Nouvelles perspectives en sciences sociales
Revue internationale de systémique complexe et d'études relationnelles

Proposition de protocole pour l'analyse des données textuelles : pour une démarche expérimentale en lexicométrie

Proposal to Add a Protocol to Textual Data Analysis: For an Experimental Procedure in *Lexicométrie*

Jean-Marc Leblanc

Volume 11, Number 1, November 2015

Sur le thème de l'analyse de données textuelles informatisée

URI: https://id.erudit.org/iderudit/1035932ar
DOI: https://doi.org/10.7202/1035932ar

See table of contents

Publisher(s)

Prise de parole

ISSN

1712-8307 (print)
1918-7475 (digital)

Explore this journal

Cite this article

Leblanc, J.-M. (2015). Proposition de protocole pour l'analyse des données textuelles : pour une démarche expérimentale en lexicométrie. *Nouvelles perspectives en sciences sociales*, *11*(1), 25–63. https://doi.org/10.7202/1035932ar

Article abstract

This paper presents a methodological reflection comparing various automated data processing techniques, approaches, and methods. It promotes the use of various types of software while working on textual data, explains how results of software can be used as data in different settings and promotes an experimental approach in atextometrical or lexicometrical field.

Proposition de protocole pour l'analyse des données textuelles : pour une démarche expérimentale en lexicométrie

Jean-Marc Leblanc
Université Paris-Est – Créteil Val de Marne

Introduction

Les analyses automatisées du discours sont fréquemment mobilisées en sciences sociales, qu'il s'agisse de veille scientifique, de positionnements concurrentiels ou marketing, de e-réputation, de dépouillements d'enquêtes, de questionnaires, d'études de marché ou de satisfaction, de web sémantique, d'analyses textuelles ou discursives. Repérer des régularités, analyser un style discursif, révéler des tendances, examiner le lexique d'un auteur, étudier le discours politique, telles sont les applications les plus répandues de ces méthodes.

Les médias et autres réseaux sociaux s'emparent en outre fréquemment de quantifications des données textuelles et lexicales, produisant listes, tableaux et visuels parfois commentés de manière aléatoire, le plus souvent dans le domaine politique.

Entre une utilisation médiatique croissante et des usages du web faisant appel à la quantification des données (*open data*, nuages de mots, mais aussi entreprise plus évoluée comme celle

de Google[1] la lexicométrie, mais aussi les statistiques textuelles plus largement ont une place à revendiquer, opposant à ces commentaires de visuels et de tableaux et à ces « nouvelles » pratiques une démarche éprouvée et raisonnée.

Il s'agira ici de présenter une réflexion méthodologique, en matière de traitement automatisé des données textuelles et de défendre une démarche expérimentale en lexicométrie ou textométrie, notions que nous définirons dans les pages qui suivent.

Nous illustrerons par des exemples d'analyses portant sur un corpus de discours politiques rituels une certaine conception de la lexicométrie, privilégiant le travail sur la forme graphique, n'excluant pas la catégorisation morphosyntaxique ou sémantique mais ne la systématisant pas, privilégiant le retour au texte, dans une démarche prenant appui sur les tendances observables du corpus pour en saisir la dimension micro-textuelle, saisissant la boussole statistique pour explorer chaque empan textuel. Il s'agira par ailleurs de valider la démarche lexicométrique sur des petits corpus à une époque où la tendance est à l'exploration de gros voire très gros corpus, mais aussi de défendre une démarche portant sur des données présentant une cohérence de la situation d'énonciation et s'inscrivant dans un genre.

Nous dresserons une typologie des outils et des méthodes, en distinguant les outils Lexicométriques « traditionnels » longitudinaux et contrastifs, des outils structurants, permettant de faire émerger d'un corpus ses structures signifiantes saillantes, puis replacerons ces outils dans leurs contextes disciplinaires en explicitant les postulats méthodologiques qui sont à l'origine de leur développement.

Nous montrerons en quoi le croisement des outils permet d'interroger le corpus avec profit en présentant un protocole d'observation tiré d'un cas d'étude concret où nous privilégierons l'expérimentation, dans une conception large de la lexicométrie.

---

1 Voir Étienne Brunet, « Au fond du goofre, un gisement de 44 milliards de mots », dans *Actes des Journées internationales d'analyse statistique des données textuelles*, JADT, 2012, p. 7-21.

Enfin nous tenterons de montrer comment l'expérimentation lexicométrique nous a conduits à nous intéresser à la visualisation des données textuelles, au développement de nouveaux outils, et à nous interroger sur le rapport qu'il convient de faire aujourd'hui entre lexicométrie et sciences des textes, *data visualisation*, TAL, etc.

Approches statistique des données textuelles

Le principe lexicométrique de base consiste à envisager le dépouillement lexical des données textuelles sous un angle probabiliste, qui permet d'associer aux divers dénombrements un jugement en probabilité. Une démarche classique en lexicométrie consiste à examiner, sur un corpus indexé et partitionné à partir d'un nombre limité de variables, les caractéristiques des différentes partitions, à vérifier sa pertinence statistique, ses propriétés zipféennes, la répartition des grandes masses de vocabulaire dans l'index hiérarchique d'ensemble (très hautes, hautes, moyennes, basses fréquences, hapax), l'accroissement du vocabulaire soit dans une perspective[2] stylométrique, soit pour mettre en évidence des ruptures, une modification de vocabulaire pouvant révéler des revirement thématiques[3].

D'autres peuvent intégrer des notions telles que la longueur des phrases et ses indices de variation, voire la complexité des phrases, l'enchâssement, l'articulation. Un troisième mouvement

---

2 La mise en œuvre de calculs tels que la richesse, la diversité, la spécialisation,ne peut raisonnablement s'appliquer à notre corpus, ne serait-ce qu'en raison de sa taille. Selon *Hyperbase*, le calcul de la richesse du vocabulaire (sur les hapax) montrerait une large avance chez de Gaulle puis Mitterrand, et un net recul chez Giscard et surtout Chirac. Nous pourrions relier cette donnée à une volonté de simplicité chez les uns, à l'appartenance de Chirac à une ère de la communication, à une tradition plus rhétorique chez de Gaulle, mais pour établir fermement notre interprétation il eût fallu pousser beaucoup plus loin l'analyse.

3 « Le concept de style désignerait l'ensemble des traits distinctifs – autres que la fréquence d'emploi des mots qui *caractériseraient un auteur et le distinguerait des autres* » (Dominique Labbé et Denis Monière, « Essai de stylistique quantitative. Duplessis, Bourassa et Lévesque », dans Annie Morin et Pascale Sébillot (dir.), *VI^e Journées Internationales d'analyse des données textuelles*, Saint-Malo, 13-15 mars 2002, Rennes, IRISA-INRIA, 2002, n° 2, p. 561-569, https://halshs.archives-ouvertes.fr/halshs-01019903, site consulté le 5 octobre 2015.

consiste à mesurer en probabilité les attirances entre les mots, leur voisinage plus ou moins étroit, les ensembles de mots privilégiés dont sont constitués les textes ou les parties de textes.

Notre démarche s'inscrit fondamentalement dans cette approche lexicométrique, telle que l'a définie Maurice Tournier, pour s'ouvrir vers d'autres perspectives, tout en mettant en place une approche plus expérimentale envisageant d'autres éclairages que le strict dépouillement *lexico*-statistique, testant les catégorisations, les ontologies ou les univers sémantiques à l'aide des outils principaux de la lexicométrie.

À la lumière des diverses approches multidimensionnelles que nous avons pratiquées dans le cadre de recherches précédentes, nous avons été amenés à nous interroger, non tant sur les fondements mathématiques de celles-ci, que nous laissons à la compétence des spécialistes[4], mais à leur utilisation dans un cadre expérimental et à leurs possibles améliorations ergonomiques à travers les pratiques et les hypothèses des utilisateurs. C'est ce qui nous a conduits notamment à développer un outil d'exploration des données textuelles, *TextObserver*[5].

La démarche lexicométrique

La démarche lexicométrique ou textométrique permet de comparer le lexique de plusieurs auteurs, de plusieurs textes, de plusieurs parties d'un même texte, de mesurer ce lexique sur la base d'un corpus représentatif, cohérent, dont les différentes parties sont comparables, sur le plan du genre, de la situation d'énonciation, mais aussi du matériau discursif lui-même, c'est-à-dire de la taille des différentes parties, déterminée par le nombre des mots de ces parties.

L'exemple qui suit illustre cette démarche contrastive. Sur un corpus constitué des messages de vœux des présidents de la Cinquième République (ici 1959-2012), nous avons appliqué la méthode des spécificités afin d'examiner la distribution de la première personne du singulier en termes de sous emplois (spé-

4 Jean-Paul Benzecri, Ludovic Lebart, André Salem, etc.

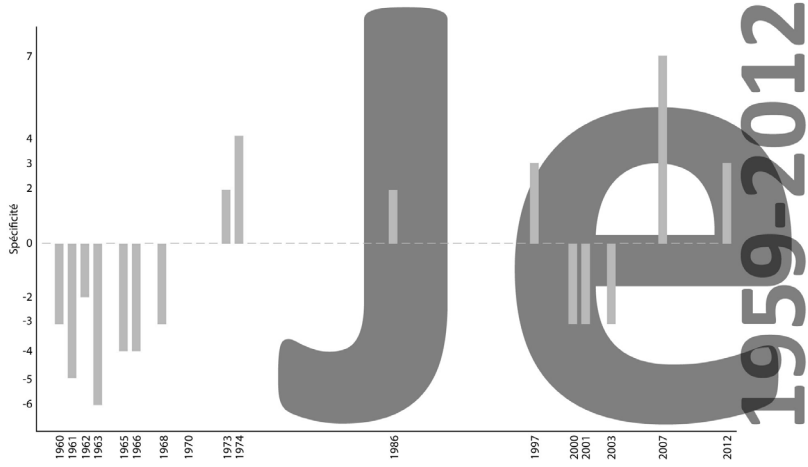
5 http://textopol.u-pec.fr/textobserver.

cificité négatives) ou du sur emploi (spécificité positive). Les valeurs notées comme nulles sur l'histogramme ci-dessous correspondent à un diagnostic de banalité : la forme considérée n'est pas absence, mais n'est ni sur employée ni sous employée. Sur la base d'un calcul de fréquence probabilisée, on considère que la fréquence observée (fréquence réelle) correspond à la fréquence théorique (fréquence attendue).

Ce calcul implique que le corpus soit doté de partitions puisqu'il s'agit de comparer le vocabulaire des différentes parties du corpus. Dans l'exemple présenté la partition en année revient à une partition en discours (un discours par an).

Figure 1

Exemple d'histogramme des spécificités de la forme « Je » dans le corpus vœux (1959-2014). Partition année



On note ici que Sarkozy, pour son premier message de vœux en décembre 2007, utilise le « je » plus que n'importe qui avant lui. Dans les années qui suivent, l'emploi du « je » chez Sarkozy devient banal... Il faudra attendre le message de 2012 (Hollande) pour que la première personne du singulier soit de nouveau en suremploi par rapport à l'ensemble du corpus.

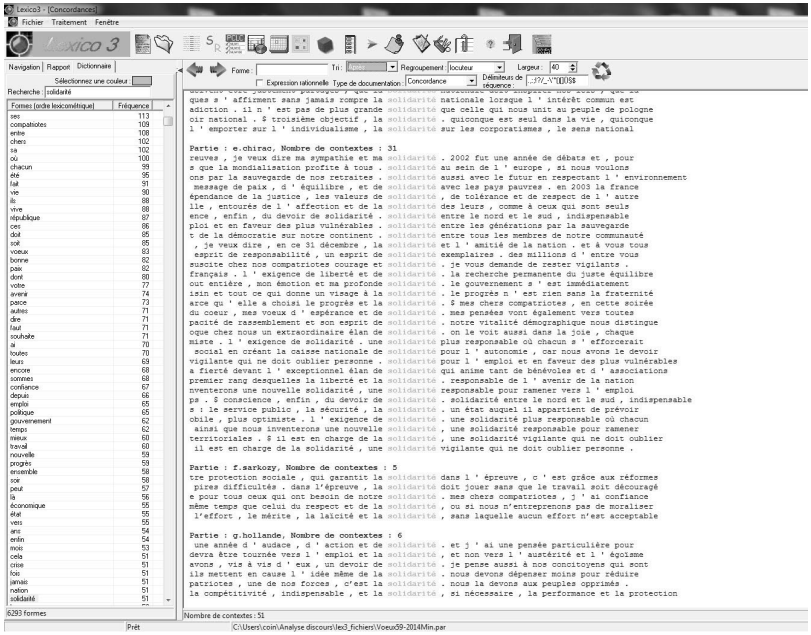
Cet histogramme montre également un profil énonciatif particulier, celui du Général de Gaulle chez qui les marques de la première personne du singulier sont, chaque année ou presque, systématiquement sous-représentées.

Ces commentaire statistiques n'ont aucun sens si l'on ne retourne pas au texte pour analyser les différents emplois de ce « je » (je veux, je voudrais, je pense, je sais...).

Ce retour au texte peut s'effectuer au moyen d'une concordance, fonctionnalité courante des logiciels lexicométriques et qui découle directement des méthodes distributionnalistes en linguistique. Quelques concordances célèbres sont d'ailleurs fort anciennes : comme les concordances de la bible par Saint Chef au XIII^{e} siècle. Dans les années 1970, bon nombre de chercheurs se sont intéressés aux concordances d'un mot, d'une notion, d'un concept, que ce soit en littérature ou en sciences politiques.

Figure 2

Concordances de la forme « solidarité » corpus vœux (1959-2014). Partition locuteur



Il s'agit donc dès l'origine de déterminer le sens d'un mot par ces différents emplois, en le replongeant dans toutes ses réalisations, en le remettant en contexte. Une autre approche, cooccurrentielle, peut aussi nous éclairer. Il s'agit de déterminer les cooccurrents principaux d'un mot, d'un groupe de mots ou d'une expression. Plusieurs calculs sont envisageables ainsi que plusieurs représentations, et il s'agit là plutôt de donner une vision plus synthétique que ne le ferait l'approche contextuelle mentionnée ci-dessus. Quelles sont les attirances entre les termes, avec quoi un mot est-il le plus souvent employé, quelles notions se rejettent ou s'attirent? Autant de questions qui peuvent être résolues par ce type de mesure.

Problèmes terminologiques: lexicométrie, textométrie, logométrie, stylométrie

D'abord statistique linguistique[6], puis statistique lexicale[7], et parallèlement analyse statistique des données textuelles[8] – ou linguistiques –, la lexicométrie est apparue en France dans les années 1970, initiée par Pierre Lafon, Maurice Tournier et Robert-Léon Wagner. Pour autant, quelques nuances sont à noter autour de ces dénominations, telles que « textométrie[9] » (André Salem), « logométrie » (Damon Mayaffre), « Analyse de discours assistée par ordinateur » (Pascal Marchand), « Traitement automatisé du discours » (Pierre Fiala).

Pour Damon Mayaffre, il y a bien un « glissement de la lexicométrie originelle vers une logométrie pleine et entière,

6 Pierre Guiraud, *Problèmes et méthodes de la statistique linguistique*, Paris, Presses universitaires de France, 1960.

7 Charles Muller, *Essai de statistique lexicale. L'illusion comique de P. Corneille*, Paris, Klincksieck, 1964.

8 Voir Jean-Paul Benzécri, *L'analyse des données. Leçons sur l'analyse factorielle et la reconnaissance des formes et travaux du laboratoire de statistique de l'Université de Paris VI*, Paris, Dunod, 1973; Jean-Paul Benzécri, *Histoire et préhistoire de l'analyse des données*, Paris, Dunod, 1982; et Jean-Paul Benzécri, *Pratique de l'analyse des données*, Paris, Dunod, 1980.

9 La dénomination de « textométrie » tend à se généraliser.

susceptible de renouveler la discipline[10] ». Cette logométrie naît donc du renouvellement des pratiques, rendu possible grâce aux progrès quantitatifs des lemmatiseurs et catégoriseurs[11]. Et c'est par cette approche plus plénière du discours, qui ne saisit plus la matérialité du texte par la seule forme graphique mais bien à travers la multiplicité des dimensions linguistiques (parties du discours, lemmes, structures syntaxiques, sans exclure cependant la forme graphique) que se définit la logométrie selon Mayaffre.

Cette définition ne nous satisfait pas pleinement: notre entrée reste la forme graphique et la lemmatisation ou la catégorisation peuvent intervenir en second lieu, en complément de l'analyse. C'est d'ailleurs en partie ce qui justifie que l'on tienne à la dénomination de lexicométrie. Par ailleurs les catégorisations ne peuvent être mises en œuvre sans une minutieuse vérification des données et l'illustration de la logométrie nous semble aujourd'hui trop attachée à un outil unique. Enfin, cette dénomination semble – à ce jour – trop globalisante et laisse entendre que l'on appréhende le discours sous tous les angles, que l'on mène une « mesure du discours », alors que ces dimensions bien que complémentaires ne peuvent être mises sur le même plan.

Un glissement terminologique vers la textométrie[12] s'explique essentiellement par la diversité des unités textuelles prises en compte dans les dernières tendances qui ne sont plus uniquement la forme graphique et le lexique mais des zones, paragraphes, segments, employées dans une nouvelle approche du texte,

10 Damon Mayaffre, « De la lexicométrie à la logométrie », *L'Astrolabe*, 2005, p. 1-11, https://halshs.archives-ouvertes.fr/hal-00551921/document, site consulté le 5 octobre 2015.

11 Les annotations effectuées par *Cordial* ne sont pas exemptes d'erreurs. Le dispositif *Hyperbase Cordial* ou *Weblex Cordial*, *TXM Cordial* n'est donc pleinement satisfaisant qu'une fois réalisées les corrections qui s'imposent. Le catégoriseur *Treetagger* ne produit pas de meilleurs résultats mais a l'avantage d'être en accès libre.

12 Antérieure à la notion de textométrie, on rencontre fréquemment chez André Salem la dénomination de statistique textuelle, témoignant peut-être déjà d'une volonté de ne pas se voir contraint au seul lexique comme unité minimale. Il nous faut ajouter que cette terminologie tend à se généraliser.

comme les phénomènes de topographie textuelle, de résonance[13], d'alignement multilingues[14].

Lexicométrie ou stylométrie?

Si la stylométrie consiste en diverses mesures statistiques du style d'un auteur, les recherches menées dans le domaine qui nous occupe ici en comportent nécessairement quelques aspects même si tous les indicateurs statistiques en rapport avec le style ne sont pas toujours mobilisés. Dominique Labbé et Denis Monière considèrent en effet que « le concept de style désignerait l'ensemble des traits distinctifs – autres que la fréquence d'emploi des mots qui caractériseraient un auteur et le distinguerait des autres[15] ».

Labbé et Monière proposent plusieurs pistes d'étude:

- la diversité du vocabulaire;
- la spécialisation;
- la richesse;
- la longueur et la structure des phrases,
- la densité des catégories grammaticales[16].

13 André Salem, « Introduction à la résonance textuelle », *VII^e Journées internationales d'analyse statistique des données textuelles*, Louvain-La-Neuve, dans *JADT*, 2004, http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_096.pdf, site consulté le 5 octobre 2015.

14 Autour des traitements multilingues on notera la création récente du groupe GADT réunissant des spécialistes de cette problématique, notamment de l'alignement de corpus, bilingues, comparables. La méthodologie est assez proche de ce qui se pratique en matière de résonance textuelle (voir André Salem, *ibid.*) et fait appel aux méthodes cooccurrentielles. Voir sur le sujet Maria Zimina, « Alignement textométrique des unités lexicales à correspondances multiples dans les corpus parallèles », *VII^e Journées internationales d'Analyse statistique des Données Textuelles*, Louvain-La-Neuve, JADT, 2004, http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_118.pdf, site consulté le 5 octobre 2015.

15 Dominique Labbé et Denis Monière, *op. cit.*

16 *Ibid.*

Des outils construits selon des histoires et des postulats méthodologiques différents

Le choix de l'outil dépend de l'objet de recherche, des hypothèses, du corpus, et non de la gratuité, de la disponibilité de l'outil, ou d'un quelconque mouvement de mode... Il convient en outre de connaître les choix qui ont présidé à la conception de tel ou tel outil statistique ou logiciel.

Ainsi, des chercheurs comme Brunet ou Muller se sont plus particulièrement intéressés à la littérature en considérant qu'un écrivain disposait d'un lexique privilégié et qu'il était possible d'analyser à travers le vocabulaire de ses œuvres sa disposition et sa manière d'utiliser la langue. Ils ont construit des méthodes, des logiciels qui visaient ce type d'étude.

À partir du milieu des années 1960, on s'est intéressé à l'intertextualité[17] en considérant que les discours étaient surtout produits par résonance avec d'autres discours. On a alors décrit la circulation d'éléments parfois plus longs que les simples formes et ces postulats méthodologiques ont permis de développer des outils qui indexaient et quantifiaient alors plutôt des segments ou des cooccurrences.

Des chercheurs comme Max Reinert, travaillant plutôt dans le domaine de la psychanalyse, ont développé des méthodes cooccurrentielles, considérant que deux mots pouvaient apparaître ensemble dans les mêmes phrases. C'est ainsi qu'ils ont développé des outils comme *Alceste* permettant de faire émerger d'un texte les structures signifiantes saillantes au moyen d'un calcul de cooccurrences inter-énoncés[18].

Ainsi la préoccupation majeure de chacun de ces groupes de concepteurs de logiciels doit être explicitée, car ils n'ont jamais évoqué ces préoccupations de manière distinctive; ils ont plutôt

17 Voir, entre autres, Mikhaïl Bakthine, *La poétique de Dostoïevski*, Paris, Seuil, coll. « Points », 1970 [1929] et Michel Foucault, *Dits et écrits* (tome 1, 1954-1975, et tome 2, 1976-1988), Paris, Gallimard, 2001 [1994].

18 Max Reinert, « Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars », *Langage et société*, vol. 66, nº 1, 1993, p. 5-39.

toujours travaillé dans l'évidence de leurs propres préoccupations (psychanalyse pour les uns, discours politique pour les autres ou encore littérature). Pour le chercheur pourtant, il est essentiel de connaître l'histoire de ces outils et le champ disciplinaire dans lequel ils s'inscrivent.

Typologie des outils

Il existe de nombreuses approches quantitatives des textes souvent liées aux recherches linguistiques, historiques ou sociologiques, comme *Prospéro*[19], *SpadT*[20], *Sphinx*[21] , *Leximappe*[22], *Astartex*[23],

19 Francis Chateauraynaud, *Prospéro: une technologie littéraire pour les sciences humaines*, Paris, CNRS, 2003.

20 *SpadT* est un logiciel d'analyse des données textuelles, plus particulièrement utilisé pour le traitement des questions ouvertes. Parmi ses fonctionnalités : lemmatisation, contextes, analyse multidimensionnelle des tableaux lexicaux (analyse discriminante).

21 *Sphinx Lexica* intègre différentes fonctions pour les enquêtes et les analyses des données : analyse de contenu, importation de corpus et un module « lexicométrique » évolué, essentiellement composé d'un lemmatiseur, d'un outil permettant de grouper des champs lexicaux, mais aussi des mesures « d'intensité lexicale », de richesse, de banalité du vocabulaire. Voir sur le sujet Jean Moscarola, « Balladur, Chirac, Jospin, les mots d'une campagne. Quelques exemples d'analyse lexicale avec *Le Sphinx* », *Journées internationales d'analyse statistique des données textuelles*, JADT, 1995.

22 *Leximappe* est un système d'organisation de corpus documentaire, fondé sur la méthode des mots associés, qui permet d'effectuer une classification des contenus en identifiant les mots qui présentent le plus fort indice d'association. Le logiciel produit des cartographies présentant les différents objets thématiques et les indices d'association qui les lient. Développé à l'origine par le Centre de Sociologie de l'innovation de l'École des mines et le Centre de documentation scientifique et technique du CNRS, *Leximappe* a été repris par le Centre de recherches infométriques. On se reportera à Vololona Rabeharisoa (*L'analyse Leximappe de la presse grand public : le cas de la controverse sur le changement climatique global*, Centre de sociologie de l'innovation, École des mines, 2005, htts//web.upmf-grenoble.fr/adest/seminaires/volo.html, site consulté le 5 octobre 2015) ainsi qu'à Michel Callon *et al.* (*La scientométrie*, Paris, Presses universitaires de France, 1993).

23 Le logiciel a été conçu par Jean-Marie Viprey. Voir http://textopol.u-pec.fr/?tag=astartex.

NooJ[24], *TXM*[25], *Tropes*, *Iramuteq*[26], pour ne citer que les plus connus. Trois types d'outils sont plus particulièrement mis en œuvre et éprouvés dans la présente démarche[27].

Les outils contrastifs ou longitudinaux : *Lexico 3*, *Hyperbase*, *TXM, Weblex*[28]

Logiciels lexicométriques standards ou « classiques », ces outils travaillent sur la base d'un tableau lexical, après réorganisation de la séquence textuelle et segmentation en unités minimales. Ils introduisent la notion de partition, sur laquelle portent des analyses contrastives et des mesures de ventilation du stock lexical dans les sous-parties du corpus (bornes chronologiques, locuteurs…). Les fonctions documentaires (concordances, contextes), statistiques (spécificités), analyses multidimensionnelles (analyse factorielle des correspondances arborées) constituent les fonctionnalités essentielles de ces outils. *Weblex,* qui n'est plus utilisé aujourd'hui présentait des caractéristiques similaires aux logiciels de type lexicométrique mais intégrait des fonctionnalités évoluées de recherche de cooccurrences reposant sur un modèle probabiliste (cooccurrences, *lexico*grammes simples et récursifs, associés ou non à une forme pôle). En outre, un langage d'interrogation, CQP, permettait de rechercher des motifs généraux, à partir de schémas. On retrouve cette fonctionnalité dans *TXM* mais aussi dans *TextObserver*. Parmi les analyses contrastives, *TXM* et *Lexico* emploient le même algorithme, celui des spécificités, de même qu'*Hyperbase*, bien que la présentation diffère quelque peu et

24 Max Silberztein et Agnès Tutin, « NooJ, un outil pour l'enseignement des langues. Application pour l'étude de la morphologie lexicale en FLE », *Apprentissage des langues et système d'information et de communication*, vol. 8, nº 2, 2005, p. 123-134, https://alsic.revues.org/336, site consulté le 5 octobre 2015.

25 Développant les travaux collaboratifs d'une ANR portée par l'ENS de Lyon.

26 Développé par Pascal Marchand et Pierre Ratinaud. Voir http://www.iramuteq.org/.

27 Textopol (http://textopol.u-pec.fr.), site d'enseignement et de recherche pour l'analyse informatisée du discours politique, chantier du CEDITEC, propose un accès raisonné à ces outils logiciels, à l'aide d'exercices, et d'autres ressources documentaires (corpus balisés, bibliographie, etc.).

28 Un grand nombre des fonctionnalités de *Weblex* sont implémentées dans la plate-forme en accès libre *TXM* développée par Serge Heiden.

prenne souvent l'apparence de l'écart réduit. Il n'est pas utile de revenir ici sur les modèles de loi normale ou de loi hypergéométrique. La méthode dite des spécificités permet de porter un jugement sur la répartition des formes dans les parties d'un corpus. Ce jugement s'exprime en termes de suremploi (spécificité positive) et de sous-emploi (spécificité négative). Selon le modèle hypergéométrique, une forme est notée spécifiquement positive si sa fréquence dans une partie est supérieure à la fréquence théorique attendue, et spécifiquement négative si cette fréquence est inférieure au seuil retenu. Ces fréquences probabilisées s'appuient sur la comparaison de quatre données : le nombre des occurrences du corpus, le nombre des occurrences dans la partie, la fréquence de chaque forme dans le corpus, et la fréquence de chaque forme dans la partie. Les indices indiquent le degré de spécificité de chaque forme et représentent la valeur absolue de l'exposant de probabilité. Un exposant de valeur 2 exprime une probabilité de l'ordre du centième, un exposant de valeur 3, du millième… L'absence d'exposant indique que l'usage ne présente pas de caractéristique remarquable. On dit que la forme est banale pour la partie considérée (voir l'exemple présenté dans les pages qui précèdent sur le *je* présidentiel).

Lexico 2 et 3, conçus initialement pour un dépouillement statistique des formes graphiques[29], n'ont pas été spécialement développés pour des données étiquetées. *Hyperbase*, depuis la version 5.5 permet de travailler sur un texte catégorisé et lemmatisé par *Cordial*, ou par *Treetagger*, ou *Winbrill*, qui ouvre des perspectives expérimentales, même si le choix du catégoriseur n'est pas offert. Il offre par ailleurs le choix le plus important de fonctionnalités lexicométriques à l'utilisateur, même si les procédures de traitement restent relativement lourdes, en cas d'expériences notamment.

29 Les développements de *Lexico 3* sont plutôt orientés vers l'ouverture au multilinguisme avec le passage à l'unicode et vers la navigation textométrique avec l'amélioration des Types généralisés.

Outils structurants (*Alceste*, *Iramuteq*, *Astartex*)

Ces outils[30] visent à faire émerger les structures saillantes d'un texte ou d'une partie du texte en opposition aux mesures constatives ou longitudinales qui comparent la composition lexicale de plusieurs parties d'un même corpus.

La méthodologie *Alceste*[31] présente des particularités originales. Son algorithme ne repose pas sur une segmentation pré-établie mais constitue des classes d'énoncés indépendamment des grandes divisions du corpus. Alors que la procédure lexicométrique standard consiste à segmenter le texte en unités minimales, à produire un tableau lexical et à croiser les unités minimales et les parties du corpus, *Alceste* construit un tableau qui croise les énoncés et les vocables et les compose en termes de présence-absence. Cette démarche inductive, fondée sur une analyse statistique distributionnelle de type harrissien met en évidence les grandes articulations du corpus, ses « mondes lexicaux[32] », en classant les énoncés du texte en fonction de la distribution de leur vocabulaire[33]. Le texte, considéré comme un ensemble

30 La généralisation de certains calculs comme le principe de la cooccurrence généralisée ou la mise en commun de certaines fonctionnalités complique une classification logicielle raisonnée. Ainsi les aspects structurants que nous évoquons ici pourraient fort bien s'appliquer à certaines fonctionnalités d'*Hyperbase*, mais aussi à *Astartex*, dans certains de leurs emplois. C'est peut-être en termes de fonctionnalités qu'il faudra désormais réfléchir, à la suite de Pincemin *et al.* par exemple (Bénédicte Pincemin *et al.*, « Fonctionnalités textométriques : proposition de typologie selon un point de vue utilisateur », *X^e Journées internationales d'analyse de données textuelles*, Rome, 8-11 mars 2010, JADT, 2010, p. 341-353, https://halshs.archives-ouvertes.fr/halshs-00856446, site consulté le 5 octobre 2015).

31 Analyse des Lexèmes Cooccurrents dans les Énoncés Simples d'un Texte.

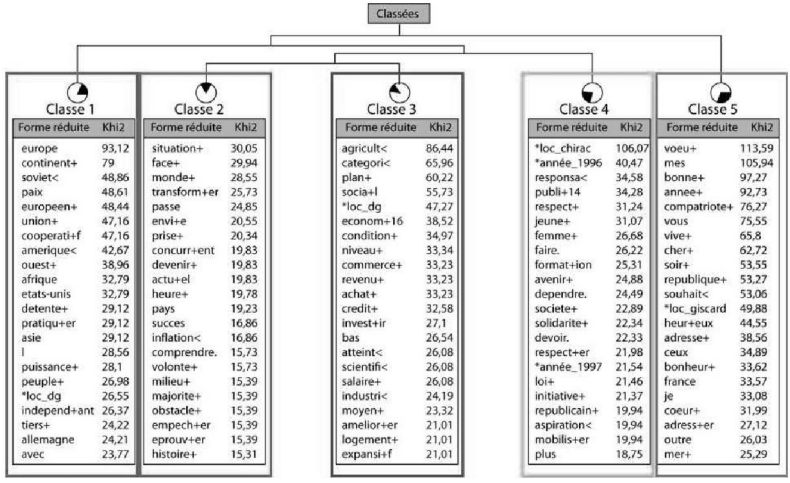
32 Voir Max Reinert, «Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars», *op. cit.* et Max Reinert, « Quelques interrogations à propos de l'"objet" d'une analyse de discours de type statistique et de la réponse "Alceste" », *Langage et société*, vol. 90, n° 1, p. 57-70, 1999.

33 On peut exprimer quelques réserves quant aux formulations présentées sur le site de la société *Image*, distributrice du logiciel, quand il y est affirmé que la méthodologie Alceste (la classification descendante hiérarchique) « n'exige aucune connaissance *a priori* sur le texte à analyser » et que cette analyse statistique des données textuelles permet de « décrire, classer, assimiler,

d'énoncés, est découpé en unités de contexte (U.C.E). Le logiciel effectue ensuite un repérage des unités lexicales, identifiées au moyen d'un dictionnaire, puis procède à la lemmatisation de ces formes. Les énoncés sont ensuite triés en fonction de la présence/absence des formes qui les composent puis classés selon la méthode de classification descendante hiérarchique. On obtient ainsi des classes de mots les plus représentatifs de ces énoncés, triés selon leur coefficient d'association à la classe par la méthode du khi2. Cette méthodologie est désormais implémentée dans un outil logiciel libre, *Iramuteq* développé et conçu par Pascal Marchand et Pierre Ratinaud. *Alceste* produit des résultats qui, plus encore que l'analyse lexicométrique classique, sont à considérer avec précaution, en raison de leur apparente évidence.

Figure 3

Les mondes lexicaux du corpus vœux (1959-2001) – *Alceste*



synthétiser automatiquement un texte ». Ces termes, notamment celui d'assimilation, paraissent abusifs d'un point vue scientifique. En tout état de cause, le point essentiel dans une démarche lexicométrique bien comprise est la connaissance préalable approfondie du corpus que l'on livre à l'analyse, à moins de procéder à une simple fouille extensive des textes, sans objectif herméneutique.

La méthodologie *Alceste*, appliquée à notre corpus de vœux (ici 1959-2001) permet d'identifier des « mondes lexicaux » qui peuvent être considérés, après un minutieux retour au texte, comme les principales thématiques abordées dans les discours de vœux.

La classe 5 regroupe les énoncés du rituel (formules d'adresses, formules finales, expression d'une « empathie présidentielle »; « Mes chers compatriotes », « je vous adresse [...] », « bonne année »...). Le mot étoilé *loc_giscard* indique que ces énoncés sont essentiellement, mais pas exclusivement, constitutifs des discours de VGE. La classe 1 (Europe, continents, Amérique, soviétique, puissances, Allemagne) représente les énoncés relevant de la politique internationale. La présence de la variable *loc_DG* explique le caractère quelque peu daté de certaines formes lexicales (entente, détente, union soviétique). On en déduit également que dans cet état du corpus vœux la politique internationale est plus particulièrement présente dans les discours de vœux du premier président de la Cinquième République.

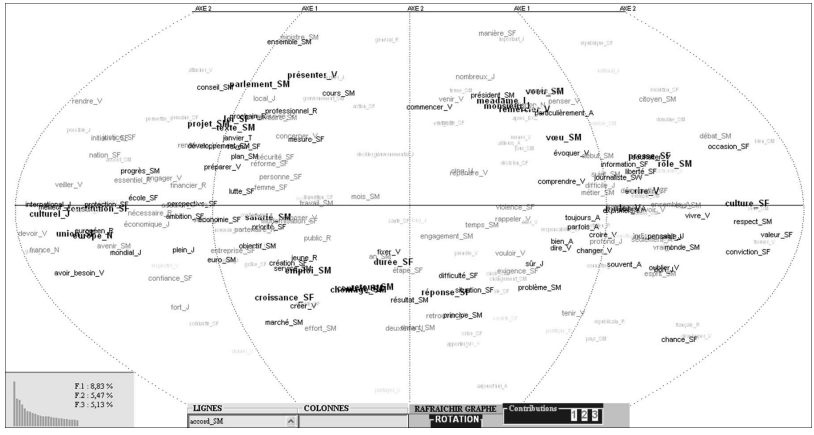
La méthode des cooccurrences généralisées, conçue par Jean-Marie Viprey et implémentée dans *Astartex*, fournit des résultats assez semblables à ceux produits par *Alceste*. Il s'agit d'extraire le profil cooccurrentiel, dit isotropie, des unités non séquentielles, à partir d'une matrice statistique[34]. Le logiciel extrait les *items* les plus fréquents (205 ou 400 selon le seuil choisi) en ôtant au préalable les mots-outils. Après extraction de ces formes les plus fréquentes (des substantifs pour l'essentiel), une matrice de cooccurrences est établie. Pour chaque forme retenue, on cherche les cooccurrents les plus fréquents. On obtient un tableau à double entrée qui croise ces termes et leurs cooccurrents, tableau que l'on soumet à une analyse factorielle, laquelle permet de mettre en évidence la relation, ou les réseaux associatifs qui se

34 Voir Margareta Kastberg et Jean-Marc Leblanc, « Extraction des isotopies d'un corpus textuel : analyse systématique des structures sémantiques et des coocurrences à travers différents logiciels textométriques », *Revue Texto*, vol. XVII, n° 3, 2012, coordonné par Christophe Cusimano, 2012, http://www.revue-texto.net/docannexe/file/3059/texto_kastberg_leblanc.pdf, site consulté le 20 septembre 2015.

construisent entre les différentes formes. Des zones sont alors identifiables qui peuvent s'interpréter comme autant de champs sémantiques construits à partir de formes cooccurrentes. *Astartex* propose aussi une visualisation originale prenant la forme d'un planisphère.

Figure 4

Projection de l'AFC sur trois axes ou micro distribution des 250 items les plus fréquents du corpus sous le mode « géodésique » – *Astartex*



Outre l'ajout du troisième axe factoriel, ce système de projection permet une visualisation différente des faits lexicaux, organisant en pôles les « constellations » de cooccurrents et des zooms permettent de se focaliser sur des pôles particuliers. Dans ses versions récentes, *Hyperbase* reprend le principe des cooccurrences généralisées dans les fonctionnalités de corrélat.

Catégoriseurs et évaluateurs sémantiques (*Tropes*, *Cordial*, etc.)

Tropes se présente comme un logiciel d'analyse de contenu, qui procède par catégorisation grammaticales et ontologies sémantiques *a priori*. Les catégories sémantiques projetées correspondent aux grands thèmes du texte analysé. Une syntaxe rudimentaire, diverses classification automatique des mots du texte, contraction et séquenciation du texte, détection des contextes, séries chrono-

logiques en constituent les fonctionnalités essentielles. Le logiciel offre la possibilité de visualiser les relations entre les thèmes dominants, sous forme de graphes, et de revenir aux contextes d'emplois des mots entrant dans la classe d'équivalence.

Cordial, d'abord catégoriseur morphologique, aide à la rédaction et correcteur orthographique, est aussi un outil d'analyse morphosyntaxique relativement fiable. Il offre des fonctionnalités documentaires comme les index et décomptes d'occurrences, les concordances, les segments répétés, l'extraction des phrases ou de paragraphes autour d'une forme pôle; il propose des dictionnaires, un conjugueur, un module d'évaluation et de comparaison stylistiques, des fonctionnalités sémantiques reposant sur des classifications. Il fournit en outre une série de mesures statistiques qui peuvent intervenir à différents moments de l'analyse. Les statistiques sur les corpus étiquetés livrent des indices sur les ambiguïtés rencontrées, sur la proportion des types grammaticaux, des genres, des nombres, des temps verbaux, sur les adverbes (négatifs, comparatifs, superlatifs), sur les fonctions grammaticales (sujet, complément d'objet...). Les statistiques complètes sur le texte offrent un éventail plus large encore de données qui portent sur le texte courant : moyennes grammaticales, phrases, ponctuation, morphologies (proportion d'articles indéfinis, de possessifs, d'indéfinis, de pronoms personnels à la première personne, de substantifs, d'adjectifs…), temps des verbes, nature des propositions, types grammaticaux, mais aussi domaines thématiques, niveau d'abstraction stylistique. D'autres données sont produites à travers la comparaison des textes analysés par rapport à un corpus de référence interne au logiciel composé de 2 600 ouvrages classés par genres (littéraire, journalistique, technique, juridique, commercial). Mais on peut également caractériser le texte courant sur la base de classifications plus fines et juger du nombre total de mots de notre texte, de la proportion de verbes ou de substantifs, de thèmes, des types de proposition, en le comparant à la poésie lyrique, au roman sérieux, au roman policier, au récit de voyage, aux fables, aux comédies... Le logiciel établit par ailleurs un diagnostic stylistique

du texte courant et fournit un indice de lisibilité, traduisant le niveau d'abstraction et de difficulté du texte. La notion de mot clé fournit, pour chaque texte, une liste de mots signifiants classés selon leur indice de sur-présence dans le texte. Différents indices sont livrés comme le taux de répétition ou le rapport entre occurrences mots et hapax occurrences. On mentionnera également les notions de phrases clés, de mots délaissés et de concepts délaissés. Quant aux concepts fondamentaux, ils reposent sur des taxinomies sémantiques. Les thèmes saillants du texte analysé sont dotés d'un coefficient produit à partir d'une comparaison avec un corpus de référence. Le taux est calculé à partir de l'écart réduit[35].

Les données statistiques fournies par *Cordial* qui portent sur des traits syntaxiques, grammaticaux, morphologiques, sémantiques peuvent être utilisées pour dresser des formes de typologies, en soumettant par exemple les tableaux (statistiques globales) à l'analyse factorielle. Ces préoccupations se rapprochent des travaux de Douglas Biber[36], mais sont également illustrées par Thomas Beauvisage[37].

Mesures contrastives, multidimensionnelles

L'analyse factorielle des correspondances (AFC) figure parmi les outils de statistique multidimensionnelle fondamentaux en lexicométrie, au même titre que les classifications automatiques et autres représentations arborées qui en sont complémentaires. Il s'agit, pour l'AFC, et dans le strict cadre des corpus textuels,

35 Nous faisons la distinction entre logiciels de type lexicométrique, cooccurrenceurs, analyseurs sémantiques, mais pourrions tout aussi bien distinguer les logiciels issus du domaine de la recherche universitaire (et qui y sont restés) des logiciels purement commerciaux, tant il est vrai que la philosophie en est parfois différente.

36 Douglas Biber, « On the Complexity of Discourse Complexity: A Multidimensional Analysis », *Discourse Processes*, vol. 15, nº 2, 1992, p. 133-163; et Douglas Biber, « Using Register-Diversified Corpora for General Language Studies », *Computational Linguistic*, vol. 19, nº 2, 1993, p. 243-258.

37 Thomas Beauvisage, « Exploiter des données morphosyntaxiques pour l'étude statistique des genres. Application au roman policier », *Traitement automatique des langages*, vol. 42, nº 2, 2001, p. 579-608.

de représenter globalement, au plus juste les grandes oppositions qui sous-tendent un corpus, d'en repérer les faits saillants en termes de proximités. On se reportera aux travaux de Philippe Cibois[38] et à ceux de Ludovic Lebart et de ses collaborateurs[39] pour d'éventuels approfondissements. Selon les logiciels, ces analyses ne portent pas nécessairement sur les mêmes données. L'analyse factorielle peut porter sur le tableau lexical et produire des représentations traduisant dans sa globalité la répartition du stock lexical entre les différentes parties du corpus. On parle de tableau lexical entier lorsque la totalité des formes graphiques est prise en compte ou d'un sous-ensemble de ce tableau lorsque l'on applique un seuil et qu'on ne tient compte que des formes atteignant une certaine fréquence. Ce tableau à double entrée croise, en colonnes les parties du corpus (locuteurs, bornes chronologiques, chapitres…c'est-à-dire les partitions sur lesquelles porte l'analyse contrastive) et, en lignes, les mots qui composent le corpus, éventuellement les segments répétés.

Le calcul de l'AFC est identique quel que soit le logiciel utilisé (bien qu'il existe d'autres mesures factorielles comme les analyses en composantes principales…). Toutefois le contenu du tableau est variable. Sous *Lexico* ce tableau est composé des fréquences absolues des mots du dictionnaire. Sous *Hyperbase* les fréquences sont pondérées par le calcul de l'écart réduit, en particulier pour compenser d'éventuels déséquilibres entre le poids des lignes et celui des colonnes. Sous *Alceste*, enfin, le tableau est d'une toute autre nature puisqu'il ne croise plus les mots et les parties du corpus mais les énoncés et les mots qui les composent.

Une autre utilisation de l'analyse factorielle proposée par *Hyperbase* est la représentation factorielle de listes qui porte de fait sur un tableau plus restreint.

On voit donc que plusieurs représentations factorielles plusieurs résultats sont envisageables, qu'il convient d'utiliser selon la nature du corpus et les hypothèses de recherche.

38 Philippe Cibois, *L'analyse factorielle*, Paris, Presses universitaires de France, 1994.

39 Ludovic Lebart *et al.*, *Statistique exploratoire multidimensionnelle*, Paris, Dunod, 2000.

Les analyses multidimensionnelles ne fournissent que des tendances globales qu'il convient d'affiner par la suite, des pistes de recherche, des indices, mais en aucun cas elles ne doivent être considérées comme une synthèse infaillible du corpus considéré. La multiplicité des résultats (néanmoins très approchants) démontre l'utilité de confronter différents états d'un même corpus, la nécessité de recouper les résultats de vérifier, d'affiner, de s'interroger, d'en éprouver la validité[40], la pratique idéale étant d'adopter un outil informatique en fonction d'une problématique, de son adéquation aux hypothèses et des stratégies de recherche, et non en fonction de sa disponibilité.

La connexion des vocabulaires, ou distance lexicale, consiste à l'origine à considérer le vocabulaire intégral d'un texte en ne tenant compte pour un mot donné que de sa présence ou absence dans les sous-parties. Ainsi deux textes sont proches s'ils ont en commun un nombre important de mots. La formule de la distance lexicale revient à calculer la part commune du vocabulaire et la part privative. La notion de présence/absence a cependant tendance à privilégier les raretés de vocabulaire sans apprécier le dosage des fréquences. Ainsi la distance intertextuelle intègre désormais cette préoccupation fréquentielle.

Les classifications automatiques, complémentaires de l'analyse factorielle, sont de plusieurs types : la classification ascendante hiérarchique; la classification descendante (utilisée sous *Alceste*). Outre la distinction principale qu'il convient d'opérer sur la base des tableaux lexicaux, on voit que tous ces outils ne disposent pas des mêmes fonctionnalités.

Tous les logiciels de lexicométrie sont issus peu ou prou de l'analyse des données textuelles de Jean-Paul Benzécri, *Alceste* en particulier, mais aussi de la statistique lexicale de Charles Muller. *Lexico*, *Weblex*, *TXM* sont plus proches de la tradition lexicométrique de Saint-Cloud, à la fois dans la conception et dans les études qui sont menées sur le discours politique, tandis qu'*Hyperbase*

40 Des outils commerciaux comme *XLStat* proposent des classifications automatiques et des analyses factorielles ou encore des mesures du Khi2. Tout le problème est alors de définir les données qui seront soumises à ces analyses.

a des origines plus littéraires. Les travaux d'Étienne Brunet sur le vocabulaire de Giraudoux[41], de Proust[42], de Zola[43], mais aussi sur l'œuvre de Rousseau[44] – certes plus politique – ont probablement grandement orienté sa conception de l'analyse statistique du vocabulaire, et donc le développement de l'outil qui fait la part belle aux problématiques stylistiques – voire stylométriques – comme la richesse, l'accroissement, la structure du vocabulaire. Les versions successives du logiciel offrent un choix toujours plus grand encore d'analyses multidimensionnelles, de calculs, de représentations, de pondérations. On peut supposer que trop de choix relatifs aux modes de calculs ne compliquent la tâche du chercheur qui ne serait pas statisticien, car les analyses statistiques sont nombreuses qui pourraient être appliquées aux données textuelles. L'outil lexicométrique peut aussi être considéré comme un «garde-fou» pour l'analyste qui ne maîtrise pas nécessairement le détail des méthodes statistiques, car nous pourrions tout aussi bien utiliser des interfaces purement statistiques comme *XLstat* et produire nos propres analyses, mais qui nous garantirait la pertinence de nos choix? La multiplicité des modes de calcul et des analyses multidimensionnelles fournies par *Hyperbase* suppose une connaissance approfondie en statistique générale : calcul de la distance lexicale, sur les formes ou sur les occurrences, factorielle sur V (vocabulaire, formes) ou sur N (occurrences), analyse arborée, sur N, sur V, portant, au choix, sur les distances brutes, sur le khi2 ou sur le carré des distances, en présentation radiale ou rectangulaire… Mais on peut aussi citer les histogrammes lignes colonnes portant sur les formes ou les occurrences, l'histogramme de la richesse lexicale portant sur les hapax ou sur le vocabulaire, de même que les analyses possibles sur les annota-

41 Étienne Brunet, *Le vocabulaire de Jean Giraudoux: structure et évolution. Statistique et informatique appliquées à l'étude des textes à partir des données du Trésor de la langue française*, Genève, Slatkine, 1978.

42 Étienne Brunet, *Le vocabulaire de Marcel Proust*, Genève, Slatkine, 1983.

43 Étienne Brunet, *Le vocabulaire de Zola*, Genève, Slatkine, 1985.

44 Étienne Brunet, *Index de l'Émile, XLIII-LIII*, dans *Études rousseauistes et Index des œuvres de J.J. Rousseau*, Genève, Slatkine, 1980 et Étienne Brunet, *Index des Lettres écrites de la montagne*, dans *Études rousseauistes et Index des œuvres de J.J. Rousseau*, Genève, Slatkine, 1983.

tions du corpus, l'accroissement lexical par tranches ou portant sur les divisions naturelles du texte. Tout ceci rend le logiciel *Hyperbase* moins accessible au néophyte. *Lexico 3* semble plus apte à la pratique d'une expertise rapide, *Hyperbase* représentant un dispositif moins expérimental, nécessitant que les corpus soient stabilisés.

Figure 5

Arborée de la distance lexicale sur le corpus 1959-2014 – Hyperbase V9

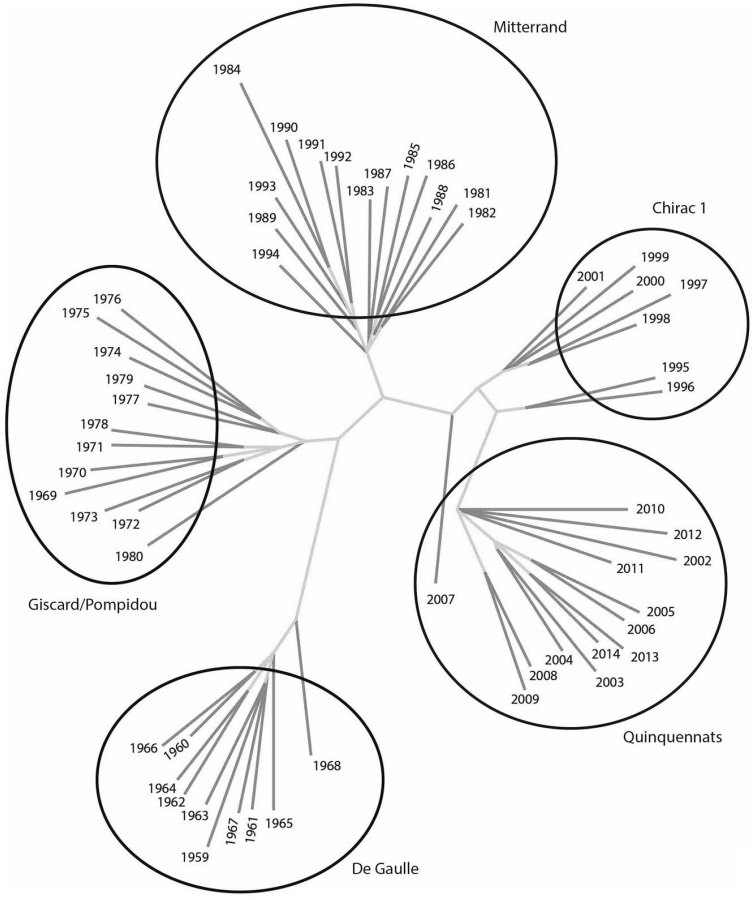
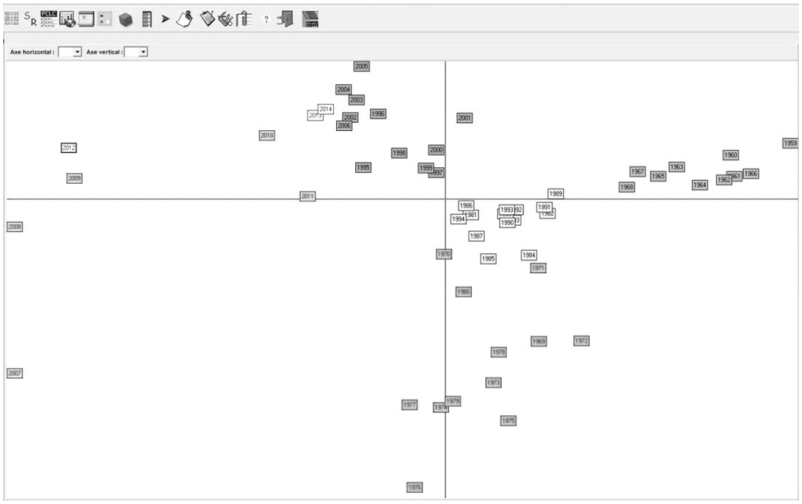


Figure 6
Analyse factorielle - corpus vœux (1959-2014) – *Lexico 3*



Commentaire: l'analyse factorielle pratiquée ici permet de mettre au jour les distances et éloignements des textes (les discours de vœux des présidents de la Cinqième République) en termes d'emploi du stock lexical.

L'analyse arborée de la connexion lexicale, calculée au moyen d'*Hyperbase* et ne tenant compte que de la présence/absence du vocabulaire mobilise un autre calcul et offre une visualisation différente. Cette analyse arborée clarifie les positions et les contrastes et permet, notamment, d'observer une caractéristique intéressante de ce corpus de vœux présidentiels qui semble révéler une rupture entre les quinquennats et les septennats.

Précautions méthodologique et interprétatives

Les divers outils lexicométriques sont appropriés à condition qu'on ne s'en remette ni à une seule méthodologie pour commenter des sorties machine automatiques, ni à un seul corpus, quel que soit son étendue, mais qu'on fasse varier tant les outils que la dimension de données, de manière à tester, comparer et combiner ces outils, d'une part, et, d'autre part, à

tenir compte des incidences possibles de la variation des corpus dans l'établissement des résultats. C'est ce que nous développerons ci-après. Il convient aussi, plus particulièrement dans le domaine de la lexicométrie, de prendre en compte les unités de décompte. Interprète-t-on une analyse factorielle portant sur la forme graphique, sur des données lemmatisées, sur des catégories morphosyntaxiques? Cette analyse factorielle prend-elle en entrée des effectifs absolus, des données pondérées, des matrices de distance?

Problèmes de visualisation

Le principe même de l'analyse automatisée des données textuelles (lexicométrie, textométrie, analyse automatique de contenu) revient à déstructurer et à reconstruire le texte. Ainsi, l'examen d'un index hiérarchique ou d'une liste de fréquences est déjà une représentation particulière du texte, où la linéarité est visualisée sous forme d'un paradigme.

En outre les propositions de visualisations sont nombreuses en lexicométrie, bien que le plus souvent fortement dépendantes de la statistique, ce qui s'explique par la genèse de la discipline.

Le développement récent du web participe à la diffusion d'autres types de représentations que les spécialistes de la lexicométrie ne peuvent ignorer. La dimension multimédia de ces cartographies, réseaux, nuages ajoute au discours qu'elles produisent.

La presse elle-même fait appel désormais à ce type de visuel. *Le Monde* du 28 janvier 2010 illustre une rapide analyse du discours d'Obama sur l'état de l'Union par une représentation apparaissant sous la forme d'un nuage de mots qui met l'accent sur les mots *travail*, *économie* et *Américains*. Comment est généré ce nuage de mots, sur quelles bases? Quelle est sa valeur scientifique?

Ces outils sont désormais à la disposition de tous et nous avons mené au moyen du site Wordle[45] le même type d'expérience sur les vœux de Sarkozy, de décembre 2009, que nous produisons ici, et de décembre 2007[46]. Si nous nous laissions aller à une

45 http://www.wordle.net/.

46 Ces expériences figurent sur le blog de textopol : http://textopol.org/dotclear.

interprétation hâtive, et que nous cédions à l'attrait médiatique, nous pourrions commenter longuement la prégnance de l'adverbe *plus* sur le nuage qui suit et mettre ce fait en relation, sans à aucun moment avoir lu le texte, avec le fameux *travailler plus pour gagner plus* qui a nourri tant de commentaires.

Figure 7

Nuage de mots produit par Wordle, Sarkozy 2009

D'autres entreprises proposent d'analyser le web et présentent des visualisations utilisant un vocabulaire plastique analogue, mêlant mouvements, colorations et dimensions. Nous renvoyons ici aux sites tels que linkfluence[47] ou gapminder[48], parmi quelques autres.

Au-delà de ces entreprises spécifiques, le web regorge d'utilisations qui s'apparentent notamment aux réseaux de cooccurrents et autres approches statistiques et lexicométriques. C'est le cas des nuages de *Taggs* présents sur la plupart des sites et réseaux sociaux, c'est le cas aussi de sites marchands qui adaptent leur offre au profil d'achat des utilisateurs. On y indique aux inter-

47 http://fr.linkfluence.net/blog/.
48 http://www.gapminder.org/.

nautes que ceux qui ont acheté tel ouvrage ont aussi commandé tel ou tel autre.

Au sein des exemples cités ci-dessus les données sur lesquelles portent les représentations ne sont pas toujours clairement établies.

En outre, l'excès de visualisation et d'interactivité peut éloigner du sens initial, voire produire des objets dont la valeur n'est quasiment qu'esthétique.

Si les cartographies lexicométriques portent sur des données connues et des modes de représentation établis, il n'en est pas toujours de même pour les exemples que nous venons de citer. De plus ces représentations issues du web et autres réseaux sociaux multiplient les possibilités visuelles, sonores et tous ces outils du multimédia créent un sens mouvant et pluriel que le créateur d'objet multimédia se doit de construire avant de les appliquer à nos outils.

Dans toute analyse mobilisant une représentation il importe avant tout de savoir quelles sont les données qui sont représentées, comment elles ont été recueillies, mesurées, et ce que représente le graphique proposé.

Belles images

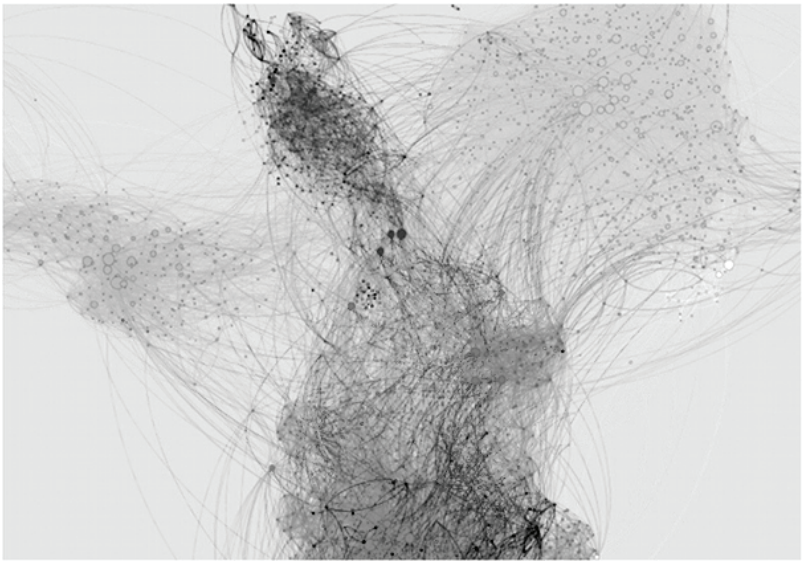
En archéologie, dans les siècles passés, les représentations architecturales et les restitutions de sites antiques prenaient comme modalités de représentation les codes que les architectes utilisaient dans leur métier (c'est-à-dire la visualisation de leur construction en cours); ces représentations ont donc été interprétées comme étant réelles et reflétant la réalité. Elles représentaient ce que l'architecte imaginait du monument antique mais non l'état réel du monument. Ces belles images, « trop réalistes », ont ainsi longtemps bloqué l'imaginaire des chercheurs en archéologie. Une représentation doit donc montrer ce qu'elle est. Si elle est une hypothèse, la visualisation doit le montrer au moyen d'un

code de représentation qui puisse exprimer le degré d'hypothèse qu'elle contient[49].

Dans l'exemple qui suit, il est impossible d'interpréter le visuel produit si l'on ne possède pas les codes, les légendes, les algorithmes mobilisés, les mesures effectuées, la nature des données[50].

Figure 8

Exemple de visualisation Gefi



La visualisation en lexicométrie et statistique textuelle

La connaissance relative des algorithmes et des données (*a contrario* d'une utilisation d'outils « boîtes noires »), des conditions de

49 Voir Jean-Marc Leblanc et Marie Pérès, « Modèles tridimensionnels pour la représentation de l'état des connaissances et propositions de visualisation pour l'analyse des corpus textuels », *XII^e Journées internationales d'analyse statistique des données textuelles*, Paris, 3-6 juin 2014, *JADT*, 2014, p. 373-384, http://lexicometrica.univ-paris3.fr/jadt/jadt2014/01-ACTES/31-JADT2014.pdf, site consulté le 5 octobre 2015.

50 Cette question est abordée en détail dans le cadre d'une journée d'étude organisée au Céditec en 2013, *Réflexion sur les visualisations en sciences humaines, quels apports pour la textométrie?* dont on trouvera les supports de communication à l'adresse suivante : http://textopol.u-pec.fr/?p=1013.

recueil et de production du corpus est indispensable à toute interprétation en lexicométrie.

Pour autant les résultats fournis par ces outils, pourtant inscrits dans une démarche méthodologique solidement éprouvée, ne sont pas à considérer comme des vérités absolues mais comme des pistes de recherche. L'interprétation d'une analyse factorielle de correspondances ne peut se limiter à la description de la position des différents points sur le plan factoriel; elle doit s'accompagner d'un retour au texte systématique. Cette interprétation doit tenir compte de l'unité minimale soumise au calcul de l'analyse factorielle[51], tenir compte des effets de corpus, de la chronologie, des éventuels écarts des différentes parties du corpus... Il convient également d'interpréter ces analyses en tenant compte de la nature du corpus. L'AFC d'un corpus formant une suite textuelle chronologique ne s'interprètera pas comme un corpus synchronique opposant différents locuteurs ou émetteurs.

Il ne s'agit pas non plus de se livrer à des simples commentaires de listes (mots les plus fréquents de tel ou tel auteur, vocabulaire spécifique de telles ou telles partie) mais bien de remettre ces faits discursifs et lexicaux en contexte.

Alceste produit des résultats qui, plus encore que l'analyse lexicométrique classique, sont à considérer avec précaution, en raison de leur apparente évidence.

Protocole d'analyse lexicométrique : pour sortir du commentaire de listes et de tableaux

Il convient ici de revenir rapidement sur le principe de la démarche lexicométrique et d'en souligner trois points essentiels.

- Premier principe, la lexicométrie est avant tout contrastive et longitudinale[52]. Elle n'a de sens que sur des textes partitionnés,

[51] La configuration factorielle portant sur la forme graphique (le mot) ou sur les catégories grammaticales n'appelle pas les mêmes remarques et ne relève pas des mêmes phénomènes.

[52] D'autres approches, que nous ne qualifierons pas de lexicométriques au sens strict, permettent de porter sur des textes non partitionnés et de les caractériser sur la base de la distribution interne de leur lexique, le plus souvent sur des profils de cooccurrences (*Alceste, Astartex*, par exemple).

c'est-à-dire munis de divisions, naturelles ou expérimentales, comme les chapitres d'un livre, les différentes livraisons d'une revue, les diverses allocutions d'un même locuteur, individuel ou collectif…

- Deuxième principe, qui découle du précédent, la lexicométrie repose avant tout, mais non exclusivement, sur une norme endogène selon laquelle le corpus est sa propre référence. On compare chaque partie à l'ensemble et à chacune des parties du corpus.
- Troisième principe, la lexicométrie repose sur la segmentation du texte en unités, que l'on dénombre, trie, classe, met en contexte, soumet à des analyses statistiques. Les unités soumises au décompte peuvent être la forme graphique (suite de caractères délimitée par un blanc, ce que nous développerons plus particulièrement ici), le segment, la cooccurrence, le lemme, la catégorie (morphosyntaxique, sémantique…), le n-gramme, ou une combinaison de ces unités.

Travail en surface (forme graphique) ou sur données catégorisées?

À cette segmentation en unités minimales peut s'ajouter un traitement supplémentaire, comme la catégorisation qui revient à projeter sur le corpus une série d'étiquettes au moyen desquelles on peut affiner la recherche de motifs textuels et les regroupements.

Figure 9

Exemple de données catégorisées par *Cordial* et par *Teetagger*

Forme	*Eti_1*	*Lemme_V*
Mes	DETPOSS	Mon
Chers	ADJMP	cher
Compatriotes		compatriote
,	PCTFAIB	,
\r\r		
Vous	PPER2P	Vous
m'	PPER1S	me
avez	VINDP2P	avoir
élu	VPARPMS	élire
,	PCTFAIB	,
en	PREP	en
mai	NCMIN	mai
dernier	ADJMS	dernier
,	PCTFAIB	,
pour	PREP	pour
que	SUB	que
nous	PPER1P	nous
construisions	VINDI1P	construire
ensemble	ADV	ensemble
une	DETIFS	un
nouvelle	ADJFS	nouveau
France	NPFS	France

A1 : Mes

	A	B	C
1	Mes	DET:POS	mon
2	chers	ADJ	cher
3	compatriote	NOM	compatriote
4	,	PUN	,
5	C'	NAM	<unknown>
6	est	VER:pres	être
7	un	DET:ART	un
8	message	NOM	message
9	de	PRP	de
10	confiance	NOM	confiance
11	et	KON	et
12	de	PRP	de
13	volonté	NOM	volonté
14	que	PRO:REL	que
15	je	PRO:PER	je
16	vous	PRO:PER	vous
17	adresse	VER:pres	adresser
18	ce	PRO:DEM	ce
19	soir	NOM	soir
20	en	PRP	en
21	vous	PRO:PER	vous
22	présentant	VER:ppre	présenter
23	mes	DET:POS	mon
24	vœux	NOM	vœu
25	pour	PRP	pour
26	la	DET:ART	le
27	nouvelle	ADJ	nouveau

Voeux2014_

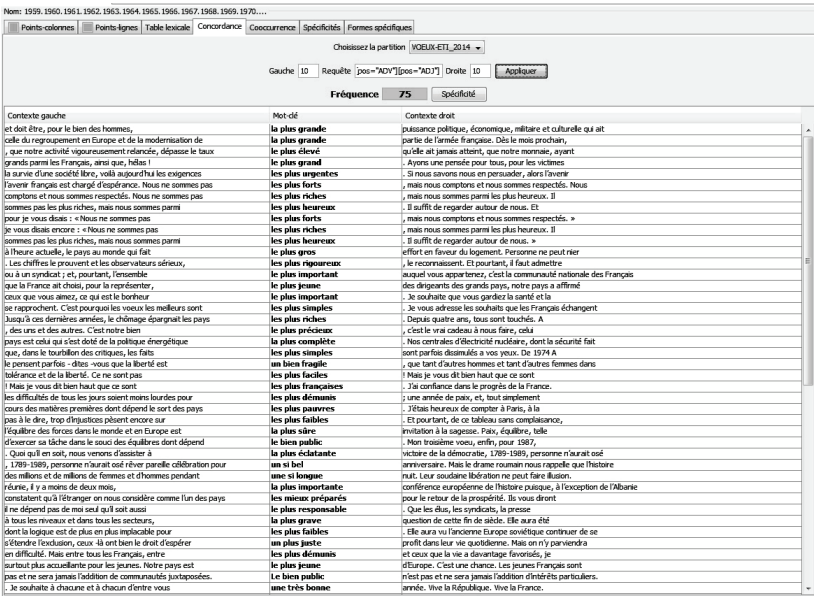
Sur la base d'une annotation morphosyntaxique, il est possible d'effectuer des retours au texte et d'extraire des empans textuels afin d'étudier des phénomènes relevant de la phraséologie.

Sur un corpus catégorisé par l'étiqueteur *Treetagger*, et soumis à *TextObserver*, nous recherchons un motif syntaxique du type : déterminant suivant d'un adverbe, suivi d'un adjectif. Le but étant de faire apparaître toutes les portions de texte correspondant à cette modélisation grammaticale. La traduction de cette requête prendra la forme suivante :

[pos=»DET:ART»][pos=»ADV»][pos=»ADJ»]

La recherche de concordances portant sur ce jeu d'étiquette permet d'isoler les réalisations ci-dessous.

Figure 10


Nom: 1999. 1960. 1961. 1962. 1963. 1964. 1965. 1966. 1967. 1968. 1969. 1970....

Points-colonnes | Points-lignes | Table lexicale | Concordance | Cooccurrence | Spécificités | Formes spécifiques

Choisissez la partition VOEUX-ETI_2014

Gauche 10 Requête pos="ADV"[pos="ADJ"] Droite 10 Appliquer

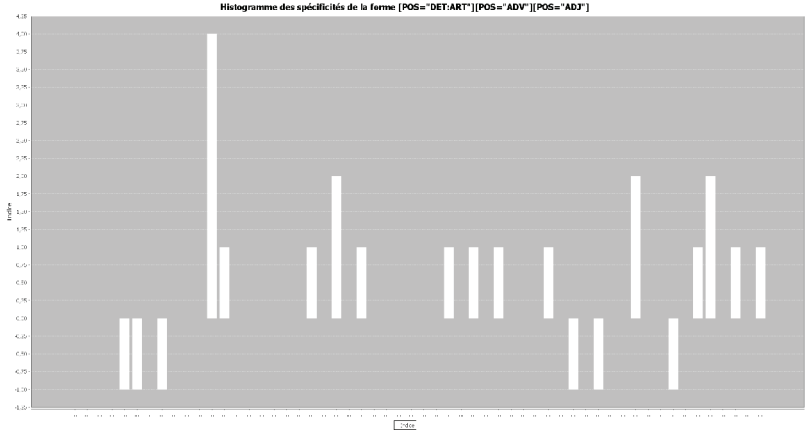
Fréquence 75 Spécificité

Contexte gauche	Mot-clé	Contexte droit
et doit être, pour le bien des hommes,	**la plus grande**	puissance politique, économique, militaire et culturelle qui ait
celle du regroupement en Europe et de la modernisation de	**la plus grande**	partie de l'armée française. Dès le mois prochain,
, que notre activité vigoureusement relancée, dépasse le taux	**le plus élevé**	qu'elle ait jamais atteint, que notre monnaie, ayant
grands parmi les Français, ainsi que, hélas !	**le plus grand**	. Ayons une pensée pour tous, pour les victimes
la survie d'une société libre, voilà aujourd'hui les exigences	**les plus urgentes**	. Si nous savons nous en persuader, alors l'avenir
l'avenir français est chargé d'espérance. Nous ne sommes pas	**les plus forts**	, mais nous comptons et nous sommes respectés. Nous
comptons et nous sommes respectés. Nous ne sommes pas	**les plus riches**	, mais nous sommes parmi les plus heureux. Il
sommes pas les plus riches, mais nous sommes parmi	**les plus heureux**	. Il suffit de regarder autour de nous. Et
pour je vous disais : « Nous ne sommes pas	**les plus forts**	, mais nous comptons et nous sommes respectés. »
je vous disais encore : « Nous ne sommes pas	**les plus riches**	, mais nous sommes parmi les plus heureux. Il
sommes pas les plus riches, mais nous sommes parmi	**les plus heureux**	. Il suffit de regarder autour de nous. »
à l'heure actuelle, le pays au monde qui fait	**le plus gros**	effort en faveur du logement. Personne ne peut nier
. Les chiffres le prouvent et les observateurs sérieux,	**les plus rigoureux**	, le reconnaissent. Et pourtant, il faut admettre
ou à un syndicat ; et, pourtant, l'ensemble	**le plus important**	auquel vous appartenez, c'est la communauté nationale des Français
que la France ait choisi, pour la représenter,	**le plus jeune**	des dirigeants des grands pays, notre pays a affirmé
ceux que vous aimez, ce qui est le bonheur	**le plus important**	. Je souhaite que vous gardiez la santé et la
se rapprochent. C'est pourquoi les vœux les meilleurs sont	**les plus simples**	. Je vous adresse les souhaits que les Français échangent
Jusqu'à ces dernières années, le chômage épargnait les pays	**les plus riches**	. Depuis quatre ans, tous sont touchés. A
, des uns et des autres. C'est notre bien	**le plus précieux**	, c'est le vrai cadeau à nous faire, celui
pays est celui qui s'est doté de la politique énergétique	**la plus complète**	. Nos centrales d'électricité nucléaire, dont la sécurité fait
que, dans le tourbillon des critiques, les faits	**les plus simples**	sont parfois dissimulés a vos yeux. De 1974 A
le pensent parfois - dites -vous que la liberté est	**un bien fragile**	, que tant d'autres hommes et tant d'autres femmes dans
tolérance et de la liberté. Ce ne sont pas	**les plus faciles**	! Mais je vous dit bien haut que ce sont
! Mais je vous dit bien haut que ce sont	**les plus françaises**	. J'ai confiance dans le progrès de la France.
les difficultés de tous les jours soient moins lourdes pour	**les plus démunis**	; une année de paix, et, tout simplement
cours des matières premières dont dépend le sort des pays	**les plus pauvres**	. J'étais heureux de compter à Paris, à la
pas à le dire, trop d'inquiétudes pèsent encore sur	**les plus faibles**	. Et pourtant, de ce tableau sans complaisance,
l'équilibre des forces dans le monde et en Europe est	**la plus sûre**	invitation à la sagesse. Paix, équilibre, telle
d'exercer sa tâche dans le souci des équilibres dont dépend	**le bien public**	. Mon troisième voeu, enfin, pour 1987,
. Quoi qu'il en soit, nous venons d'assister à	**la plus éclatante**	victoire de la démocratie, 1789-1989, personne n'aurait osé
, 1789-1989, personne n'aurait osé rêver pareille célébration pour	**un si bel**	anniversaire. Mais le drame roumain nous rappelle que l'histoire
des millions et de millions de femmes et d'hommes pendant	**une si longue**	nuit. Leur soudaine libération ne peut faire illusion.
réunie, il y a moins de deux mois,	**la plus importante**	conférence européenne de l'histoire puisque, à l'exception de l'Albanie
constatent qu'à l'étranger on nous considère comme l'un des pays	**les mieux préparés**	pour le retour de la prospérité. Ils vous diront
il ne dépend pas de moi seul qu'il soit aussi	**le plus responsable**	. Que les élus, les syndicats, la presse
à tous les niveaux et dans tous les secteurs,	**la plus grave**	question de cette fin de siècle. Elle aura été
dont la logique est de plus en plus implacable pour	**les plus faibles**	. Elle aura vu l'ancienne Europe soviétique continuer de se
s'étendre l'exclusion, ceux -là ont bien le droit d'espérer	**un plus juste**	profit dans leur vie quotidienne. Mais on n'y parviendra
en difficulté. Mais entre tous les Français, entre	**les plus démunis**	et ceux que la vie a davantage favorisés, je
surtout plus accueillante pour les jeunes. Notre pays est	**le plus jeune**	d'Europe. C'est une chance. Les jeunes Français sont
pas et ne sera jamais l'addition de communautés juxtaposées.	**Le bien public**	n'est pas et ne sera jamais l'addition d'intérêts particuliers.
. Je souhaite à chacune et à chacun d'entre vous	**une très bonne**	année. Vive la République. Vive la France.

TextObserver permet ensuite de calculer les spécificités de ce motif, sur la partition en années ou en locuteurs, c'est-à-dire les sur-emplois ou sous-emplois de ce motif.

Figure 11

Histogramme des spécificités du motif [pos=»DET:ART»] [pos=»ADV»][pos=»ADJ»]



L'exploration à partir des catégories grammaticales permet en outre d'identifier des caractéristiques individuelles, en recourant par exemple à la méthode des spécificités. L'exemple ci-dessous représente les catégories grammaticales spécifiques de Chirac sur les corpus des vœux de 1959 à 2001. Le catégoriseur utilisé dans cet exemple est *Cordial*.

Le tableau produit ci-après présente donc les étiquettes morphosyntaxiques sur-employées ou sous-employées par Chirac par rapport aux autres locuteurs de notre corpus.

Figure 12

Spécificités des étiquettes morphosyntaxiques chez Jacques Chirac (vœux 1959-2001)

Forme	Frq. Tot.	Fréquence	Coeff.	Forme	Frq. Tot.	Fréquence	Coeff.
DETIFS	350	124	10	DETDFS	1928	375	-2
VINF	1330	366	8	PREP	5468	1101	-2
NCFP	899	244	5	ADJINT	30	2	-2
NCPIG	137	48	4	VIMPP2P	28	1	-2
NCFS	2852	677	4	VINDI3S	72	8	-2
PPER1P	654	166	3	VCONP3S	43	3	-2
ADV	2388	560	3	VINDF3S	210	31	-2
VINDP1P	311	87	3	NCMP	1027	197	-2
ADJFS	670	162	2	COO	1628	315	-2
VPARPFS	141	40	2	PRI	1017	191	-2
ADJFP	298	77	2	NPMS	93	9	-3
VINDP3P	542	132	2	SUB	897	145	-5
VINDP3S	1537	357	2	NCMIN	794	122	-5
NCFIN	106	13	-2	VPARPMS	497	71	-5
				ADJNUM	236	19	-8

Là encore, sans retour au texte ces résultats n'ont aucun sens.

La première étiquette « DETIFS » (Déterminant Indéfini Féminin Singulier) renvoie à une caractéristique chiraquienne (une « France », « une Europe », « une solidarité »). La seconde étiquette « VINF » (Verbes à l'Infinitif) est indissociable de l'emploi précédent. Le retour au texte et l'examen des régularités révèlent en effet que ces emplois de l'infinitif interviennent principalement au sein de constructions déontiques du type : « au gouvernement de + *infinitif*, nous devons + *infinitif* … ». L'énoncé type identifié par cette méthode est d'ailleurs le suivant: « Nous devons *construire* une France plus accueillante ».

Une interprétation fiable des corpus en lexicométrie, et plus généralement en statistique textuelle, se fonde, selon nous, sur :

- une connaissance des postulats méthodologiques qui ont conduit au développement des outils logiciels et du protocole d'analyse;
- une compréhension des algorithmes ou du moins des méthodes statistiques appliquées (khi2, loi normale, loi hypergéométrique);
- une connaissance de la nature des données représentées (unités minimales, formes graphiques, catégories, lemmes, segmentation);

- une maîtrise des règles présidant à la lecture des visuels proposés. Alors que l'analyse factorielle, dont le calcul est pourtant connu et la méthode éprouvée depuis le début des années 1970, résiste encore parfois à l'analyse, de nouveaux visuels émergent, qui nécessitent un apprentissage préalable (voir l'exemple de *Gefi* plus haut);
- le retour au texte systématique;
- la construction de corpus s'inscrivant dans un genre ou dans une situation de communication homogène;
- une connaissance préalable du corpus (sources, données sociologiques conditions de collection et de recueil du corpus).

Enfin, nous proposons une démarche ne s'appliquant pas nécessairement aux très gros corpus, dans laquelle l'exploration s'appuie dans un premier temps sur des phénomènes typologiques et tendanciels pour atteindre la granularité du texte.

Une telle démarche pourrait se pencher sur :

- les caractéristiques quantitatives du corpus (taille du corpus et des sous-parties);
- une analyse typologique (analyse factorielle ou classification permettant de juger de la proximité entre les différentes parties de ce corpus en termes d'emploi du stock lexical);
- une analyse des spécificités (vocabulaire sur-employé ou sous-employé dans une partie par rapport aux autres et à l'ensemble);
- l'examen des attirances entres les mots du texte (cooccurrences associées à un pôle);
- le retour au texte sous la forme de concordances ou de contextes, ou la recherche d'empans textuels au moyen d'expressions régulières portant sur la forme graphique ou sur des catégories morphosyntaxiques.

Conclusion : pour une démarche expérimentale en lexicométrie

En complément de cette recherche, nous suggérons qu'une démarche expérimentale permet de fiabiliser l'interprétation.

Cette méthode consiste à croiser les outils et les fonctionnalités (on confrontera, par exemple, l'analyse factorielle du tableau lexical à une analyse arborée des distances) ou l'on utilisera un outil de type longitudinal comme *Lexico 3*, *TXM*, *TextObserver* en complément d'un outil « structurant » comme *Alceste*, *Iramuteq* ou *Astartex*. Enfin on fera varier les corpus, qu'il s'agisse de comparer deux états différents comme les discours de vœux de 1959 à 2001 et ces mêmes discours de 1959 à 2015 ou de compléter l'analyse menée sur l'ensemble des discours de vœux à des sous-corpus (les vœux du général de Gaulle, les vœux de Mitterrand, les vœux de Sarkozy ou de Hollande).

La variation peut aussi porter sur des éléments du lexique ou du discours, comme par exemple la suppression des marques personnelles ou du vocabulaire rituel[53].

[53] Nous présentons cette méthode, implémentée dans l'outil *TextObserver* dans Amani Daknou et Jean-Marc Leblanc, « TextObserver/WebObserver : Propositions ergonomiques pour l'exploration et l'exploitation des données textuelles multidimensionnelles», dans Stéphane Chaudiron, Madjid Ihadjadene et Bernard Jacquemin (dir.), *Dispositifs numériques : contenus interactivité et visualisation*, Actes du 16e colloque international sur le document électronique (CIDE 16), Lille, 21-22 novembre 2013, Paris, Europia; et dans Christine Barats et Jean-Marc Leblanc, « Exploration de corpus multimodaux pour l'analyse d'un processus de médiatisation : l'exemple du "classement de Shanghai" dans la presse francophone (2003-2010) et de son incidence sur la présentation de soi des universités sur leurs pages Web », *XIe Journées internationales d'analyse statistique des données textuelles*, Liège, 13-15 juin 2012, JADT, 2012, p. 81-93, http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Barats,%20Christine%20et%20al.%20-%20Exploration%20de%20corpus%20multimodaux.pdf, site consulté le 6 octobre 2015.

Bibliographie

Bakthine, Mikhaïl, *La poétique de Dostoïevski*, Paris, Seuil, coll. « Points », 1970 [1929].

Christine Barats et Jean-Marc Leblanc, « Exploration de corpus multimodaux pour l'analyse d'un processus de médiatisation : l'exemple du "classement de Shanghai" dans la presse francophone (2003-2010) et de son incidence sur la présentation de soi des universités sur leurs pages Web », *XIe Journées internationales d'analyse statistique des données textuelles*, Liège, 13-15 juin 2012, *JADT*, 2012, p. 81-93, http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Barats,%20Christine%20et%20al.%20-%20Exploration%20de%20corpus%20multimodaux.pdf, site consulté le 6 octobre 2015.

Beauvisage, Thomas, « Exploiter des données morphosyntaxiques pour l'étude statistique des genres. Application au roman policier », *Traitement automatique des langages*, vol. 42, no 2, 2001, p. 579-608.

Benzécri, Jean-Paul, *L'Analyse des données. Leçons sur l'analyse factorielle et la reconnaissance des formes et travaux du laboratoire de statistique de l'Université de Paris VI*, Paris, Dunod, 1973.

Benzécri, Jean-Paul, *Histoire et préhistoire de l'analyse des données*, Paris, Dunod, 1982.

Benzécri, Jean-Paul, *Pratique de l'analyse des données*, Paris, Dunod, 1980.

Biber, Douglas, « On the Complexity of Discourse Complexity; a Multidimensional Analysis », *Discourse Processes*, vol. 15, no 2, 1992, p. 133-163.

Biber, Douglas, « Using Register-Diversified Corpora for General Language Studies », *Computational Linguistic*, vol. 19, no 2, 1993, p. 243-258.

Brunet, Étienne, « Au fond du goofre, un gisement de 44 milliards de mots », dans *Actes des Journées internationales d'analyse statistique des données textuelles*, JADT, 2012, p. 7-21.

Brunet, Étienne, *Index de l'Émile, XLIII-LIII*, dans *Études rousseauistes et Index des œuvres de J.J. Rousseau*, Genève, Slatkine, 1980.

Brunet, Étienne, *Index des Lettres écrites de la montagne*, dans *Études rousseauistes et Index des œuvres de J.J. Rousseau*, Genève, Slatkine, 1983.

Brunet, Étienne, *Le vocabulaire de Jean Giraudoux: structure et évolution. Statistique et informatique appliquées à l'étude des textes à partir des données du Trésor de la langue française*, Genève, Slatkine, 1978.

Brunet, Étienne, *Le vocabulaire de Marcel Proust*, Genève, Slatkine, 1983.

Brunet, Étienne, *Le vocabulaire de Zola*, Genève, Slatkine, 1985.

Callon, Michel *et al.*, *La scientométrie*, Paris, Presses universitaires de France, 1993.

Chateauraynaud, Francis, *Prospéro: une technologie littéraire pour les sciences humaines*, Paris, CNRS, 2003.

Cibois, Philippe, *L'analyse factorielle*, Paris, Presses universitaires de France, 1994.

Daknou, Amani et Jean-Marc Leblanc, « TextObserver/WebObserver : Propositions ergonomiques pour l'exploration et l'exploitation des données textuelles multidimensionnelles», dans Stéphane Chaudiron, Madjid Ihadjadene et Bernard Jacquemin (dir.), *Dispositifs numériques : contenus interactivité et visualisation*, Actes du 16e colloque international sur le document électronique (CIDE 16), Lille, 21-22 novembre 2013, Paris, Europia.

Foucault, Michel, *Dits et écrits* (tome 1, 1954-1975, et tome 2, 1976-1988), Paris, Gallimard, 2001 [1994].

Guiraud, Pierre, *Problèmes et méthodes de la statistique linguistique*, Paris, Presses universitaires de France, 1960.

Kastberg, Margareta et Jean-Marc Leblanc, « Extraction des isotopies d'un corpus textuel : analyse systématique des structures sémantiques et des coocurrences à travers différents logiciels textométriques », *Revue Texto*, vol. XVII, n° 3, Coordonné par Christophe Cusimano, 2012, http://www.revue-texto.net/docannexe/file/3059/texto_kastberg_leblanc.pdf, site consulté le 20 septembre 2015.

Labbé, Dominique et Denis Monière, « Essai de stylistique quantitative. Duplessis, Bourassa et Lévesque », dans Annie Morin et Pascale Sébillot, *VIe Journées Internationales d'analyse des données textuelles*, Saint-Malo, 13-15 mars 2002, Rennes, IRISA-INRIA, 2002, no 2, p. 561-569, https://halshs.archives-ouvertes.fr/halshs-01019903, site consulté le 5 octobre 2015.

Lebart, Ludovic *et al.*, *Statistique exploratoire multidimensionnelle*, Paris, Dunod, 2000.

Leblanc, Jean-Marc et Marie Pérès, « Modèles tridimensionnels pour la représentation de l'état des connaissances et propositions de visualisation pour l'analyse des corpus textuels », *XIIe Journées internationales d'analyse statistique des données textuelles*, Paris, 3-6 juin 2014, *JADT*, 2014, p. 373-384, http://lexicometrica.univ-paris3.fr/jadt/jadt2014/01-ACTES/31-JADT2014.pdf, site consulté le 5 octobre 2015.

Mayaffre, Damon, « De la lexicométrie à la logométrie », *L'astrolabe*, 2005, p. 1-11, https://halshs.archives-ouvertes.fr/hal-00551921/document, site consulté le 5 octobre 2015.

Moscarola, Jean, « Balladur, Chirac, Jospin, les mots d'une campagne. Quelques exemples d'analyse lexicale avec *Le Sphinx* », *Journées internationales d'Analyse statistique des Données Textuelles*, *JADT*, 1995.

Muller, Charles, *Essai de statistique lexicale. L'illusion comique de P. Corneille*, Paris, Klincksieck, 1964.

Pincemin, Bénédicte et *al.* « *Fonctionnalités textométriques : proposition de typologie selon un point de vue utilisateur », X^e Journées internationales d'analyse de données textuelles*, Rome, 8-11 mars 2010, *JADT*, 2010, p. 341-353, https://halshs.archives-ouvertes.fr/halshs-00856446, site consulté le 5 octobre 2015.

Rabeharisoa, Vololona, *L'analyse Leximappe de la presse grand public : le cas de la controverse sur le changement climatique global*, Centre de sociologie de l'innovation, École des mines, 2005, htts//web.upmf-grenoble.fr/adest/seminaires/volo.html, site consulté le 5 octobre 2015.

Reinert, Max, « Les "mondes lexicaux" et leur "logique" à travers l'analyse statistique d'un corpus de récits de cauchemars », *Langage et société*, vol. 66, no 1, 1993, p. 5-39.

Reinert, Max, « Quelques interrogations à propos de l'"objet" d'une analyse de discours de type statistique et de la réponse "Alceste" », *Langage et société*, vol. 90, no 1, 1999, p. 57-70.

Salem, André, « Introduction à la résonance textuelle », *VII^e Journées internationales d'analyse statistique des données textuelles*, Louvain-La-Neuve, *JADT*, 2004, http://lexicometrica.univ-paris3.fr/jadt/jadt2004/pdf/JADT_096.pdf, site consulté le 5 octobre 2015.

Silberztein, Max et Agnès Tutin, « NooJ, un outil pour l'enseignement des langues. Application pour l'étude de la morphologie lexicale en FLE », Apprentissage des langues et système d'information et de communication, vol. 8, no 2, 2005, p. 123-134, https://alsic.revues.org/336, site consulté le 5 octobre 2015.