Lexicométrie : quels outils pour les sciences humaines et sociales ?

Cyril Labbé, Dominique Labbé

Laboratoire Printemps (Université de Versailles)
Association Internationale des Sociologues de Langue Française

Journées d'étude
Usages de la lexicométrie en sociologie
12-13 juin 2013

Lexicométrie : quels outils pour les sciences humaines et sociales ?

Cyril Labbé
Laboratoire d'Informatique de Grenoble - Université Joseph Fourier
(cyril.labbe@imag.fr)

Dominique Labbé
Laboratoire PACTE (CNRS - Institut d'Etudes Politiques de Grenoble)
(dominique.labbe@iep.grenoble.fr)

Résumé

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.

Abstract

Lexicometry is the alliance of linguistics, mathematics and computer science. It processes large sets of texts (corpus), establishing their vocabulary, classifying the words types in terms of their frequency, their distribution, their grammatical categories. It establishes the contexts of a word type and the most frequent combinations in which it appears. These collocations are used to determine the meanings of the terms. Lexicometry also highlights the main themes of a corpus, its gender and style. It segments the corpus according to the main thematic and stylistic discontinuities. To obtain these results, some pre-processing is necessary: text labeling, correction and standardization of spellings, lemmatization. Then the text can be placed into an electronic library available to researchers.

Le texte – manuscrit, transcription de l'oral, archive, imprimé – est le principal matériau des sciences humaines et sociales. Ce constat amène logiquement deux questions.

Premièrement, quels sont les objectifs des chercheurs quand ils analysent ce matériel textuel ? Quels outils utilisent-ils et en sont-ils satisfaits ?

Deuxièmement, quelle aide peut leur apporter l'alliance des sciences du langage – notamment la lexicologie et la lexicographie -, des mathématiques appliquées et de l'informatique ? c'est-à-dire la « statistique lexicale » (Muller 1977) (ou « lexicométrie » dans la suite de cet exposé).

Evidemment, la seconde question dépend de la réponse à la première. Or, la traditionnelle « analyse qualitative » (au mieux « analyse du contenu ») semble suffire à la grande majorité des chercheurs en sciences humaines et sociales. La seconde question n'aurait donc pas de pertinence ? Le succès des deux journées d'étude organisées par le laboratoire Printemps et le GT16 de l'AISLF montre que tel n'est pas le cas et que la seconde question ne peut être ignorée même si elle ne concerne qu'une petite minorité parmi les chercheurs concernés.

Notre communication commencera par énumérer rapidement les principales questions posées à propos des textes et les réponses apportées par la lexicométrie. Une seconde partie évoquera succinctement les principales opérations réalisées sur les « données textuelles » pour passer de celles-ci à une véritable lexicométrie.

Pour illustrer notre propos nous tirerons quelques exemples d'une vaste bibliothèque électronique comportant au total 6 470 textes et 27,6 millions de mots dont 2,8 sont des transcriptions de l'oral et intéressent plus particulièrement les sociologues (voir annexe 1). Ce sera aussi l'occasion de présenter les principales notions (en gras dans ce texte) et quelques travaux de notre équipe (bibliographie à la fin de cette communication), spécialement, une exploitation secondaire encore inédite de 44 entretiens semi-directifs réalisés auprès d'électeurs isérois dans le cadre du suivi de la campagne présidentielle de 2007 (enquête « Formation du jugement politique » *FJP07* : Denni & Al. 2007). On pourra comparer ces résultats à ceux obtenus avec les mêmes traitements sur les discours des candidats à cette même élection (Labbé & Monière 2008a).

I. QUELQUES QUESTIONS DES CHERCHEURS EN FACE DES TEXTES

De manière générale, les questions portent d'abord sur le vocabulaire d'un texte - ou d'un groupe de textes (corpus) - et sur les singularités de ce vocabulaire. La réponse à ces premières questions fait généralement surgir une foule de nouvelles interrogations…

Quel vocabulaire ?

La lexicométrie répond à une première question évidente : une unité de lexique est-elle présente dans le corpus analysé et, dans l'affirmative, avec quelle densité ?

L'index alphabétique

L'information de base sur un texte ou une collection de textes est la liste alphabétique (**index**) de toutes les unités du lexique ou **vocables** (**entrées du dictionnaire** et **catégories grammaticales** correspondantes) avec leurs **effectifs** (c'est-à-dire le nombre de leurs **occurrences**). Cette liste donne, pour chaque vocable, les **formes graphiques** sous lesquelles il s'actualise.

Prenons l'exemple de *être*, verbe le plus utilisé dans tout corpus en langue française (tableau 1). Dans notre bibliothèque, les formes les plus fréquentes de ce verbe sont la troisième personne de l'indicatif présent (*est*), l'infinitif (*être*), le participe passé (*été*). Dans les entretiens, comme dans la majorité des textes de la bibliothèque, la première personne du singulier de l'indicatif présent (*suis*) arrive en quatrième position. Il en est pratiquement toujours ainsi, en français, dès que le texte est suffisamment long. Or ces quatre formes sont ambigües…

Tableau 1. Extraits de l'index alphabétique de la bibliothèque électronique (effectifs absolus, total 27,398 millions de mots)

(…)	
est (nom masculin)	1 108
(…)	
étai (nom masculin)	
étais	2
(…)	
été (nom masculin)	
été	1971
étés	43
(…)	
être (verbe)	
est	355 243
étais	6 865
été	36 410
être	47 355
fût	2 408
soit	16 571
sommes	11 366
suis	29 091
(…)	
être (nom masculin)	
être	1 219
êtres	764
(…)	
fût (nom masculin)	
fût	8
(…)	
soit (conjonction)	4 293
(…)	
somme (nm)	
sommes	95
(…)	
somme (nf)	
sommes	574
(…)	
sommer (verbe)	
somme	14
sommes	1
(…)	
suivre (verbe)	
suis	493
(…)	

Le tableau révèle que plusieurs formes graphiques du verbe *être* sont « **homographes** » avec d'autres formes notamment des substantifs, d'autres verbes, une conjonction…. Ce problème est bien connu : en français, la majorité des mots usuels peuvent être rattachés à plusieurs entrées de dictionnaire.

Bien sûr, certaines collisions sont rares. Par exemple, le substantif *est* représente 0,4% de toutes les apparitions de la forme correspondante et le verbe *être* les 99,6% restants, de telle sorte qu'on peut toujours imaginer commettre une erreur négligeable en rattachant toutes les formes *est* au seul verbe. Au vu du tableau 1, on pourrait appliquer le même raisonnement à l'infinitif, au participe passé ou à *suis*. Mais, ce faisant, on se condamne à passer à côté de faits intéressants. Par exemple, jusqu'à la chute du mur de Berlin, les politiques français – non-communistes – utilisaient « l'est » pour désigner *l'Union Soviétique* et ses satellites sans les nommer… Ou encore, quelques écrivains - spécialement J.-M. Le Clézio, prix Nobel de littérature 2008 -, éprouvent une prédilection marquée pour la belle saison (*l'été*). De même, dans la poésie – spécialement chez V. Hugo – il est beaucoup question de *l'être* (nom masculin). Ces quelques exemples suffisent pour comprendre que les mots rares ne peuvent être négligés a priori : le **réseau sémantique** dans lequel ils s'insèrent peut leur donner une grande importance ; une faible fréquence en langue ne se retrouve pas forcément au niveau de tous les genres, de tous les corpus et de tous les locuteurs.

Enfin, dans certains cas, ce postulat de l'erreur négligeable ne tient pas. Ainsi la conjonction « soit » représente une occurrence sur cinq de la forme graphique (les 4 autres étant le verbe homographe). Dans ce cas, non seulement, on se condamne à ne pas pouvoir étudier la conjonction en français, mais on commet une erreur évidente en rattachant toutes les formes au verbe !

Ces exemples rappellent aussi que certains vocables sont répartis d'une manière plutôt homogène alors que d'autres sont localisés dans quelques textes et quelques auteurs. L'index associe donc aux effectifs un **indice de répartition** qui rend compte de cette information (Hubert & Labbé 1990).

Naturellement, ce qui intéresse principalement le chercheur, ce sont les vocables les plus fréquents (**index hiérarchique**)

L'index hiérarchique

Dans l'**index hiérarchique**, les principaux vocables sont classés par catégorie grammaticale et **rangés** par ordre décroissant de **fréquences –** effectifs divisés par le nombre de mots (ou **longueur**) du corpus - afin de pouvoir comparer des corpus de longueurs inégales. Les fréquences sont exprimées en « pour mille mots » et non en pour cent, du fait des faibles effectifs. L'annexe 2 reproduit le début de l'index hiérarchique pour les substantifs, adjectifs, pronoms et verbes employés par les enquêtés FJP07 comparés aux discours des candidats à l'élection présidentielle (Labbé & Monière 2008a).

Dans tout discours oral improvisé, *chose* figure parmi les substantifs les plus utilisés, car il peut se substituer à tout objet dont le nom n'est pas immédiatement présent à l'esprit du locuteur. Son rang beaucoup plus modeste chez les candidats à cette élection de 2007 (18^{e}) indique que la grande majorité de leurs propos n'étaient pas improvisés – même dans leurs entretiens et leurs débats radiotélévisés. De même, chez les enquêtés, « gens » remplace « Français » et « peuple », chez les hommes politiques.

Les adjectifs « *vrai, sûr, clair, normal, possible* » (combinés avec *être*) sont à la fois la marque de l'oral – ils sont sur-employés par rapport au français écrit – et sont les qualités essentielles qu'attendent les enquêtés, à la fois du discours politique et des élus ; ceux-ci en revanche les emploient beaucoup moins. On remarque également que, dans tout texte en français, les adjectifs *petit* et *grand* sont les plus employés. Ici les enquêtés privilégient *petit*

(problème), les candidats : *grand* et *nouveau* (politique, problème)… Enfin, *social* est nettement plus employé par les candidats que par les enquêtés.

Pour les verbes, dans tout texte en français, *être*, *avoir*, *faire* occupent les trois premières places dans cet ordre. A l'oral, *dire* vient généralement en 4e position. La suite du classement est plus révélatrice. Elle se compose toujours des principaux **auxiliaires modaux** (Labbé et Labbé 2010b) qui donnent au discours une orientation particulière vers la volonté (*vouloir*), le possible (*pouvoir*), l'obligation (*devoir*), la nécessité (*falloir*), la connaissance (*savoir*), le constat (*voir*), le futur immédiat (*aller*), le passé proche (*venir*).

Pour les enquêtés, le classement est : futur proche (*aller voter* : le scrutin est la principale borne temporelle), possible, connaissance, nécessité, volonté ; pour les hommes politiques : volonté, possible, futur proche, obligation, nécessité, connaissance. Chez les électeurs, *le possible, la connaissance et la nécessité sont plus importants que la volonté*. En revanche dans le discours des candidats, *la volonté rend les choses possibles* !

Quand un vocable est attesté dans un corpus, le chercheur pose habituellement deux questions complémentaires : quels contextes d'emploi ? et quel(s) sens le ou les auteurs donnent-ils à ce vocable ?

Contexte d'un vocable

En ce qui concerne la première question, les **concordances** donnent un contexte minimal standard (une ligne d'imprimante), et offrent la possibilité d'élargir cette fenêtre quand le cas est jugé intéressant. A titre d'exemple, l'annexe 3 donne le début de ces concordances pour le verbe *suivre* (sous la forme *suis*) dans un corpus d'entretiens réalisés par S. Pionchon (2001) à propos des *femmes et la politique* et révèle que certaines femmes – contrairement à la plupart des hommes - avouent volontiers ne pas *suivre l'actualité*, spécialement les *informations politiques*…

Naturellement, le programme (**concordancier**) doit retrouver toutes les attestations du vocable – sans en oublier -, mais uniquement celles-ci…

Ces concordances sont la matière première du lexicographe (Blumenthal & Al. 2006). C'est grâce à elles qu'il peut établir les différents sens d'un vocable et sélectionner les exemples les plus caractéristiques. Mais, en français, plus du tiers des mots sont homographes (notamment tous les verbes usuels, comme *est, été, fût, soit, sommes*…) Les concordances établies sur les simples formes graphiques sont donc difficiles à exploiter. Par exemple, comment le lexicographe pourrait-il retrouver les mille substantifs « est » dans une liste de plus de 356 000 lignes ?

Dès que la concordance dépasse quelques centaines de lignes, la seconde question est inévitable : comment synthétiser cette information trop abondante ? C'est-à-dire avant tout, avec quel(s) autre(s) mot(s) le vocable recherché est-il le plus souvent associé ?

Pour répondre à cette question, on utilise les « **collocations** » (Blumenthal & Al. 2006) et les « segments répétés » (Salem 1987). Mais le recensement de ces combinaisons de formes est rendu difficile par les homographies, les flexions des verbes et par certains « mots outils » (pronoms, conjonctions, prépositions, articles, adverbes) qui s'intercalent dans ces constructions. Les **syntagmes répétés** dépassent ces difficultés (associations des vocables par catégories grammaticales : Pibarot et Al 1998). Par exemple, appliquée aux combinaisons des

verbes, cette méthode révèle une particularité largement méconnue du français : l'importance des « **modalités verbales** » ou « pseudo-auxiliaires » qui régissent un autre verbe employé à l'infinitif (Labbé & Labbé 2010b).

L'annexe 4 donne les principales modalités verbales utilisées par les enquêtés FJP07 et les candidats à l'élection présidentielle. Enquêtés et candidats partagent la même préférence pour « vouloir dire », mais les seconds l'utilisent 3.5 fois moins que les premiers parce qu'ils se débrouillent pour se trouver le moins possible en situation de véritable interaction verbale. Alors que les enquêtés utilisent surtout ce syntagme pour préciser leur propos, les seconds l'emploient pour affirmer leur détermination à prendre position par le discours…

On peut également mesurer la productivité de chacun des auxiliaires modaux. Cela permet de montrer la prédominance de la volonté dans le discours électoral et celle du possible – et du futur immédiat - dans les propos des électeurs.

Enfin, la dernière ligne des deux tableaux en annexe montre que les 20 premières combinaisons sont deux fois moins employées chez les candidats que chez les enquêtés. Parce que leurs discours sont soigneusement préparés, mais aussi parce qu'ils sont des professionnels de la parole, les hommes politiques parviennent à diversifier leur propos et à éviter les chevilles qui émaillent la conversation quotidienne.

L'analyse porte également sur les combinaisons « nom + nom » (ou nom + adjectif) qui permettent de préciser les principaux thèmes du corpus.

En général, ces listes des combinaisons les plus fréquentes (et les plus singulières) font surgir de nouvelles interrogations, notamment à propos du sens que le ou les locuteurs donnent aux mots les plus utilisés.

Le sens d'un vocable ?

Il s'agit de reconstituer le **réseau sémantique** du vocable concerné, c'est-à-dire *l'ensemble des relations de synonymie, d'hyperonymie et d'antonymie qui unissent ce vocable à tous les autres constituant le vocabulaire du corpus considéré*, voire à tous les vocables d'un lexique de spécialité ou du genre auxquels appartient le texte ou le corpus étudiés. En quelque sorte, il s'agit d'automatiser l'atelier du **lexicographe** – rédacteur de dictionnaires - en demandant à l'ordinateur de reconstituer l'équivalent d'un article de dictionnaire (Labbé D. 2010).

Voici quelques illustrations de cette lexicographie assistée par ordinateur :

- dans Leselbaum et Labbé 2002, le sens du mot « banque » dans la presse économique de la fin des années 1990 ;

- dans Arnold 2008, on verra que, chez T. Blair, Premier ministre anglais (de 1997 à 2007), le contenu de mots aussi courants que *people* ou *Europe* était assez éloigné de celui qu'un Français leur attribuerait spontanément. *People* est toujours employé au pluriel (*people are*) et signifie « les gens » jamais « le peuple ». Dans le discours politique anglais, l'entité politique souveraine est la nation pas le peuple (au sens français). Le logiciel donne les **phrases caractéristiques** (celles qui contiennent une proportion anormalement forte de vocables associés au mot recherché) – un peu à la manière des citations illustratives d'un article de dictionnaire. Voici la phrase la plus caractéristique de l'Europe chez T. Blair :

> What's fascinating about Europe is these new countries coming in from central and eastern Europe are countries that totally share Britain's view of Europe not as a super state, but as a Europe of nations, Europe standing strong with the United States of America and the transatlantic alliance, Europe reforming itself economically (31 mai 2003).

Cette phrase de 59 mots en contient 30 qui sont associés à *Europe*, au premier rang desquels : *country(ies), nation(s), America, transatlantic alliance*… L'*Europe* de T. Blair c'est une association de puissants Etats-nations, le refus d'une administration européenne, l'alliance privilégiée avec les Etats-Unis, le leadership britannique et les réformes économiques libérales.

Autre exemple : la conception de la *France* chez Marine Le Pen comparée avec les autres candidats à l'élection présidentielle de 2012 (Labbé & Monière (2013). Car la méthode permet de comparer plusieurs locuteurs (Labbé D. 1998).

Ce calcul permet aussi de répondre à une autre question qui vient généralement dans le prolongement de la précédente :

2. Quel est le vocabulaire caractéristique ?

Le **vocabulaire caractéristique** - *d'un texte, d'un corpus, d'un locuteur ou d'un groupe de locuteurs - est constitué des vocables et catégories grammaticales que ce ou ces textes ou locuteurs sur-emploient ou sous-emploient par rapport à un groupe de référence*. Cette définition soulève au moins deux questions préalables : l'ensemble de référence et la méthode de mesure.

Quel ensemble de référence ?

Ecartant les critères *a-priori* (genre, catégories socioprofessionnelles, classes d'âge, etc.), on recherche le « meilleur » ensemble de référence possible pour effectuer cette recherche.

La **classification** assistée par ordinateur permet d'ordonner une grande collection de textes en quelques groupes (composés des textes les plus proches possibles et les plus contrastés entre eux). Elle nécessite une réflexion préalable sur la mesure de la **distance** entre textes – « **intertextuelle** » - et sur les méthodes de classification les plus adaptées. Cette question est notamment discutée dans Labbé & Labbé 2004 ; 2006 ; 2011a. Par exemple, la méthode proposée se révèle particulièrement utile lorsque le sociologue doit analyser un grand nombre d'entretiens (Bergeron & Labbé 2000 ; Labbé & Labbé 2001 ; Labbé D. 2002a) ou les réponses aux questions ouvertes dans un sondage (Labbé & Labbé 2012a).

La même méthode répond à beaucoup d'autres questions intéressantes comme :

- l'identification des plumes de l'ombre des politiciens et la mesure d'une éventuelle influence des premiers sur les seconds (Monière & Labbé 2006) ;

- l'attribution à un auteur connu de textes anonymes ou douteux (Merriam 2002, 2003, 2005 ; Labbé D. 2007b) ;

- la détection des plagiats et au sein de ceux-ci des principaux passages copiés (Labbé & Labbé 2012c) ;

- les fausses communications scientifiques (Labbé & Labbé 2012b)

Etc.

Une fois défini le « bon » ensemble de référence comment déterminer les singularités d'un texte ou d'un groupe de textes ?

Caractéristiques du vocabulaire d'un individu par rapport à l'ensemble de référence

La question est connue sous le nom de « spécificités du vocabulaire » (Lafon 1984 ; Labbé & Labbé 1994). Ce calcul est implanté dans la plupart des logiciels d'analyse des données textuelles. Mais il comporte plusieurs défauts généralement méconnus (Labbé & Labbé 1994 ; Monière & Labbé 2012). Une méthode neutralise ces problèmes (Monière & Al. 2005 ; Monière & Labbé 2012). L'expression **vocabulaire caractéristique** souligne que ce calcul n'est pas celui des « spécificités du vocabulaire ».

Cette méthode permet de comparer, pour un même locuteur, plusieurs sous-ensembles et de répondre ainsi à des questions intéressantes. Par exemple, les hommes politiques tiennent-ils le même discours quand ils sont candidats et quand ils sont au pouvoir (Savoy 2010, Monière & Labbé 2010a) ou quand ils président successivement un gouvernement majoritaire puis un gouvernement minoritaire, dont la survie dépend de l'opposition ? (cette situation s'est produite plusieurs fois au Canada : Monière et Al. 2005).

Ces réponses font habituellement surgir encore d'autres interrogations que l'on va évoquer.

3. Quels sont les thèmes, le style, le genre… ?

La détection des principaux thèmes est une question soulevée quand on étudie de vastes corpus avec un grand nombre de locuteurs. A l'opposé, les questions portant sur le style ou le genre se posent plutôt pour les corpus plus réduits avec un ou un petit nombre de locuteurs.

Thématique

L'**analyse automatique des thèmes** parvient à deux résultats principaux :

1. Le passage d'un vocabulaire de plusieurs milliers de vocables à quelques thèmes, l'identification du contenu de chacun de ces thèmes, la mesure de leurs poids relatifs, c'est-à-dire de l'importance que leur accorde chaque locuteur. Citons deux applications récentes :

- la campagne présidentielle française de 2012 : plusieurs milliers de messages comptant 1,7 millions de mots – et un vocabulaire de 19 600 vocables - peuvent être réduits à une quarantaine de thèmes dont on mesure, pour chaque candidat, le contenu, le poids relatif et l'évolution au cours de la campagne (Labbé & Monière 2013) ;

- dans les 1 500 réponses à une question ouverte dans un sondage, la méthode isole 8 thèmes principaux et reclasse chaque réponse en fonction de l'adhésion ou du rejet de l'enquêté face à chacun de ces thèmes (Labbé & Labbé 2012a).

2. La **localisation des ruptures thématiques** et la **segmentation automatique** (découpage) d'un corpus en fonction de ces ruptures. Cette méthode est empruntée à la climatologie (Hubert, Carbonnel & Chaouche 1989) : quand peut-on considérer qu'une variable - précipitations, températures, débit d'un fleuve… - s'écarte de l'état stationnaire et à partir de quel moment peut-on considérer qu'un nouveau palier est atteint ? L'adaptation de ce calcul aux textes a été présentée dans : Labbé, Labbé & Hubert 2004.

Parmi les nombreuses applications, voici deux exemples :

- les discours des huit premières années du Premier ministre anglais T. Blair : la méthode met en valeur le tournant décisif induit par la guerre en Irak (Arnold 2005).

- D. Monière a rassemblé tous les discours d'ouverture des sessions annuelles du parlement québécois depuis l'institution de la Confédération canadienne (1867). La méthode a permis de découper automatiquement ce vaste corpus en une série de périodes homogènes et de caractériser le vocabulaire et les thèmes dominants de chacune de ces périodes (Monière & Labbé 2010). Cette dernière publication présente également une amélioration de la segmentation automatique.

Style et genre

De nombreuses questions sont aussi posées, notamment par les littéraires. Par exemple :

- caractériser et comparer les **styles** de plusieurs auteurs ou locuteurs (Monière & Labbé 2002 ; Monière, Labbé & Labbé 2008).

- les structures et le contenu des **phrases** en fonction de leurs longueurs (Labbé & Labbé 2010a).

- les **genres** et les caractéristiques de chacun de ces genres. Par exemple, le français oral (Labbé D. 2007a), la correspondance (Labbé & Labbé 2009), la poésie (Labbé & Labbé 2011b) ou le discours politique (Labbé & Labbé 2011c).

Deux remarques pour conclure cette rapide présentation.

Premièrement, nous n'avons pu aborder toutes les analyses possibles et encore moins illustrer la richesse des conclusions. Chacun pourra ajouter de nouvelles recherches en fonction de ses intérêts personnels. Il suffit de formuler la question d'une manière qui soit « programmable » afin de la traduire en modèles, tests statistiques, algorithmes et programmes informatiques… Car pour chaque question, on procède un peu de la même manière : modélisation du phénomène, formulation d'hypothèses, constitution de corpus, enregistrement des variables et tests.

Deuxièmement, il faut travailler sur les unités de lexique (les vocables) et non pas sur les formes graphiques. Cela nécessite quelques opérations qui vont maintenant être décrites.

II. DU TEXTE A LA BIBLIOTHEQUE ELECTRONIQUE

Les 44 entretiens FJP07 sont des transcriptions de qualité ayant fait l'objet d'une correction orthographique et d'une relecture par l'enquêteur. Ces transcriptions suivent les règles courantes de la « sténographie » en tenant compte des hésitations, des reprises mais pas des prononciations singulières. Ces principes avaient été retenus pour la constitution de la section orale du British National Corpus (Nelson 1997). Ils permettent de comparer l'oral et l'écrit mais ils laissent de côté les propositions plus ambitieuses formulées par Blanche-Benveniste et Al. (1987 & 1990).

Outre la relecture soignée, trois opérations interviennent avant les traitements lexicométriques (Labbé D. 1990 et 2002b).

Trois opérations préalables

Balisage

Trois balises sont placées en tête de chaque texte électronique (Tableau 2 ci-dessous).

Tableau 2. Balises en-tête des fichiers de la bibliothèque (Entretiens FJP 07)

<FJP07-01 – Entretien sur la campagne électorale – Réalisé par… - Mars 2007 - Enquête sur la formation du jugement politique – PACTE – IEP Grenoble> <Transcription par …. (nom et institution) le… (date de la transcription)>
<Balisage – correction et standardisation orthographiques – lemmatisation : Dominique Labbé - PACTE Grenoble – 2010>

Remarques :

- le nom de l'enquêté n'apparait pas. Il est remplacé par un code (ici « FJP07-01 » qui renvoie à un autre fichier – non communiqué celui-ci – dans lequel figurent toutes les caractéristiques de l'enquêté(e) : sexe, âge, profession, situation familiale, niveau d'étude, commune de résidence, intention de vote, etc. Comme il est d'usage, le nom et l'adresse ont été détruits dès la fin de l'enquête… Naturellement, on a effacé dans l'entretien tout ce qui permettrait de reconnaître des personnes privées (par opposition aux personnages publics).

- les noms des personnes qui ont effectué l'entretien (première ligne), la transcription (deuxième ligne) et les opérations qui vont maintenant être décrites. Ces noms n'apparaissent que si ces personnes sont d'accord pour que leurs identités soient communiquées. Dans ce cas, toute publication issue d'une exploitation secondaire doit mentionner ces noms. En effet, certains opérateurs préfèreront rester dans l'ombre mais d'autres peuvent légitimement souhaiter que leur travail bénéficie d'un minimum de reconnaissance...

D'autres balises isolent les séquences du texte : remarques, questions et réponses, incidents et « didascalies » (ci-dessous : tableau 3).

Tableau 3. Le balisage des textes de la bibliothèque (entretien FJP07)

<Question 1 Alors ma première question sera la suivante: vous habitez dans la commune d'Echirolles - pouvez vous me dire ce que vous pensez de l'évolution de votre cadre de vie depuis que vous êtes ici? > <Réponse 1> Oh je pense que ça a bien marché. <rires> Bon c'est vrai qu'on… Je regrette un petit peu, euh, un petit peu le développement urbain. <Question 2> <inaudible sonnerie du téléphone interruption> <Réponse 2>

Le bénéfice de l'opération est évident : on peut étudier uniquement les réponses — c'est habituellement l'objet d'une exploitation secondaire — ou les questions, si l'on s'intéresse à la technique de l'entretien… etc.

Cette opération ne supprime rien, au contraire, elle rend disponible toute l'information recueillie.

Standardisation des graphies

Pour faire comprendre l'intérêt de ces opérations, voici quelques-unes des fluctuations graphiques rencontrées dans les textes originaux de notre bibliothèque :

- *M., Mr., Monsieur, monsieur*… Un automate peut reconnaître le même mot dans les trois dernières formes mais la première doit être corrigée à la main : *monsieur, Marcel, Maurice, Marie, mètre(s)*…? De même, dans les entretiens FJP07, pour *S. Royal* et N. Sarkozy (qui apparaissent également avec leur prénom en entier), etc.

- *Khadafi, Kadhafi, Gadhafi, Gaddafi, al-Gaddafi*…. Il y a probablement d'autres manières encore de transcrire ce nom. Pour Mao, dans les textes qui ont alimenté notre modeste base, il y a 8 graphies différentes, sans compter les fautes d'orthographe. Par exemple, dans FJP07, il y a aussi (entre autres) : Ségolène *Royale*, Nicolas *Sarkosy*, Simone *Veille*… Or, il est essentiel de pouvoir signaler à l'usager de la bibliothèque électronique tous les textes qui parlent d'une même personne, sans en omettre ni l'obliger à imaginer toutes les variantes graphiques, voire les fautes d'orthographe possibles…

- *19H, 19h., 19 h., 19 heures, dix-neuf heures, évènement, événement, puis, peux,* etc.

Dans tout texte imprimé, ces **variantes graphiques** concernent plus d'un mot sur dix (sans compter les fautes d'orthographe, les majuscules initiales de phrases, les noms communs affublés à tort d'une majuscule). Si l'on opère des calculs sur le texte « brut » sans aucune standardisation des graphies – comme le font les logiciels usuels d' « analyse des données textuelles » - cela signifie que les résultats obtenus sont affectés d'une incertitude d'au moins 10%... D'où des questions évidentes : à quoi sert-il d'opérer des calculs sophistiqués sur des données dont l'enregistrement comporte une telle incertitude ? Comment comparer les corpus entre eux ? Etc.

Enfin, *moi-même, aujourd'hui, parce que, La Fontaine, c'est-à-dire* forment un seul mot et non pas deux ou quatre. Il n'est donc pas possible d'utiliser le tiret, l'apostrophe ou l'espace comme « caractères délimiteurs » de mot, du moins si l'on veut étudier les pronoms, les adverbes, les conjonctions – c'est-à-dire le lexique, matière première de la lexicométrie ! - ou retrouver tous les textes parlant du célèbre fabuliste… De plus, ces problèmes de délimitation concernent une proportion significative des mots, ce qui soulève les mêmes questions que celles déjà posées à propos des fluctuations graphiques.

Une fois les textes balisés et les graphies standardisées, on procède à leur étiquetage.

Etiquetage

Cette troisième opération préalable est la plus complexe et la plus décisive. Pour en comprendre la portée considérons deux "textes" :

(*A*) La secrétaire lui dit : «je suis le président ».

(*B*) «Je suis le président », lui dit le président.

Apparemment, les deux vocabulaires sont identiques à un mot près (secrétaire/président).

Cependant, (A) est ambiguë. On peut comprendre que la secrétaire déclare qu'elle *suit* le président (elle fait partie des personnes qui l'accompagnent). Mais il se peut aussi que cette personne assure à la fois le secrétariat et la présidence. En effet, en France, certaines fonctions sont désignées par le masculin, même lorsque le titulaire est une femme (Madame *le ministre*,

Madame *le président*)… En général, le contexte permet de lever ces ambiguïtés. Mais quand on consulte des listes de mots (tableau 1), le contexte est perdu…

Dans l'exemple ci-dessus, en suivant l'ordre alphabétique, il faut résoudre les difficultés suivantes :

- *dit* : verbe dire (troisième personne du singulier de l'indicatif présent ou du passé simple ou participe passé masculin singulier) ; adjectif masculin singulier (*un homme dit de droite*) ; nom masculin singulier (*un dit*) ;

- *Je* n'est pas un nom propre (majuscule initiale), mais le pronom personnel placé en début de phrase. L'analyse doit donc le traiter comme s'il avait une minuscule initiale (c'est sa **graphie standard**) ;

- *la* : article défini féminin singulier (le) ; pronom relatif troisième personne féminin singulier (le) ; substantif féminin (note de musique) ;

- *le* : article défini masculin singulier ; pronom relatif troisième personne masculin singulier ;

- *lui* : verbe *luire* au participe passé masculin singulier ; pronom personnel de la troisième personne du singulier (des deux genres quand il est employé comme complément) ;

- *président* : verbe *présider* à l'indicatif présent troisième personne du pluriel ; nom masculin singulier.

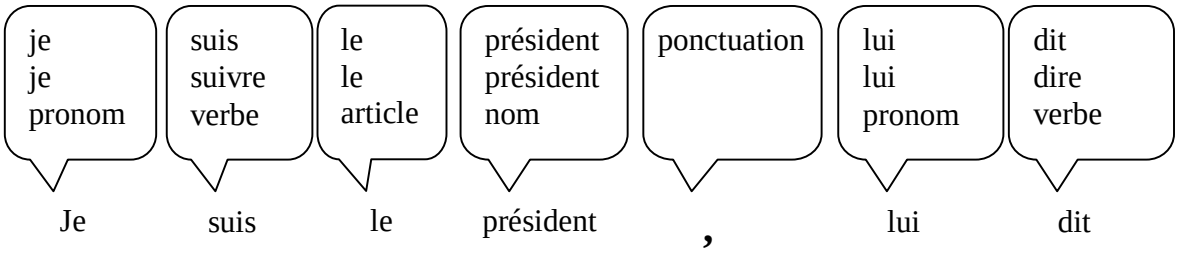
- *secrétaire* : nom féminin singulier ; nom masculin singulier ;

- *suis* : verbe *être* à l'indicatif présent première personne du singulier ; verbe *suivre* à l'indicatif présent première ou seconde personnes du singulier.

Cette **nomenclature** fait consensus parmi les lexicographes. Par exemple, en français, toutes les flexions d'un verbe (modes, temps, personnes) sont regroupées sous l'infinitif de ce verbe ; de même les substantifs sont identifiés par leur genre (président/présidente). Ainsi, "un secrétaire" peut être un homme ou un meuble mais pas une femme...

Quand l'ordinateur a commencé à être utilisé pour le dépouillement des textes, il a été proposé d'utiliser cette nomenclature, notamment par Muller dans un article de 1963 et dans son manuel de 1977. Nous avons réalisé l'implémentation de ces normes dans des algorithmes et des programmes informatiques (Labbé D. 1990). L'opération consiste à ajouter une étiquette à chacun des mots du texte (tableau 4 ci-dessous).

Tableau 4. Exemple d'étiquettes attachées à chacun des mots du texte



L'étiquette vient s'ajouter au texte sans l'altérer. Elle comporte trois informations :

- la **graphie standard** : majuscule initiale des mots communs ramenée en minuscule (comme pour "Je"), réduction des formes multiples à une graphie unique.

- puis le **vocable**, c'est-à-dire l'**entrée** où se trouve la graphie standard dans le dictionnaire et la catégorie grammaticale, telle qu'elle figure en seconde position dans cette entrée de dictionnaire.

Naturellement,

- ces étiquettes s'ajoutent au texte original qui n'est pas modifié. On ne perd donc aucune information ; on en gagne beaucoup…

- les opérations de standardisation des graphies et de lemmatisation sont confiées à des automates qui résolvent la quasi-totalité des cas et qui, pour les quelques ambiguïtés restantes, proposent à l'opérateur les différentes solutions possibles.

La nomenclature des mots, apprise à l'ordinateur, est systématique, exhaustive (tous les mots y trouvent leur place), univoque (une seule entrée par mot), elle exclut tout double compte, elle ne comporte pas de catégorie ad hoc, ou fourre-tout, etc.

Enfin l'opération est réversible : *on peut retrouver le texte original, sans altération, à partir du fichier étiqueté.*

Pour bien comprendre les opérations présentées ci-dessous, il faut se souvenir qu'un texte est une succession de **mots** (en anglais « **word tokens »**, emplacements**)** – dont le nombre total donne la **longueur** – ces mots étant issus d'un **vocabulaire** nécessairement plus restreint puisque certains **vocables** (en anglais « **word types** ») sont employés plusieurs fois dans le texte. Par exemple, "le", "les", "la", "l´" – et leurs équivalents avec une majuscule initiale - sont les différentes **graphies** sous lesquelles l'article ou le pronom "le" apparaissent dans un texte. "le, article" et "le, pronom" sont des **vocables** (ou "entrées de dictionnaire"). Chacune des **occurrences** de ces deux vocables – sous les formes "le", "la", "les", "l´", "Le", "La", "Les", "L´" – constitue un mot du texte.

La bibliothèque électronique

Une fois ces opérations achevées, le texte peut entrer dans la bibliothèque électronique. Comme son équivalent « manuel », cette collection de documents (6 470 à ce jour) est divisée en sections et sous-sections – les « corpus » - qui sont dotées de divers « catalogues » à la disposition des chercheurs pour faciliter leurs recherches à l'aide d'un certain nombre d'outils.

Les données

En, l'occurrence, chaque **texte** original est balisé et standardisé, puis il est doté de :

- un fichier "image" (lemmatisé) comportant pour chaque mot du texte original, les étiquettes décrites ci-dessus (graphie standard, entrée de dictionnaire et catégorie grammaticale). Autrement dit, chaque emplacement (« vocable ») du fichier image pointe vers l'emplacement correspondant (« mot ») du texte original ;

- un **index** (dont le tableau 1 ci-dessus donne une idée) : les vocables classés par ordre alphabétique avec le nombre de leurs occurrences. Cet index fournit une « porte d'entrée » permettant d'aller directement là où se trouve le mot (ou le groupe de mots) recherché par l'opérateur ;

- Une **fiche « auteur ».** Par exemple, la fiche « auteur » pour la première déclaration gouvernementale suivant la démission du général de Gaulle (février 1945) :

Gouin Félix : Homme politique français – SFIO – Chef du gouvernement provisoire de la République française (janvier – juin 1946). Texte présent dans la bibliothèque : Déclaration d'investiture devant l'Assemblée constituante (23 janvier 1946).

- Des **liens** vers tous documents pertinents. Par exemple, les autres textes du même auteur présents dans la bibliothèque, les caractéristiques du corpus, les études réalisées à l'aide de ces fichiers.

Les outils d'exploitation

Outre les logiciels de lemmatisation, d'indexation, de recherche (« fouille ») et d'édition des concordances, un certain nombre d'outils sont à la disposition du chercheur pour étudier un texte ou groupe de textes :

- **vocabulaire** total et vocabulaire caractéristique de ce texte ou groupe de texte par rapport à un ensemble de référence choisi par le chercheur ;
- **syntagmes** les plus fréquents et/ou les plus caractéristiques ;
- différents **sens** d'un vocable (ou article de dictionnaire) et **phrases** les plus caractéristiques de ces significations ;
- **ruptures** thématiques et stylistiques ; segmentation des corpus ;
- **thèmes** : présence, contenu et poids relatifs ;
- **phrases** : longueurs, structures, contenu ;
- **style** d'un auteur, **énonciation** de la subjectivité, etc.
- **genre**, **auteur**, **époque** d'un texte ou d'un groupe de textes…

Trois remarques pour conclure cette présentation de la bibliothèque électronique.

Premièrement, les étiquettes s'ajoutent au texte original qui est intégralement conservé. Ces étiquettes sont celles de la nomenclature standard de la lexicographie française – nomenclature connue de tous les locuteurs de cette langue – qui est scrupuleusement respectée ;

Deuxièmement, les matériaux de base de cette bibliothèque existent – les fichiers originaux, les fichiers corrigés, standardisés, balisés ; les fichiers lemmatisés ; les fiches auteurs et corpus ; les index des textes, des corpus, des sections de la bibliothèque ; les logiciels… - mais il manque une interface de consultation et de gestion de cet ensemble, de telle sorte qu'il n'est pas consultable en ligne ;

Troisièmement, ces textes ont été rassemblés dans des buts spécifiques – comme l'étude d'un auteur, d'un genre, d'une campagne électorale, d'une fonction (les présidents français, les Premiers ministres québécois, canadiens ou français), etc. Nous ne prétendons pas que cet ensemble puisse fournir un échantillon représentatif de la langue française mais simplement qu'il suggère combien un tel échantillon pourrait être utile à la connaissance du français moderne.

Conclusions

Cet exposé aura montré ce qui différencie la **lexicométrie** de l'**analyse de données textuelles**. Cette dernière repose sur deux postulats :

- les techniques exploratoires (« analyse des données ») sont suffisantes pour obtenir une bonne représentation du matériel. La modélisation, la formulation d'hypothèses, les tests sont inutiles ;

- les segments du texte original – découpés automatiquement à l'aide de caractères délimiteurs - comportent toute l'information pertinente et ce texte n'a besoin d'aucune préparation avant son traitement en machine (d'où l'adjectif « textuel »).

En revanche, la lexicométrie signifie que :

- les graphies du texte original sont soigneusement corrigées et standardisées afin de rendre ces textes comparables entre eux ;

- chaque mot du texte est rattaché à un vocable, c'est-à-dire à sa place dans le système de la langue (d'où « lexical » opposé à « textuel ») ;

- les outils des linguistes (spécialement les lexicologues et les lexicographes) sont traduits en modèles, en algorithmes et en calculs ; ces modèles permettent de formuler des hypothèses et de les tester.

Dans notre esprit, les deux méthodes ne sont pas opposées mais plutôt complémentaires (par exemple : Brugidou & Labbé 2000) mais il est clair que seule la seconde est une lexicométrie…

Cette présentation plaide pour un archivage des entretiens sociologiques dans une véritable base de données (si possible standardisés, balisés et étiquetés…)

Quelle utilité aurait une telle archive ?

— avant de lancer une nouvelle enquête, elle permettrait un débroussaillage du problème, une réflexion sur les questionnaires (ou les guides d'entretien), et sur la sélection des enquêtés ;

— elle fournirait une base de comparaison pour l'analyse des résultats de la nouvelle enquête. Qu'apporte-t-elle ? Que confirme-t-elle ? Quelles "nouveautés" annonce-t-elle ? Avec la possibilité d'un regard rétrospectif, cette base serait un moyen précieux pour identifier les tendances lourdes au sein de la population, ce qu'une enquête isolée a plus de mal à faire. On dit souvent que les enquêtes intégrant le temps, grâce aux panels ou aux techniques longitudinales, offrent une sorte de film alors que l'enquête ponctuelle se contente d'un instantané, nécessairement figé…

- en l'absence d'enquête d'usage récente - prolongeant l'étude pionnière de Gougenheim (1956 et 1958) — cette base pourrait offrir un outil de connaissance de la langue française telle qu'on la parle effectivement.

Enfin, les grandes bibliothèques électroniques, ainsi conçues, peuvent apporter une aide précieuse aux linguistes et aux lexicographes. L'outil pourrait également trouver des applications dans de nombreuses activités allant de la terminologie à la critique littéraire, en passant par la traduction assistée par ordinateur ou la recherche d'informations sur la toile…

Bibliographie

Les travaux publiés par notre réseau de recherche sont consultables en ligne (notamment sur le site hal.archives-ouvertes.fr). Une liste plus complète peut être consultée sur la page personnelle de D. Labbé.

Arnold Edward (2005). Le discours de Tony Blair (1997-2004). *Corpus*, 4, p. 55-77.

Arnold Edward (2008). Le sens des mots chez Tony Blair (people et Europe). 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 1, p 109-119.

Berger Guy, Leselbaum Nelly dir. (2002*), La prévention des toxicomanies en milieu scolaire : éléments pour une évaluation*, Montpellier, CNDP.

Bergeron Jean-Guy & Dominique Labbé (2000). L'évaluation de la négociation raisonnée par les acteurs. Une analyse lexicométrique. *Communication au XVIe 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.

Bergeron Jean-Guy & Dominique Labbé (2004). 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.

Blanche-Benveniste Claire & Jeanjean Colette (1987). *Le français parlé. Transcription et édition*. Paris : Didier.

Blanche-Benveniste Claire & Jeanjean Colette (1990). *Le français parlé. Etude grammaticale*. Paris : CNRS.

Blumenthal Peter & Hausmann Franz J. Eds (2006). "Collocations, corpus, dictionnaires". *Langue française*, 150, juin 2006.

Brugidou Mathieu & Labbé Dominique (2000). 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 5e journées internationales d'analyse des données textuelles*. Lausanne : Ecole polytechnique fédérale, 2000, vol 1, p. 85-94.

Burnard Lou (1995). *Users Reference Guide for the British National Corpus*. Oxford : Oxford University Computing Services.

Denni Bernard, Caillot Philippe, Moine Michèle, Roux Guillaume, Salomon Annie-Claude, Jessica Sainty & Claire Brachet (2007). *Formation du Jugement Politique*. Grenoble : Pacte, mars et juin 2007.

Gougenheim Georges, en collaboration avec Michea René, Rivenc Paul, Sauvageot Aurélien (1956). *L'élaboration du français élémentaire : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris : Didier. Réédition augmentée en 1964 sous le titre : *L'élaboration du français fondamental. Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Paris : Didier.

Gougenheim Georges (1958). *Dictionnaire fondamental de la langue française*. Paris : Didier. Nouvelle édition revue et augmentée, Didier, Paris, 1977.

Hubert Pierre, Carbonnel Jean-Pierre & Chaouche Ahmed (1989). Segmentation des séries hydrométéorologiques - Application à des séries de précipitations et de débits de l'Afrique de l'Ouest. *Journal of hydrology*, 110, 349-367.

Hubert Pierre & Labbé Dominique (1990). La répartition des mots dans le vocabulaire présidentiel. *Mots*, n° 22, mars 1990, p. 80-88.

Labbé Cyril & Labbé Dominique (1994). *Que mesure la spécificité du vocabulaire ?* Grenoble : CERAT, décembre 1994. Reproduit dans *Lexicometrica*, 3, 2001.

Labbé Cyril & Labbé Dominique (2001). 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.

Labbé Cyril & Labbé Dominique (2003). La distance intertextuelle. *Corpus*, 3, p. 95-118.

Labbé Cyril, Labbé Dominique et Hubert Pierre (2004). "Automatic Segmentation of Texts and Corpora". Journal of Quantitative Linguistics, 11-3, p 193-213.

Labbé Cyril & Labbé Dominique (2005). How to measure the meanings of words ? Amour in Corneille's work. *Language Resources Evaluation*, 39, p. 335-351.

Labbé Cyril & Labbé Dominique (2006). A Tool for Literary Studies: Intertextual Distance and Tree Classification. *Literary and Linguistic Computing*. 21-3, 2006, p. 311-326.

Labbé Cyril & Labbé Dominique (2009). Existe-t-il un genre épistolaire ? Hugo, Flaubert et Maupassant. *Communication aux Xe Journées de l'ERLA*. Banks David. *Le texte épistolaire du XVIIe siècle à nos jours*. Paris : l'Harmattan, 2013, p. 53-86.

Labbé Cyril & Labbé Dominique (2010a). 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..

Labbé Cyril & Labbé Dominique (2010b). La modalité verbale en français contemporain. *Communication aux XIe Journées de l'ERLA*. Brest : 19 novembre 2010.

Labbé Cyril & Labbé Dominique (2011a). 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.

Labbé Cyril & Labbé Dominique (2011b). Baudelaire, Rimbaud et Verlaine. In Banks David (Ed). *Aspects diachroniques du texte poétique*. Paris, l'Harmattan, 2011, p. 17-45.

Labbé Cyril & Labbé Dominique (2011bc). Existe-t-il un langage propre à la politique ? *Communication aux XIIe Journées de l'ERLA*. Brest : 18-19 novembre 2011.

Labbé Cyril & Labbé Dominique (2012a). 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.

Labbé Cyril & Labbé Dominique (2012b). Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science? *Scientometrics*. Published on line : 22 june 2012.

Labbé Cyril & Labbé Dominique (2012c). 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, p.537-551.

Labbé Cyril, Labbé Dominique & Hubert Pierre (2004). Automatic Segmentation of Texts and Corpora. *Journal of Quantitative Linguistics*, december 2004, 11-3, p. 193-213.

Labbé Dominique (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble : Cahiers du CERAT.

Labbé Dominique (1998). 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.

Labbé Dominique (2001). Normalisation et lemmatisation d'une question ouverte. Les femmes face au changement familial. *Journal de la Société Française de Statistique*. 142-4, décembre 2001, p. 37-57.

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

Labbé Dominique (2002b). 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.

Labbé Dominique (2007a). Coordination et subordination en français oral. In Banks David (Ed.). *La coordination et la subordination dans le texte de spécialité*. Paris : L'Harmattan, 2007, p. 161-182.

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

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

Labbé Dominique & Monière Denis (2003). *Le vocabulaire gouvernemental. Canada, Québec, France (1945-2000).* Paris : Champion.

Labbé Dominique & Monière Denis (2006). 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.

Labbé Dominique & Monière Denis (2008a). 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.

Labbé Dominique & Monière Denis (2008b). *Les mots qui nous gouvernent Le discours des premiers ministres québécois : 1960-2005*. Montréal : Monière-Wollank Editeurs, 2008.

Labbé Dominique & Monière Denis (2012). 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.

Labbé Dominique & Monière Denis (2013). *La campagne présidentielle de 2012. Votez pour moi !* Paris : l'Harmattan (collection logiques politiques), 2013.

Lafon Pierre (1984). *Dépouillements et statistiques en lexicométrie*. Genève-Paris : Slatkine:Champion.

Merriam Thomas (2002). "Intertextual Distances between Shakespeare Plays, with Special Reference to *Henry V* (verse)". *Journal of Quantitative Linguistics*. 9-3, 260-273.

Merriam Thomas (2003a). "An Application of Authorship Attribution by Intertextual Distance in English". *Corpus*. 2, p 167-182.

Merriam Thomas (2003b). "Intertextual Distances, Three Authors". *Literary and Linguistic Computing*, 18-4, 379-388.

Merriam Thomas (2005). *The Identity of Shakespeare in Henry VIII*. The Renaissance Institute, Tokyo.

Monière Denis & Labbé Dominique (2002). Essai de stylistique quantitative. Duplessis, Bourassa et Lévesque. In Morin Annie et Sébillot 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.

Monière Denis, Labbé Cyril & Labbé Dominique (2005). Les particularités d'un discours politique : les gouvernements minoritaires de Pierre Trudeau et de Paul Martin au Canada. *Corpus*, 4, 2005, p. 79-104.

Monière Denis, Labbé Cyril & Labbé Dominique (2008). 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.

Monière Denis & Labbé Dominique (2010a). 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.

Monière Denis & Labbé Dominique (2010b). 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.

Muller Charles (1963). Le mot, unité de texte et unité de lexique en statistique lexicologique. Reproduit dans : *Langue française et linguistique quantitative*. Genève-Paris: Slatkine-Champion, 1979, p 125-143.

Muller Charles (1977). *Principes et méthodes de statistique lexicale*. Paris : Hachette.

Nelson Gerald (1997). "Standardizing Wordforms in a Spoken Corpus". *Literary and Linguistic Computing*, 12, 2 , p 79-85.

Pionchon Sylvie (2001), *Les Françaises et la politique*, Thèse pour le doctorat de science politique, Institut d'Etude Politique, Grenoble.

Pibarot André, Picard Jacques & Labbé Dominique (1998). Les syntagmes répétés dans l'analyse des commentaires libres. In Mellet Sylvie (ed). 4e Journées d'analyse des données textuelles. Nice, 1998, p. 507-516.

Salem André (1987). *Pratique des segments répétés. Essai de statistique textuelle*. Paris : Klincksieck.

Savoy Jacques (2010). "Discours électoral et discours présidentiel". In Bolasco Sergio et al. (Eds). *Proceedings of 10th International Conference Statistical Analysis of Textual Data*. Rome : Edizioni Universitarie di Lettere Economia Diritto, Vol 2, p. 827-838.

Silberztein Max (1993), *Dictionnaires électroniques et analyse automatique des textes : le système INTEX*, Paris, Masson.

Annexe 1

Bibliothèque électronique du français moderne (1er juin 2013)

	Longueur (mots)	Vocabulaire
Discours politique	11 075 846	42 496
Présidents français (1958-2013)	3 351 713	25 242
Premiers ministres canadiens (1867-2012)	1 098 161	13 514
Premiers ministres québécois (1867-2012)	2 993 823	22 458
Premiers ministres français (1945-2012)	288 526	7 952
Littérature (XVIIe –XXe siècles)	9 589 021	56 044
Romans et nouvelles	5 846 986	48 265
Théâtre	2 571 497	15 551
Poésie	675 187	18 810
Correspondance	345 542	11 070
Romans policiers	548 682	17 274
Presse	1 773 818	34 842
Economie	645 341	13 964
Hydrologie	129 173	7 197
Français oral*	2 730 136	18 429
Total	27 618 508	100 691

*** Le français oral**

Les Français(es) et la politique (Pionchon 2001) :
32 entretiens : 345 752 mots, 6 540 vocables différents

La négociation raisonnée au Québec (Bergeron & Labbé, 2000, 2004) :
61 entretiens : 409 225 mots, 6 591 vocables différents

La prévention des toxicomanies en milieu scolaire (Berger & Leselbaum 2002) :
15 entretiens : 92 992 mots, 4 255 vocables différents

Confort électrique EDF réalisé par les sociologues du Grets en six enquêtes (Labbé & Labbé, 2001 ; Labbé, 2002a) : 201 entretiens : 1 270 307 mots, 10 904 vocables différents

Formation du jugement politique lors de la campagne présidentielle de 2007 en Isère (Denni & Al, 2007) : 44 entretiens semi-directifs : 389 952 mots, 8 089 vocables différents

Questions ouvertes dans un sondage auprès des femmes divorcées réalisé par l'INED (Labbé, 2001) :
3000 enquêtés : 56 107 mots, 2 786 vocables différents

Questions ouvertes dans un sondage auprès des citoyens belges sur la droite et la gauche :
1000 enquêtés : 22 294 mots , 1 706 vocables différents

Questions ouvertes dans deux sondages sur l'intérêt pour la campagne présidentielle de 2007 (Denni & Al. 2007 ; Labbé & Labbé 2012) 1 467 enquêtés répondants : 142 294 mots, 3 712 vocables différents

Divers :
7 entretiens : 115 494 mots, 4 922 vocables différents

Total transcriptions de l'oral :
361 entretiens et trois sondages : 2 730 136 mots, 18 429 vocables différents

Annexe 2

Index hiérarchique des entretiens FJP07 et des discours des candidats à la présidentielle

Substantifs

Entretiens (FJP 2007)			Candidats (présidentielle 2007)	
Rang	Vocable	Fréquence (‰)	Vocable	Fréquence (‰)
1	chose	3.33	pays	2.56
2	gens	3.24	politique (nf)	1.76
3	problème	1.86	travail	1.65
4	an	1.53	enfant	1.56
5	niveau	1.50	monde	1.52
6	fait	1.30	république	1.52
7	temps	1.14	état	1.44
8	monde	1.13	entreprise	1.42
9	pays	1.07	an	1.29
10	truc	1.03	vie	1.28
11	fois	1.02	homme	1.23
12	vie	0.99	droit	1.20
13	enfant	0.96	valeur	1.16
14	personne	0.95	emploi	1.14
15	rapport	0.93	école	1.09
16	travail	0.88	société	1.08
17	côté	0.87	peuple	1.01
18	façon	0.84	chose	0.98
19	moment	0.83	famille	0.89
20	heure	0.82	besoin	0.88

Adjectifs

Entretiens (FJP 2007)			Candidats (présidentielle 2007)	
Rang	Vocable	Fréquence (‰)	Vocable	Fréquence (‰)
1	vrai	2.10	grand	1.72
2	petit	2.08	social	1.39
3	bon	1.34	français	1.32
4	sûr	0.86	public	1.00
5	grand	0.74	nouveau	0.98
6	important	0.53	politique	0.91
7	social	0.53	seul	0.89
8	pareil	0.50	national	0.86
9	politique	0.49	européen	0.70
10	seul	0.48	économique	0.62
11	gros	0.41	présidentiel	0.57
12	français	0.34	petit	0.51
13	normal	0.32	jeune	0.46
14	clair	0.32	possible	0.46
15	difficile	0.31	cher	0.41
16	européen	0.30	juste	0.39
17	possible	0.30	bon	0.38
18	différent	0.27	important	0.37
19	cher	0.26	vrai	0.37
20	nouveau	0.26	professionnel	0.36

Verbes

	Entretiens (FJP 2007)		Candidats (présidentielle 2007)	
Rang	Vocable	Fréquence (‰)	Vocable	Fréquence (‰)
1	être	46.00	être	32.07
2	avoir	33.23	avoir	18.75
3	faire	8.70	faire	5.80
4	dire	8.44	vouloir	5.64
5	aller	7.70	dire	4.14
6	pouvoir	4.40	pouvoir	3.79
7	savoir	4.00	aller	2.73
8	falloir	3.82	devoir	2.61
9	vouloir	3.72	falloir	2.27
10	voir	3.45	savoir	1.76
11	penser	3.33	croire	1.30
12	trouver	2.03	voir	1.25
13	parler	1.77	donner	1.22
14	voter	1.52	mettre	1.12
15	mettre	1.44	prendre	1.09
16	passer	1.32	venir	1.03
17	arriver	1.26	vivre	1.03
18	travailler	1.22	parler	1.03
19	prendre	1.20	penser	0.92
20	venir	1.10	proposer	0.92
	Total	139.65		90.47

Pronoms

	Entretiens (FJP 2007)		Candidats (présidentielle 2007)	
Rang	Vocable	Fréquence (‰)	Vocable	Fréquence (‰)
1	je	36,65	je	18,32
2	ce	29,36	qui	15,61
3	on	18,95	ce	13,86
4	y	14,94	il	11,97
5	il	14,90	nous	7,65
6	ça	14,83	on	6,92
7	qui	12,93	se	6,64
8	que	7,04	que	6,27
9	moi	6,38	le	6,07
10	ils	6,37	vous	5,73
11	se	6,3	ils	4,45
12	vous	5,07	celui	3,82
13	le	4,74	y	3,70
14	quoi	4,33	cela	2,29
15	en	3,65	en	2,03
16	nous	2,71	tout	1,91
17	tout	2,58	moi	1,21
18	rien	1,33	autre	1,17
19	lui	1,27	lequel	1,09
20	autre	1,20	dont	1,08
		195,53		121,79

Annexe 3

Un exemple de concordance : « suis » du verbe suivre dans un corpus d'entretiens sur les femmes et la politique (Pionchon)

Alice		
se, en fait. Mais c'est vrai que bon, c'est vrai que je, je	suis	, je suis pas de près. Par exemple, si vous me demandie
it. Mais c'est vrai que bon, c'est vrai que je, je suis, je	suis	pas de près. Par exemple, si vous me demandiez au gouve
es chefs, et encore voyez Chirac, je le savais même pas. Je	suis	pas, je suis pas, comment dire ? Je m'intéresse pas de
encore voyez Chirac, je le savais même pas. Je suis pas, je	suis	pas, comment dire ? Je m'intéresse pas de près quoi, en
e que, oui si je crois en lui, j'ai confiance en lui, je le	suis	. Mais je le suivrais pas pour un parti, tu vois il peu
dire ? C'est pour ça que je te dis que moi la politique, je	suis	pas. Y a des personnes que j'aime bien. Lui il y a des
Angele		
ions. Moi je sais pas, moi… Oui mais pas, oui, mais pas, je	suis	pas. Quand ça vient les informations. Voyez. Mais bon,
Audrey		
avoir un bon travail, c'est pour ça que je travaille, que je	suis	des études, c'est pour essayer d'avoir un bon travail,
Cathy		
Puis dans la mesure où bon, dans le cadre de mon travail je	suis	beaucoup la réglementation, donc, j'en ai un petit peu
x mesures qui ont été annoncées sur l'emploi des jeunes, je	suis	un petit peu les, les éléments quoi, donc quelque chose
rtain nombre de points. Donc, dans la mesure où après, j'en	suis	un petit peu l'application, je suis de près … Bon dans
a mesure où après, j'en suis un petit peu l'application, je	suis	de près … Bon dans la mesure où, en plus, moi je travai
i que là, tout ce qui est lié à la politique locale, bon je	suis	beaucoup plus, alors là je suis dedans quoi, donc c'est
Chantal		
c'est ce que, c'est ce que je pense. Mais c'est vrai que je	suis	, je suis pas tellement les débats politiques à la télé
que, c'est ce que je pense. Mais c'est vrai que je suis, je	suis	pas tellement les débats politiques à la télé. Enfin ça
Christiane		
c'est l'utopie… Ouais. Ouais quand même. Ben ouais j'ai, je	suis	les infos … Euh … Pfff … La façon dont je regarde la po
… Pfff… La façon dont je regarde la politique ? Ouais, je la	suis	par les informations, c'est tout. Ouais la télé, en mêm
je sais pas… Ouais, ouais, là je regarde déjà… Ouais ben, je	suis	un petit peu. Par exemple la réouverture du dossier, be
Edith		
, quarante pour cent de l'autre… Suivant les gens, leur… Je	suis	, bon d'une manière générale, bon il en a, y a l'histoire
Germaine		
ministres de telle ou telle branche, ou telle, moi ça me… Je	suis	pas, ma pauvre … Quand vous dites l'actualité, oui mais

Annexe 4
Principales combinaisons *auxiliaire modal + complément à l'infinitif*

Enquêtés FJP07

Auxiliaire	Complément	Fréquence (pour 10 000 mots)	% Auxiliaire	% Complément
vouloir	dire	14,5	39,1	17,2
aller	dire	4,6	6,0	5,4
aller	faire	4,3	5,6	5,0
pouvoir	faire	3,5	8,0	4,1
aller	voter	3,2	4,2	21,2
pouvoir	être	2,7	6,2	0,6
pouvoir	dire	2,6	5,8	3,0
falloir	faire	2,5	6,6	2,9
aller	être	2,4	3,2	0,5
aller	voir	1,5	2,0	4,4
devoir	être	1,5	16,8	0,3
falloir	être	1,5	4,0	0,3
aller	aller	1,5	1,9	1,9
vouloir	faire	1,5	3,9	1,7
pouvoir	avoir	1,4	3,1	0,4
être	dire	1,4	0,3	1,6
entendre	parler	1,4	22,1	7,7
aller	falloir	1,2	1,5	3,1
pouvoir	aller	1,1	2,5	1,4
falloir	arrêter	0,9	2,3	23,4
Total des vingt premiers		55,2	7,3	5,3

Lecture : chez les enquêtés, on rencontre en moyenne 14.5 fois la combinaison « vouloir dire » pour 10 000 mots, cela représente 39% des emplois du verbe *vouloir* et 17% de ceux de *dire*.

Candidats à la présidentielle de 2007

Auxiliaire	Complément	Fréquence (pour 10 000 mots)	% Auxiliaire	% Complément
vouloir	dire	3,8	6,7	9,1
vouloir	être	3,1	5,4	1,0
devoir	être	3,0	11,5	0,9
pouvoir	être	1,8	4,8	0,6
pouvoir	faire	1,6	4,2	2,8
vouloir	faire	1,5	2,6	2,5
falloir	faire	1,2	5,1	2
vouloir	parler	1,0	1,8	9,8
aller	faire	1,0	3,5	1,6
devoir	faire	0,9	3,4	1,5
faire	vivre	0,9	1,5	8,2
être	faire	0,8	0,2	1,3
devoir	avoir	0,7	2,7	0,4
aller	devoir	0,7	2,5	2,6
pouvoir	avoir	0,7	1,8	0,4
faire	respecter	0,7	1,1	10,9
pouvoir	continuer	0,7	1,7	19,5
être	être	0,7	0,2	0,2
aller	falloir	0,6	2,3	2,8
pouvoir	vivre	0,6	1,6	6,0
Total des vingt premiers		25,6	3,2	4,2