

Explorer, mesurer, contextualiser. Quelques apports de la textométrie à l'analyse des discours

Damon Mayaffre, Bénédicte Pincemin, Céline Poudat

▶ **To cite this version:**

Damon Mayaffre, Bénédicte Pincemin, Céline Poudat. Explorer, mesurer, contextualiser. Quelques apports de la textométrie à l'analyse des discours. Langue française, 2019, Les outils informatiques au service des linguistes, 203, pp.101-115. 10.3917/lf.203.0101 . hal-02419199

HAL Id: hal-02419199
https://hal.science/hal-02419199

Submitted on 19 Dec 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License

Paru dans *Langue française*, 203, septembre 2019
numéro coordonné par Max Silberztein
« Les outils informatiques au service des linguistes »
p.101-115.

Explorer, mesurer, contextualiser
Quelques apports de la textométrie à l'analyse de discours

Exploration, Measure, Contextualisation
What Textometry can contribute to Discourse analysis

Damon Mayaffre

Université Côte d'Azur, CNRS, BCL (UMR 7320)

Bénédicte Pincemin

CNRS, Université de Lyon, Institut d'Histoire des Représentations et des Idées dans les Modernités - IHRIM (UMR 5317)

Céline Poudat

Université Côte d'Azur, CNRS, BCL (UMR 7320)

Résumé

Notre recherche vise à décrire la parole publique du président français Macron par rapport à celle de ses prédécesseurs de la V^{e} République (1958-2019). L'article montre comment la textométrie est mobilisable pour effectuer différents types d'investigations linguistiques : irrégularités de répartition des mots, évolution diachronique du vocabulaire, relevés systématiques et organisés d'attestations en contexte, synthèse statistique des contextes syntagmatiques locaux par cooccurrence, visualisations de la structure lexicale globale du corpus par analyse factorielle des correspondances et analyse arborée. Notre objectif est double : il s'agit d'une part d'analyser le discours macronien et d'autre part de montrer comment la textométrie permet de répondre à des questionnements linguistiques classiques en corpus, en mobilisant les logiciels Hyperbase Web et TXM.

Mots-clés

Analyse de Données Textuelles, Textométrie, Linguistique de corpus, Discours politique

Abstract

This research aims at describing the French President Macron's public speeches, in comparison with those of the former French Fifth Republic presidents (1958-2019). Our paper bears testimony to the extent to which textometric methods are relevant to investigate different types of linguistic phenomena: irregularities in word distribution patterns, vocabulary change

over time, systematic and organized overviews of attested forms in context, statistical summary of local syntagmatic contexts using cooccurrences, visualizations of corpus structure using correspondence analysis and additive tree clustering. Our goal is twofold: providing an analysis of Macron's speeches on the one hand, and showing how textometry enables linguists to answer standard questions investigating corpora on the other hand. In that respect, we will resort to two software programs: Hyperbase Web and TXM.

Keywords

Textual Data Analysis, Textometry, Corpus Linguistics, Political Discourse

1. Introduction

Le développement de grands corpus numériques s'est accompagné de méthodes d'interrogation pour les explorer, mettant en évidence des phénomènes réguliers inédits tout en renouvelant les parcours interprétatifs et les pratiques (Brunet 2016). Ainsi, le travail du linguiste s'en est trouvé profondément modifié (Habert 2006). Parmi les dispositifs méthodologiques disponibles, nous nous concentrons sur l'approche textométrique, et plus spécifiquement sur les méthodes qu'elle propose pour investiguer et décrire un corpus particulier, celui des discours présidentiels sous la V^{e} République avec un intérêt particulier pour ceux encore peu décrits d'Emmanuel Macron. Notre objectif est donc double : il s'agit d'une part d'explorer le discours macronien et de contribuer à sa description et d'autre part de montrer comment la textométrie permet de mettre en évidence des phénomènes réguliers à travers trois niveaux de questionnement correspondant à des problématiques classiques en matière d'exploration linguistique de corpus.

La textométrie, dite aussi analyse des données textuelles (ADT)[1], provient à l'origine de deux champs de recherche qui ont articulé leurs méthodes : la statistique lexicale d'une part (voir par exemple Muller, 1977), et l'analyse multidimensionnelle lexicale d'autre part, qui a connu d'importants développements en France sous l'impulsion de Benzécri. Si la première approche a développé un ensemble de mesures visant à décrire le vocabulaire d'un ou plusieurs textes, la seconde a mis au point un ensemble de méthodes dédiées au traitement statistique des tableaux de données dont l'analyse factorielle des correspondances (Benzécri 1973), applicable aux tables lexicales recensant la répartition des mots parmi les textes d'un corpus. D'abord connue sous le nom de *lexicométrie*, la démarche a été renommée *textométrie* (ou *logométrie*) au début des années 2000 pour rendre compte du fait qu'elle allait au-delà de la seule analyse du lexique, en considérant aussi la morphosyntaxe, l'enchaînement des mots, les structures textuelles et intertextuelles, etc.

Dans le domaine de l'analyse de corpus outillée, la textométrie se caractérise par la place centrale du texte tout au long de l'analyse (Valette 2016). Le principe même des traitements est la caractérisation systématique des unités (lexicales, textuelles) par leurs contextes au sein du corpus (l'apport de dictionnaires ou de connaissances extérieures est secondaire et facultatif) ; et les résultats des calculs sont interprétés

[1] Voir de très nombreux exemples d'analyse de tous types de textes – discours politiques ou autres – dans les actes des *Journées internationales d'Analyse statistique des Données Textuelles* (JADT) disponibles en ligne sur le site Lexicometrica (http://lexicometrica.univ-paris3.fr/jadt/index.htm).

en observant les contextes d'emploi des mots au sein des textes (le *retour au texte*). Cette prévalence du texte différencie la textométrie de la linguistique de corpus anglo-saxonne (*Corpus Linguistics*), qui s'attache plutôt à l'analyse d'un marqueur linguistique décrit en usage : on s'intéressera par exemple aux marqueurs exprimant le positionnement (*stance*, voir Englebretson, 2007) privilégiés par les femmes dans les forums, avec un intérêt particulier pour les collocations, ces associations privilégiées entre éléments du discours. La question du texte et de ses catégorisations reste ainsi secondaire alors même qu'elle est centrale pour la textométrie, ce qui entraîne des différences importantes dans le développement des méthodes et des outils. Par exemple, l'*analyse des correspondances*, centrale pour la textométrie, est absente des outils classiques de la linguistique de corpus, comme AntConc, WordSmith ou Sketch Engine, et les parcours méthodologiques qui règlent l'exploration des données sont bien distincts.

Dans cet article nous étudions la parole publique du président Macron de façon contrastive, dans la lignée des précédents présidents français de la V^{e} République, en analysant le corpus ÉLYSÉE. Réalisé par le laboratoire BCL (Mayaffre 2012a), ce corpus rassemble actuellement les versions écrites de 700 discours présidentiels, et représente un volume de 2,7 millions de mots. La partie correspondant à Macron compte 27 discours, pour un peu plus de 200 000 mots. Il a été annoté automatiquement en morphosyntaxe par le logiciel TreeTagger (Schmid 1994) en utilisant le fichier de paramètres d'Achim Stein pour le français (Stein & Schmid 1995). Nous avons choisi d'interroger le corpus à l'aide de deux logiciels de textométrie, Hyperbase (Brunet 2011a) principalement dans sa version Web (http://hyperbase.unice.fr)[2] et TXM (http://textometrie.ens-lyon.fr) (Heiden *et al.* 2010)[3]. Tous deux sont développés en contexte académique et diffusés gratuitement, et implémentent les principaux calculs textométriques (concordance, spécificités, cooccurrences, AFC, présentés dans les sections suivantes), mais ils sont complémentaires pour certains calculs spécifiques et pour les propriétés des sorties graphiques.[4]

Le corpus ÉLYSÉE est directement disponible dans la bibliothèque de corpus d'Hyperbase Web. Il est exportable depuis Hyperbase Web, et importable (selon le format ALCESTE) dans TXM. Toutes les analyses présentées ci-après sont ainsi reproductibles par le lecteur. Pour la présentation détaillée des calculs statistiques, nous renvoyons aux publications fondatrices citées ou aux ouvrages généraux du domaine (Lebart & Salem 1994, Poudat & Landragin 2017, Lebart *et al.* 2019).

[2] Hyperbase a bénéficié d'une aide du gouvernement français, gérée par l'Agence Nationale de la Recherche au titre du projet Investissements d'Avenir UCAJEDI portant la référence n° ANR-15-IDEX-01.

[3] TXM a bénéficié d'aides du gouvernement français, gérées par l'Agence Nationale de la Recherche, actuellement au titre des projets Democrat (ANR-15-CE38-0008) et Antract (ANR-17-CE38-0010).

[4] Les figures 1, 4, 5 et 6 ont été générées par Hyperbase Web, 8 par Hyperbase, et les figures 2, 3 et 7 par TXM.

2. Observation méthodique de distributions contextuelles et d'attestations

2.1. Le calcul des spécificités, pour détecter des affinités contextuelles

Le calcul des spécificités (Lafon 1980) permet de repérer les mots (ou traits linguistiques) anormalement fréquents dans une partie du corpus au regard de leur fréquence dans le corpus entier. Lancé sur l'ensemble des mots, il signale des caractéristiques de diverses natures, que le linguiste aura à démêler en revenant aux contextes d'emploi (cf. § 3). Ainsi, les toutes premières caractéristiques quantitatives du discours macronien (tableau 1) concernent ici les formes d'interlocution (*nous*, *notre*/*nos*, et dans une moindre mesure *vous*), les modalités (nous *devons*, je *veux*), le registre pédagogique et argumentatif (*parce que*, *justement*), le lexique dominant (*construire*, le *quotidien*, les *territoires*, ainsi que les notions *engagement(s)*, *transformation(s)*, *innovation(s)*, *défi(s)*, *projet(s)*, *acteurs*).

Tableau 1 - Extrait de spécificités

mesurées sur la partie Macron par rapport au corpus ÉLYSÉE. Certains mots ont un sur- ou sous-emploi dans une typographie et flexion particulière (graphie, en partie gauche du tableau), d'autres sont sur- ou sous-représentés indépendamment de leurs variations flexionnelles (lemme, en partie droite). *F* note le nombre d'occurrences du mot en corpus, *f* celui chez Macron.

Graphie	***F***	***f***	***Spécif.***	***Lemme***	***F***	***f***	***Spécif.***
concitoyens	304	192	132	notre	10 864	1 669	156
ça	1 846	439	99	nous	20 845	2 472	104
intelligence artificielle	75	75	83	parce que	4 749	755	77
devons	1 040	264	66	construire	490	182	74
quotidien	11	75	52	parfois	763	217	64
veux	1 832	343	51	engagement	704	203	61
justement	341	121	47	donc	4 688	647	45
territoires	277	99	39	aussi	4 956	674	44
acteurs	159	74	38	et	53 393	5 000	44
certain	1 550	7	- 43	transformation	283	103	41
U/un certain nombre d(e)	1 070	0	- 37	innovation	192	84	41
France	9 451	436	- 33	défi	237	92	39
inflation	282	0	- 9	vous	13 654	1 482	38
probablement	179	0	- 6	durant	124	65	38
pratiquement	150	0	- 5	projet	1 056	210	35
quelquefois	145	0	- 5	problème	3 453	61	- 54

Symétriquement, le calcul pointe des sous-emplois (cf. bas du tableau 1, spécificités négatives), *toujours au regard des usages tels que représentés par le corpus*. Macron fuit l'approximation (pas de *probablement*, de *pratiquement*, d'*un*

certain nombre de), limite (comme son prédécesseur) l'explicitation de *problème(s)*. Sans le calcul, saurions-nous percevoir que les 436 occurrences du mot *France* traduisent un usage extrêmement parcimonieux (score de -33), en partie au profit de l'*Europe* (score de 10) ? Le calcul se fait aussi le révélateur des absences les plus étonnantes (dites *nullax*), telles l'*inflation* ou *quelquefois* (auquel est préféré *parfois*).

Enfin, le calcul statistique peut être employé au recensement des éléments les plus uniformément partagés, le *vocabulaire banal* (spécifique d'aucun président dans ce corpus) potentiellement signature ici des genres textuels sous-jacents. Dans nos discours présidentiels c'est le cas des *leçon(s)* (données, reçues, tirées), de la *répartition* (qui à certaines périodes se spécialise : *système/retraite/régime par répartition*), des verbes *confier* (une mission, une responsabilité), *préciser*, *présider*, etc.

Ces observations ont porté sur « l'occupation » de la « surface » globale du texte. Des calculs plus ciblés morphosyntaxiquement pourraient évaluer les sur- ou sous-emplois au sein d'une catégorie particulière ou d'un type de construction (Mayaffre 2006, Guillot *et al.* 2013, Vigier 2017).

Lorsque le corpus se divise en périodes, les spécificités fournissent un indicateur précis d'évolution quantitative. En figure 1, le graphe de gauche montre la croissance d'emploi de *territoires* par les présidents depuis Mitterrand. L'exemple de droite permet d'affiner l'observation du sur-emploi de *ça* par Macron, en comparaison avec *cela*, avec un double mouvement diachronique : place croissante de ces pronoms démonstratifs dans le discours présidentiel, décroissance relative de l'usage de *cela* au profit de *ça* (dans les discours prononcés ou/et dans la façon de les transcrire), sur-représentation de *ça* la plus marquée chez Sarkozy (Mayaffre 2012b).

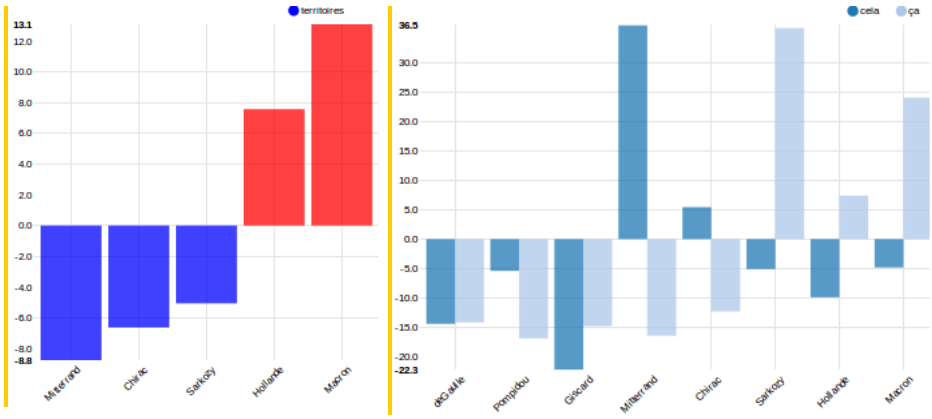


Figure 1 - Évolution quantitative du mot *territoires* (focus de Mitterrand à Macron - diagramme de gauche), et des mots *cela* et *ça* (diagramme de droite)[5]

[5] Ces graphiques représentent des valeurs de spécificités rapportées à une échelle d'écart-réduit (Brunet, 2011a : 40), alors que le tableau 1 est en spécificités brutes.

2.2. Concordance

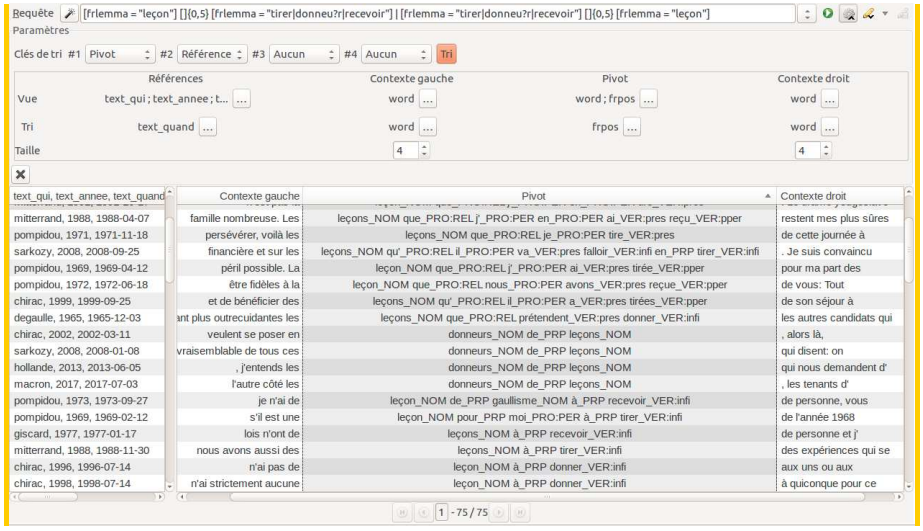


Figure 2 - Concordance sur les occurrences du prédicat *tirer/donner/recevoir* avec l'argument *leçon(s)*, triée par construction de surface (séquence de catégories grammaticales) puis par date.

L'interprétation des décomptes doit être éclairée par l'observation en contexte des faits comptabilisés. Par rapport à une recherche d'occurrences dans un traitement de texte, les concordances permettent en plus *(i)* la recherche de constructions et de *motifs* en s'appuyant sur les informations linguistiques disponibles sur les mots, *(ii)* une présentation synthétique et structurée (en une page on peut visualiser efficacement plusieurs dizaines d'attestations), *(iii)* des tris mettant en évidence des récurrences de constructions ou de contextes locaux, en gardant le lien avec les contextes globaux.

2.3. Relevé de mots ou d'expressions

Lorsqu'il s'agit d'analyser les diverses réalisations d'un motif complexe (Longrée & Mellet 2013), un relevé quantitatif présente une synthèse que l'on peut ordonner de façon à mettre en évidence les dominantes. En effet, alors que la concordance restitue chaque occurrence en contexte sur une ligne, le relevé groupe en une seule ligne toutes les occurrences présentant la même réalisation, en indiquant sa fréquence. Le jeu des réglages permet d'affiner sous quel angle effectuer les regroupements, par exemple en prenant en compte ou non les flexions, et sur tout ou sur un élément ciblé du motif.

Ainsi, nous avons vu que Macron sur-employait le nom *transformation*, mais de quelle(s) transformation(s) parle-t-il ? Nous pouvons consulter de façon systématique les qualifications du nom, et les arguments introduits par la préposition *de* (figure 3).

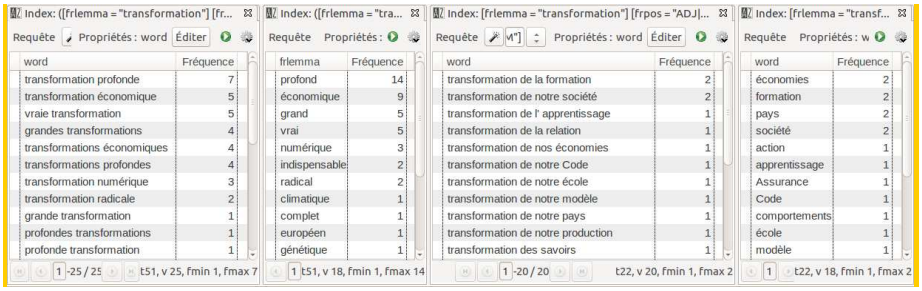


Figure 3 - Relevés de contextes du mot *transformation(s)* chez Macron (fréq. décroissante)

(i) occurrences de *transformation(s)* précédées ou suivies d'un adjectif qualificatif,

(ii) adjectifs qualificatifs épithètes (lemmatisés) de *transformation(s)* (par fréquence décroissante),[6]

(iii) occurrences de *transformation(s)* suivies d'un complément du nom introduit par *de*,

(iv) noms (fléchis) relevés dans la construction précédente.[7]

3. Sémantique lexicale et textuelle : opérationnaliser les contextes via la cooccurrence

Au-delà de son appareillage statistique hérité de Muller ou de Benzécri, la textométrie s'inscrit dans le cadre d'une linguistique contextualisante. Que l'on cherche ses échos linguistiques dans le contextualisme anglo-saxon de (Firth 1957) ou de (Halliday & Hasan 1976), dans l'analyse du discours ou la socio-linguistique de (Dubois & Sumpf 1969) ou de (Tournier 1980) ou dans la sémantique interprétative ou sémantique de corpus de (Rastier 2011), la textométrie affiche en effet pour objectif une contextualisation pertinente des unités du corpus, pour une sémantique qui devient dès lors endogène.

Le sens naît en/du contexte, et les parcours de lecture numériques que la textométrie propose doivent permettre d'aborder les mots dans leurs éco-systèmes textuels micro (le syntagme par exemple), méso (la phrase ou le paragraphe), macro (le texte ou le corpus).

Si l'acte final de contextualisation passe par la lecture des concordances et plus loin par la lecture du texte intégral, la textométrie établit l'approche statistique des cooccurrences comme un acte sémantique ou contextualisant primordial.

[6] Requête dans TXM pour *(ii)* (pour *(i)* idem sans les @) : ([frlemma = "transformation"] [frpos = "ADV"]? @[frpos = "ADJ"] | [frpos = "ADV"]? @[frpos = "ADJ"] [frlemma = "transformation"])

[7] Requête dans TXM pour *(iv)* (pour *(iii)* idem sans le @) : [frlemma = "transformation"] [frpos = "ADJ|ADV"]{0,2} [frlemma = "de|du"] [frpos = "DET.*"]? [frpos = "ADJ|ADV"]{0,2} @[frpos = "NOM|NAM"]

3.1. La cooccurrence comme forme minimale et calculable du contexte

La cooccurrence est l'association statistiquement significative de deux unités linguistiques (en général deux mots) dans une fenêtre déterminée (en général le paragraphe). La cooccurrence est chiffrée avec des indices éprouvés depuis les années 1980, et visualisée grâce à des outils dont nous ne donnerons ici que quelques exemples.

Mais par-delà l'aspect technique, l'important est d'abord de rappeler que la cooccurrence permet aux pratiques textométriques de faire un saut décisif, pour toucher les rivages du sens (Mayaffre 2014). Si l'occurrence reste en général infra-linguistique et a-sémantique, la cooccurrence représente une plus-value phraséologique, sémantique, textuelle remarquée. Constater l'occurrence de *France* chez un locuteur ne nous apprend rien. Constater la cooccurrence statistique de *France* avec *grandeur*, *nation* ou *indépendance*, nous apprend immédiatement sur le patriotisme du locuteur.

Lorsque deux mots apparaissent ensemble dans une même fenêtre syntagmatique, ils se contextualisent *de facto* mutuellement, de manière certes élémentaire mais essentielle. C'est pourquoi la cooccurrence peut être définie comme la forme minimale et calculable du contexte et, dès lors que le contexte est la condition de l'émergence du sens, comme la première molécule sémantique d'un texte (Mayaffre 2008, Brunet 2016 : 295 sq.).

3.2 La cooccurrence d'un mot-pôle : calcul et visualisation

Simplement, les pratiques cooccurrentielles consistent à sélectionner un mot-pôle et à calculer les mots qui y sont associés. L'univers lexical du mot se trouve ainsi systématiquement décrit par les vertus d'un indice statistique.

Par exemple dans le corpus ÉLYSÉE, Macron partage avec de Gaulle la sur-utilisation du mot *souveraineté*. Seulement, les mots associés (*i.e.* cooccurrents) à *souveraineté* chez l'un et chez l'autre varient et distinguent deux messages politiques différents.

Dans une visualisation simple (figure 4), le nuage des cooccurrences de De Gaulle (à gauche) dessine le cadre de souverainetés étatiques et nationales, particulièrement celui de la souveraineté française (*France*, *français*, *État*, *Tunisie*, *marocain*, etc.). Le nuage des cooccurrences de Macron (à droite) décline d'autres types de souverainetés (la souveraineté *numérique* face aux géants d'internet ou la souveraineté *économique* face au capitalisme mondialisé), et avant tout la souveraineté européenne (*Europe*, *européen*, *européenne*).



Figure 4 - Univers lexical de *souveraineté* chez de Gaulle (à gauche) *versus* Macron (à droite) (la taille des mots indique la force de l'indice de cooccurrence)

Cette préoccupation européenne essentielle autour du mot *souveraineté* se visualise de manière plus complexe et paramétrable sur la figure 5.

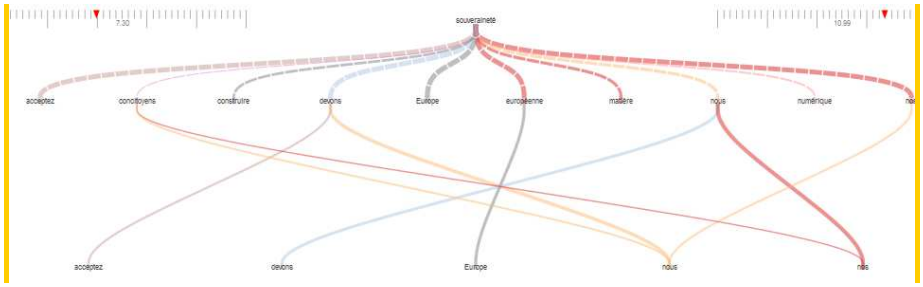


Figure 5 - Cooccurrents de premier et de deuxième niveaux de *souveraineté* chez Macron

L'épaisseur des liens entre le mot-pôle (*souveraineté*) et ses cooccurrents de premier ordre (*acceptez*, *concitoyens*,... *Europe*, *européenne*,… *numérique*,…) indique l'importance de la force d'attraction. Puis, en cascade, une fois une paire de mots établie, le calcul est réitéré afin de déterminer les cooccurrents de second niveau. De manière triviale sur ce graphe, on mesure par exemple que la *souveraineté* est associée à la notion d'obligation ou de devoir (*devons*) chez Macron. Et lorsqu'on trouve *souveraineté* et *devons* ensemble, le verbe accepter (*acceptez*) renforce le caractère grave et obligatoire de la question.

3.3. Les cooccurrences généralisées

Au-delà d'un mot-pôle déterminé, l'analyse cooccurrentielle peut être généralisée. Le traitement consiste alors à embrasser les mots du corpus dans leur ensemble et à décrire toutes les cooccurrences qui existent entre eux. C'est ainsi la trame textuelle, la textualité ou la texture qui sont approximés donnant à voir les relations réticulaires qui existent entre les unités du texte pour visualiser le système-texte dans son ensemble (Viprey 2006).

Une des possibilités est d'établir une matrice carrée mots × mots dans laquelle est reporté le nombre de fois où le mot *A* cooccurre avec les mots *B*, *C*, *D*, *E*, etc. Puis le nombre de fois où le mot *B* rencontre les mots *A*, *C*, *D*, *E*, etc. Et ainsi jusqu'à épuisement du vocabulaire (tableau 2).

Tableau 2 - Matrice mots × mots ou matrice cooccurrentielle

	Mot A	Mot B	Mot C	Mot D	Etc.
Mot A	***	*x* (cooc A_B)	*y* (cooc A_C)	*z* (cooc A_D)	…
Mot B	*x* (cooc A_B)	***	*v* (cooc B_C)	*w* (cooc B_D)	…
Mot C	*y* (cooc A_C)	*v* (cooc B_C)	***	*u* (cooc C_D)	…
Mot D	*z* (cooc A_D)	*w* (cooc B_D)	*u* (cooc C_D)	***	…
Etc.	…	…	…	…	***

Dès lors, l'AFC (Benzécri 1973) présentée en section 4 est directement adaptée à l'analyse de ce type de tableau (figure 6).

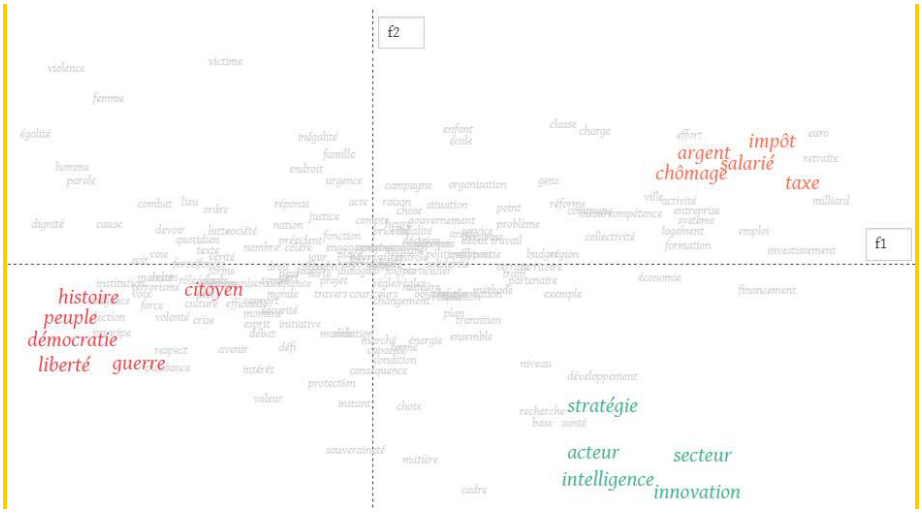


Figure 6 - Carte cooccurrentielle ou corrélats, établie sur 300 noms les plus fréquents du corpus ÉLYSÉE (les mots surlignés le sont pour faciliter la lecture)

Sur la carte cooccurrentielle (figure 6), dans les principaux quadrants ou dans des zones plus resserrées, se distinguent les grandes thématiques ou les grandes isotopies qui permettent à Macron de déployer son discours. Par exemple, une polarité lexicale autour des mots *peuple*, *démocratie*, *histoire*, etc., en bas à gauche du graphique, permet au président de construire un propos régalien qui sied à la fonction présidentielle. En bas à droite du graphique on trouve des mots plus originaux dans le discours élyséen autour de la *recherche*, de l'*innovation*, de l'*intelligence*, etc. Et précisons dans le détail du traitement que les mots rassemblés sur le graphique sont des mots pas seulement cooccurrents entre eux, mais des mots qui partagent le même profil cooccurrentiel, c'est-à-dire des mots qui ont le même comportement sémantique ou contextuel (ils ont les mêmes cooccurrents).

Dans un texte, l'item n'est jamais seul : les mots ne sont ni indépendants ni isolés. La cooccurrence que ces quelques lignes ne prétendent nullement épuiser, tant la recherche textométrique en la matière est florissante, permet d'objectiver

grâce aux statistiques des phénomènes de contextualisation décisifs pour la construction du sens et de l'interprétation.

4. Exploration de la structure d'un corpus

Lorsque l'on explore un vaste corpus textuel, il est toujours instructif de mettre en évidence sa structure (v. Poudat & Landragin 2017), à savoir les lignes d'organisation qui le sous-tendent, observables depuis les associations et les oppositions significatives entre les textes et les observations linguistiques choisies pour les décrire (mots, catégories grammaticales, etc.). Cette démarche est tout à fait classique en textométrie. En effet, notre intérêt pour le corpus comme objet d'observation faisant système, et pour les méthodes permettant le dévoilement de cette structure, s'origine en partie dans ce courant qu'on a appelé l'*analyse de données à la française* et dont Benzécri fut l'une des figures de proue.

4.1. Analyse factorielle des correspondances

Parmi les méthodes proposant des visualisations permettant de résumer de vastes ensembles de données, l'Analyse factorielle des correspondances (désormais AFC) est certainement la plus classique en textométrie. Développée par Benzécri à partir des années 1960, l'AFC a précisément été pensée pour résumer des tableaux de fréquences, dont le cas le plus usuel dans notre domaine est la table lexicale qui croise les textes d'un corpus et les mots les plus fréquents.

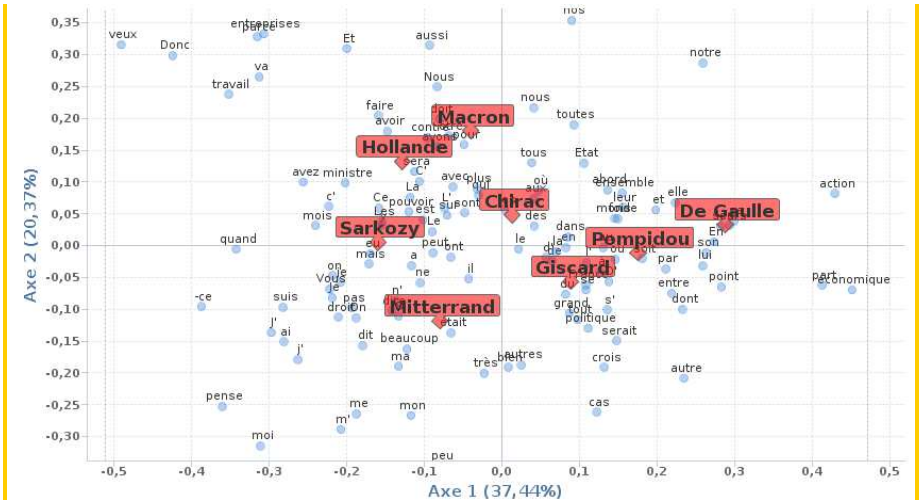


Figure 7 - AFC caractérisant les 8 présidents par les 200 mots les plus fréquents dans le corpus ÉLYSÉE

La figure 7 présente ainsi les deux premiers axes factoriels de l'AFC réalisée sur les 200 mots les plus fréquents du corpus répartis par président.[8]

Le corpus est d'abord nettement sous-tendu par une structure chronologique, opposant sur le versant positif de l'axe 1 les premiers présidents de la V^e^ République, ceux des années 1950 à 1970 (De Gaulle, Pompidou et Giscard), aux derniers présidents en date, ceux des années 2000 (Sarkozy, Hollande, Macron) sur son versant négatif. Cette structure s'accompagne d'une différence énonciative fondamentale, opposant un discours conceptuel et nominal (*action*, *politique*, *monde*, etc.) à un discours plus centré sur l'énonciateur-président qui met en scène sa parole, avec une utilisation particulière des pronoms déictiques *je*, *vous* et *on* ou encore des verbes de parole et des modaux. On pourrait ainsi dire que le geste régalien qui marquait les premiers discours est progressivement supplanté par un geste populaire privilégiant le phatique au référentiel – interprétation qui doit être nuancée par les positions particulières de Mitterrand et Chirac, qui contrarient l'agencement chrono-énonciatif décrit : le premier utilise précocement, dès les années 1980, une énonciation tendue (*je*, *vous*, *ne… pas*, *n'*), tandis le second verbalise encore dans les années 1990 un discours lointainement gaulliste.

Ce plan factoriel nous semble être une bonne illustration de ce que l'AFC apporte à l'analyste ; la structure globale mise au jour va ainsi lui permettre de cibler et d'objectiver le choix des phénomènes retenus et de contextualiser ses interprétations de phénomènes plus locaux.

4.2. Classifications et analyse arborée

L'analyse factorielle va généralement de pair avec une classification des textes, permettant de préciser les structures d'organisation du corpus en dégageant les ressemblances et les oppositions les plus significatives des textes entre eux. Dans cette perspective, la textométrie recourt notamment à l'analyse arborée (Barthélemy & Luong 1998).

La figure 8 propose une classification des différents mandats des présidents du corpus ÉLYSÉE sous forme d'arbre. Son centre topologique, défini mathématiquement sur les seules données textuelles (mots des discours), est noté par un cercle noir. Les discours sont regroupés au sein de différentes branches, correspondant aux nœuds de la classification et la visualisation de l'arbre généré nous permet d'apprécier la distance entre les mandats.

[8] À noter qu'un léger *peeling* du nuage de mots a dû être réalisé (voir Poudat & Landragin 2017) en écartant *monsieur*, ainsi que les démonstratifs *cela* (*Cela*) et *ça*, ce dernier étant très spécifique à Sarkozy (Mayaffre 2012b). Puis le graphique a été aéré en filtrant les mots les moins pertinents pour son interprétation (peu contributeurs aux axes 1 et 2 et mal représentés dans ce plan).

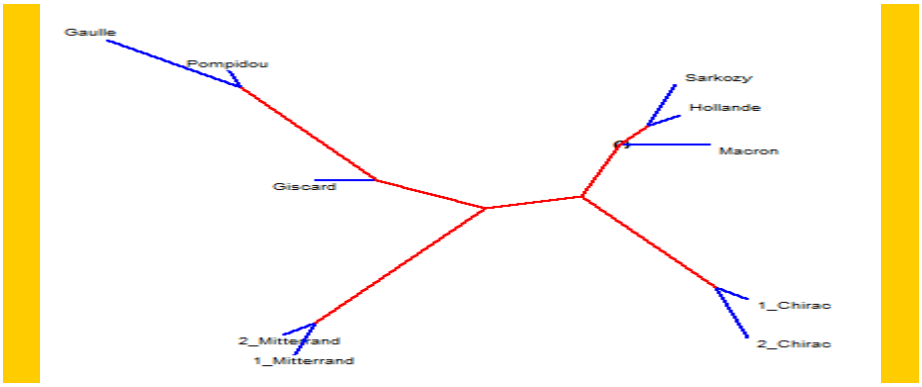


Figure 8 - Classification du corpus ÉLYSÉE avec l'analyse arborée – représentation radiale

L'analyse arborée corrobore les éléments de structure mis au jour avec l'AFC en les précisant. On voit ainsi que Mitterrand et Chirac conservent une position particulière ; leurs deux mandats respectifs sont ainsi regroupés, ce qui atteste de leur signature lexicale et discursive forte.

Surtout, les premiers présidents de la V^e^ République restent opposés aux derniers présidents des années 2000 mais on notera, dans ce cadre, la position particulière et originale de Macron, au centre topologique de l'arbre.

La raison chronologique aurait pu l'emporter et la feuille macronienne aurait pu se situer sur la branche de Sarkozy et de Hollande pour en former l'extrémité. Le poids de la chronologie est en effet généralement prégnant sur les discours. Or, du point de vue de ce calcul[9], Emmanuel Macron se distingue de ses immédiats devanciers pour se repositionner, à rebours d'une chronologie normalement déterminante, au centre topologique de l'arbre, certes à proximité de Hollande et de Sarkozy mais également attiré par Mitterrand ou Chirac, Giscard, Pompidou et de Gaulle. D'une certaine manière, le discours de Macron apparaît comme un discours médian, ou recentré, qui paye évidemment sa plus forte rançon lexicale à son époque mais qui fait montre également d'emprunts lexicaux plus lointains et plus complexes.

5. Conclusion

Nous avons proposé une exploration textométrique du corpus ÉLYSÉE avec un focus sur le discours macronien, qui a pu être caractérisé à différents niveaux de généralité, de ce qui le spécifiait aux affinités qu'il entretenait avec ses prédécesseurs. Notre propos était double : proposer sur le plan de l'analyse quelques premiers éléments de description du discours de Macron, et restituer sur les plans de la méthode et de l'heuristique les grandes questions que se pose la textométrie, et les méthodes qu'elle a développées pour y répondre.

[9] Il s'agit du nouveau calcul proposé dans Hyperbase depuis 2019, basé sur la distance d'Évrard (Brunet & Vanni 2019).

La présente contribution s'est ainsi concentrée sur les principes premiers et les fonctionnalités textométriques essentielles, et le manque de place a laissé dans l'ombre nombre d'outils et de méthodes qui ont pourtant fait la preuve de leur pertinence descriptive ou classificatoire comme les calculs de spécificités chronologiques ou le traitement d'objets linguistiques multi-niveaux comme les motifs (Longrée & Mellet 2013).

À l'heure actuelle, la textométrie poursuit ses développements en résonance avec les évolutions numériques, et doit faire face à différents défis tant au niveau des corpus que des techniques. De nouveaux parcours méthodologiques et interprétatifs doivent être mis en œuvre pour le traitement des corpus multimodaux et des corpus richement annotés tandis que de nouvelles potentialités heuristiques (*deep learning*, techniques d'intelligence artificielle) permettant de faire découvrir de nouveaux observables du texte, chevilles interprétatives inédites, ne sauraient être négligées à l'avenir (Brunet & Vanni 2019), renouvelant les parcours de lecture et offrant au linguiste de nouveaux instruments d'observation.

Bibliographie

BARTHELEMY J.-P. & LUONG X. (1998), « Représenter les données textuelles par les arbres… », *in* S. Mellet (éd.), *JADT 1998, 4es Journées internationales d'Analyse statistique des Données Textuelles*, Université de Nice, 49-71.

BENZECRI J.-P. et coll. (1973), *L'Analyse des Données*. Tome 1 : *La Taxinomie*. Tome 2 : *L'Analyse des Correspondances*, Paris, Dunod.

BRUNET É. (2011a), *Hyperbase. Logiciel hypertexte pour le traitement documentaire et statistique des corpus textuels. Manuel de référence*, Université de Nice, Laboratoire BCL. [consulté le 07-05-2019, http://ancilla.unice.fr/bases/manuel.pdf]

BRUNET É. (2016), *Tous comptes faits, Écrits choisis tome III, Questions linguistiques*, B. Pincemin (éd.), Paris, Champion.

BRUNET É. & VANNI L. (2019), « Deep learning et authentification des textes », *Texto! Textes & Cultures* XXIV, 1. [consulté le 07-05-2019, http://www.revue-texto.net/index.php?id=4194]

DUBOIS J. & SUMPF J. (éds) (1969), *Langages* 13 : *L'analyse du discours*, Paris, Didier.

ENGLEBRETSON R. (éd.) (2007). *Stancetaking in Discourse*. Amsterdam, John Benjamins.

FIRTH J. R. (1957), "A Synopsis of Linguistic Theory 1930-1955", *in* J. R. Firth (éd.), *Studies in Linguistic Analysis*, Oxford, Blackwell, 1-32.

GUILLOT C., LAVRENTIEV A., PINCEMIN B., HEIDEN S. (2013), « Le discours direct au Moyen Âge : vers une définition et une méthodologie d'analyse », in D. Lagorgette & P. Larrivée (éds), *Représentations du sens linguistique 5*, Université de Savoie, 17-41.

HABERT B. (2006), « Portrait de linguiste(s) à l'instrument », *in* C. Guillot, S. Heiden, S. Prévost (éds), *À la quête du sens. Études littéraires, historiques et linguistiques en hommage à Christiane Marchello-Nizia*, Lyon, ENS Éditions, 163-173.

HALLIDAY M. A. K. & HASAN R. (1976), *Cohesion in English*, London, Longman.

HEIDEN, S., MAGUE, J-P. & PINCEMIN, B. (2010), « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », *in* S. Bolasco, I. Chiari, L. Giuliano (éds), *Statistical Analysis of Textual Data. Proc. of JADT 2010*, Roma, Edizioni Universitarie di Lettere Economia Diritto, 1021-1032.

LAFON P. (1980), « Sur la variabilité de la fréquence des formes dans un corpus », *Mots* 1, 127-165.

LEBART L., PINCEMIN B. & POUDAT C. (2019), *Analyse des données textuelles*, Québec, Presses de l'université du Québec.

LEBART L. & SALEM A. (1994), *Statistique textuelle*, Paris, Dunod.

LONGREE D. & MELLET S. (2013), « Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours », *Langages* 189, 65-79.

MAYAFFRE D. (2006), « Faut-il prendre en compte la composition grammaticale des textes dans le calcul des spécificités lexicales ? Tests logométriques appliqués au discours présidentiel sous la Vème République », *in* J.-M. Viprey (éd.), *Actes des JADT 2006*, Presses universitaires de Franche-Comté, Besançon, 677-685.

MAYAFFRE D. (2008), « De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie », *Syntaxe & Sémantique* 9, 53-72.

MAYAFFRE D. (2012a), *Le discours présidentiel sous la V^e^ république : Chirac, Mitterrand, Giscard, Pompidou, de Gaulle*, Paris, Les Presses de Science Po.

MAYAFFRE D. (2012b), *Nicolas Sarkozy : mesure et démesure du discours, 2007-2012*, Paris, Les Presses de Science Po.

MAYAFFRE D. (2014), « Plaidoyer en faveur de l'Analyse de Données co(n)Textuelles. Parcours cooccurrentiels dans le discours présidentiel français (1958-2014) », *in* É. Née, J.-M. Daube, M. Valette & S. Fleury (éds), *JADT 2014, Proceedings of the 12th International Conference on Textual Data Statistical Analysis*, Paris, Inalco-Sorbonne nouvelle, 15-32.

MULLER Ch. (1977), *Principes et méthodes de statistique lexicale*, Paris, Hachette. Réimpression en 1992 : Paris, Champion.

POUDAT C. & LANDRAGIN F. (2017), *Explorer un corpus textuel*, Louvain-la-Neuve, De Boeck.

RASTIER F. (2011), *La mesure et le grain. Sémantique de corpus*, Paris, Champion.

SCHMID H. (1994), "Probabilistic Part-of-Speech Tagging Using Decision Trees", *in International Conference on New Methods in Language Processing*, 44–49.

STEIN A. & SCHMID H. (1995), « Étiquetage morphologique de textes français avec un arbre de décisions », *Traitement automatique des langues* 36, 1-2, 23-35.

TOURNIER M. (1980), « En souvenir de Lagado », *Mots* 1, 5-9.

VALETTE M. (2016), « Analyse statistique des données textuelles et traitement automatique des langues. Une étude comparée », in D. Mayaffre, C. Poudat, L. Vanni, V. Magri, P. Follette (éds), *JADT 2016 – Statistical Analysis of Textual Data*, Nice, Presses de FacImprimeur, 697-706.

VIGIER, D. (2017), « La préposition *dans* au XVI^e^ siècle. Apports d'une linguistique instrumentée », *Langages* 206, 105-122.

VIPREY J.-M. (2006), « Structure non-séquentielle des textes », *Langages* 163, 71-85.