits sur 10. Deux classifications (hiérarchique et arborée) aboutissent aux mêmes résultats. Une échelle standardisée de la distance intertextuelle permet d'attribuer un texte de manière simple et sûre sans avoir à reprendre l'ensemble de la procédure.

En collaboration avec LABBE Cyril. *La répartition du vocabulaire*. Grenoble : Laboratoire d'informatique de Grenoble, septembre 2017.

La répartition d'un mot dans une collection de textes (corpus) est l'ensemble des emplacements où ce vocable apparaît. Cette dimension a été peu étudiée et uniquement pour des corpus constitués d'échantillons de longueurs égales. Cette note analyse le phénomène dans les corpus de textes entiers (longueurs inégales) et propose un indice dont les propriétés sont décrites à l'aide de plusieurs corpus de grandes dimensions. Une procédure simple permet d'isoler les vocables les plus régulièrement utilisés et ceux qui sont localisés en un point du corpus. Cette dimension complète la fréquence et apporte une information supplémentaire sur le vocabulaire du corpus.

En collaboration avec ARNOLD Edward, LABBE Cyril & MONIERE Denis. *Parler pour gouverner : Trois études sur le discours présidentiel français*. Grenoble : Laboratoire d'Informatique de Grenoble, 2016.

Trois études de statistique appliquée au discours présidentiel (vocabulaire, thèmes et style). Les campagnes pour les élections présidentielles de 2002, 2007 et 2012 indiquent une personnalisation, une tension et une agressivité croissante entre les candidats. L'analyse des débats télévisés entre les deux finalistes depuis 1974 aboutit aux mêmes conclusions et montre que la conjoncture politique du moment l'emporte sur le clivage droite-gauche et sur la personnalité des candidats. Le recensement de la communication des sept présidents depuis 1958 montre qu'ils parlent beaucoup et qu'ils privilégient l'allocution plutôt que les entretiens ou les conférences de presse.

En collaboration avec BASSON Jean-Charles. *Jean Racine. Aétius, Juba, Tachmas. Tragédies inédites transcrites et présentées par Jean-Charles Basson et Dominique Labbé*. Montréal : Monière-Wollank Editeurs, 2015.

Aétius, Juba et Tachmas sont les soeurs cadettes des sept tragédies présentées par Jean Racine entre 1667 (Andromaque) et 1677 (Phèdre). Conservés aux Archives départementales de la Haute-Garonne à Toulouse, les manuscrits de ces trois tragédies inédites sont transcrits en français contemporain et reproduits dans ce livre.

Ces trois tragédies n'avaient pas été reconnues car un lecteur, même érudit, ne peut pas répondre à la question : "qui a écrit ce texte ?" Or, au XVIIe siècle, plus de la moitié des pièces de théâtre ne sont pas présentées par les écrivains qui les ont écrites mais par des intermédiaires qui financent leur production et les négocient avec les troupes. L'examen de la vie de Jean

Racine et plusieurs témoignages de contemporains montrent qu'il continue à produire des pièces, après sa prétendue retraite de 1677, et qu'il est associé à deux prête-noms : Jean de La Chapelle puis Jean-Galbert Campistron.

Au total, quatorze tragédies viennent s'ajouter au corpus racinien. Plusieurs de ces pièces ont été de grands succès et méritent d'être redécouvertes.

Qui a écrit Juba, Aétius et Tachmas ? Une attribution d'auteur par ordinateur. Rapport technique. Grenoble : Pacte, décembre 2014.

Qui a composé *Aétius*, *Juba* et *Tachmas*, pièces inédites du fonds Maniban-Campistron aux Archives départementales de la Haute-Garonne à Toulouse ? La réponse est apportée grâce à une procédure d'attribution d'auteur par ordinateur. Un calcul de distance mesure précisément la plus ou moins grande ressemblance de ces textes par rapport à un vaste corpus de référence comportant 235 pièces contemporaines. Les textes séparés par les distances les plus faibles sont écrits par un même auteur. Des procédures de classification repèrent les groupements optimaux au sein de cette population et mesurent le degré d'appartenance de chaque texte à un groupe donné. Ces procédures permettent d'identifier l'auteur de ces trois manuscrits qui est aussi celui de toutes les tragédies représentées sous le nom de Jean-Galbert Campistron et de Jean de La Chapelle.

De nombreux indices lexicaux et stylistiques viennent confirmer cette attribution, notamment les longueurs, structures et fonctions des phrases qui caractérisent le style de cet auteur par rapport à ses contemporains. Enfin, une mesure de l'influence du temps permet d'offrir une datation de ces œuvres.

La lecture de ce texte ne demande aucune connaissance en mathématique et statistique.

En collaboration avec LABBE CYRIL. *Who wrote this scientific text?* Technical report. Grenoble : Laboratoire d'Informatique de Grenoble (LIG). September 2014.

The IEEE bibliographic database contains a number of proven duplications with indication of the original paper(s) copied. This corpus is used to test a method for the detection of hidden intertextuality (commonly named "plagiarism"). The intertextual distance, combined with the sliding window and with various classification techniques, identifies these duplications with a very low risk of error. These experiments also show that several factors blur the identity of the scientific author, including variable group authorship and the high levels of intertextuality accepted, and sometimes desired, in scientific papers on the same topic.

En collaboration avec MONIERE Denis. *La campagne présidentielle de 2012. Votez pour moi !* Paris : l'Harmattan (collection logiques politiques), 2013.

Du premier janvier au 4 mai 2012, la communication des cinq principaux candidats à l'élection présidentielle de 2012 (F. Bayrou, F. Hollande, M. Le Pen, J.-L. Mélenchon et N. Sarkozy) a été scrutée quotidiennement grâce à l'analyse de contenu et à la statistique appliquée au langage (lexicométrie). Au total 2 241 messages (discours, entretiens, communiqués…), soit 1,774 millions de mots, ont été dépouillés offrant ainsi une « radioscopie » de la communication politique française contemporaine : intensité et orientation des messages, mise en valeur de soi et critique des autres, thématique et style de chaque candidat. A l'exception de F. Bayrou qui a choisi l'explication, la communication des quatre autres candidats a été dominée par la critique du ou des principaux concurrents. Ce choix a entretenu une véritable "spirale de la négativité". L'élection présidentielle de 2012 confirmerait la prédominance, dans la communication politique contemporaine, de la polémique et du dénigrement, sur le contenu positif.

Si deux et deux sont quatre Molière n'a pas écrit Dom Juan. Paris : Max Milo, 2009 (Réédition revue et augmentée de *Qui a écrit Tartuffe ?* Montréal : Monière-Wollank, 2009).

Au XVIIe siècle, le théâtre français a connu une floraison exceptionnelle. A cette époque, la majorité

des pièces ont été présentées par des "comédiens poètes". Ces acteurs achetaient des textes aux écrivains et les négociaient avec les troupes. Ils les mettaient en scène puis, en cas de succès, ils les publiaient sous leurs noms. Le livre explique les raisons de ce système du comédien poète, prête-nom d'un grand auteur, et la manière dont les troupes de théâtre fonctionnaient. Molière était le plus célèbre de ces comédiens poètes mais aussi : favori du roi et riche financier. Il ne se comportait pas comme un écrivain. Ses contemporains le considéraient comme un comédien talentueux et non comme un homme de plume. Enfin, ce livre démontre que P. Corneille était la plume de l'ombre de Molière. Il a composé le *Misanthrope*, le *Tartuffe*, *Dom Juan*, *l'Avare*… et toutes les grandes pièces présentées sous le nom de Molière.

En collaboration avec MONIERE Denis. *Les mots qui nous gouvernent Le discours des premiers ministres québécois : 1960-2005*. Montréal : Monière-Wollank Editeurs, 2008. Ouvrage couronné par l'Assemblée nationale du Québec (premier prix de la présidence – journée du livre politique – 2009).

Application de la statistique lexicale aux discours des Premiers ministres Québécois entre 1960 et 2005 (Lesage, Johnson, Bertrand, Bourassa, Lévesque, Parizeau, Bouchard, Landry, Charest). On analyse d'abord l'évolution des thématiques qui balisent l'histoire de la société québécoise depuis la révolution tranquille. Le style discursif de chaque Premier ministre est décrit grâce au vocabulaire, aux catégories grammaticales, au maniement des verbes et des noms, à la longueur et la structure des phrases. Cette comparaison montre des différences significatives entre les Premiers ministres et révèle la stratégie de communication privilégiée par chacun d'eux.

Corneille dans l'ombre de Molière. Histoire d'une recherche. Bruxelles : Les impressions nouvelles, 2003.

Pierre Corneille a écrit les principales pièces de Molière, notamment : L'Ecole des femmes, Tartuffe, Dom Juan, Misanthrope, l'Avare, les Femmes savantes… La collaboration entre les deux hommes a duré pendant plus de 15 ans. Elle a abouti à la création de 18 pièces qui forment la première "comédie humaine" des Temps modernes. Molière l'a mise en scène, s'est battu pour elle et a permis qu'elle parvienne jusqu'à nous. Ce livre présente la solution d'une énigme scientifique : comment identifier l'auteur d'un texte inconnu ou d'origine douteuse ? Il expose enfin les bouleversements qu'apporteront les mathématiques appliquées et l'informatique dans les sciences humaines.

En collaboration avec MONIERE Denis. *Le vocabulaire gouvernemental. Canada, Québec, France (1945-2000).* Paris : Champion, 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.

Analyse des représentations du confort électrique à partir d'un corpus d'entretiens. Rapport pour le GREST-EDF. Grenoble : CERAT, juin 2002.

Analyse secondaire de 200 transcriptions d'entretiens — réalisés au cours des 8 dernières années, autour des usages de l'électricité et du confort lié à ces usages — comportant au total plus de 1,2 millions de mots. Tous ces textes ont été balisés, les graphies ont été normalisées puis étiquetées. Enfin, une série de traitements lexicométriques révèlent un partage entre les utilisateurs et les "prescripteurs" (agents EDF, installateurs, gestionnaires de HLM, travailleurs sociaux). Pour ces derniers le confort renvoie à des questions d'appareillage et d'abonnement : il est à la portée des usagers pourvu qu'ils fassent les bons choix techniques. Environ un dixième des usagers partagent cette vision des choses. En revanche, l'écrasante majorité met plutôt l'accent sur les contraintes engendrées par le système de tarification, leur poids sur les rythmes de vie, les exigences de confinement de l'habitation, les difficultés d'utilisation des gestionnaires. Une minorité importante place au premier plan le coût de cette énergie et la contrainte budgétaire.

Cette expérience démontre l'intérêt des grandes bases de données d'enquêtes et de leur exploitation secondaire. Mais ces bases doivent obéir à des règles strictes en matière de transcription, de balisage et d'étiquetage.

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.

Normes de saisie et de dépouillement des textes politiques. *Cahier du CERAT n° 7*. Grenoble : CERAT-IEP, 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.

Le vocabulaire de François Mitterrand. Paris : Presses de la Fondation Nationale des Sciences Politiques, 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.

En collaboration avec SERANT Daniel et THOIRON Philippe (direction d'ouvrage collectif). *Etudes sur la richesse et la structure lexicales*. Genève-Paris : Slatkine-Champion, 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.

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 se 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.

Le discours communiste. Paris : Presses de la Fondation Nationale des Sciences Politiques, 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 ?

II - ARTICLES et COMMUNICATIONS depuis 1981

(classement par ordre chronologique inverse)

Réponse à Florian Cafiero et Jean-Baptiste Camps. Why Molière most likely did write his plays. *Science Advances.* 5. 27 November 2019. Grenoble : PACTE, 27 novembre 2019.

Dans cet article, MM. Cafiero et Camps prétendent apporter la preuve que P. Corneille n'a écrit aucune des pièces présentées par Molière. Ils utilisent pour cela 6 "caractérisques" (lemmes, formes, mots outils, rimes, affixes, n-grams) couplées avec des classifications automatiques. En fait, les auteurs fournissent peu d'informations précises sur ces méthodes et aucune donnée chiffrée. Les quelques informations, notamment dans les annexes en ligne, suffisent pour soulever beaucoup de doutes. Par exemple, la liste des "mots outils" comporte de nombreuses étrangetés qui ne peuvent s'expliquer simplement par des maladresses. De même, ils ont opéré un tri dans les pièces de Molière, retirant de l'expérience 24 des 33 pièces. Parmi ces pièces écartées : *Psyché* qu'il ne fallait surtout pas retirer ! Enfin, le détail des classifications (publié dans une annexe en ligne séparée de l'article) montre un échec total. Leurs méthodes se révèlent incapables de reconnaître : Boursault, Chevalier, Dancourt, Donneau de Visé, Gillet de la Tessonnerie, Pierre Corneille, Thomas Corneille, La Fontaine, Ouville, Quinault, Régnard, Rotrou… et Molière.

En collaboration avec LABBE Cyril. *Le calcul du sens des mots. Les univers lexicaux*. Rapport de recherche. Grenoble : Laboratoire d'Informatique de Grenoble. mai 2019.

Procédure de recherche du sens d'un mot dans un grand corpus. Résultats obtenus sur le mot "Europe" dans les interventions du président Chirac (1995-2007) : caractéristiques principales de l'univers, densités des catégories grammaticales, vocables sur-employés et sous-employés avec *Europe* (classés par catégories grammaticales et par fréquence) ; vocabulaire commun ; phrases les plus caractéristiques. Publié en annexe de : Cyril Labbé et Dominique Labbé. Le sens des mots. L'Europe dans le vocabulaire de Jacques Chirac. *Document numérique*. 2019/1, volume 22, p. 31-61.

En collaboration avec LABBE Cyril. Le sens des mots. L'Europe dans le vocabulaire de Jacques Chirac. *Document numérique*. Volume 22, numéro 1-2, 2019, p.31-61.

Présentation du corpus des interventions de J. Chirac durant ses deux présidences (1995-2002 et 2002-2007) : poids des différents modes de communication, description du vocabulaire, vocables les plus utilisés. On constate une stabilité de la communication sur les 12 ans et un poids considérable de l'Europe dans les principaux thèmes. Mais quel sens donnait J. Chirac à ce mot ? Pour répondre à cette question, on reconstitue le réseau d'associations, d'oppositions et de substitutions qui relie ce mot aux autres vocables du corpus. Les principales caractéristiques de cet univers montrent que le président Chirac avait une attitude distanciée vis-à-vis de la "construction de l'Europe".

En collaboration avec MONIERE Denis. Analyse comparée du discours gouvernemental au Canada et au Québec. *Document numérique*. Volume 22, 1/2019, p. 85-105.

Analyse comparée du discours politique au Canada et au Québec à l'aide de quatre corpus : les discours du trône (Canada) et inauguraux (Québec), de 1867 à nos jours, et les discours de circonstance tenus par les Premiers ministres en dehors de la chambre (1995-2010). Après un rappel des normes de constitution des corpus, la comparaison porte sur des indices stylistiques puis lexicométriques. Une nouvelle méthode d'étude des caractéristiques lexicales de deux

corpus sensiblement de même longueur. Les différences entre Canada et Québec reflètent le partage des compétences entre les niveaux fédéral et provincial, mais aussi diverses conceptions du pouvoir, de l'action politique, et des relations aux citoyens.

Soixante-ans de discours présidentiels français (1958 – 2018). Qu'est-ce qui singularise Emmanuel Macron ? *Séminaire Mathématique et société*. Université de Neuchâtel, 17 mai 2019.

Présentation des méthodes statistiques pour la détermination du vocabulaire caractéristique d'un corpus de textes comparé à un ensemble de référence. L'ensemble des discours des présidents français depuis 1958 sert à déterminer le vocabulaire caractéristique des interventions d'E. Macron durant les 20 premiers mois de son quinquennat (2017-2018). Ce président s'inscrit en rupture par rapport à ses prédécesseurs. Il privilégie la politique internationale et parle moins de la France et des Français. Il néglige certains thèmes privilégiés par les autres (l'économie, l'emploi, le progrès, la croissance, les revenus, le social). Il assume peu ses propos et utilise plus le "nous" que le "je". Son discours a une visée pédagogique, mais il est souvent lourd et abstrait. Les phrases sont longues et compliquées, parfois obscures.

Naissance de l'industrie moderne du spectacle. Les comédiens poètes. Séminaire Linguistique du français moderne: Linguistique de corpus. Université de Neuchâtel, 19 mai 2019.

Présentation d'un corpus de pièces de théâtre françaises du XVIIe siècle qui seront mises en ligne. Ces pièces ont été transcrites en français contemporain, étiquetées et indexées. Les principales caractéristiques de ce corpus seront analysées. Ce sera surtout l'occasion de redécouvrir les débuts de l'industrie moderne du spectacle et notamment le rôle des "comédiens poètes", rouage essentiel de l'économie des troupes, qui présentèrent près de la moitié des pièces de l'époque dont certaines seront redécouvertes à cette occasion.

Apports de la lexicométrie à l'analyse des entretiens. Le cas de l'enquête « Les Français et la politique ». Avec une note d'Etienne Schweisguth et une réponse de l'auteur. Communication au colloque "*Des instruments au service de la recherche en sciences sociales"*, Paris, 28 septembre 2018.

Dépouillement lexicométrique des 64 entretiens de l'enquête « les Français et la politique » pilotée par Etienne Schweisguth en 1983 : balisage des textes, standardisation des graphies, étiquetage des mots, longueur des textes, principales caractéristiques de leurs vocabulaires décrites à partir du tableau lexical, des index hiérarchiques, des concordances et des syntagmes répétés. Le calcul des distances entre textes et leur classification révèlent l'influence de l'enquêteur sur les propos de l'enquêté. En conclusion, un appel à constituer une grande base de transcriptions d'entretiens. En annexe, une note d'E. Schweisguth et une réponse de l'auteur.

En collaboration avec MONIERE Denis. Le vocabulaire des campagnes électorales. In Iezzi Domenica F., Celardo Livia, Misuraca Michelangelo. *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*. Roma: UniversItalia, 2018, p. 522-540.

Après un premier mandat présidentiel, C. de Gaulle, V. Giscard d'Estaing, F. Mitterrand, J. Chirac et N. Sarkozy ont été candidats à un deuxième mandat. On compare leurs discours électoraux avec leurs discours présidentiels à l'aide des spécificités du vocabulaire. Il apparaît

que ces spécificités dépendent surtout des catégories grammaticales et des effectifs des mots. On présente des modifications du calcul classique qui permettent de neutraliser l'influence des catégories grammaticales et, au moins partiellement, celle des fréquences. Le discours électoral privilégie le verbe au détriment du nom, il est plus personnalisé que le discours au pouvoir, il se centre sur le pays et ses habitants, le reste du monde passant au second plan. Enfin, ces dernières années, la dimension polémique devient prédominante.

En collaboration avec LABBE Cyril. Les phrases de Marcel Proust. In Iezzi Domenica F., Celardo Livia, Misuraca Michelangelo. *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data*. Roma: UniversItalia, 2018, p. 400-410.

Analyse des longueurs de phrases dans *A la recherche du temps perdu* de Marcel Proust. Présentation des normes de dépouillement et des différentes mesures possibles. Durant la majorité de sa lecture, le lecteur se trouve confronté à des phrases très longues et syntaxiquement complexes. Une comparaison avec un large panel d'écrivains montre qu'il s'agit d'un phénomène exceptionnel mais pas unique et que certaines caractéristiques se retrouvent dans quelques œuvres dont certaines sont citées dans la *Recherche du temps perdu*.

Réponses à M. Forestier et à ses amis. Grenoble : Laboratoire PACTE, mai 2017

Quatre réponses à M. G. Forestier et à ses amis à propos de l'attribution à Pierre Corneille des principales pièces parues sous le nom de Molière. Trois réponses datent de 2003 et répondent aux principales objections faites à l'époque. Depuis lors, M. Forestier a multiplié les pages internet qui concernent rarement nos recherches (bien qu'il affirme le contraire). Ces pages montrent que M. Forestier est incapable de dire à l'aide de quelles caractéristiques observables, mesurables, reproductibles et falsifiables il reconnaît l'auteur d'un texte. De plus, il ne respecte pas les règles du débat scientifique et multiplie les injures et les calomnies. Nos réponses donnent les principaux liens afin que le lecteur puisse se faire une opinion impartiale.

Jean Racine, plume de l'ombre ? *Séminaire Linguistique du français moderne: Linguistique de corpus*, Université de Neuchâtel, 28 février 2017.

J. Racine a produit - sous son nom - onze tragédies dont certaines continuent à être jouées aujourd'hui. En 1677, il entre au service du roi et abandonne le théâtre ne sortant de sa réserve qu'à deux reprises (Esther et Athalie). Pourtant, en 1695, le Père Colonia - lui-même dramaturge - affirme que J. Racine continue à produire des pièces et que J.-G. Campistron est son prête-nom. L'attribution d'auteur assistée par ordinateur permet de vérifier la réalité de ces affirmations. Au total, 14 tragédies viennent s'ajouter au corpus racinien. Certaines ont été de très grands succès et méritent d'être redécouvertes.

En collaboration avec MONIERE Denis. *Tel père, telle fille ? Le discours de Jean-Marie et Marine Le Pen*. Grenoble : PACTE (CNRS), 2016.

En 2011, Marine Le Pen a succédé à son père Jean-Marie à la présidence du Front national qui est devenu le premier parti de France. Elle aurait réussi à changer l'image de son parti, en modifiant son discours, et à élargir ses assises électorales. A l'aide d'indices statistiques, les discours électoraux du père, prononcés durant les campagnes de 2002 et 2007, sont comparés à ceux de sa fille en 2012 afin d'identifier les changements entre les deux et leurs singularités par rapport aux autres dirigeants politiques français contemporains. M. et J.-M. Le Pen utilisent tous les deux un vocabulaire plus étendu que les autres leaders, mais la fille a réduit la longueur de la période oratoire. Elle tourne ses propos vers l'avenir et non vers le passé comme son père. Elle tient un discours plus explicatif, mais aussi plus personnalisé et se situe

sur le plan des idées et des valeurs. Elle se centre mieux sur quelques thèmes : la France, les Français, la critique du système, la crise économique, l'Europe et l'immigration-insécurité.

En collaboration avec LABBE Cyril. 57 ans de communication présidentielle (1958-2015)**.** In Edward Arnold, Cyril Labbe & Denis Monière. *Parler pour gouverner : Trois études sur le discours présidentiel français*. Grenoble : Laboratoire d'Informatique de Grenoble, 2016, p. 38-53.

Recensement de tous les textes disponibles et bilan de la communication présidentielle depuis 1958. Le général de Gaulle a inventé cette communication et ses successeurs (G. Pompidou, V. Giscard d'Estaing, F. Mitterrand, J. Chirac, N. Sarkozy et F. Hollande) ont suivi les usages que le Général avait établis sans rien modifier d'essentiel. Les allocutions restent le vecteur privilégié avec les entretiens. Les conférences de presse et les messages complètent cette communication. L'étude conduit à s'interroger sur l'exceptionnelle intensité de cette communication et sur la fonction présidentielle sous la Ve République.

En collaboration avec ARNOLD Edward. Vote for me. Don't vote for the other one. *Journal of World Languages*. Routledge, 2015, p. 1-18.

The French presidential election takes place in two ballots. The second round opposes the two leading candidates at the end of the first. Between the two ballots, since 1974, the two finalists take part in a TV debate along the lines of the US presidential debates. This presentation analyses the texts of these six debates (136,000 words). A library of more than 6000 political texts – and nearly 13 million words – provides some benchmarks. This paper presents the statistical indices proposed for the analysis of the communication within a situation of interaction. These indices are derived from theories concerning the presentation of actants in the speech, the expression of the speaker's subjectivity and the speech modalization. The application of these indices allows to bring a new perspective on these debates and it defines, for each of these indices, its scope, limitations and possible improvements. The first part analyses the tendency of the speakers to personalize. These indices are broken down into the following dimensions: the relative importance given to the speaker, to the other and to the real message recipients (the listeners). The second part measures the fundamental choice in favour of the verb and, within this part of speech, between the accomplished ones (verbs to be and to have) and modal verbs (possible, desirable, obligation, knowledge). Finally, the greater or lesser density of the negation highlights the real scope of discourse. The study leads to interesting conclusions about electoral discourse and the evolution of French political discourse over the last 40 years. Finally, it emphasizes the usefulness of large corpuses of texts and of lexicometry for language studying and teaching.

En collaboration avec LABBE Cyril et PORTET François. Detection of Computer-Generated Papers in Scientific Literature. In Degli Esposti Mirko, Altmann, Eduardo G., Pachet François (Eds.). *Creativity and Universality in Language*. *Lecture Notes in Morphogenesis*. Special issue. Springer, 2016, p 105-121.

Meaningless computer generated texts can be used in several ways. For example, they have allowed Ike Antkare to become one of the most highly cited scientists. Such fake publications are also appearing in real scientific conferences and, as a result, in the bibliographic services (Scopus, ISIWeb of Knowledge, Google Scholar). More than 120 papers have been withdrawn from subscription databases of two high-profile publishers, IEEE and Springer, because they were computer generated thanks to the SCIgen software designed to generate randomly computer science research papers. We present different methods (like Probabilistic Context Free Grammar) that can be used to generate such meaningless texts. We also discuss the fact that such generators can be automatically detected mainly because they are behaving like

authors that would have very caracteristic features. This shows that quantitative approaches, like intertextual distance, provide effective tools to characterize originality (or banality) in language.

En collaboration avec MONIERE Denis. Ne votez pas l'autre ! La spirale de la négativité. In Gerstlé Jacques & Magni Berton Raul (dir.). *2012, La campagne présidentielle*. Paris : l'Harmattan, 2014, p. 195-209.

Synthèse des analyses de contenu et lexicométrique des communiqués et des discours des principaux candidats à la présidentielle de 2012 et de leurs partis (Bayrou, Hollande, Le Pen, Mélenchon, Sarkozy). Sauf Bayrou qui privilégie la présentation de soi, de son programme et de son soutien, les autres candidats mettent au premier plan la critique de l'autre, spécialement les deux finalistes (Hollande et Sarkozy), déclenchant une spirale de la négativité qui a culminé entre les deux tours. Cette négativité s'est considérablement accrue par rapport à 2007. Enfin, la fin de la campagne est dominée par les thèmes imposés par Le Pen (la France, les Français, l'immigration).

En collaboration avec LABBE Cyril. Was Shakespeare's Vocabulary the Richest? In Née Emilie, Daube Jean-Michel, Valette Mathieu, Fleury Serge (dir.). *Proceedings of the 12th International Conference on Textual Data Statistical Analysis*. Paris: June 3-6 2014, p 323-336.

It is generally assumed that the vocabulary of W. Shakespeare is exceptionally rich and his work contains a very large number of different words. We present a method to compare the extent of the vocabularies of several authors' works of unequal length. Applied to the theater of Shakespeare's time, it shows that the vocabulary of Shakespeare is not exceptional and that some or his contemporaries – like B. Jonson or T. Dekker – used a larger vocabulary.

Il est généralement admis que le vocabulaire de W. Shakespeare est remarquablement riche. Son œuvre contiendrait un très grand nombre de mots différents. On présente une méthode qui permet de comparer l'étendue du vocabulaire utilisé dans des textes de longueur différente. Appliquée au théâtre de l'époque de Shakespeare, elle montre que le vocabulaire de cet auteur n'a rien d'exceptionnel et que certains contemporains – comme B. Jonson ou T. Dekker – utilisaient un vocabulaire plus étendu.

En collaboration avec MONIERE Denis. Un siècle et demi de discours gouvernemental au Canada. Contribution de la lexicométrie à l'Histoire politique. In Née Emilie, Daube Jean-Michel, Valette Mathieu, Fleury Serge (dir.). *Proceedings of the 12th International Conference on Textual Data Statistical Analysis*. Paris: June 3-6 2014, p. 485-494.

Le Premier ministre canadien ouvre chaque session du parlement d'Ottawa par un discours du "trône" soit, depuis l'origine de ces institutions (1867), 128 discours comportant 260 836 mots. La segmentation automatique de ce corpus met en lumière deux tournants majeurs délimitant trois périodes dont les thèmes propres sont déterminés grâce à leurs vocabulaires caractéristiques. Le premier tournant correspond à la seconde guerre mondiale et à l'après-guerre où les activités du gouvernement central s'étendent considérablement ; le deuxième survient en 1968 avec l'arrivée au pouvoir de Trudeau qui veut bâtir une nation canadienne et un pouvoir fédéral fort. Au sein de ces trois périodes principales, on distingue plusieurs épisodes secondaires importants dont on établit également les vocabulaires caractéristiques. Ainsi, la lexicométrie fournit des outils intéressants pour la périodisation de l'histoire politique.

En collaboration avec ARNOLD EDWARD. Votez pour moi… Ne votez pas pour l'autre (Les débats télévisés entre-deux-tours de l'élection présidentielle française depuis 1974). *Communication à la 59e conférence annuelle. The International Linguistic Association*. Paris, 22-24 mai 2014.

L'élection présidentielle française se déroule en deux scrutins. Le second tour oppose les deux candidats arrivés en tête au premier. Entre les deux tours, depuis 1974, un débat télévisé oppose les deux finalistes sur le modèle des débats présidentiels aux Etats-Unis. Notre communication utilisera les textes de ces 6 débats (136 000 mots). Une bibliothèque de plus de 6 000 textes politiques offre des points de comparaison.

Cette communication présente des indices statistiques construits pour l'analyse de cette communication en situation d'interaction. Ces indices sont issus des théories concernant la présentation des actants du discours, l'énonciation de la subjectivité du locuteur et de la modalisation du discours. L'application de ces indices permet d'apporter un éclairage neuf sur ces débats mais surtout de définir, pour chacun de ces indices, sa portée, ses limites et les améliorations possibles.

La première partie analyse la tendance à la personnalisation propre à chaque orateur et la décompose dans les dimensions suivantes : l'importance relative donnée à l'orateur, à l'autre et aux véritables destinataires du message (les auditeurs). La seconde partie mesure le choix fondamental en faveur du verbe et, au sein de celui-ci, entre l'accompli (densité des verbes, de *être* et *avoir*) et les modalités (possible, souhaitable, volonté, obligation, connaissance). Enfin, la densité plus ou moins importante de la négation mesure la portée polémique du discours.

L'étude conduit à des conclusions intéressantes concernant les discours électoraux et l'évolution du discours politique français depuis 40 ans. Elle souligne enfin l'utilité des vastes corpus de textes et de la lexicométrie pour l'étude de la langue et son enseignement.

Identification de l'auteur d'un texte (Hugo, Lamartine, Musset et Vigny). Conférence invitée au séminaire *L'œuvre et son auteur : problèmes d'attribution*. Lille : Université de Lille-Nord de la France, 21 mai 2014.

La statistique lexicale permet-elle d'identifier l'auteur d'un texte ? En 1988, E. Brunet avait répondu par la négative en utilisant des pièces de théâtre, des romans et des poésies d'Hugo, Lamartine et Musset. Nous proposons de revisiter cette expérience : débarrassée de ses biais et de ses présupposés, elle montre que ces trois auteurs sont clairement identifiables tant au niveau de leurs vocabulaires que de leurs styles. Cela permet de répondre à la question « qu'est-ce qu'un auteur ? » en identifiant les caractéristiques particulières de son vocabulaire et de son style par rapport à ses contemporains. L'introduction de Vigny permet en outre de mettre en valeur des proximités et des influences entre auteurs.

Les plumes de l'ombre. Molière a-t-il écrit ses pièces ? Conférence invitée. *Université Inter-Ages du Dauphiné*. Grenoble : 18 février 2014.

Description du système de la plume de l'ombre. Cette pratique était indécelable. Ainsi personne n'a reconnu R. Gary derrière E. Ajar. Les auteurs qui ont écrit les pièces de théâtre produites par le Cardinal Richelieu ne sont toujours pas identifiés. Nul n'a pu dire quelle part T. Corneille a prise dans l'écriture des pièces présentées par Montfleury ou Hauteroche. Or au XVIIe siècle, la majorité des pièces de théâtre n'ont pas été présentées par les écrivains qui les ont écrites mais par des intermédiaires qui les négociaient avec les troupes. Une méthode d'attribution d'auteur par ordinateur permet de résoudre ces énigmes. La fiabilité de la méthode a été rigoureusement testée. Elle attribue à P. Corneille toutes les pièces en vers

présentées par Molière ainsi que *Dom Juan, l'Avare, le Bourgeois gentilhomme* et *le Malade imaginaire*. De nombreux indices historiques confirment cette attribution.

En collaboration avec LABBE Cyril. Le chiffre dans le discours politique français contemporain. V. Giscard d'Estaing et les autres présidents. *Communication aux XIVe Journées de l'ERLA*. Brest : 15-16 novembre 2013. In BANKS David. *La quantification dans le texte de spécialité*. Paris : L'Harmattan, 2016, p. 53-75.

Dans leurs discours, les politiques utilisent beaucoup la quantification (nombres et dates). C'est ce que révèle la comparaison des discours tenus par un grand nombre de responsables politiques français quand on les compare à un vaste échantillon du français moderne comportant 30 millions de mots, tous étiquetés. Les chiffres et les dates servent à donner aux discours un ancrage dans le temps, la réalité économique et sociale. Une illustration est donnée avec V. Giscard d'Estaing (1974-1981) comparé aux cinq autres présidents de la République entre 1958 et 2012.

En collaboration avec LABBE Cyril. L'intertextualité dans les publications scientifiques. Conférence invitée. *Séminaire du Laboratoire de Linguistique et Didactique des langues*. Grenoble, 28 juin 2013.

La base de données bibliographiques de l'IEEE contient un certain nombre de duplications avérées avec indication des originaux copiés. Ce corpus est utilisé pour tester une méthode d'attribution d'auteur. La combinaison de la distance intertextuelle avec la fenêtre glissante et diverses techniques de classification permet d'identifier ces duplications avec un risque d'erreur très faible. Cette expérience montre également que plusieurs facteurs brouillent l'identité de l'auteur scientifique, notamment des collectifs de chercheurs à géométrie variable et une forte dose d'intertextualité acceptée voire recherchée.

En collaboration avec LABBE Cyril. Lexicométrie : quels outils pour les sciences humaines et sociales ? *Communication aux Journées d'étude Usages de la lexicométrie en sociologie*. Université de Versailles, 12-13 juin 2013.

La lexicométrie est l'alliance des sciences du langage, des statistiques et de l'informatique. Elle permet de traiter de vastes ensembles de textes (corpus), d'établir leur vocabulaire, de classer les vocables en fonction de leur fréquence, de leur répartition, de leurs catégories grammaticales. Elle établit les contextes d'emploi d'un vocable et les combinaisons les plus fréquentes dans lesquelles il entre, ce qui permet de déterminer le ou les sens de ce vocable. Elle retrouve les principaux thèmes présents dans un corpus, son genre et son style. Elle segmente ce corpus en fonction des ruptures thématiques ou stylistiques. Pour obtenir ces résultats, des traitements préalables sont nécessaires : balisage des textes, correction et standardisation orthographiques, étiquetage des mots. Le texte peut alors entrer dans une bibliothèque électronique à la disposition des chercheurs.

En collaboration avec LABBE Cyril. L'ordinateur peut-il écrire ? *Communication au séminaire Mathématiques et société*. Neuchâtel, novembre 2012.

Il existe des "générateurs" d'articles scientifiques (notamment en informatique, physique, mathématique ou philosophie). La conférence explique comment marchent ces générateurs et présente une méthode qui permet de détecter les faux articles réalisés avec eux. Pour l'instant, ces articles sont assez faciles à identifier car ils sont dépourvus de sens et surtout parce qu'ils n'ont pas la diversité de vocabulaire et de syntaxe des humains. Pourtant, plusieurs fausses publications se sont faufilées dans des congrès et dans les grandes bases de données bibliographiques payantes qui servent à évaluer les chercheurs. Enfin, il existe plusieurs voies pour améliorer ces générateurs qui seront peut-être, dans le futur, des auxiliaires utiles pour la rédaction des textes.

En collaboration avec LABBE Cyril. *Réponses à MM. Bernet et Brunet*. Grenoble : Laboratoire PACTE et LIG, octobre 2012.

> Il y a trois ans, paraissaient dans un volume d'hommages à Charles Muller, deux prétendues réfutations – par MM. Bernet et Brunet - de notre attribution à Corneille des principales pièces parues sous le nom de Molière. Contrairement aux usages en matière de publication scientifique, nous n'avons eu connaissance de ces textes qu'en juin 2011 (Brunet) et octobre 2012 (Bernet). Loin d'infirmer nos conclusions, ces deux textes les renforcent.

En collaboration avec LABBE Cyril. Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science? *Scientometrics*. Published on line: 22 june 2012.

> Two kinds of bibliographic tools are used to retrieve scientific publications and make them available online. For one kind, access is free as they store information made publicly available online. For the other kind, access fees are required as they are compiled on information provided by the major publishers of scientific literature. The former can easily be interfered with, but it is generally assumed that the latter guarantee the integrity of the data they sell. Unfortunately, duplicate and fake publications are appearing in scientific conferences and, as a result, in the bibliographic services. We demonstrate a software method of detecting these duplicate and fake publications. Both the free services (such as Google Scholar and DBLP) and the charged-for services (such as IEEE Xplore) accept and index these publications.
>
> Deux types d'outils bibliographiques sont utilisés pour retrouver les publications scientifiques et pour les rendre disponibles en ligne. Le premier type fournit un accès gratuit à l'information disponible en ligne. Le second est payant car ces bases stockent l'information fournie par les principaux éditeurs de revues scientifiques. Les premières sont facilement manipulables mais il est généralement admis que les secondes garantissent la qualité de l'information qu'elles vendent. Malheureusement, des articles totalement ou partiellement copiés et des faux sont acceptées dans certaines conférences scientifiques et, en conséquence, ces documents sont indexés dans les bases bibliographiques gratuites comme payantes. Nous présentons des outils informatiques pour détecter ces copies et ces faux articles.

En collaboration avec MONIERE Denis. Le vocabulaire caractéristique du Premier ministre du Québec J. Charest comparé à ses prédécesseurs. In Dister Anne, Longrée Dominique, Purnelle Gérald (éds). *Proceedings of the 11th International Conference on Textual Data Statistical Analysis*. Liège : LASLA - SESLA, 2012, p.737-751.

> Comment mesurer les singularités du vocabulaire d'un locuteur ? On répond à l'aide d'un exemple : les discours de Jean Charest (Premier ministre du Québec depuis 2003). Pour minimiser l'influence du contexte de l'énonciation, on utilise un corpus de référence contenant les discours émis par ses prédécesseurs (les Premiers ministres du Québec depuis le début du XXe siècle). Le calcul des spécificités du vocabulaire est rappelé, plusieurs améliorations sont proposées. Outre les caractéristiques des discours de J. Charest, l'expérience montre que plus un vocable est employé, plus il a de chances d'être spécifique.

En collaboration avec LABBE Cyril. Detection of Hidden Intertextuality in the Scientific Publications. In Dister Anne, Longrée Dominique, Purnelle Gérald (éds). *Proceedings of the 11th International Conference on Textual Data Statistical Analysis*. Liège : LASLA - SESLA, 2012, p.537-551.

> Intertextuality is the presence of one text contained within another. Hidden intertextuality is a problem for scientific publications. We propose a detection method combining calculation of intertextual distance and several classifications with the technique of the "sliding window" (to pinpoint any

duplicated excerpts). This method is tested using a group of texts extracted from the IEEE bibliographic database.

En collaboration avec MONIERE Denis. *Radioscopies de la campagne présidentielle*. Onze analyses publiées du 10 février au 4 mai 2012 sur le site www.trielec2012.fr (également consultables sur le site hal.archives-ouvertes.fr).

- I. *La pré-campagne (1er janvier - 4 février).*

  Les communiqués des candidats et de leurs partis sont particulièrement révélateurs des stratégies de communication dans une campagne électorale. L'analyse de ce corpus pour les cinq principaux candidats pendant les cinq premières semaines de 2012 montre que l'UMP est pleinement engagée dans la campagne électorale en diffusant plus de communiqués que ses concurrents. A l'exception de F. Bayrou dont les communiqués sont plus ancrés « dans la réalité », tous les candidats ou leur parti orientent en priorité leur discours vers la critique de l'adversaire. Dans ce registre, la palme de l'agressivité revient à l'UMP qui concentre la quasi totalité de ses critiques contre François Hollande et le Parti socialiste et ignore totalement F. Bayrou.

- II. *Les mots et les thèmes de la pré-campagne (1e janvier -11 février 2012)*

  Analyse lexicométrique des communiqués et des discours des cinq principaux candidats à l'élection présidentielle durant les 6 premières semaines de 2012. L'essentiel de la précampagne n'a pas consisté à présenter des propositions et des programmes mais à critiquer l'autre : le président sortant pour ses quatre challengers ; F. Hollande et le PS pour l'UMP au nom de N. Sarkozy pas encore officiellement entré en campagne). Les candidats ont choisi des relations très différentes aux Français et au pays. F. Bayrou recherche une relation inclusive et familière avec ses auditoires et avec le peuple. F. Hollande entretient une relation distante avec les Français et centre son discours sur sa propre personne. N. Sarkozy, encore président, se situe à peu près à mi-chemin de F. Hollande et de F. Bayrou. Quant à M. Le Pen, elle développe un propos tendu mais dépersonnalisé. On présente enfin les principaux thèmes développés par chacun.

- *III. L'entrée en scène du Président sortant (5-18 février).*

  Les dernières semaines ont été marquées par le discours de N. Sarkozy le 29 janvier et surtout par son entrée en campagne officielle le 15 février. Ces deux événements correspondent à des évolutions importantes dans la tonalité des communiqués de presse de l'UMP qui deviennent plus positifs. Occupés à valoriser leur candidat et ses politiques, ses porte-paroles ont réduit la place consacrée à la critique des adversaires. Dans le même temps, la candidature du président a provoqué des ajustements stratégiques chez ses concurrents qui ont eu tendance à intensifier leurs interventions et à se montrer plus offensifs. Ce changement a surtout été manifeste pour le candidat du Modem qui a marqué sa différence par rapport à N. Sarkozy dont il a dénoncé les propositions. F. Hollande s'est montré plus pugnace avec l'entrée en campagne de son principal adversaire. Durant ces deux semaines, c'est M. Le Pen qui connaît la plus forte baisse de régime, non seulement ses communiqués sont moins nombreux, mais leur contenu est aussi moins combatif comme si sa campagne connaissait une pause.

IV. *Sale mec ou gentil garçon ? Portraits croisés des principaux candidats (1er janvier-25 février 2012).*

L'analyse lexicométrique de la campagne présidentielle montre que la communication de N. Sarkozy, de F. Hollande, et de leurs partisans respectifs, est plus consacrée à la critique de l'adversaire qu'à la mise en valeur du candidat. M. Le Pen et le Front National ont choisi de valoriser la candidate et de concentrer l'essentiel de leurs attaques sur N. Sarkozy. Curieusement, le Front de Gauche, et J.-L. Mélenchon, ont choisi comme cible principale M. Le Pen et le FN et non pas N. Sarkozy ou F. Hollande. Enfin, F. Bayrou et le Modem, ont fait jusqu'à maintenant des choix assez différents des autres : le candidat est moins cité et moins mis en valeur. La critique des concurrents a occupé – avant la mi-février – une place plus faible que chez les autres. L'analyse montre également que F. Hollande, N. Sarkozy et leurs soutiens respectifs utilisent les mêmes procédés de péjoration et de valorisation des personnes. Elle souligne enfin l'agressivité croissante de leur campagne.

- V. *Faire campagne contre les autres ? (19 février-3 mars)*

  Les deux semaines (19 février au 3 mars), ont été marquées par le démarrage de la campagne de N. Sarkozy. Il a exposé ses valeurs, sa relation avec la France, sa gestion de la crise, son désir de valoriser le travail et les thèmes de sa campagne axée sur l'éducation, la sécurité, l'immigration qui ont ensuite été relayés par les messages de ses partisans. Les autres candidats semblent s'être un peu effacés, intervenant moins en public. A l'inverse, certains partis ont intensifié leur communication. L'UMP et le PS ont émis plus de communiqués En revanche, le Modem, le FN, et le Front de Gauche ont réduit leur communication. Au total, depuis le 1er janvier, les 5 principaux candidats et leurs partisans ont émis plus 1160 messages comportant au total près de 700.000 mots. Dans cet océan de mots l'analyse de contenu met en lumière les stratégies de communication choisies par chacun des candidats et par leurs partis respectifs.

- *VI. L'effet Villepinte (4-17 mars).*

  Le grand rassemblement de Villepinte organisé par N. Sarkozy le dimanche 12 mars a-t-il été un tournant dans la campagne présidentielle ? L'analyse de contenu de la communication des candidats durant la première quinzaine de mars et l'analyse assistée des thèmes par ordinateur apportent quelques réponses. Deux candidats se détachent pour l'intensité de leur communication : N. Sarkozy et F. Hollande. Les stratégies de communication de tous les candidats ont connu quelques infléchissements. N. Sarkozy, M. Le Pen, et J.-L. Mélenchon mettent un peu plus l'accent sur les dimensions positives de leurs messages, alors que F. Bayrou et F. Hollande se montrent plus offensifs. Le candidat du front de gauche change de cible et s'en prend désormais plus à N. Sarkozy qu'à la candidate du Front national. Sauf chez F. Bayrou, les thèmes favoris des candidats sont la critique de leur principal rival, puis la mise en valeur de leur personne et de leurs propositions. Ensuite, les thèmes économiques et sociaux dominent dans la communication de tous, suivis des problèmes de société, notamment l'éducation et l'immigration. L'entrée en lice de N. Sarkozy a tendu la campagne électorale et fait reculer la plupart des thèmes de campagne, excepté autour de l'immigration et de la critique de l'Europe qui étaient déjà les thèmes privilégiés par M. Le Pen.

- *VII. Les attentats de Montauban et de Toulouse : un tournant dans la campagne électorale ? (18-24 mars 2012)*

  Les événements de Toulouse ont eu quelques effets immédiats sur la communication des candidats. F. Hollande M. Le Pen et N. Sarkozy et leurs soutiens ont ralenti pendant trois jours le rythme de celle-ci. En revanche, F. Bayrou et J.-L. Mélenchon ont poursuivi leur campagne. N. Sarkozy et F. Hollande ont appelé à l'unité. Le premier – en tant que chef de l'Etat – a appelé à l'unité de la France autour des institutions, le second (F. Hollande) à l'union autour des valeurs de la République. Après le 22 mars, tous deux ont intensifié leur campagne. Le président a mis l'accent sur les thèmes qu'il avait lancés à Villepinte : respect de l'autorité en France, critique de l'Europe, notamment sur le plan de l'immigration. F. Hollande a repris ses thèmes traditionnels. F. Bayrou insiste d'avantage sur les valeurs de tolérance, de solidarité et sur le produire en France. En définitive, les événements de Toulouse ont probablement été une parenthèse dans la campagne présidentielle mais peut-être aussi l'occasion d'une radicalisation de celle-ci.

- *VIII. La spirale de la négativité (25 mars-7 avril).*

  Après une courte accalmie, la campagne a repris son rythme normal tant sur le plan de l'intensité que de la négativité. Les deux principaux candidats ont alimenté la spirale de la négativité en se dénonçant réciproquement. Ils ont pratiqué la "chasse en meute" en mobilisant simultanément plusieurs de leurs plumes pour dénigrer une déclaration ou une prise de position de leur adversaire. En seconde partie, cette note décrit la thématique des candidats depuis le début de la campagne. Chacun a développé une vision particulière de la France. Tous ont axé leur campagne sur la situation économique et sur l'Europe. Enfin, cette note révèle leurs thèmes spécifiques.

- *IX. La course de fond des candidats à l'élection présidentielle (jusqu'au 14 avril)*

  Les thèmes et le style des principaux candidats ont-ils changé depuis leur entrée en lice ? L'analyse conduit à trois conclusions principales. Premièrement, les principaux candidats et leurs équipes ont

privilégié la simplicité. A l'exception de M. Le Pen, ils ont adopté, dans leur discours et leurs entretiens, un vocabulaire peu diversifié et assez général, réservant le vocabulaire spécialisé aux communiqués. Deuxièmement, tous les candidats se sont tenus aux thèmes qu'ils s'étaient fixés au départ, sans y apporter de modifications importantes en cours de campagne. F. Bayrou a présenté son programme fin janvier et l'a développé sans y apporter de modifications ou de nouveauté majeure. F. Hollande a développé pendant six mois, les thèmes présentés lors de la convention d'investiture d'octobre 2011, accentuant simplement les critiques contre N. Sarkozy en fin de campagne. N. Sarkozy a profité des vœux que le président adresse en début d'année aux différents corps constitués – puis de la crise de l'Euro - pour esquisser son programme pour un second septennat. En dehors de son discours de Villepinte, il n'a rien apporté de neuf à cette thématique. Troisièmement, F. Hollande est le seul à présenter des fluctuations stylistiques importantes, comme s'il avait eu du mal à trouver son style de communication ou comme si ses rédacteurs avaient changé à plusieurs reprises. Les autres font preuve d'une stabilité remarquable qui suggère que leur communication a été confiée à un très petit nombre de collaborateurs sous une direction unique.

- *X. La dernière ligne droite (8-21 avril).*

  Durant la dernière quinzaine de la campagne pour le premier tour de l'élection présidentielle, F. Bayrou, F. Hollande, M. Le Pen, et leurs partis respectifs, ont augmenté l'intensité de leur communication alors que N. Sarkozy et l'UMP diminuaient la leur. Malgré une légère modération de F. Hollande, durant les quinze derniers jours, la fin de la campagne a été dominée par les attaques entre candidats et par l'agressivité. Aucun changement notable n'est observé dans les stratégies de communication adoptée par les candidats. F. Bayrou et F. Hollande ont choisi de couvrir un plus grand nombre de thèmes que N. Sarkozy et M. Le Pen. M. Le Pen a axé sa campagne sur la France, les Français, le peuple, la nation et contre l'Europe. N. Sarkozy a adopté ces mêmes thèmes durant les deux derniers mois (après son discours de Villepinte). En dehors de ce choix premier, la campagne a été dominée par les thèmes économiques et financiers, l'éducation et l'immigration. A part l'Europe, la politique étrangère a été absente de cette campagne.

- *XI. La finale (22 avril-4 mai).*

  La campagne du second tour a été dominée par la spirale de la négativité. F. Hollande et N. Sarkozy ont consacré une proportion croissante de leur communication à critiquer l'autre, plutôt qu'à mettre en valeur leur candidature et leurs projets. Comme au premier tour, N. Sarkozy, et surtout l'UMP, ont été nettement plus critiques. F. Hollande et le PS ont donné une place plus importante à la campagne électorale et à la mobilisation des électeurs. La situation économique et financière, le chômage ont occupé la première place dans la communication des deux candidats. Ils ont réduit la place de certains thèmes qu'ils privilégiaient au premier tour : l'éducation, la culture, les PME, le travail. F. Hollande a aussi moins parlé des retraites, de la formation, du logement, de la situation des classes moyennes et populaires. N. Sarkozy a fait l'impasse sur la jeunesse. Les deux candidats ont beaucoup plus parlé de l'immigration. Pour F. Hollande, c'est l'un des principaux échecs de N. Sarkozy. Celui-ci a souligné l'importance des frontières et le danger communautaire. Les deux candidats se sont opposés à propos du droit de vote des étrangers non-communautaires aux élections municipales. Enfin, si N. Sarkozy s'adressait beaucoup au peuple. F. Hollande a préféré éviter ce thème. En revanche, au second tour par rapport au premier, les deux candidats ont donné beaucoup plus d'importance au pays et aux Français. C'était les thèmes privilégiés par M. Le Pen.

En collaboration avec LABBE Cyril. Analyser les questions ouvertes dans les sondages. *Journée d'étude : Comment convaincre ? Analyse scientifique de la campagne électorale 2012*. Grenoble : Institut d'études politiques de Grenoble, 9 Mars 2012.

Présentation d'une méthode d'analyse des réponses aux questions ouvertes dans les enquêtes d'opinion. On décrit d'abord la transcription, la codification et le traitement des réponses. Les caractéristiques particulières de ces réponses (longueur et vocabulaire) se prêtent mal à l'analyse « textuelle » traditionnelle. Grâce aux univers lexicaux et à la classification automatique, on repère les

principaux thèmes présents dans les réponses. Cela permet de traiter les questions ouvertes comme les questions fermées.

En collaboration avec LABBE Cyril. Existe-t-il un langage propre à la politique ? *Communication aux XIIe Journées de l'ERLA*. Brest : 18-19 novembre 2011. In BANKS David. Aspects linguistiques du texte politique. Paris : L'Harmattan, 2014, p. 7-28.

Après avoir présenté le corpus du discours politique français et la bibliothèque du français moderne, on recherche les caractéristiques singulières du discours politique. Le cas des présidents de la Ve République depuis 1958 est particulièrement étudié. Les vocables usuels ne sont pas propres à la politique mais leurs fréquences d'emploi et leurs combinaisons dessinent plusieurs univers singuliers. Le discours politique est assez dépersonnalisé avec une tonalité nettement pédagogique. Il est orienté vers le présent et le futur. Il privilégie les généralités et les abstractions plutôt que l'action. L'expérience montre l'utilité de la statistique et des vastes corpus de référence pour la linguistique appliquée.

En collaboration avec MONIERE Denis. Les discours de René Lévesque au regard de la statistique lexicale. Communication au colloque de la Fondation René Lévesque (Montréal - 4 novembre 2011). In Alexandre Stephanescu et Eric Bédard. *René Lévesque. Homme de la parole et de l'écrit*. Montréal : VLB éditeur, 2012, p. 45-65.

Analyses lexicale, stylistique et thématique des discours prononcés par R. Lévesque lorsqu'il était premier ministre du Québec (1976-1985) et comparaison avec ses prédécesseurs et successeurs. L'observation de l'accroissement du vocabulaire isole trois périodes principales. Le vocabulaire caractéristique montre une forte réticence à utiliser les vocables liés à la nation, un discours tendu et orienté vers l'action mais assez impersonnel et difficilement assumé. La phrase de R. Lévesque est longue et très complexe. Enfin, le thème de l'indépendance est très peu présent.

En collaboration avec LABBE Cyril. *Are the major bibliographic databases reliable?* Technical Report. Grenoble : LIG, novembre 2011.

It is generally assumed that fee paying bibliographic databases guarantee the quality of the data they sell. This technical report shows that this assertion can be challenged. It demonstrates a software method of detecting duplicate and fake publications. The charged-for services (such as IEEEXplore) accept and index these kinds of publications.

Comédiens et écrivains au XVIIe siècle. A la redécouverte des frères Corneille. Séminaire de stylistique française. Université de Cologne. Jeudi 9 juin 2011.

Après avoir rappelé les indices historiques et les mesures statistiques qui établissent la paternité de Pierre Corneille sur toutes les pièces en vers et sur au moins 3 pièces en prose parues sous le nom de Molière, on présente les premiers résultats d'une recherche en cours sur l'œuvre de son frère Thomas. Ses relations avec Montfleury et Hauteroche sont confirmées. Il a également collaboré à certaines œuvres présentées par Quinault puis par Regnard.

Lettre ouverte aux animateurs de la revue Lexicometrica *à propos de l'article : Stephan Vonfelt*, Le graphonaute ou Molière retrouvé. Grenoble : Laboratoire PACTE, avril 2011.

La revue en ligne Lexicometrica a publié un article de M. S. Vonfelt intitulé « le graphonaute » qui prétend « réfuter » notre attribution à P. Corneille des principales pièces parues sous le nom de Molière. En fait, l'indice de M. Vonfelt est l'inverse de la longueur des textes et non pas, comme le prétend l'auteur, la distance entre le vocabulaire des textes. De plus, il y a 11% d'erreurs dans les prétendues attributions et l'auteur a opéré plusieurs manipulations sur les

données et les graphiques. Enfin les responsables de la revue et l'auteur n'ont pas respecté les règles élémentaires de la discussion scientifique.

En collaboration avec LABBE Cyril. La classification des textes. Comment trouver le meilleur classement possible au sein d'une collection de textes ? *Images des mathématiques. La recherche mathématique en mots et en images*. (http://images.math.cnrs.fr/La-classification-des-textes.html). 28 mars 2011.

On mesure la distance entre deux textes grâce au nombre de mots différents qu'ils contiennent. Quatre facteurs expliquent cette distance : le genre, l'auteur, le thème et l'époque. On estime le poids de chacune de ces variables. Pour retrouver l'auteur d'un texte inconnu ou d'origine douteuse, il faut le comparer à d'autres, écrits à la même époque et dans un même genre, par des auteurs incontestables. Présentation d'une expérience en aveugle qui valide ce principe. Application au théâtre français du XVIIe siècle. Paternité des œuvres présentées par Molière.

Corneille nell'ombra di Molière. Conférence invitée. Scuola Superiore di Lingue Moderne per Interpreti e Traduttori, Università di Trieste (Italia), 21 gennaio 2011 (Traduzione **:** Irene Borsato). Publié dans : *Rivista Internazionale di Tecnica della Traduzione*. 12-2010, p. 117-138.

Molière ha davvero composto le opere teatrali comparse con il suo nome? Molti indizi storici dimostrano che non è così. Nel XVII secolo le commedie satiriche dei grandi autori venivano presentate dai "comédiens poètes", ovvero attori poeti. Molière non si è comportato da autore e nessuno dei suoi contemporanei l'ha mai trattato come tale. Anzi, giravano diverse voci, alcune delle quali indicavano in P. Corneille l'autore delle opere presentate da Molière. Queste voci sono confermate da cinque indici statistici: la distanza intertestuale, le classificazioni automatiche, i segmenti ripetuti comprendenti verbi comuni, il senso delle parole più frequenti e la lunghezza delle frasi.

Corneille dans l'ombre de Molière. Comment identifier un auteur ? Conférence invitée. Cercle philologique. Université de Padoue (Italie), 19 janvier 2011.

Plusieurs indices historiques montrent que Molière n'a pas composé les pièces représentées sous son nom. Au XVIIe siècle, les comédies satiriques des grands auteurs étaient présentées sous le nom de "comédiens poètes". Molière ne s'est pas comporté en écrivain et aucun de ses contemporains ne l'a traité comme tel. Au contraire, de nombreuses rumeurs ont couru, certaines désignant P. Corneille comme l'auteur des pièces de Molière. Cinq indices statistiques confirment ces rumeurs : la distance intertextuelle, les classifications automatiques, les combinaisons des verbes usuels, le sens des mots les plus fréquents, la longueur des phrases.

En collaboration avec LABBE Cyril. La modalité verbale en français contemporain. Les hommes politiques et les autres. *Communication aux XIe Journées de l'ERLA*. Brest : 19 novembre 2010. In BANKS David. *La modalité, le mode et le texte spécialisé*. Paris : L'Harmattan, 2013, p. 33-61.

Etude de la modalité verbale en français. Cette construction associe un auxiliaire modal suivi d'un complément à l'infinitif, comme "pouvoir faire", "vouloir dire", etc. Ces constructions sont plus fréquentes que celles associant un participe passé précédé d'un auxiliaire avoir ou être. Un test statistique permet de mesurer la singularité de chaque locuteur. Cette étude est complétée par trois indices de tension. Application à de grands corpus du discours politique, de la littérature et de la langue orale.

Le calcul du sens des mots. La lexicologie assistée par ordinateur. Communication au séminaire Mathématiques et société. Neuchâtel, 3 novembre 2010.

L'étude du langage peut tirer partie des ordinateurs et de la statistique. On présente de nouvelles méthodes pour calculer le sens des mots chez un auteur, dans un lexique spécialisé ou dans la langue.

Ces méthodes peuvent apporter des outils intéressants pour la lexicologie, la lexicographie ou pour les études littéraires comme le montrent deux exemples d'actualité : le vocabulaire de Le Clézio et le sens du mot « banque » dans le vocabulaire économique et social français contemporain.

Qui a écrit Dom Juan ? Molière est-il l'auteur des pièces parues sous son nom ? Communication devant la Société jurassienne d'émulation. Porrentruy, 2 novembre 2010.

Cette conférence expose les raisons pour lesquelles Molière n'a pas composé les pièces présentées sous son nom. Après avoir présenté le dossier historique et la pratique consistant à représenter les comédies satiriques des grands auteurs sous le nom de "comédiens poètes", cette conférence expose les méthodes statistiques qui attribuent à Corneille les comédies les plus connues de Molière. A cette occasion, sont présentés les résultats d'une expérience inédite effectuée en aveugle sur des romans français du XIXe siècle. La statistique appliquée au langage apporte aux études littéraires des outils précieux.

En collaboration avec MONIERE Denis. Segmentation des corpus chronologiques : 143 ans de discours gouvernemental au Québec. In Bolasco Sergio, Chiari Isabella, Giuliano Luca (Eds). *Proceedings of 10th International Conference Statistical Analysis of Textual Data*. Rome : Edizioni Universitarie di Lettere Economia Diritto, 2010, Vol 2, p. 805-816.

Méthode originale pour segmenter un corpus chronologique en périodes homogènes. On calcule l'accroissement du vocabulaire et son ajustement par une tendance. Un algorithme de segmentation associé à des tests de validité donne le découpage optimal du corpus. Une série d'indicateurs mesure l'ampleur des mouvements de vocabulaire caractérisant chacune des périodes. Application aux déclarations du gouvernement québécois à l'ouverture de chaque session du parlement provincial de 1867 à 2009.

Ce que disent les phrases de Corneille et Molière. Communication devant les Xe Journées Internationales d'Analyse des Données Textuelles. Rome : 11 juin 2010.

Dans le théâtre français du XVIIe siècle, les longueurs de phrases singularisent chacun des auteurs, sauf Corneille et Molière. Cette proximité est confirmée par d'autres indices : distances entre textes, classifications, combinaisons des mots fréquents, sens des vocables usuels. On rappelle ensuite le témoignage de plusieurs contemporains qui indiquent clairement que Molière n'est pas l'auteur des pièces qu'il présentait et qui désignent P. Corneille comme étant sa plume de l'ombre dans trois cas (*le Dépit amoureux*, le *Bourgeois gentilhomme*, *Psyché*). Il s'agissait d'un système : à cette époque, neuf comédies sur dix n'ont pas été présentées par leur auteur mais par un comédien poète, comme Molière.

En collaboration avec LABBE Cyril. Ce que disent leurs phrases. In Bolasco Sergio, Chiari Isabella, Giuliano Luca (Eds). *Proceedings of 10th International Conference Statistical Analysis of Textual Data*. Rome : Edizioni Universitarie di Lettere Economia Diritto, 2010, Vol 1, p. 297-307.

Etude de la phrase dans le théâtre du XVIIe siècle (Corneille, Mairet, Molière, Quinault, Racine). Les valeurs centrales (moyenne, mode, médiane et médiale) et la distribution des fréquences révèlent des caractéristiques propres à chacun des auteurs et les contraintes très fortes que font peser les vers alexandrins et les règles du théâtre classique. On distingue quatre type de phrases (en fonction de leur longueur) qui remplissent des fonctions différentes : interpeller, dialoguer, exposer voire soliloquer.

En collaboration avec MONIERE Denis. Quelle est la spécificité des discours électoraux? Le cas de Stephen Harper. *Canadian Journal of Political Science / Revue canadienne de science politique*, 43:1, (March/ mars 2010), p. 69–86.

Cette étude de cas démontre que le discours électoral possède des caractéristiques propres. On donne

l'exemple de Stephen Harper dont les discours tenus lors des élections de 2008 se différencient de ceux qu'il a prononcés à titre de chef de gouvernement. Le discours électoral est plus ancré socialement. C'est aussi un discours qui valorise le collectif national : le locuteur privilégie l'emploi du «nous», plutôt que de se présenter comme principal responsable des choix collectifs. Comparativement aux discours gouvernementaux, le discours électoral est aussi moins abstrait et plus orienté vers l'action, comme l'indique la prédominance du groupe verbal sur le groupe nominal. La forte présence de la construction négative et la désignation des adversaires (noms propres) soulignent le caractère polémique du discours électoral.

Qui a écrit Dom Juan ? Molière est-il l'auteur des pièces parues sous son nom ? *Communication devant le séminaire Mathématiques et société*. Université de Neuchatel : 9 décembre 2009.

A l'occasion de la parution de son ouvrage "Si deux et deux sont quatre, Molière n'a pas écrit Dom Juan" (Editions Max Milo), Dominique Labbé expose les raisons pour lesquelles Molière n'a pas composé les pièces présentées sous son nom. Après avoir discuté les méthodes statistiques qui attribuent à Corneille les comédies les plus connues de Molière, il explique pourquoi les deux hommes ont collaboré et il montre que cette collaboration n'avait rien d'exceptionnelle.

En collaboration avec LABBE Cyril. Existe-t-il un genre épistolaire ? Hugo, Flaubert et Maupassant. *Communication aux Xe Journées de l'ERLA*. Brest : 20-21 novembre 2009. Publié dans : BANKS David. *Le texte épistolaire du XVIIe siècle à nos jours*. Paris : L'Harmattan, 2013, p. 53-85.

Existe-t-il un genre épistolaire particulier ? Qu'est-ce qui différencie ce genre du reste de la littérature? On répond à ces deux questions en utilisant les correspondances de Victor Hugo, de Flaubert et Maupassant comparées au reste de leurs œuvres (romans, théâtre, poésie). Les lettres utilisent plus le vocabulaire usuel. Elles sont centrées sur la relation je-vous. Elles sont ancrées dans le temps, dans l'espace géographique et social. Elles privilégient le groupe verbal (verbes, pronoms et adverbes) au détriment du groupe nominal. Elles sont un substitut à la conversation.

En collaboration avec MONIERE Denis. Maurice Duplessis orateur : vocabulaire, style et axes de communication du chef de l'Union nationale. In MONIERE Denis (Ed.). *Maurice Duplessis vous parle. Discours recueillis et présentés par Denis Monière*. Québec : Société du patrimoine politique du Québec, 2009, p. 217-234. Repris dans GELINAS Xavier & FERRETTI Lucia (Eds.). *Duplessis, son milieu son époque*. Québec : Septentrion, 2010, p. 117-135.

52 discours ont pu être retrouvés dans diverses archives. Ils couvrent 30 ans de vie politique (1929-1959) et permettent de dresser un portrait lexical, grammatical et stylistique de M. Duplessis. A partir de 1938, son discours ne se renouvelle plus. Les discours de M. Duplessis étaient fortement ancrés dans l'espace national et dans son époque. Souvent teintés de religiosité, ils étaient peu personnalisés mais fortement polémiques. Ils étaient dominés par le conservatisme politique et social et par le souci de défendre les droits du Québec dans la fédération canadienne.

Les déclarations gouvernementales sous la Ve République (1959-1997). In AUTIN Jean-Louis et WEILL Laurence (Eds). *Le Droit figure du politique. Etudes offertes au professeur Michel Miaille*. Montpellier : Université de Montpellier I, 2008, tome I, p. 843-865.

Conférence prononcée à la Faculté de Droit de Montpellier le 3 décembre 1998. Présentation des mesures de distance intertextuelle, des classifications automatiques et arborées : méthodes, intérêt et fiabilité des mesures. Application aux discours inauguraux des Premiers ministres de la Ve République. Ces textes ne se partagent pas selon la coupure entre droite et gauche mais selon la conjoncture

politique du moment et selon la solidité plus ou moins grande des majorités parlementaires.

En collaboration avec MONIERE Denis. Des mots pour des voix : 132 discours pour devenir président de la République française. *Revue Française de Science Politique*. 58, 3 (2008), p. 433-455.

Analyse lexicométrique des discours des principaux candidats à l'élection présidentielle de 2007 : vocabulaire, thèmes, phrases, styles et stratégies de communication. S. Royal privilégie trois thèmes – pacte présidentiel, valeur travail et services publics – et attaque N. Sarkozy, alors que ce dernier donne priorité au pouvoir d'achat, à l'histoire de France et à l'identité nationale, et ne nomme pas ses adversaires jusqu'au second tour. Pour S. Royal, la politique est obligation morale et compétence ; pour N. Sarkozy, c'est affaire de conviction et de volonté. Enfin, les phrases de S. Royal sont plus longues, plus complexes, avec un vocabulaire peu spécialisé, contrairement à N. Sarkozy, dont les phrases sont très courtes, répétées, mais avec un vocabulaire adapté au thème traité.

En collaboration avec MONIERE Denis et LABBE Cyril. Les styles discursifs des premiers ministres québécois de Jean Lesage à Jean Charest. *Canadian Journal of Political Science / Revue canadienne de science politique*. 41:1, mars 2008, p. 43-69.

La forme, le style des hommes politiques sont révélateurs de leurs personnalités et de leurs stratégies de communication. Pour analyser les styles des premiers ministres québécois, un corpus comprenant 789 discours officiels prononcés de 1960 à 2005 a été soumis à une analyse stylistique quantitative. Nous avons analysé les catégories grammaticales, le maniement des verbes et des noms, la longueur et la structure des phrases. Cette comparaison statistique montre des différences significatives entre les premiers ministres et révèle également la stratégie de communication privilégiée par chacun d'eux.

En collaboration avec LABBE Cyril. Peut-on se fier aux arbres ? In HEIDEN Serge et PINCEMIN Bénédicte (Eds). *9e Journées internationales d'analyse statistique des données textuelles (Lyon, 12-14 mars 2008)*. Lyon : Presses universitaires de Lyon, 2008, volume 2, p. 635-645.

La distance intertextuelle fournit une solution simple et intéressante pour mesurer les proximités et les oppositions dans un grand corpus de textes. Ses propriétés en font un bon outil pour la classification des textes, spécialement pour l'analyse arborée qui est présentée et discutée. Deux indices sont proposés pour mesurer la qualité de ces classifications. La méthode fournit un outil efficace pour les études littéraires et l'attribution à des auteurs connus de textes d'origine douteuse ou inconnue, ainsi qu'il est démontré grâce à une expérience en aveugle.

En collaboration avec MONIERE Denis. Je est un autre ? In HEIDEN Serge et PINCEMIN Bénédicte (Eds). *9e Journées internationales d'analyse statistique des données textuelles (Lyon, 12-14 mars 2008)*. Lyon : Presses universitaires de Lyon, 2008, volume 2, p. 647-656.

Comparaison des fréquences d'un même mot dans deux corpus. On présente une méthode pour décider si les écarts constatés entre ces deux fréquences sont statistiquement significatifs. Les tests issus de la loi normale sont comparés avec les variations constatées dans l'utilisation de la première personne par les 4 principaux candidats à l'élection présidentielle française de 2007 (F. Bayrou, J.-M. Le Pen, S. Royal, N. Sarkozy). Il apparaît que la densité de la première personne n'est pas déterminée par des personnalités ou des styles propres aux individus mais par les stratégies de communication, le sujet traité et la situation d'énonciation.

En collaboration avec MONIERE Denis. Des mots pour des voix. *Communication aux 6e journées de la Société d'Etudes des langages du politique (SELP)*. Nice : Faculté des Lettres de Nice (29-30 novembre 2007).

Présentation du corpus des discours des 4 principaux candidats à l'élection présidentielle française de 2007 : F. Bayrou, J.-M. Le Pen, S. Royal, N. Sarkozy (dans cette communication, les exemples sont principalement tirés des corpus des deux finalistes). L'analyse porte d'abord sur le vocabulaire : les mots les plus employés, le sens de ces mots les plus fréquents, les principaux thèmes. Ensuite, l'étude des pseudo-auxiliaires, des pronoms, des groupes nominaux et verbaux révèle les tactiques et les stratégies de communication des candidats. Enfin, leur style particulier est défini par quatre dimensions : diversité et spécialisation du vocabulaire, longueur et complexité de la phrase. D'autres recherches sont signalées en conclusion : caractéristiques particulières du discours de campagne, tournants dans les campagnes électorale, influence des plumes de l'ombre. Enfin, on évoque un programme de recherche à propos des effets des discours électoraux sur les électeurs.

En collaboration avec LABBE Cyril. Baudelaire, Rimbaud et Verlaine. *Communication aux VIIIe Journées de l'ERLA*. Brest : 16-17 novembre 2007. Publié dans BANKS David (Ed). *Aspects linguistiques du texte poétique*. Paris, l'Harmattan, 2011, p. 17-45.

Existe-t-il un genre poétique particulier ? Qu'est-ce qui différencie ce genre poétique du reste de la littérature ? On répond à ces deux questions en utilisant d'abord les œuvres de Baudelaire, Rimbaud et Verlaine. L'analyse est ensuite étendue à l'ensemble de la littérature de la seconde moitié du XIXe siècle. La poésie en vers se distingue de la prose par des densités d'emplois différentes des parties du discours. La poésie privilégie le groupe nominal (substantifs, adjectifs et déterminants) ; la prose utilise plus de verbes, d'adverbes et de pronoms. On présente enfin le vocabulaire et les thèmes propres à la poésie versifiée.

En collaboration avec LABBE Cyril. Corneille a écrit 16 pièces représentées sous le nom de Molière. Réponses à : VIPREY Jean-Marie et LEDOUX Claude-Nicolas, 'About Labbé's "Inter-textual Distance". Grenoble : PACTE-IEP, 2007.

Un article paru en décembre 2006 dans le *Journal of Quantitative Linguistics*, sous le titre 'About Labbé's "Inter-textual Distance"', remet en cause l'attribution à P. Corneille de toutes les pièces en vers de Molière et de deux de ses pièces en prose (Dom Juan et l'Avare). Cet article présente des "expériences" qui ne concernent pas Corneille et Molière. Les calculs comportent de nombreuses erreurs. Par exemple, les ponctuations sont comptées comme des mots. Les exemples choisis se situent en dehors des limites de validité de l'échelle des distances. Les graphiques sont manifestement erronés. La distance intertextuelle sort renforcée et la dernière de ces "expériences" confirme que P. Corneille a bien écrit les pièces en vers de Molière ainsi qu'au moins deux de ses pièces en prose.

Experiments on Authorship Attribution by Intertextual Distance in English. *Journal of Quantitative Linguistics*. 14-1, 1, April 2007, p. 33-80.

How can it be said that texts are "near to" or "distant from" one another ? Are different texts by a single author more similar than texts by different authors ? To answer these questions, a method is proposed by calculating intertextual distance. A blind test and some additional experiments show that this calculation offers an interesting tool for non-traditionnal authorship attribution.

Compte rendu de "Michel Pinault. *La science au Parlement.*". *Histoire & mesure*. vol. XXII -1, 2007, p. 194-197.

Compte-rendu de : *Michel Pinault, La science au Parlement. Les débuts d'une politique des recherches scientifiques en France*, Paris, CNRS éditions, 2006. Texte intégral accessible sur le site de la revue.

En collaboration avec LABBE Cyril. La diachronie dans le discours politique. Le général de Gaulle. *Communication aux VIIe Journées de l'ERLA*. Brest : 17-18 novembre 2006. Publié dans BANKS David (Ed). *Aspects diachroniques du texte de spécialité*, Paris, l'Harmattan, 2010, p. 129-148.

Après avoir rappelé les règles de normalisation et de lemmatisation nécessaires pour le traitement informatique des textes en langue française, on présente une méthode capable de localiser avec précision les ruptures thématiques dans un texte ou un corpus. Une fois les sous-parties délimitées, on établit le vocabulaire et les thèmes caractéristiques de chacune d'elles. La méthode est d'abord appliquée à l'entretien accordé à M. Droit par le général de Gaulle entre les deux tours de l'élection présidentielle de 1965 puis à l'ensemble des interventions radiotélévisées du Général entre 1958 et 1969.

En collaboration avec LABBE Cyril. A Tool for Literary Studies: Intertextual Distance and Tree Classification. *Literary and Linguistic Computing*. 21-3, 2006, p. 311-326.

How to measure proximities and oppositions in large text corpora? Intertextual distance provides a simple and interesting solution. Its properties make it a good tool for text classification, and especially for tree-analysis which is fully presented and discussed here. In order to measure the quality of this classification, two indices are proposed. The method presented provides an accurate tool for literary studies -as is demonstrated by applying it to two areas of French literature, Racine's tragedies and an authorship attribution experiment.

En collaboration avec MONIERE Denis. L'influence des plumes de l'ombre sur les discours des politiciens. In Condé Claude et Viprey Jean-Marie. *Actes des 8e Journées internationales d'Analyse des données textuelles*. Besançon : 19-21 avril 2006, II, p. 687-696.

Par la comparaison des discours de deux premiers ministres québécois et par le calcul des distances intertextuelles, les auteurs tentent d'identifier les discours qui sont attribuables à une plume de l'ombre. Les auteurs montrent qu'il y a une forte différenciation entre les discours écrits et les discours oraux et qu'il n'y a pas qu'une seule plume à l'œuvre. Les discours des chefs de gouvernement amalgament une diversité d'influences et combinent des styles oraux et écrits. Ils contrôlent leurs discours même s'ils ne les écrivent pas eux-mêmes.

En collaboration avec LABBE Cyril. How to measure the meanings of words ? Amour in Corneille's work. *Language Resources Evaluation*. 2005, 39, p. 335-351.

We present a new method to describe the contextual meaning of a key word in a corpus. The vocabulary of the sentences containing this word is compared to that of the entire corpus in order to highlight the words which are significantly overutilized in the neighbourhood of this key word (they are associated in the author's mind) and the ones which are significantly underutilized (they are mutually exclusive). This method provides an interesting tool for lexicography and literary studies as is shown by applying it to the word amour (love) in the work of Pierre Corneille, the most famous French playwright of the 17th century.

En collaboration avec MONIERE Denis et LABBE Cyril. Les particularités d'un discours politique : les gouvernements minoritaires de Pierre Trudeau et de Paul Martin au Canada. *Corpus*, 4, 2005, p. 79-104.

Depuis la seconde guerre mondiale, l'État fédéral canadien a été gouverné à sept reprises par des équipes sans majorité au parlement : en 1957, 1962, 1963, 1965, 1972, 1979 et 2004. Dans cette étude, nous avons retenu deux cas : celui du gouvernement Trudeau 1972-1974 et celui du gouvernement Martin 2004 parce que dans ces deux cas, le même parti et le même chef ont assumé le pouvoir en situation majoritaire puis en situation minoritaire. Le vocabulaire caractéristique de ces deux équipes est mis en lumière grâce au calcul des "spécificités du vocabulaire". On présente ce calcul en détail : utilisation de la loi hypergéométrique et pondération en fonction de la catégorie grammaticale du mot étudié. Il apparaît que le discours gouvernemental minoritaire est nettement impersonnel ; il sous-utilise les verbes et se réfugie dans un catalogue d'actions de court terme.

En collaboration avec LABBE Cyril et HUBERT Pierre. Automatic Segmentation of Texts and Corpora. *Journal of Quantitative Linguistics*, december 2004, 11-3, p. 193-213.

Le découpage des grands corpus de textes est l'une des questions cruciales posées aux études littéraires. Il est proposé une double méthode. L'analyse de la croissance du vocabulaire (type-token ratio) met en lumière les principaux changements de rythme. Ces résultats sont complétés par l'étude de la diversité du vocabulaire. Un algorithme de segmentation, associé à un test de validité, indique le découpage optimal. La méthode est appliquée aux œuvres de Racine, Corneille et aux discours du Général de Gaulle.

En collaboration avec LAPIERRE Francis. L'évangile de Marc : un carrefour linguistique sémitico-grec. *Communication aux Ve Journées de l'ERLA*. Brest : 19-20 novembre 2004. Publiée dans BANKS David. *La langue, la linguistique et le texte religieux*. Paris : L'Harmattan, 2008, p. 85-100.

Analyse littéraire et lexicométrique des 4 évangiles. L'impact des différentes traductions en français est examiné. L'évangile de Marc contient un grand nombre de « doublets » correspondant à deux couches rédactionnelles. Cet évangile comporte un « canevas » d'origine sémitique de 250 versets environ (que l'on retrouve également dans les évangiles de Matthieu et de Luc). Un commentaire grec plus tardif a été ajouté à ce récit primitif.

Romain Gary et Emile Ajar. Grenoble : Cerat-IEP, mai 2004.

Note adressée à Benoît Peeters et Michel Lafon pour leur livre *Nous est un autre*. Application au cas Gary-Ajar des méthodes statistiques d'attribution d'auteur (distance intertextuelle, classification arborée, syntagmes répétés). Gary avait codifié un vocabulaire et un style qui était propres à Ajar. Mais l'existence d'un auteur unique ne fait pas de doute et démontre qu'il est difficile de déguiser son écriture du moins sur plusieurs centaines de pages.

THOMAS Jérémie. *Architecture de* Lexicométrie *et développement de son interface*. Rapport de stage sous la direction de LABBE Cyril et LABBE Dominique. Grenoble : Polytech'Grenoble et Cerat-IEP, août 2004.

Réalisation d'une interface permettant de gérer un ensemble de plusieurs milliers de textes en français (œuvres littéraires, discours politiques, articles de presse, transcriptions de l'oral, etc.) qui ont déjà fait l'objet d'un traitement préalable (normalisation graphique et lemmatisation). Il s'agit de réfléchir aux moyens de transformer cet ensemble en une "base de données" et de réaliser un logiciel pour la gérer : architecture, interrogation, modification. Le stage a porté sur la recherche d'un mot dans tout ou partie de cette base à la manière dont on consulte un dictionnaire en donnant la possibilité de combiner plusieurs critères de recherche : entrées de dictionnaire, catégories grammaticales, différentes orthographes. Le logiciel devrait également offrir plusieurs affichages possibles permettant à l'utilisateur de visualiser les contextes qui l'intéressent.

Deux réponses à Etienne Brunet à propos de Corneille et Molière. Grenoble : CERAT-IEP, avril 2004.

Deux réponses à des interventions de M. Brunet contestant notre attribution à Corneille des principales pièces parues sous le nom de Molière. Début 2003, avec la complicité de quelques littéraires et journalistes, M. Brunet s'est livré à une supercherie destinée à me discréditer mais dont les véritables victimes sont Baudelaire et Rimbaud. En 2004, il a fait paraître deux articles qui déforment sciemment nos travaux, prétendent faussement utiliser nos méthodes et contiennent plusieurs erreurs qui discréditent l'ensemble de sa prétendue démonstration.

Corneille in the shadow of Molière. French Department Research Seminar. Dublin : University of Dublin (Trinity College), April 6 2004.

Calculation of inter-textual distance. This euclidian metric allows unbiased classifications and the calibration of a distance scale for French texts. A large number of trials have validated this method. Applied to the works of Corneille and Molière, it reattributes to Corneille all the masterpieces in verse by Molière and two comedies in prose by the same: *Dom Juan* and *l'Avare*. Many stylistic and historic evidence corroborates this attribution. At least, we answer the main objections: possible "unconscious" imitation of Corneille by Molière, the influence of prosody on distances, the possible "quarrel" about *l'Ecole des femmes*.

Corneille et Molière. Table ronde 7e Journées d'Analyse des Données Textuelles. Louvain-la-Neuve 11 mars 2004. Grenoble : CERAT-IEP, 2004.

Compte-rendu des interventions de l'auteur lors de la table ronde, sur Corneille et Molière, réunie pendant des Journées d'Analyse statistique des Données Textuelles à Louvain-la-Neuve le 11 mars 2004. Un des participants ayant affirmé qu'il faut utiliser les "collocations" pour l'attribution d'auteur, il est démontré que les collocations des mots les plus fréquents confirment la paternité de Corneille sur les pièces en vers de Molière, ainsi que sur Dom Juan et l'Avare. En réponse aux interventions, on rappelle les limites de validité de l'échelle de la distance intertextuelle et, dans ces limites, l'indépendance de cette distance par rapport aux différences de longueur entre les textes.

En collaboration avec BERGERON Jean-Guy. Analyser les entretiens sociologiques. In PURNELLE Gérald, FAIRON Cédrick et DISTER Anne (Eds). *Le poids des mots. Actes des 7e journées internationales d'analyse des données textuelles*. Louvain-la-Neuve : Presses Universitaires de Louvain, 2004, p. 136-147.

On examine les problèmes posés par l'analyse des entretiens sociologiques à l'aide d'un corpus d'une soixantaine d'interviews à propos des relations industrielles dans les entreprises du Québec : normalisation orthographique, balisage et lemmatisation des textes. Il faudrait pouvoir comparer les enquêtés avec la population générale. Nous donnons un exemple de cette démarche à l'aide d'un corpus de plus de 300 entretiens réalisés avec des Français (plus de 2 millions de mots). Il apparaît que les Québécois préfèrent le groupe nominal et qu'ils utilisent moins d'adverbes et de conjonctions et que leurs propos sont nettement moins tendus que ceux des Français.

Corneille et Molière. *Séminaire du Groupe Langues Informations Représentations*. Université de Paris XI-Orsay : Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), 13 janvier 2004.

Comment calculer la distance entre plusieurs textes composant un corpus ? Cette distance euclidienne permet d'effectuer des classifications sans biais. Cette procédure montre que Corneille est l'auteur de toutes les comédies en vers et d'au moins deux en prose (Don Juan et l'Avare). On présente ensuite les tests qui ont été effectués pour valider cette procédure. Enfin, on examine les éléments historiques qui vont dans le sens de cette attribution et les objections qui ont été formulées par quelques critiques.

Coordination et subordination en français oral. IVe journées de l'ERLA *Coordination/subordination dans le texte de spécialité*. Brest : 14-15 novembre 2003. Reproduit dans : BANKS David (Ed.). *La coordination et la subordination dans le texte de spécialité*. Paris : L'Harmattan, 2007, p. 161-182.

Comment analyser le français oral ? On présente les principales étapes à l'aide d'un corpus de plus de 300 entretiens comptant au total 2,3 millions de mots : règles de transcription, normalisation orthographique, balisage des textes, étiquetage des mots. Pour décrire les caractéristiques propres à l'oral, il faut connaître celles du français écrit. Nous donnons un exemple à l'aide d'un corpus de plus de 2000 textes comprenant plus de 7 millions de mots. Les différences entre l'oral et l'écrit sont considérables : vocabulaire, syntaxe, structure des phrases. L'examen de la coordination et de la

subordination fait apparaître un système assez éloigné de celui que décrivent traditionnellement les grammaires du français. Des enquêtes d'usage et des corpus représentatifs sont indispensables pour une étude scientifique de la langue.

En collaboration avec LABBE Cyril. La distance intertextuelle. *Corpus*, 2, 2003, p. 95-118.

Présentation de l'indice de la distance intertextuelle et de ses propriétés. Discussion des limites du calcul : influence des décimales et des différences de taille entre textes comparés. Examen de la contribution à la distance des vocables classés en fonction de leurs catégories grammaticales et de leurs fréquences. L'indice de la distance intertextuelle fournit un outil intéressant pour la mesure des ressemblances et des dissemblances au sein des grandes bases de textes.

RUHLMANN Mathieu. *Analyse arborée. Représentation arborée par la méthode des groupements*. Rapport de stage sous la direction de LABBE Cyril et LABBE Dominique. Grenoble : Polytech'Grenoble et Cerat-IEP, août 2003.

Méthode pour représenter exactement sur un plan les dissimilarités entre les individus composant une population donnée. Transposition à l'analyse des textes des méthodes d'analyse arborée d'usage courant en biologie et en génétique. Calcul de la distance intertextuelle, méthode de groupement d'après les travaux de X. Luong, tassement de la matrice des données, procédure d'affichage, lecture des arbres. Application à deux corpus : le théâtre de Jean Racine et les discours présidentiels du Général de Gaulle.

PAEQUIN Gaétan. *Segmentation automatique des corpus. D'après les travaux de P. Hubert, J.-P. Carbonnel et A. Chaouche sur la segmentation des séries hydrométéorologiques pour application à la segmentation automatique des textes*. Rapport de stage sous la direction de C. Labbé et D. Labbé. Grenoble : Polytech'Grenoble et Cerat-IEP, août 2003.

Objectif du stage : rédiger un programme inspiré d'un algorithme de segmentation automatique inventé par trois hydrologues pour le traitement des séries climatiques longues et tester certaines améliorations à cet algorithme. Variable utilisée : la diversité du vocabulaire. Application à deux corpus : le théâtre de J. Racine et les discours présidentiels du Général de Gaulle.

En collaboration avec MONIERE Denis. Le vocabulaire gouvernemental en France, au Canada et au Québec : 1944-2000. *Etudes canadiennes*. 52, 2002, p. 103-116.

Cette analyse lexicométrique compare le vocabulaire utilisé dans les déclarations de politique générale de trois pays afin de déterminer si les différences de régimes et de cultures politiques induisent des choix lexicaux spécifiques ou si les gouvernements ont tendance à partager un univers lexical commun. On traite ici plus spécifiquement du vocabulaire servant à désigner le locuteur et ses destinataires ainsi que l'emploi des diverses catégories grammaticales. Grâce à des tests statistiques, nous tentons d'évaluer les effets liés à la période historique, aux contraintes institutionnelles, aux orientations idéologiques et à la personnalité des chefs de gouvernement.

La lexicométrie appliquée au discours politique. Le général de Gaulle. Séminaire ARCATI. Paris : décembre 2002.

Présentation des principaux outils de la statistique lexicale appliqués aux 79 interventions radio-télévisées du général de Gaulle entre juin 1958 et avril 1969. Les graphies sont standardisées et l'on attache à chaque mot une étiquette indiquant son entrée de dictionnaire et sa catégorie grammaticale (lemmatisation). Ces opérations préalables rendent aisée la consultation des textes, la mesure des principales caractéristiques du vocabulaire, la recherche du sens des mots, la localisation des ruptures thématiques et stylistiques. On présente ensuite les outils de classification indispensables pour le traitement d'un grand nombre de textes. Enfin, l'exemple des trois entretiens télévisés de décembre

1965 montre que les corpus étiquetés permettront des recherches inédites en science politique aussi bien qu'en linguistique ou en stylistique.

Le général de Gaulle en campagne. Communication aux IIIe Journées de l'ERLA *Aspects linguistiques du texte de propagande*. Brest : 15-16 novembre 2002. Reproduit dans : BANKS David (Ed.). *Aspects linguistiques du texte de propagande*. Paris : L'Harmattan, 2005, p. 213-233.

Alors que le général de Gaulle préparait toujours avec soin ses allocutions radiotélévisées, 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.

Qui a écrit quoi ? L'attribution d'auteur et la distance intertextuelle. Grenoble : CERAT, juillet 2002, 18 p.

Compte rendu d'une expérience en double aveugle réalisée avec Etienne Brunet qui a constitué un corpus de 50 textes anonymés afin d'éprouver notre méthode d'attribution d'auteur (détermination de l'auteur d'un texte d'origine douteuse ou inconnue). La distance intertextuelle, combinée à deux méthodes de classification automatique, se révèle un outil efficace.

La lemmatisation des grandes bases de textes. Un exemple : Corneille, Molière et Racine. Communication au colloque *L'édition électronique en littérature et dictionnairique, 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

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)*. Rennes : IRISA-INRIA, 2002, vol. 2, 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.

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)*. Rennes : IRISA-INRIA, 2002, Vol. 2, 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. Cela permet de repérer les différents sens possibles d'un mot, de réaliser des paraphrases, comparables à des définitions de dictionnaires, et d'isoler les phrases les plus caractéristiques de chacun de ces sens. 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 HUBERT Pierre et LABBE Cyril. Segmentation automatique des corpus. Voyages de l'autre côté de J.-M. Le Clezio. In MORIN Annie et SEBILLOT Pascale (Eds). *VIe Journées Internationales d'Analyse des Données Textuelles (Saint-Malo 13-15 mars 2002)*. Rennes : IRISA-INRIA, 2002, Vol. 1, 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 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 neighbourhood 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 Moliere's masterpieces.

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, p. 37-57.

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. Discrimination et classement au sein d'un groupe d'entretiens. Le cas du confort électrique. . *Communication aux journées d'études du CIDSP*. Grenoble : 9 mars 2001.

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.

En collaboration avec BERGERON Jean-Guy. L'évaluation de la négociation raisonnée par les acteurs. Une analyse lexicométrique. *Communication au XVI^e Congrès international de l'Association internationale des sociologues de langue française*. Québec : juillet 2000. Reproduit dans BERNIER Colette et Al. *Formation, relations professionnelles à l'heure de la société-monde*. Paris-Québec : L'Harmattan - Les Presses de l'Université Laval, 2002, p. 239-252.

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 panel 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 MONIERE Denis. La connexion intertextuelle. Application au discours gouvernemental québécois. In RAJMAN Martin et CHAPPELIER Jean-Cédric (Eds). *Actes des 5^e 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 BRUGIDOU Mathieu. Le vocabulaire syndical français à la lumière de l'analyse des données textuelles et de la statistique lexicale. In RAJMAN Martin et CHAPPELIER Jean-Cédric (Eds). *Actes des 5^e 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.

Analyse des données textuelles et Statistique lexicale (Textual Data Analysis and Lexical Statistics). *Conférence introductive aux 5^e 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.

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.

Compte rendu de VILLONE Massimo et ZULIANI Alberto (dir.). « *L'attività dei governi della Republica italiana (1948-1994)". Mots*. 62, mars 2000, p. 117-119.

Remarquable exemple de coopération interdisplicinaire 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.

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.

La richesse du vocabulaire politique : de Gaulle et Mitterrand. In MELLET Sylvie et 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 France chez de Gaulle et Mitterrand. In FIALA Pierre et 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.

En collaboration avec PIBAROT André et PICARD Jacques. Les syntagmes répétés dans l'analyse des commentaires libres. In MELLET Sylvie (dir.). *IVe journées internationales d'analyse statistique des données textuelles*. Nice : Université de Nice-Sophia Antipolis, 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 HUBERT Pierre. La connexion des vocabulaires. In MELLET Sylvie. *IVe journées internationales d'analyse statistique des données textuelles*. Nice : Université de Nice-Sophia Antipolis, 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 FABRE Cécile et 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.

En collaboration avec HUBERT Pierre. Vocabulary Richness. *Lexicometrica*. n° 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 circumtances. 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.

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 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.

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 et 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 coocurrences 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.

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 LABBE Cyril. *Que mesure la spécificité du vocabulaire ?* Grenoble : CERAT, décembre 1994 et juin 1997. Reproduit dans *Lexicometrica*, 3, 2001.

Cette note présente la formule 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.

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 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.

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.

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.

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.

Bibliographie des études en langue française sur le "discours socialiste". *Mots*. n° 22, mars 1990, p. 105-106.

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.

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 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.

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.

En collaboration avec HUBERT Pierre. Note sur l'approximation de la loi hypergéométrique par la formule de Muller. In LABBE Dominique, SERANT Daniel et THOIRON Philippe. *Etudes sur la richesse et la structure lexicales*. Genève-Paris : Slatkine-Champion, 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. Un modèle de partition du vocabulaire. In LABBE Dominique, SERANT Daniel et THOIRON Philippe. *Etudes sur la richesse et la structure lexicales*. Genève-Paris : Slatkine-Champion, 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 appararaître 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".

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.

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.

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.

La France et moi, Jacques Chirac (Analyse de la déclaration de politique générale du 9 avril 1986). Rapport de recherche. Grenoble : CERAT-IEP, avril 1986.

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.

Nous les communistes... *Mots*. n°10, mars 1985, p. 133-146.

Le dépouillement des résolutions politiques des congrès du PCF depuis 6O 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.

Structure de l'idéologie communiste : le cas du parti communiste français. *ECPR (European Consortium of political research)*. Freiburg : mars 1983.

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.

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.

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.