Université de Neuchâtel

Institut de Mathématiques

Institut d'Informatique

SEMINAIRE MATHEMATIQUES ET SOCIETE

Vendredi 17 mai 2019

Soixante-ans de discours présidentiels français (1958 – 2018)

Qu'est-ce qui singularise Emmanuel Macron ?

En hommage à Jacques Bernoulli (Bâle 1654 - 1705)

Dominique LABBE
(PACTE – CNRS – Grenoble)
dominique.labbe@umrpacte.fr

Résumé

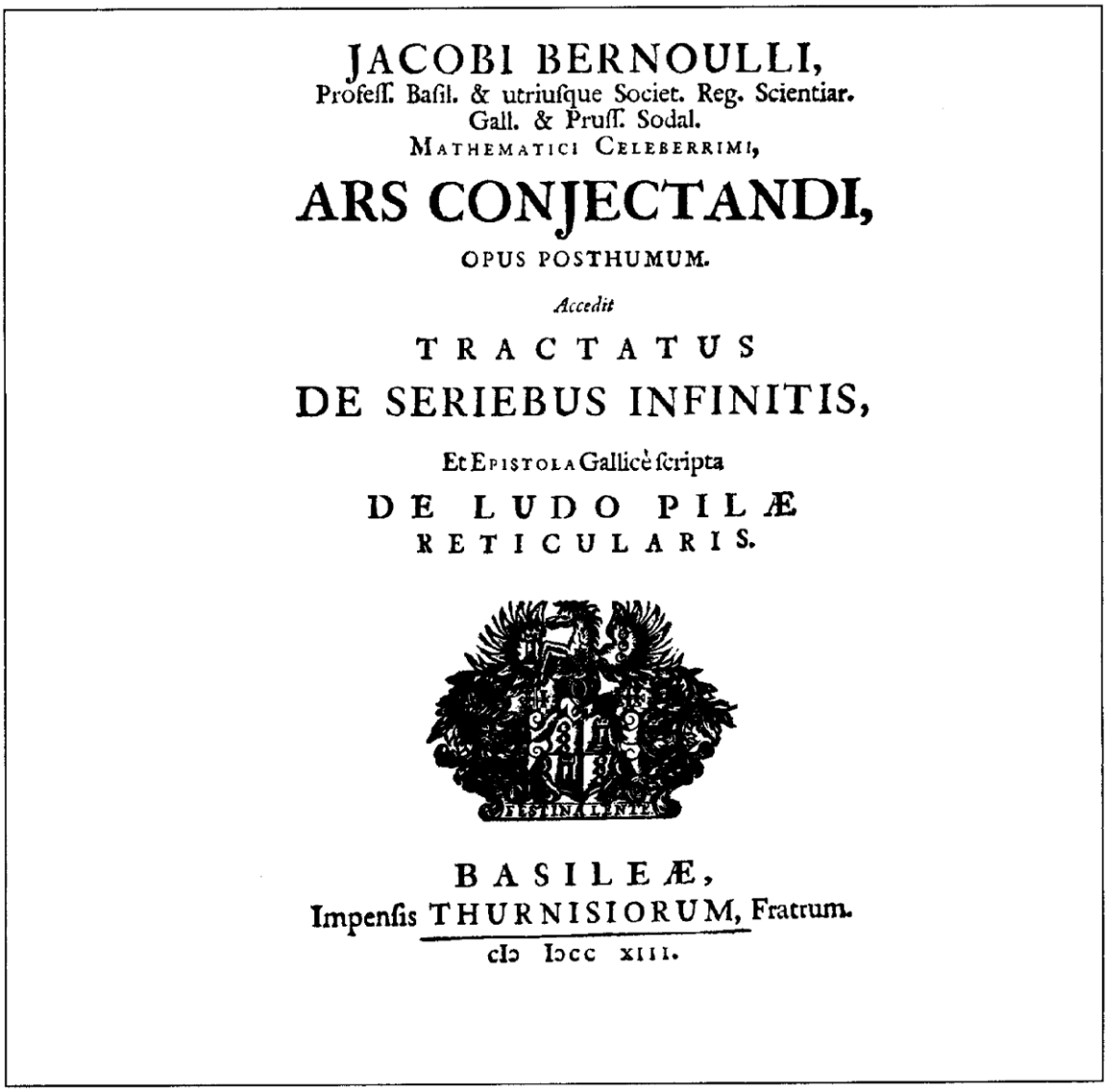
Présentation des méthodes statistiques pour la détermination du vocabulaire caractéristique d'un corpus de textes comparé à un ensemble de référence. L'ensemble des discours des présidents français depuis 1958 sert à déterminer le vocabulaire caractéristique des interventions d'E. Macron durant les 20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

Abstract

Presentation of the statistical methods for the determination of the characteristic vocabulary of a corpus of texts compared to a set of reference. The set of speeches by the French Presidents since 1958 is used in order to determine the characteristics of E. Macron's vocabulary as seen in his interventions during the first twenty months of his five-year term (2017-2018). This president is very distinct from his predecessors. He prefers subjects on international politics and speaks less about France and the French people. He neglects many themes of his predecessors (economics, employment, progress, growth, income, society). He takes little ownership of what he says. For example, he uses the pronoun "we" more than the "I". His speeches have a pedagogical aim, but are often very compact and abstract. The sentences are long and complicated, and are sometimes obscure.

Il y a neuf ans, dans ce même séminaire, ont été évoqués les travaux d'un Suisse qui a profondément influencé la science du XXe siècle (Ferdinand de Saussure).

Aujourd'hui, l'hommage s'adresse à une famille de mathématiciens et de physiciens suisses : les Bernoulli et particulièrement à Jacques Bernoulli (né à Bâle en 1654 et mort dans la même ville en 1705). Son ouvrage majeur « Ars conjectandi » - publié de manière posthume en 1713[1] - pose les principes du calcul des probabilités qui va être utilisé lors de ce séminaire.



JACOBI BERNOULLI,
Profeſſ. Baſil. & utriuſque Societ. Reg. Scientiar.
Gall. & Pruſſ. Sodal.
MATHEMATICI CELEBERRIMI,

ARS CONJECTANDI,

OPUS POSTHUMUM.

Accedit

TRACTATUS
DE SERIEBUS INFINITIS,

Et EPISTOLA Gallicè ſcripta
DE LUDO PILÆ
RETICULARIS.

BASILEÆ,
Impenſis THURNISIORUM, Fratrum.
cIↃ IↃcc xIII.

[1] Pour les curieux, il existe deux traductions partielles du traité de Bernoulli en français : Meusnier 1987 et 1992. Sur la postérité de J. Bernoulli, un article en ligne : Schaffer 1996 ; pour une présentation simple en français des principales lois de probabilité : Dodge 1993.

Le 1er janvier 2019, la France a fêté le soixantième anniversaire de la Ve République. C'est une durée exceptionnelle dans l'histoire française contemporaine. Seule la IIIe république (1875 – 1940) a eu une longévité légèrement supérieure. Dans la Ve République, le président de la république est la « clef de voûte » de la constitution voulue par le général de Gaulle qui en a exercé le premier la charge (1958-1969)[2]. Lui ont ensuite succédé : Georges Pompidou (1969-1974), Valéry Giscard d'Estaing (1974-1981), François Mitterrand (1981-1995), Jacques Chirac (1995-2007), Nicolas Sarkozy (2007-2012), François Hollande (2012-2017) et enfin Emmanuel Macron, actuellement au pouvoir depuis 2 ans (les textes traités vont jusqu'au 31 décembre 2018).

Depuis quarante ans, une petite équipe collecte les interventions des présidents, ce qui permet de tracer une sorte de portrait "discursif" de chacun d'eux et celui de la communication présidentielle depuis 1958[3].

Les discours présidentiels de mai 1958 (retour au pouvoir du général de Gaulle) à décembre 2018) représentent, à la date de cette conférence, 8 748 textes comptant au total : 18 431 088 mots (voir annexe 1). Que représente un tel volume ? Par exemple, les trois tomes des *Mémoires de guerre* du général de Gaulle comptent 372 664 mots ; les plus longs romans en français : *Les Misérables* (Hugo), 564 301 mots ; *Les Mystères de Paris* (E. Sue), 578 933 mots. L'édition originale de *A la recherche du temps perdu* (Proust), chez Gallimard (1913-1927) compte 11 volumes de 300 pages en moyenne, pour 1 327 859 mots. Si l'on imprimait à ce format toutes les interventions des présidents, il faudrait 152 volumes, 45 760 pages et plus de quatre mètres de rayonnages pour les ranger.

Au sein de cet ensemble, le "corpus Macron" représente : 1 037 988 mots – soit 5,36% de cette section de la bibliothèque électronique - pour un vocabulaire de 14 819 vocables. Quelles sont les principales caractéristiques et les singularités du discours de l'actuel président comparé à ses sept prédécesseurs ?

I. Le recensement du vocabulaire

Dix-huit millions et demi d'individus en plus de 8 700 textes, c'est une vaste population et la connaître avec précision exige des méthodes rigoureuses d'observation. Il ne s'agit pas, comme on l'entend parfois, d'"enfourner les textes dans la machine et de voir ce qu'il en sort"[4]. Ces textes doivent subir des traitements préalables avant d'être soigneusement rangés, à leur place, dans une bibliothèque électronique où les programmes informatiques pourront les retrouver aisément (annexe 1). A ce prix, en suivant la méthode de Bernoulli, on peut inférer quelques conclusions générales à partir d'un nombre d'observations limitées.

[2] A la suite de l'ouvrage de J.-M. Cotteret et R. Moreau (1969), il est de tradition d'inclure dans le corpus de Gaulle les interventions qu'il a prononcées comme président du conseil (juin-décembre 1958). Jusqu'en 2002, le mandat était de sept ans, après de cinq. C. de Gaulle, F. Mitterrand et J. Chirac ont exercé deux mandats.

[3] Arnold et al. 2016. Cette collection est accessible en ligne sur le site du Centre de Linguistique de Corpus de l'Université de Neuchâtel.

[4] C'est la philosophie de la "statistique textuelle" qui travaille sur les formes graphiques et qui est exploratoire, contrairement à la statistique lexicale que nous allons présenter (Sur la première voir : Lebart et Salem 1992).

Traitements préalables

Le traitement préalable consiste d'abord à corriger les fautes d'orthographe[5] puis à standardiser les graphies. Par exemple, *Abou Dabi* (*Emirats Arabes Unis*) : 5 mots ou deux ? De plus, ces mots apparaissent dans le corpus Macron sous les formes *Abou-Dabi, Abou Dhabi, Abu Dabi* et *Abu Dhabi*, *Emirats arabes unis*… Une seule solution : adopter une graphie standard (celle des dictionnaires) et enregistrer toutes les variantes comme un seul mot.

Cette solution est coûteuse en temps mais elle présente deux intérêts.

1/ Ne pas recenser des fantômes ou compter plusieurs fois le même individu comme des personnes différentes… En français, dans tout texte, en moyenne plus d'un mot sur 10 est susceptible d'avoir plus d'une graphie. De plus, tout mot commun peut parfois recevoir une majuscule initiale, parce qu'il figure en début de phrase ou pour en souligner la majesté (l'Etat, le Président, l'Université)…

2/ C'est le seul moyen de permettre à l'usager de la base de retrouver tous les emplois de ce mot et d'être certain qu'aucun n'est oublié.

Une fois corrigé, chaque texte a été traité de la manière suivante.

Une fiche indique l'auteur, sa fonction, la date et le lieu d'émission ainsi que la nature du texte (par exemple, pour le discours politique : allocution, entretien, conférence de presse, message…). La fiche comporte également la source du document, la date du traitement et le nom de l'opérateur.

A l'intérieur du texte, des balises isolent le texte du "para-texte" (par exemple : les questions des journalistes), afin de ne traiter que les propos de l'auteur étudié, tout en conservant le "para-texte" qui doit être fourni à toute personne consultant la bibliothèque.

Puis, chaque mot du texte est doté d'une étiquette comportant sa "graphie standard" (opération importante pour les noms propres) et son "entrée de dictionnaire" (mot "vedette" et catégorie grammaticale). Par exemple, le féminin et le pluriel d'un adjectif sont groupés sous le masculin singulier de celui-ci, ou encore toutes les flexions d'un même verbe sont groupées sous l'infinitif, etc. Ces conventions ont été présentées par C. Muller (1963 et 1977) et ont été implémentées sur ordinateur (Labbé 1990). Elles épousent, au plus près, les conventions en usage dans la lexicographie française (la science du dictionnaire) - c'est-à-dire celles communes aux locuteurs du français - et sont confiées à des automates qui réalisent la quasi-totalité de l'étiquetage. Il est important de souligner que cet étiquetage est sans erreur (par rapport aux conventions retenues) et que toutes ces informations s'ajoutent au texte proprement dit auquel on ne touche pas.

[5] Même à la présidence de la République française, l'orthographe est souvent malmenée. Un exemple, parmi beaucoup d'autres : le 27 juin 2018, devant les caméras, le chef de l'Etat promulgue la loi de réforme de la SNCF. Il fait une déclaration – probablement à l'aide d'un prompteur mais il a aussi un papier à la main. L'Elysée ne met en ligne que la vidéo, mais une transcription est divulguée dans laquelle on lit : "des trains moins chère", des "garanti social", le "caractère publique", les "titres détenues", le "domaine publique", "celle qui voudrais", etc., etc. Ces erreurs doivent impérativement être corrigées, sinon elles se répercutent sur l'analyse syntaxique, l'étiquetage des mots et, in fine, sur les résultats statistiques. Nous revenons en conclusion sur cette étrange communication d'E. Macron.

Aperçu du vocabulaire présidentiel

La première opération consiste à établir le vocabulaire de chaque texte (son "index") et, par agrégation de ces listes, le vocabulaire des différents corpus, puis de la bibliothèque entière. Cela permet de connaître notamment les vocables les plus fréquents (tableau 1).

Tableau 1. Les vocables les plus employés dans les interventions d'E. Macron (fréquences exprimées en pour mille mots)

Rang	Vocable	Effectifs	Fréquence (‰)
1	le (dét)	109 557	105,55
2	de (pré)	80 928	77,97
3	être (v)	32 460	31,27
4	et (cj)	30 978	29,84
5	à (pré)	26 679	25,70
6	avoir (v)	21 274	20,50
7	un (dét)	18 784	18,10
8	nous (pro)	16 224	15,63
9	ce (dét)	15 837	15,26
10	qui (pro)	14 969	14,42

Les mots les plus fréquents jouent un rôle essentiellement syntaxique et n'acquièrent un contenu qu'avec le contexte de l'énonciation. Le cas le plus typique est sans doute le pronom "nous" (8^{e} rang pour une fréquence de 15,63 pour mille mots) : simple pluriel de majesté pour désigner le président en style soutenu ? "Moi et mon équipe" ? "Moi et vous m'écoutez" ? "Moi et les Français" ?... Le premier substantif – le nom propre *France* – se trouve au 41e rang avec 2 847 occurrences (soit 2,79 ‰ de la surface des textes). Avec une fréquence inférieure à 0,3%, "France" est donc 38 fois moins fréquent que l'article "le", bien que beaucoup plus important pour l'analyse de la communication présidentielle.

Les 27 vocables les plus utilisés couvrent la moitié du texte et, dans ce groupe, ne figure qu'un "mot plein" : le verbe *faire* (au 27^{e} rang avec une fréquence de 5,26 ‰). A l'opposé les 14 791 vocables restants se partagent l'autre moitié du texte avec des densités faibles. Parmi eux, tous les substantifs, adjectifs et verbes (sauf *être*, *avoir* et *faire*). C'est-à-dire l'essentiel du message. Autrement dit, le vocabulaire d'un corpus est une grande collection d'événements rares très inégalement distribués, dont les plus fréquents ne sont pas, apparemment, les plus intéressants pour l'analyse du discours. Ce phénomène d'inégale distribution des fréquences a été mis en lumière par Zipf 1935 (voir également Mandelbrot, 1957). Il complique évidemment l'analyse statistique.

La comparaison avec les autres présidents porte à la fois sur le rang et sur la fréquence. Le rang est intéressant parce qu'il suggère la hiérarchie que donne le locuteur aux vocables, donc aux objets qu'il traite, spécialement quand il s'agit d'entités identifiées par des noms propres. Ainsi E. Macron partage avec ses prédécesseurs les cinq premiers noms propres dans le même ordre (tableau 2). Continuité relative ? Ou poids des réalités qui s'imposent à tout président français ? Mais cette continuité relative ne doit pas masquer les variations importantes dans les

effectifs (pour comparer des corpus de longueurs inégales, on utilise les fréquences).

Tableau 2. Les dix noms propres les plus employés par E. Macron comparés aux autres présidents.

Rang Macron	Rang Présidents	Vocable	Effectif	Fréquence (‰)	Macron/ Présidents (%)
1	1	France	2 897	2,79	-32,0
2	2	Europe	1 982	1,91	+4,4
3	3	Français	588	0,57	-39,4
4	4	Paris	544	0,52	+30,0
5	5	Afrique	388	0,37	-5,1
6	7	Union Européenne	358	0,34	+21,4
7	87	Sahel	281	0,27	+831,0
8	11	Chine	277	0,27	+68,8
9	6	Allemagne	202	0,19	-36,7
10	8	Etats-Unis	191	0,18	-35,7

La lecture de ce tableau est horizontale : *France* occupe le premier rang aussi bien chez E. Macron que chez ses prédécesseurs. L'actuel président l'a prononcé 2 897 fois soit 2,79 fois pour mille mots. Par rapport à la moyenne de ses prédécesseurs, cette densité est inférieure de 32% (dernière colonne).

Une baisse encore plus forte (-39%) affecte *Français* qui arrive cependant toujours en troisième position.

Chez tous les présidents, depuis 1962, le second nom propre est toujours *Europe*. Sa présence est d'autant plus forte qu'il faut y ajouter *Union Européenne* (en forte augmentation : +21,4%). Toujours en cinquième position : l'*Afrique*.

Une interrogation fondamentale demeure : ces variations sont-elles "anormales" ? En effet,

Au sein d'une vaste population – que l'on suppose homogène - un caractère donné n'est jamais distribué de manière uniforme chez tous les individus composant cette population. Il y a une variabilité normale comprenant la quasi-totalité de la population et quelques individus qui sortent par le haut ou par le bas de cet intervalle standard.

Le calcul des probabilités permet de définir cet intervalle de fluctuation normale et de repérer les individus "exceptionnels".

Considérons le mot *France*, il faut répondre à deux questions : l'usage qu'a fait E. Macron du mot "France" au cours des 20 premiers mois de sa présidence s'écarte-t-il *significativement* (en plus ou en moins) de celui de ses prédécesseurs ? Combien a-t-on de chances de se tromper en adoptant une réponse : oui/non ?

Probabilités et vocabulaire

J. Bernoulli nous permet d'apporter une réponse à ces questions.

A notre connaissance, l'idée d'appliquer ce calcul au vocabulaire revient à C. Muller (1964 et 1977). Mais les ordinateurs de l'époque obligeaient à utiliser une approximation (l'écart réduit). P. Lafon (1980 et 1984) a présenté, sous le nom de "spécificité du vocabulaire", le raisonnement qu'on va lire ci-dessous mais il l'avait appliqué aux "formes graphiques" et dans une présentation peu commode[6].

Il s'agit de choisir – avec un risque d'erreur qu'il faut minimiser - entre deux hypothèses contradictoires :

- H_0 (hypothèse nulle) : la densité d'emploi de *France* par E. Macron ne diffère pas significativement de celle de tous les autres ;

- H_1 : E. Macron présente une propension à utiliser *France* différente des autres.

Accepter une hypothèse (avec une certaine incertitude) ne signifie pas qu'elle est "vraie" mais simplement que l'on peut écarter l'hypothèse contraire (Desrosières 1988). Par exemple, accepter H_0 (le phénomène est le même) signifie que les différences constatées sont considérées comme des fluctuations propres à tout phénomène naturel (rejet de H_1). A l'inverse, accepter H_1 revient à dire que l'écart entre les observations et les valeurs attendues est si fort que l'on peut rejeter H_0 avec un risque d'erreur minime et chiffré.

La procédure qui permet de valider l'une ou l'autre des deux hypothèses et d'estimer le risque d'erreur est la suivante. Soit une "urne de Bernoulli" - appelée "Présidents" – contenant :

- N_p le nombre total de mots du discours présidentiel (18 413 088)

- F_{ip}, l'effectif du vocable i dans le discours présidentiel (pour *France* 75 493)

On tire au hasard de cette urne un échantillon dénommé "Macron" :

- N_m : nombre de mots contenus dans l'échantillon M (1 038 899)

- F_{im}: le nombre de fois que, dans l'échantillon M, se trouve le vocable i (pour *France* : 2 897).

Soit la variable X représentant le nombre de fois que le vocable i se trouve dans un échantillon de N_m mots. La probabilité d'obtenir $X = F_{im}$ dépend de la combinaison de deux événements :

- le nombre de possibilités différentes de sortir N_m mots de l'urne P :

$$C_{N_m}^{N_p} = \frac{N_p\,!}{N_m\,!\,\left(N_p - N_m\right)\,!} = \binom{N_p}{N_m}$$

- le nombre de possibilités différentes de sortir F_{im} mots dans un effectif total de F_{ip} mots :

$$C_{F_{im}}^{F_{ip}} = \frac{F_{ip}\,!}{F_{im}\,!\,\left(F_{ip} - F_{im}\right)\,!} = \binom{F_{ip}}{F_{im}}$$

La probabilité composée de ces deux événements suit une loi hypergéométrique de paramètres : N_p, N_m, F_{ip}, F_{im}

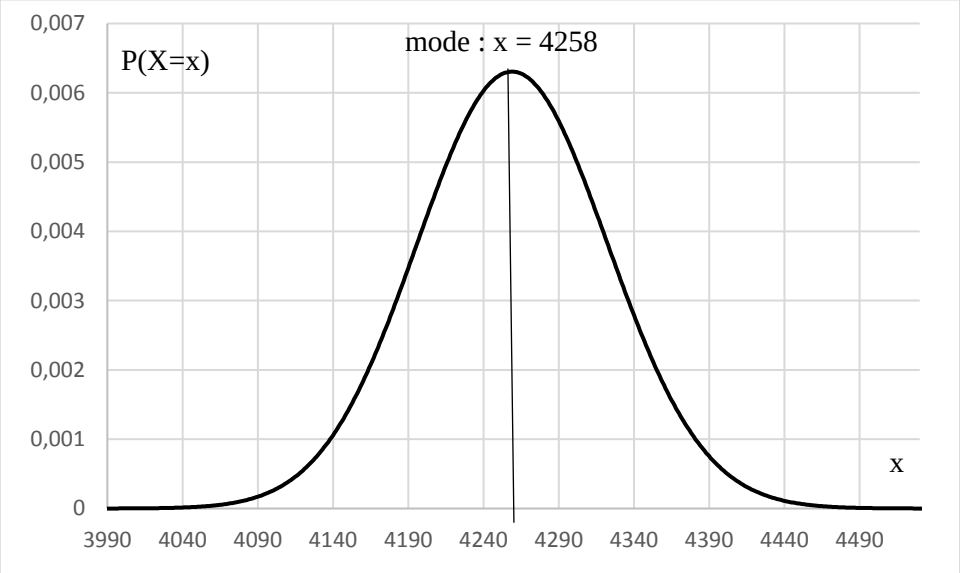
$$P(X = F_{im}) = \frac{\binom{F_{ip}}{F_{im}}\binom{N_p - F_{ip}}{N_m - F_{im}}}{\binom{N_p}{N_m}} \qquad (1)$$

[6] Pour une discussion : Labbé et Labbé 1994. Pour une application du même genre au discours présidentiel : Savoy 2010.

Pour des grands effectifs comme *France* chez les présidents, cette formule n'est praticable qu'à l'aide des logarithmes qui remplacent les multiplications par des sommes. Avec cette aide, la réponse est que la probabilité de l'événement "2 897 fois le mot *France* dans 1 038 899 mots tirés de l'urne P" est inférieure à 1^{e-32} (infiniment petite)[7].

Cependant, avant de conclure, observons comment se comporte P(X). Pour cela, dans la formule (1), F_{im} est remplacé par une variable *x* (entier naturel) : l'ordinateur calcule toutes les valeurs possibles de *P* entre le minimum concevable (x = 0 : aucun *France* dans l'échantillon) et le maximum x = 75 493 (tous les *France* contenus dans l'urne sont présents dans l'échantillon). Les résultats sont portés sur la Figure 1 ci-dessous.

Figure 1. Histogramme des valeurs de P(X=x) calculées avec la formule (1) et appliquée au vocable *France* chez E Macron comparé aux autres présidents.



Afin de rendre l'étalement mieux visible, l'origine des abscisses est placée à la valeur de la variable (x = 3 989) pour laquelle P(X) dépasse pour la première fois 1/10 milliards (1^{e-8}), de même pour la dernière valeur (x = 4 530) de P(X) dépassant ce même seuil qui est justifié plus bas. Si l'on avait retenu $P(x) > 0{,}1^{e-32}$ l'étendue aurait été [3509 … 5010]. Le phénomène central d'étalement aurait été peu visible. Cependant, dans cet intervalle très large, on ne peut affirmer que l'événement est strictement impossible…

La forme de la courbe dite "en cloche" est celle que l'on obtient toujours dans ce type d'expérience. *P(X = x)* passe par un maximum appelé "mode" pour x = 4 258.

Notons :
- Le taux d'échantillonnage :

$$U = N_m / N_p = 1038899/18413088 = 0{,}0564 \text{ ou } 5{,}64\%$$

[7] Le calcul est limité à 32 décimales. Dans la formule (1), les valeurs peuvent être très grandes (ou très petites), ce qui pose un problème de précision du résultat que les calculateurs contemporains permettent de maîtriser assez bien. Nous revenons plus bas sur cette question (commentaire sous la Figure 1).

- le nombre de mots *France* attendus dans le corpus Macron en fonction de sa fréquence dans P (ou effectif théorique) :

$$Ftheo_{im} = F_{ip} * U = 75\ 493*0{,}0564 = 4\ 258 \qquad (2)$$

Cet effectif théorique correspond effectivement au sommet de la courbe.

Toutefois, même pour ce mode, la probabilité (axe vertical) demeure extrêmement faible (inférieure à 0,007). Dès lors, comment répondre aux deux questions ci-dessus, sachant que les tables usuelles de probabilités ne vont pas jusqu'aux chiffres que nous manipulons aujourd'hui ?

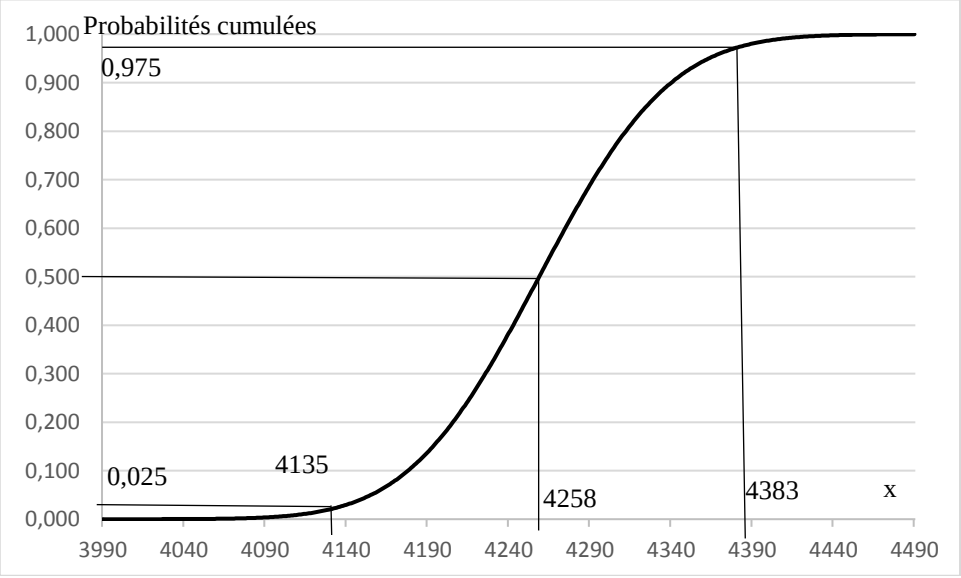
Calcul du vocabulaire caractéristique

Voici une solution simple. Soit l'indice C. Il cumule les probabilités, calculées avec la formule (1) ci-dessus, en faisant varier x de 0 jusqu'à la valeur observée dans l'échantillon. Soit, pour le vocable *i* (survenant F_{im} fois dans l'échantillon M), l'indice C_i sera :

$$C_i = \sum_0^{F_{im}} P(X \leq F_{im}) \qquad (3)$$

Comme précédemment remplaçons, dans la formule (3), F_{im} par *x* (entier naturel) et faisons varier *x* de 0 à F_{ip}. (Figure 2). Pour permettre la comparaison avec le graphique précédent, les abscisses sont les mêmes.

Figure 2. Variation de C en fonction de x (France chez E. Macron comparé à ses prédécesseurs).



Notons d'abord que C atteint 1.0 pour x = 4 590 et ne bouge plus ensuite mais que la neuvième décimale est non nulle. Comme cette courbe est le résultat du cumul de 1500 valeurs de P(X) calculées avec la formule (1) à l'aide des logarithmes[8], on en conclut que la précision du calcul est satisfaisante, mais qu'il ne faut pas exprimer l'indice C avec plus de 8 décimales.

Sur le graphique, trois bornes indiquent des valeurs remarquables :

C atteint 0,025 pour x = 4 135

C atteint 0,5 pour 4258, c'est-à-dire $Ftheo_i$. C'est également le point d'inflexion de la courbe (le trait vertical est l'axe de symétrie de la courbe),

C atteint 0,975 pour x = 4 383

Autrement dit, 95% du total des valeurs de l'indice sont comprises dans l'intervalle [4135 … 4383]. On en tire que, si E. Macron utilisait *France* comme ses prédécesseurs, il y aurait moins de 5% de chances de rencontrer ce vocable dans ses discours moins de 4 135 fois ou plus de 4 383. Par convention ce risque d'erreur est noté α. L'intervalle [4135 … 4383] est la "plage de variation normale" (avec α = 5%). Pour toute valeur comprise dans cet intervalle, il est impossible de rejeter l'hypothèse nulle selon laquelle l'usage que fait E. Macron du vocable considéré ne diffère pas de celui de ses prédécesseurs.

On trouvera dans le tableau 3 ci-dessous, ces intervalles de confiance pour différentes valeurs de α.

Tableau 3. Intervalles de confiance associés aux effectifs attendus du vocable *France* chez E. Macron en fonction de la densité de ce vocable dans l'ensemble du corpus présidentiel et du risque d'erreur (α)

α	Borne inférieure (α)	Borne supérieure (1-α)
0,05	4 135	4 383
0,01	4 097	4 421
0,001	4 051	4 468
(…)		
0,000 000 001	3 910	4 609

E. Macron n'utilise *France* que 2 897 fois, nettement moins que la borne la plus basse (3 910). Il y a donc moins d'une chance sur un milliard de se tromper en affirmant qu'il a moins utilisé ce vocable que ses prédécesseurs. On peut encore dire qu'il a eu une propension nettement plus faible à prononcer ce mot.

Ces propriétés sont d'un grand intérêt. Elles permettent notamment de comparer la distribution d'un caractère dans une ou plusieurs populations, de définir des plages de "fluctuations normales", de repérer les individus qui se situent en dehors de ces plages, etc.

Répétons l'opération pour les autres noms propres les plus fréquents (tableau 4)

[8] Le cumul commence quand P(X) atteint $0,1^{e-32}$ (x=3509) et devrait théoriquement s'achever à x=5010.

Tableau 4. Effectifs constatés (F_{im}) et attendus (F_{theo}) pour les 10 noms propres les plus fréquents dans le corpus Macron comparé au corpus présidentiel total

Vocable	F_{im}	F_{theo}	Ecart absolu ($F_{theo} - F_{im}$)	Ecart relatif (%)	Intervalle	$\alpha = 0,05$
France	2 897	4 252	-1355	-47	4135 - 4383	-
Europe	1 982	1 903	79	4	1819 - 1988	≈
Français	588	974	-386	-66	912 - 1035	-
Paris	544	415	129	24	376 - 454	+
Afrique	388	408	-20	-5	369 - 447	≈
Union Européenne	358	292	66	18	259 - 326	+
Sahel	281	30	251	89	19 - 41	+
Chine	277	170	107	39	144 - 196	+
Allemagne	202	316	-114	-57	282 - 351	-
Etats-Unis	191	291	-100	-53	258 - 325	-
Somme	7 708	9 052	-1344	-17	8902 - 9201	-

Le tableau se lit horizontalement. Par exemple, la deuxième ligne indique que E. Macron a employé 1 982 fois *Europe* alors que la pratique moyenne de ses prédécesseurs n'en laissait attendre que 1 903, soit un écart de +79 (ou encore il y en a 79 "en plus" dans les propos d'E. Macron), soit un excédent de 4%. Mais l'effectif constaté (1 982) est légèrement inférieur à la borne supérieure de l'intervalle de confiance (1 988), par conséquent, il est impossible, avec moins de 5% de chances de se tromper, d'écarter l'hypothèse nulle selon laquelle E. Macron n'utilise pas différemment *Europe* par rapport à ses prédécesseurs (dernière colonne du tableau). Même conclusion pour *Afrique*.

Pour les huit autres, les écarts sont significatifs et parfois très significatifs. Ainsi, il y aurait moins d'une chance sur un milliard de se tromper en affirmant que E. Macron emploie plus *Paris*, *Sahel* et *Chine* et moins *France*, *Français*, *Allemagne* et *Etats-Unis*.

Avant d'aller plus loin dans le commentaire, remarquons que cette démarche peut se voir opposer certaines réserves.

Trois réserves

Premièrement, le modèle implique que, plus les effectifs d'un vocable sont élevés dans le vocabulaire total (l'urne), plus ses effectifs dans l'échantillon tendront vers la valeur attendue. On parle de "convergence en probabilité". Par exemple, pour *France*, l'intervalle d'incertitude autour de la valeur attendue ($F_{theo} = 4532$) est de 248 mots [4383 … 4135] soit 5,8% de $F_{theo.}$; pour *Etats-Unis*, il est de 23% (67/291). Dans la pratique, le resserrement de l'intervalle relatif, en fonction de l'augmentation des effectifs, est moins fort que ne le laisse attendre le modèle, de telle sorte que plus un vocable est fréquent, plus il a de chance de figurer dans le vocabulaire caractéristique d'une partie d'un ensemble.

La réponse à cette première objection consiste à "durcir" les seuils pour les vocables les plus fréquents.

Deuxième réserve : une certaine lourdeur. Il serait très fastidieux de répéter cette opération plusieurs milliers de fois, or le vocabulaire d'E. Macron comporte 14 819 vocables…

La réponse consiste à lire directement C.

Comme l'indique la figure 2, l'indice varie harmonieusement entre 0 et 1 sans saut pour certaines valeurs. L'intervalle de confiance découle de la valeur de α choisie. Trois situations sont possibles.

- '-' lorsque C est inférieur à $0{,}5\alpha$ (par exemple 0,025 pour un risque d'erreur de 5% ; 0,005 pour α = 1%, etc.) : le vocable est significativement moins employé que dans le corpus de référence ;

- '+' lorsque l'indice est supérieur à $1\text{-}05\alpha$ (0,975 pour un risque d'erreur de 5% ; 0,995 pour 1%, etc.) : le vocable est sur-employé par rapport au corpus de référence ;

- '≈' lorsque l'indice est compris dans l'intervalle $[0{,}5\alpha \ldots 1 - 0{,}5\alpha]$, on ne peut écarter l'hypothèse selon laquelle la densité du vocable dans le corpus sous revue ne diffère pas significativement de celle dans l'ensemble de référence. Au centre de cet intervalle, lorsque l'effectif observé est égal à celui attendu, C est égal à 0,5.

L'indice donne ainsi directement le renseignement désiré, son interprétation est aisée et sans ambiguïté. Avant d'en présenter les résultats sur les discours d'E. Macron, voyons la dernière objection qui concerne le poids des catégories grammaticales.

Poids des catégories grammaticales

L'examen des 10 principaux noms propres utilisés par E. Macron a suggéré que ce président emploierait nettement moins de noms propres que ses prédécesseurs (dernière ligne du tableau 4) et que cette différence a une chance infinitésimale d'être due au hasard. Cette caractéristique particulière est-elle confirmée sur l'ensemble du vocabulaire et, dans ce cas, comment adapter le calcul ?

Rappelons que, lors des traitements préalables, chaque mot a été doté d'une étiquette indiquant son entrée de dictionnaire et sa catégorie grammaticale. Cela rend possible une comparaison de l'utilisation qu'en fait chaque président (annexe 2)

La lecture du tableau en annexe 2 est horizontale : pour tous les mots d'une catégorie grammaticale donnée, la première colonne donne son poids chez les présidents ayant précédé E. Macron, la deuxième colonne : la densité de cette même catégorie chez E. Macron ; la troisième l'écart entre les deux ; dans la dernière colonne, le signe répond à la question : la propension d'E. Macron à utiliser cette catégorie est-elle plus faible, plus forte ou sensiblement égale à la moyenne des autres (au risque d'erreur choisi) ?

La dernière ligne du tableau 4 suggère que l'actuel président utilise 17% de noms propres en moins par rapport à ses prédécesseurs (estimation sur les dix noms propres les plus utilisés). Le calcul sur l'ensemble de cette catégorie (annexe 2), confirme que, par rapport aux autres, E. Macron utilise 16,7% de noms propres en moins. L'indice en dernière colonne étant inférieur à 0,0025 : il y a moins de 1 chance sur mille de se tromper en affirmant qu'E. Macron utilise moins de noms propres que ses prédécesseurs.

Logiquement, avec les formules (1) à (3), la majorité des noms propres d'E. Macron apparaissent en C- à l'inverse de la majorité des noms communs et surtout des adjectifs qui sont C+ parce qu'il a tendance à les utiliser plus que la moyenne des autres présidents, etc. Si l'on admet qu'il doit y avoir un relatif équilibre entre les caractéristiques positives et négatives, on peut dire que la procédure classique – le calcul des "spécificités" – comporte un biais[9].

Pour éliminer ce biais, le calcul doit prendre en compte les catégories grammaticales (g)[10].

Soit N_{gp} et N_{gm} le nombre de mots appartenant à la catégorie grammaticale (g) respectivement dans l'ensemble du discours présidentiel et chez E. Macron, les formules (1) et (2) deviennent :

$$P_i(X = F_{im}) = \frac{\binom{F_{ip}}{F_{im}}\binom{N_{gp} - F_{ip}}{N_{gm} - F_{im}}}{\binom{N_{gp}}{N_{gm}}} \qquad (4)$$

$$F_{tbeo} = F_{ip} * U \; avec \; U = \frac{N_{gm}}{N_{gp}} \qquad (5)$$

Autrement dit, les tirages aléatoires ne s'effectuent plus dans une urne unique mais dans neuf urnes correspondant aux principales catégories grammaticales (verbes, noms propres, substantifs, adjectifs, pronoms, adverbes, déterminants, préposition et conjonctions).

Les formules (4) et (5) aboutissent à un équilibre relatif, au sein de chaque catégorie, entre les C+ et les C- . Ces formules neutralisent donc la liaison entre caractéristiques et densités des catégories grammaticales. Comme indiqué dans Monière & Labbé 2012, cette modification change notablement la liste des "mots caractéristiques". Par exemple, pour les noms propres d'E. Macron (tableau 5), *Europe*, *Afrique* et *Italie* passent de C≈ à C+ ; *Nations Unies* et *Liban* passent de C- à C≈.

Le même test statistique peut être appliqué aux locutions, groupes figés de mots que la statistique textuelle a baptisés "segments répétés" (et les informaticiens "N-Grams") : "chef de l'état", "premier ministre", "mener une politique", "porter un projet", etc. Il peut également servir à révéler les préférences du locuteur pour telle ou telle catégorie grammaticale (voir annexe 2), cela permet alors de caractériser le vocabulaire et le style d'un auteur.

La seconde partie de cet exposé illustrera ces potentialités à propos de la singularité d'E. Macron.

[9] On parle de "biais" lorsqu'apparaît une différence systématique entre les valeurs de l'espérance d'un estimateur et les observations qu'il est censé estimer.

[10] La modification est présentée dans : Monière, Labbé, Labbé 2005 ; Monière, Labbé 2012 et Monière, Labbé 2018.

II. Le vocabulaire et le style d'E. Macron

Une approche des principaux thèmes est donnée par les noms propres alors que les verbes indiquent plutôt une certaine relation du locuteur par rapport à son auditoire et à son action.

Noms propres

En choisissant un seuil rigoureux (risque d'erreur inférieur à 1‰), l'indice C permet de déterminer les principaux sujets abordés par le président Macron au cours de ses premiers mois de présidence (tableau 5).

Tableau 5. Indice de caractéristique des principaux noms propres d'E. Macron comparés à l'ensemble du discours présidentiel.

Rang Macron	Rang Présidents	Vocable	Effectif	F_{tbeo}	Indice ($\alpha < 0.001$)
1	1	France	2 897	3 575	-
2	2	Europe	1 982	1 600	+
3	3	Français	588	819	-
4	4	Paris	544	349	+
5	5	Afrique	388	343	+
6	7	Union Européenne	358	246	+
7	87	Sahel	281	25,1	+
8	11	Chine	277	142,9	1
9	6	Allemagne	202	266	-
10	8	Etats-Unis	191	245	-
11	10	Nations Unies	183	163	≈
12	23	Syrie	157	82,8	+
13	1734	Libye	149	15,2	+
14	34	Corse	145	59,5	+
15	13	Liban	126	115	≈
16	21	Méditerranée	120	83	+
17	337	G5	108	7,6	+
18	64	Guyane	100	33,9	+
19	26	Italie	94	76	+
20	128	Union Africaine	93	13,1	+

Ce tableau met en lumière le poids de la conjoncture, et l'importance des deux guerres dans lesquelles la France se trouvait engagée à l'arrivée au pouvoir d'E. Macron : la *Syrie* (+58%) et surtout le *Sahel* (+831 %) auquel se rattachent : *Libye*, *Union Africaine* et *G5* : acronyme très polysémique qui désigne chez E. Macron le groupe des 5 pays engagés aux côtés de la France dans la guerre au Sahel (groupe créé par F. Hollande en 2016 et désigné 7 fois sur 10 par "G5

Sahel"), mais aussi la coopération policière contre le terrorisme entre 5 pays européens lancée par N. Sakozy au début des années 2000, etc. Dans le même ordre d'idée, le tableau indique l'importance de la crise migratoire (*Italie*) ou le retour de la question *corse* sur le devant de la scène politique intérieure.

Au-delà de la conjoncture, le tableau révèle des choix fondamentaux peu décelables à la simple écoute du discours présidentiel, notamment la place plus importante – par rapport à ses prédécesseurs – accordée à l'*Europe* (et l'*Union Européenne*) ou à la *Chine* et, en contrepartie, celle nettement moindre accordée par E. Macron à la *France*, aux *Français* et aux relations avec l'*Allemagne* et les *Etats-Unis* (que privilégiaient les précédents présidents).

Trois remarques complètent ce tableau.

Premièrement, il faut prendre garde aux changements intervenus dans les dénominations. Par exemple, en fonction de l'usage moyen des présidents depuis 1958, on attend 76 fois "Union Soviétique" chez E. Macron alors qu'il ne prononce pas ce mot pour des raisons évidentes. En revanche, le sous-emploi de *Russie* en devient beaucoup plus significatif : ce vocable passe du 12^{e} rang (chez les autres présidents) au 30^{e} (chez E. Macron) avec une fréquence en recul de 34%). La remarque inverse peut être faite à propos de l'*Union Européenne* – appelée d'abord "Marché Commun" ou "Communauté Economique Européenne" jusqu'en 1992, puis "Communauté Européenne" jusqu'en 2009. Comparé aux seuls N. Sarkozy et F. Hollande, E. Macron emploie autant *Union Européenne* mais nettement plus *Europe*.

Deuxièmement, il ne faut pas considérer les vocables isolés mais les familles auxquelles ils appartiennent. Par exemple, chez E. Macron, les reculs de *France* et de *Français* s'accompagnent d'une baisse équivalente des substantifs "pays" (-25%) et "nation" (-5%), des adjectifs "français" (-11%) et "national" (-18%). Dans le même ordre d'idées, E. Macron utilise moins les noms de peuples *Américain* et *Allemand* ainsi que les adjectifs "américain" ou "allemand". A l'inverse, l'augmentation de *Europe*, s'accompagne de celle des adjectifs *européen* (+46%) et *commun* (+ 80% essentiellement dans "politique commune" et "projet commun"). De même, les adjectifs *africain* et *chinois* figurent dans les suremplois caractéristiques d'E. Macron.

Troisièmement, pour l'actuel président, la grande affaire est la *géopolitique* (densité multipliée par 12 par rapport à ses prédécesseurs). Pourtant, il a une répugnance manifeste à nommer certains pays – au premier rang desquels l'*Allemagne*, l'*Amérique* et la *Russie* -, ou certains dirigeants et il utilise des métonymies. Par exemple à propos des relations franco-allemandes, il parle de *Paris* et *Berlin* ou de la *chancelière* (plutôt que de Mme Merkel).

La même analyse peut être conduite sur chaque catégorie grammaticale. Par exemple, que peut-on dire des verbes[11] d'E. Macron ?

Verbes

Le tableau en annexe 2 indique que, là où ses prédécesseurs utilisaient 100 verbes, il n'y en a que 98,2 chez E. Macron, la différence peut sembler faible mais elle est significative au seuil de $\alpha = 0{,}000\ 000\ 001$ (maximum de précision du calcul). Chose intéressante, cette moyenne est

[11] C'est le point aveugle de la statistique "textuelle" car les flexions et les homographies sont trop nombreuses pour permettre de retrouver les verbes en l'absence de lemmatisation. Or le verbe est l'"opérateur de la phrase" et joue un rôle clef dans la communication (Dubois 1969).

la résultante de mouvements internes très significatifs quant aux temps utilisés (Benveniste 1959) :

- préférence pour le futur (ce qui peut sembler logique puisqu'il est en début de mandat), et surtout pour le participe présent et l'infinitif qui sont les formes verbales les plus proches du nom (qui a la préférence d'E. Macron pour des raisons qu'on découvrira plus loin).

- en revanche, il évite autant que possible le conditionnel – expression d'un vœu ou d'un doute - et les temps du passé.

Les verbes favoris sont donnés dans l'annexe 3. La plupart des écarts sont significatifs au seuil de 1‰. Le sous-emploi de "avoir" s'explique par celui du passé. Celui inverse de *faire* est à rattacher à la préférence pour le nom ; par exemple, plutôt que "déclarer", il dira "faire une déclaration".

Les verbes "pseudo-auxiliaires" sont particulièrement éclairants (Benveniste 1965 ; Gross 1999 ; Labbé, Labbé 2013). Ce sont des constructions verbales du type "verbe modalisateur + verbe à l'infinitif" (comme *vouloir faire*, *pouvoir dire*, etc.). Par rapport à ses prédécesseurs, E. Macron délaisse les modalités du possible (*pouvoir*) et de la nécessité (*falloir*) au profit de l'obligation morale ou légale (*devoir*) et de la volonté (*vouloir et souhaiter*). Quant à *aller* et *venir* il ne s'agit pas d'un déplacement dans l'espace mais dans le temps. Combinés à un autre verbe à l'infinitif, ils remplacent respectivement le futur (*aller faire*) et le passé (*venir de dire*) en plaçant l'action dans une sorte d'extension du présent de l'énonciation.

Les écarts les plus forts signalent des évolutions des manières de parler autant que des tics de langage du président. Par exemple, on dit "construire" (multiplication par 3,6) - pour concevoir, élaborer -, voire "co-construire", *conduire* (x 2,3), *porter* (x 2,2), *acter* une *politique*, un *projet*, une *position*. De même *accompagner* (x 3,3) est devenu un quasi-synonyme de *aider*, *soutenir*, *secourir*, etc.

Pour accéder plus commodément aux principales caractéristiques, les vocables sont classés en fonction de leur indice. Les annexes 4 et 5 donnent les vocables les plus sur-employés et sous employés qui offrent un aperçu des singularités du président.

Personnalisation singulière

La surprise principale vient de la présence dans les caractéristiques négatives les plus significatives des pronoms "je" (11^{e} ligne avant la fin dans l'annexe 5) et "il" (4^{e} vocable le plus sous-employé). Non seulement, E. Macron utilise significativement moins de pronoms personnels par rapport à ses prédécesseurs (-7%, annexe 2) mais de plus, il évite manifestement de dire "je" (-17% par rapport à ce qui était attendu en tenant compte de la faible propension à utiliser cette catégorie) ou "il" (-30%). A l'inverse, il privilégie "nous" (+60% par rapport à l'usage moyen).

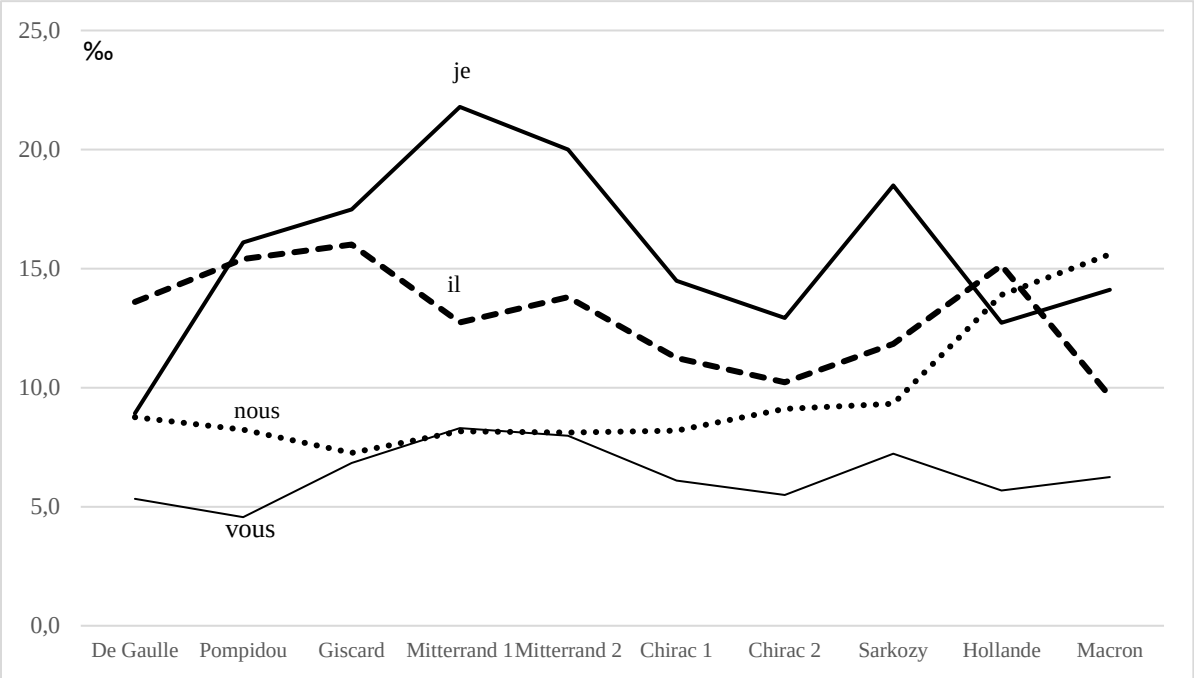
La figure 3 ci-dessous résume ces mouvements en distinguant chaque président.

Comme l'indique E. Benveniste (1956 et 1958), ces pronoms ont pour première caractéristique de prendre un sens particulier à chaque énonciation. Cependant, dans une communication assez contrainte, comme l'est la communication présidentielle, le "je" est toujours le président et le "vous" le destinataire du message (et au-delà les Français, car sinon à quoi bon parler ?). En revanche, le "nous", comme le "il", peut avoir des significations assez

différentes en fonction du reste de la phrase (contexte du mot) et de la situation dans laquelle le discours est prononcé.

Aucun président n'est parvenu à dire aussi peu "je" que C. de Gaulle. Certes, le Général laisse le souvenir d'un "pouvoir personnel" fort mais, à l'époque, cette impression s'expliquait par les circonstances de la guerre d'Algérie et par le contraste avec les traditions parlementaires antérieures. En fait, ce sont ses trois successeurs qui ont imposé cette personnalisation, spécialement, F. Mitterrand qui a fait preuve d'un véritable "égotisme" qu'aucun autre président n'a approché après lui (sauf N. Sarkozy qui dépasse V. Giscard d'Estaing mais ne parvient pas à égaler F. Mitterrand).

Figure 3. Fréquences des pronoms personnels chez les présidents depuis 1958



En ce qui concerne E. Macron, la faiblesse relative de la première personne surprend parce qu'il avait comparé la fonction présidentielle à Jupiter[12] et que les observateurs s'accordent pour estimer que, jamais sous la Ve République, le pouvoir n'a été aussi concentré entre les mains du président.

A l'inverse, le "nous", même s'il est parfois une autre manière de dire "je", est le plus souvent "moi et d'autres", c'est-à-dire une tentative d'inclure le destinataire dans le propos pour l'inciter à assumer ce que dit l'orateur (Labbé 1997). A partir de F. Mitterrand, une proportion non négligeable des "nous" sont des pluriels de majesté équivalent à "moi parlant au nom de la France". L'idée selon laquelle le président est la "voix de la France" explose littéralement avec E. Macron.

Enfin, comme l'indique E. Benveniste, la troisième personne ("il") n'est pas un "impersonnel" mais désigne un tiers absent de l'interlocution qui peut avoir été nommé précédemment (anaphore) ou être mentionné implicitement (souvent équivalent à "on"). Ce tiers peut ne pas être une personne physique et prendre toutes les nuances allant de l'impératif ("il faut") au simple constat ("il pleut"). Chez le Général, la majorité de ces pronoms

[12] Entretien accordé au magazine *Challenges*, octobre 2016.

renvoyaient à la réalité à laquelle la France devait s'adapter. G. Pompidou puis V. Giscard d'Estaing se sont inscrits dans ce prolongement.

En fait, la montée du "nous" et le recul du "je" commencent à la fin du mandat de N. Sarkozy, quand les difficultés ont commencé à s'amonceler. Le "je" s'efface et passe la responsabilité à la collectivité. Chez F. Hollande le principal sujet devient "il" – la fatalité – alors que chez E. Macron c'est le *nous*. Par exemple, plus bas, on lira dans l'encadré : « je *souhaite* que *nous puissions* continuer cette stratégie ».

Les suremplois les plus caractéristiques (annexe 4) montrent qu'il s'agit d'un trait essentiel de son discours.

Réseaux sémantiques

En tête des suremplois, le pronom démonstratif (ce) - *c'est, ce que* et *ce qui* - que le logiciel isole comme étant le trait le plus caractéristique du nouveau président. La suite montrera qu'il ne s'agit pas d'un simple tic de langage, mais d'un choix de communication.

Plus au fond : il s'agit de *refonder* (C+), de *mener* (C+), de *construire* (C+) une *politique*, une *action*, de *porter* une *réforme*, des *projets* suivant une *autre méthode*, des *axes structurants* (tous termes en C+). Le président ne dit pas *falloir* (C-) mais *souhaiter* (C+), il évite *penser* ou *savoir* et privilégie *croire* (C+).

Affirmer que l'on *souhaite* (C+) une *chose* (C-) n'équivaut pas à annoncer qu'on va la *réaliser* (C- division par 2,5). La *réalisation* (C-, division par 2,3) ou *non* (C-) de cette *chose* n'ont pas la même *portée* (C-, division par 2,2). Si l'on s'est contenté de la *souhaiter* (C+) et qu'elle n'*advient* (C+) pas, ce n'est pas *grave* (C- division par 2,5), surtout si l'*exécution* (C-) du *souhait* (C+) dépend de *certains* (C+ multiplication par 1,9). En revanche, si le locuteur *ne parvient* pas (C-) *à réaliser* (C- ÷ 2,5) cette *volonté* (C-), l'*échec* (C-) est *personnel* (C- division par 1,7).

De même, privilégier *croire* évite d'avoir à se justifier contrairement à la *pensée* (C-) ou à la *connaissance* (C-) qui doivent s'exposer *rationnellement* (C-). E. Macron se contente d'énoncer des *évidences* (C+) qui n'ont pas à être *discutées* (C-). Comparé à ses prédécesseurs, les deux formules préférées d'E. Macron sont "il est indispensable" et la "France a besoin". Sans surprise, l'adverbe préféré d'E. Macron est "évidemment" (8^{e} rang des caractéristiques positives, 3,6 fois plus employé que chez les autres).

Ces condensés forment la trame des messages d'E. Macron et résument sa communication des 20 premiers mois. Pour les retrouver, il a suffi de considérer les mots usuels dont les fréquences s'écartent le plus significativement de ce qui est attendu et de retrouver leurs associations les plus fréquentes, c'est-à-dire le réseau que forment ces mots entre eux.

A titre d'exemple voici un autre réseau sémantique - moins fréquent mais très caractéristique d'E. Macron - autour de "religion" qu'il emploie 143 fois alors que l'usage des autres en laissait attendre seulement 62 (multiplication par 2,3). Dans ce réseau, le calcul révèle d'abord l'adjectif "islamiste" dont la fréquence est multipliée par 8 par rapport aux prédécesseurs. Puis viennent le *catholicisme* et les *catholiques* (n. m.) dont les densités sont multipliées par 6, suivis par la locution "religion catholique" (x4). Les *protestants* (n.m. et adj.) voient également leurs densités multipliée par 4 (et le *protestantisme* par 3). E. Macron parle également des *chrétiens* (n. m.) 2,5 fois plus que la moyenne de ses prédécesseurs ou encore du *pape* (x 2,2). En revanche, dans les interventions d'E. Macron, les densités de *juif* (n. m. et adj.) et de *musulman*

ne s'écartent pas de celles enregistrées chez les précédents présidents. Enfin, les substantifs *dialogue* et *croyance* sont également beaucoup plus employés par E. Macron.

Dernier exemple d'un réseau sémantique intéressant : E. Macron parle beaucoup plus de *francophonie* et de *langue (française)*… alors qu'il parsème ses discours de mots étrangers (spécialement anglais : plus d'un mot sur 200 et une multiplication par 5 par rapport à la moyenne de ses prédécesseurs). Attention, il ne s'agit pas des très nombreux mots étrangers intégrés au lexique français (leur entrée dans les dictionnaires récents en atteste). Voici quelques exemples de ces mots entrés récemment dans le lexique français et qu'E. Macron affectionne : "start-up" (nom féminin qui figure au 22^{e} rang de ses vocables les plus caractéristiques), "business" (nom masculin), "cluster" (pour groupe), "mix" (pour mélange), "task force" (n. f.), "testing" (pour action de tester), "tweet" (n. m.), etc, etc. Parmi les 198 mots étrangers repérés chez E. Macron, la plupart ne sont employés qu'une fois. Voici ceux qu'il utilise au moins 5 fois : *planet* (56), *one* (55), *summit* (33), *make* (12), *our* (11) *again, for*, *great*, *small* (10), *group* (9), *and* (7), *climate*, *good*, *the* (6), *name*, *belt*, *road*, *shame*, *to* (5).

"Aujourd'hui se tient à Paris ce qu'on a voulu appeler en bon français la « Paris Digital Week »" (12 novembre 2018).

Est-ce la conscience de cet envahissement qui peut expliquer l'insistance sur la défense du français ? Ou, plus fondamentalement, conçoit-il la francophonie comme un atout dans sa "stratégie géopolitique" ?

Sous-emplois caractéristiques

Le sous-emploi caractéristique commence à la situation suivante : E. Macron n'emploie pas le vocable alors que l'effectif attendu (F_{theo}) est supérieur à 4. En pratique, étant donné que le corpus Macron pèse 5,36% du total, seuls les vocables présents plus de 90 fois chez les autres et absents chez Macron ont un indice C- (avec $\alpha = 0,01$). Par conséquent, seuls les vocables les plus fréquents dans l'ensemble des discours des présidents, peuvent apparaître sous-employés par l'un d'eux. Ce sous-emploi peut avoir deux causes : un choix du locuteur ou l'inactualité du vocable. Comme indiqué plus haut à propos de l'*Union Soviétique*, il faut éliminer cette seconde cause pour pouvoir retrouver les choix de l'orateur.

La dernière ligne de l'annexe 5 réserve une surprise : le sous-emploi le plus remarquable concerne l'article "le". Il en manque près de 4 000 chez Macron par rapport à ce qui serait attendu s'il avait suivi l'usage moyen des autres. Comment un tel écart peut-il se produire sur un mot outil si banal ? Il s'agit d'un choix stylistique consistant à remplacer, devant le nom, cet article par d'autres déterminants à chaque fois que c'est possible. Au premier rang des déterminants qui se substituent à "le" (par ordre d'importance) : l'article démonstratif "ce" (multiplication par 1,6) – qui est à mettre en relation avec l'usage du pronom démonstratif "ce" déjà signalé -, puis le possessif "notre" (multiplication par 1,7) - à mettre en relation avec la préférence de l'actuel président pour le pronom "nous" - puis les indéfinis *tout*, *chaque*, *plusieurs* et *quelque*. Autrement dit, le "déficit" en article "le" n'est pas un accident mais le produit logique de plusieurs propensions particulières à E. Macron et notamment celle de montrer à son auditoire. L'annexe 2 confirme cette propension : chez lui la densité des adjectifs démonstratifs est de 63% plus élevée que chez les autres présidents. Mais il y aussi une forte propension à substituer de l'indéfini au défini qui est également intéressante quand elle est rapprochée du faible nombre de chiffres et de dates (annexe 2 : -23,7%). En effet, les chiffres (et les noms propres) sont les principaux ancrages du discours dans le temps, l'espace, la société. Leur faible présence accentue le caractère général et abstrait des propos.

Parmi les mots les plus sous-employés (annexe 5), certains sont remarquables parce qu'ils suggèrent combien ce président s'écarte des habitudes de ses prédécesseurs. En particulier, il évite soigneusement : *poser une question, soulever un problème*. D'ailleurs, avant janvier 2018, la plupart des rencontres avec les journalistes (ou le public) se limitent à une déclaration liminaire qui s'achève par "je vous remercie", manière polie de congédier les assistants sans qu'ils puissent interroger le président.

Quant au fond, parmi les thèmes privilégiés par ses prédécesseurs et quasiment ignorés par E. Macron, on trouve : la *modernité* (et l'adjectif *moderne* : fréquences divisées par 8), la *technique* (et l'adjectif *technique* : division par 7), la *croissance*, le *progrès* ou le *chômage* (division par trois), *chômeur* (division par 1,9), *emploi* et *impôt* (division par 1,7) ou très significativement sous-employés comme *avenir* (-40%) *développement* (-38%), *social* (adj -25%) et *économie* (-10%). L'absence des questions d'emploi est d'autant plus remarquable que E. Macron parle beaucoup de l'*assurance-chômage* (densité trois fois plus forte), mais il s'agit de *réformer* cette institution et de *lutter* contre les *abus* et non de créer des emplois.

La présence des adverbes "ne… pas" dans les sous-emplois les plus significatifs est également révélatrice. En effet, nier quelque chose c'est déjà faire une place dans son propos à la parole contraire, donc reconnaître son existence. Or E. Macron ne mentionne pas les oppositions ou les critiques, sauf parfois grâce à un pronom indéfini comme "certain" (multiplication par 1,8).

Vocabulaire commun

L'annexe 6 donne les vocables les plus fréquents pour lesquels l'usage du président ne s'écarte pas significativement de celui des autres (au seuil $\alpha = 0,05$).

Le vocabulaire commun est composé de tous les vocables pour lesquels l'indice se situe entre les deux bornes de l'intervalle de confiance choisi (0,025 – 0,975). Si l'on considère le seuil bas, le calcul ne porte que sur les vocables présents au moins 90 fois dans l'ensemble du vocabulaire présidentiel et présents au moins 4 fois dans le corpus Macron. Le seuil supérieur concerne tous les vocables présents 5 fois dans Macron et dont la fréquence théorique est supérieure à 1 occurrence, soit 37 occurrences dans le corpus total. Le calcul concerne potentiellement près de 6 500 vocables, parmi lesquels seulement 1 716 vocables sont communs. Plusieurs méritent d'être mentionnés - comme : *choisir, agricole, alliance* (*atlantique*), *conviction, courage, logement,* etc. – car ils indiquent des continuités qui ne sont pas négligeables.

Toutefois le poids respectifs des trois vocabulaires (+, -,≈) est loin de ce que laisse attendre l'hypothèse nulle (tableau 6).

Les vocables pris en considération dans le calcul représentent 43,8% du vocabulaire mais ils couvrent la quasi-totalité du texte (99,4%). On remarque également que les sous-emplois équilibrent à peu près les suremplois (grâce notamment à la prise en compte des catégories grammaticales).

Le calcul isole 2 441 vocables sur-employés par E. Macron par rapport à ses prédécesseurs, soit 16,5 de son vocabulaire. Au total, ces 2 441 vocables représentent 442 934 occurrences (mots) – soit 42,7% du corpus total - alors que l'hypothèse nulle en laisse attendre 312 246 soit un écart de 130 687 mots en excédent ou encore 12,6% de la surface totale du corpus Macron (alors qu'on en attend 2,5% dans le cas d'une distribution normale). Les 2 330 vocables sous-employés représentent des proportions à peu près semblables.

Tableau 6. Poids respectifs des trois vocabulaires (C+, C- et C≈) et écarts par rapport à l'hypothèse nulle.

	N	% vocabulaire	ΣF_{im}	% texte ($\Sigma F_{im}/N_m$)	ΣF_{theo}	Ecart absolu (ΣFim - ΣFtheo)	% texte (Ecart/N_m)
C+	2 441	16,5	442 934	42,7	312 246	130 687	12,6
C-	2 330	15,7	554 548	53,4	668 918	124 370	12,0
$C_{\approx}$	1 716	11,6	34 788	3,4	35 077	289	0,0
	6 487	43,8	1 032 270	99,4	1 016 241	16 029	1,5

Du coup, le vocabulaire commun semble résiduel (11,6 % des vocables et 3,4% des mots). Cependant, en ne prenant en compte que les écarts par rapport à l'hypothèse d'homogénéité des présidents (dernière colonne du tableau), c'est 25% du texte qui serait caractéristique alors que l'on en attend que 5%. Cet écart est important et s'explique de deux manières. Statistiquement, la convergence en probabilité ne se produit pas aussi rapidement que le modèle le postule, de telle sorte que les vocables les plus fréquents apparaissent rarement dans le vocabulaire commun. D'autre part, E. Macron s'écarte beaucoup des thèmes de ses prédécesseurs et de la manière de parler qui était en usage auparavant à la présidence de la République.

En conclusion de cette analyse du vocabulaire, même en reconstituant les combinaisons de mots les plus fréquentes et les familles auxquelles ces mots appartiennent, ces listes demeurent assez abstraites, car il manque les contextes larges et les situations dans lesquelles ils sont employés. Pour disposer d'exemples éclairants, il est demandé à l'ordinateur de trouver les phrases les plus caractéristiques d'E. Macron comparé aux autres présidents.

Phrases caractéristiques

Pour trouver ces exemples, le programme relit l'ensemble des interventions d'E. Macron et classe chacune des phrases en fonction de la densité (absolue et relative) des C+ et C- y figurant. L'encadré ci-dessous donne les deux phrases longues et courtes qui contiennent le plus de vocables C+ (et le moins de C-). Ces phrases peuvent être considérées comme les citations canoniques que tout dictionnaire donne à l'appui de ses définitions. Ici, le choix n'est pas fait arbitrairement par le lexicographe mais il est effectué objectivement, de telle sorte que ces phrases sont certainement les plus caractéristiques du locuteur étudié.

Les phrases caractéristiques

Les deux phrases longues les plus caractéristiques :

« Nous étions hier dans un lycée formidable, le lycée Michel Rocard, exemplaire et j'y ai vu toute la richesse de l'archipel et je souhaite que nous puissions continuer cette stratégie éducative qui est le pilier indispensable pour que chacune et chacun y trouve sa place, pour que toute la jeunesse, toute la jeunesse puisse avoir accès à la meilleure éducation et à la meilleure formation, pour qu'on puisse dans l'école enseigner les savoirs fondamentaux, les meilleurs

enseignements mais aussi les cultures, les langues qui font la tradition et la richesse de l'archipel et pour qu'on puisse par ce travail aussi construire ce qui est l'un des défis de cette stratégie qui est devant nous, c'est celle d'une société et d'un archipel plus équilibré, équilibré sur le plan territorial et de son économie, je sais combien vous y tenez et c'est je crois l'un des grands acquis des vingt dernières années, sans doute accéléré ces dernières années et sur lequel nous devons aller plus loin, équilibré en donnant une place à chacun et en étant sûr que chaque enfant aura les mêmes opportunités où qu'il soit né, dans quelque endroit et dans quelque famille ; plus équilibré en s'assurant que les femmes auront autant d'opportunités que les hommes et seront autant respectées que les hommes c'est cette société là que nous devons dessiner parce qu'elle sera exemplaire dans tout le Pacifique, exemplaire par sa vision, exemplaire par ses équilibres, exemplaire parce qu'elle aura tout à la fois le meilleur de l'histoire ancestrale de la Nouvelle-Calédonie et de ce que doit produire l'histoire et l'actualité de la république française. » (Score absolu : 127 mots caractéristiques. Nouméa, 5 mai 2018)

« Il y aura l'année prochaine une saison culturelle entre nos deux pays qui permettra de mettre en avant la vigueur de cette relation et de son histoire, mais aussi des artistes contemporains, les liens forts, académiques, culturels entre nos deux pays, nous nous y engagerons, le programme est en train d'être parachevé, et je reviendrai donc, au lancement de cette saison en Roumanie pour m'engager personnellement et pour, durant l'année 2018, pouvoir venir faire un geste symbolique auquel je tiens beaucoup, c'est de pouvoir venir planter un nouveau chêne ; non pas le chêne de Berthelot, planté quelque dix ans après le geste fondateur qui avait été le sien aux côtés de la Roumanie et qui, aujourd'hui, est ce chêne centenaire, puisqu'il avait pris soin de planter un chêne de dix ans d'âge, mais revenir en 2018 pour planter un nouveau chêne qui sera celui de l'avenir à construire entre nos deux pays, et de la force de cette nouvelle relation ; et cette saison culturelle partagée sera un élément important pour renforcer, sur le plan artistique, linguistique, ces liens, et nous y travaillons. » (Score absolu : 95 mots caractéristiques. Allocution à la communauté française en Roumanie – Ambassade de France à Bucarest - 24 août 2017).

Deux phrases courtes les plus caractéristiques

« Et ça, nous allons le faire et nous allons le faire là ». (22 février 2018, tous les mots sont caractéristiques).

« C'est légitime et donc c'est ce que nous allons faire » (25 janvier 2018, 10 mots caractéristiques sur 12).

Ces phrases proviennent toutes de transcriptions mises en ligne sur le site de l'Elysée et correspondent à ce qui a été prononcé. Elles illustrent le style très particulier d'E. Macron.

Style

En parlant, tout orateur effectue des choix fondamentaux dont le principal concerne la relation qu'il établit avec son auditoire et avec ce qu'il dit. Dans leur étude sur le général de Gaulle, J.-M. Cotteret et R. Moreau (1969) avaient ainsi mis en lumière les deux registres principaux du discours politique : la polémique (le discours de combat qui fait surtout appel à l'affect) et la pédagogie (le discours programmatique s'adressant à la raison). Toute

communication politique recourt, en proportion variable, à ces deux registres (Arnold, Labbé 2015).

Le polémiste interpelle, donc il personnalise : son discours est dominé par la tension entre "je" et "vous" ; il utilise beaucoup de noms de personne ; il oriente son propos vers l'action donc mobilise une proportion importante de verbes, spécialement ceux de la volonté et de la nécessité. Il fait des phrases assez courtes et recourt volontiers au slogan. Les discours sont en général assez brefs.

Le pédagogue expose et explique. Il utilise plus de noms, moins de verbes dans des phrases plus longues. Il dépersonnalise le propos avec peu de pronoms personnels et en fusionnant le *je* (l'orateur) et le *vous* (l'auditoire) dans le *nous*. Dans un discours à dominante pédagogique, il y aura moins de noms de personnes, de lieux, de dates et de chiffres (plus leurs densités sont faibles, plus le propos est abstrait). L'orateur montre à son auditoire (pronoms et articles démonstratifs). Enfin, le pédagogue fait des discours plus longs que le polémiste.

Toutes les caractéristiques du discours pédagogique dominent dans les propos d'E. Macron. La longueur est particulièrement remarquable. Non seulement, ce président parle longuement mais ses phrases sont anormalement longues : moyenne 33 mots. C'est plus que le général de Gaulle et G. Pompidou (moyenne 30 mots), mais surtout, E. Macron rompt avec un mouvement séculaire de simplification de la phrase qui avait amené cette longueur à 24 mots chez F. Hollande et 22 chez N. Sarkozy (Arnold 2019). La moitié du temps, l'auditeur d'E. Macron est confronté à des phrases de plus de 42 mots et, pendant un dixième du temps, à des phrases de plus de 90 mots. Le décryptage de tels monstres est déjà difficile à l'écrit, il est évidemment impossible à la simple écoute.

Cela ne signifie pas qu'E. Macron a renoncé aux discours de combat mais que, en tant que président, il répugne à cette posture et préfère nettement celle du pédagogue. Cependant la leçon est en partie perdue puisque, pour un auditeur, au moins la moitié du temps, le propos est difficile, voire impossible à saisir…

Le deuxième choix que doit effectuer un locuteur concerne la construction de la phrase. En effet, dès que la phrase s'allonge, elle devient complexe et l'orateur se trouve devant une alternative. Il peut empiler les propositions les unes après les autres en les coordonnant – soit avec des virgules, soit avec des conjonctions de coordination - ou les imbriquer les unes dans les autres en les subordonnant (essentiellement à l'aide de conjonctions de subordination ou de pronoms relatifs) mais cela suppose que les idées soient hiérarchisées. E. Macron pratique rarement l'imbrication et préfère l'empilement de propositions mises sur le même plan et souvent peu charpentées. Cela se traduit par un excédent de coordinations, spécialement "et" (+19%), et par un déficit en conjonctions de subordination : -17% (en premier lieu : *que* et *si*). Autrement dit, le propos est faiblement problématisé et ressemble un peu à un convoi de marchandises où l'ordre des wagons obéit à une logique mystérieuse.

Ajouté à la longueur et à l'abstraction, ce choix stylistique rend le propos parfois obscur, du moins à l'audition (les deux phrases longues dans l'encadré ci-dessous en fournissent de bons exemples : il y en a comme cela des centaines chez E. Macron).

Conclusions

Le discours d'E. Macron, durant les 20 premiers mois de son mandat (mai 2017-décembre 2018), s'inscrit en assez profonde rupture par rapport à ses prédécesseurs. Il a privilégié la politique internationale – en premier lieu : l'*Europe*, l'*Afrique* et la *Chine* - et il a nettement

moins parlé de la *France* et des *Français*. Il a semblé négliger la plupart des thèmes classiques comme la modernité, l'économie, l'emploi, le progrès, la croissance, les revenus, le social, les impôts… Il assume peu ses propos et privilégie le *nous*. Son discours se veut pédagogique, mais il est souvent lourd et abstrait. Les phrases sont longues et compliquées, parfois obscures.

L'analyse de la communication d'E. Macron présente une difficulté supplémentaire. Contrairement à ses prédécesseurs, le site actuel de l'Elyséene met en ligne des transcriptions que pour environ la moitié des interventions signalées (mais une partie de ses interventions n'est même pas mentionnée sur l'agenda)[13]. Pour une partie des autres, seule la vidéo est mise en ligne (on y voit parfois le chef de l'Etat se référer à son papier ou regarder le prompteur, ce qui prouve qu'il y a un texte). Il parle souvent devant des micros de presse, mais on n'entend jamais les questions des journalistes ni les réponses du président. Quand il tient une conférence de presse conjointe avec un autre chef d'Etat (ou de gouvernement), sur le site de l'Elysée, on n'entend qu'E. Macron, jamais son homologue, même quand celui-ci parle français[14]. D'autres fois encore, il n'y a que des extraits dans les dépêches d'agence ou quelques secondes sur tweeter. Quant aux transcriptions que l'on trouve sur certains sites (notamment les ministères concernés ou le site viepublique.fr), elles n'ont évidemment pas l'aval de l'Elysée même si la présidence en est souvent la source.

Le contenu du discours, fortement décalé par rapport à ses prédécesseurs, la lourdeur du style et l'abstraction du propos, le flou et la superficialité de la communication présidentielle – où l'image et le symbole sont toujours privilégiés sur le contenu –, tout cela explique probablement en partie l'incompréhension de beaucoup de Français envers leur président, telle qu'elle ressortait des enquêtes d'opinion à la fin 2018.

Naturellement, ces premières conclusions seront sujettes à révision aussi longtemps qu'E. Macron restera à la tête de la république française, du moins si l'on peut continuer à disposer d'un nombre suffisant de transcription de ses interventions.

Au-delà du cas Macron, nous espérons avoir montré combien la statistique appliquée au langage est un outil intéressant pour extraire de l'information dans une masse de textes équivalant ici à 150 forts volumes dont aucune technique "manuelle" ne peut venir à bout, spécialement pour identifier les principaux réseaux sémantiques, les thèmes privilégiés, les choix stylistiques et de communication.

Grâce aux probabilités – et particulièrement, à Bernoulli -, l'on peut, à partir d'un nombre limité d'observations, bien organisées, inférer des conclusions générales concernant un phénomène. Ces analyses ont évidemment un prix. Les formules mathématiques sont traduites en algorithmes informatiques soigneusement testés. Les observations suivent un protocole rigoureux, notamment quant à la collecte des données (correction orthographique, standardisation des graphies, lemmatisation). Enfin la démarche est entièrement reproductible. Cela implique notamment un accès libre et complet aux données. De cette dernière condition dépend le progrès des connaissances scientifiques, notamment dans les sciences sociales et humaines (Camerer 2018).

[13] Par exemple, il manque trois entretiens télévisés : 17 décembre 2017 (avec L. Delahousse, France 2), 16 avril 2018 (avec E. Plenel et J.-J. Bourdin sur BFM) et 12 octobre 2018 (France 24, sommet d'Erevan sur la francophonie).

[14] Par exemple, la conférence de presse conjointe avec Justin Trudeau, Premier ministre du Canada, le 16 avril 2018. Il n'y a que la déclaration liminaire d'E. Macron. Les très nombreux chefs d'Etat africains que rencontre E. Macron ont droit au même effacement.

Des travaux ultérieurs, utilisant les mêmes méthodes appliquées à ces corpus présidentiels, permettront d'affiner les portraits lexicaux et stylistiques du général de Gaulle et de ses successeurs. En plaçant nos fichiers dans le domaine public, nous espérons susciter d'autres recherches en ce domaine.

Remerciements

Paul Jolissaint, animateur du séminaire "Mathématique et société" a organisé cette séance.

Les corpus des interventions des présidents français (1958-2018) sont en ligne sur le site du Centre de Linguistique de Corpus (Université de Neuchâtel). La plupart des interventions des présidents français depuis V. Giscard d'Estaing sont sur le site http://www.vie-publique.fr/. Les logiciels sont disponibles sur demande auprès de Cyril et Dominique Labbé.

Toutes nos recherches ont été réalisées sans aide publique.

Bibliographie

Arnold Edward (2019). Le vocabulaire et le style du général de Gaulle. A paraître dans *Document numérique* (Lavoisier). Numéro spécial : *Les corpus politiques*.

Arnold Edward, Labbé Cyril, Monière Denis (2016). *Parler pour gouverner : Trois études sur le discours présidentiel français*. Grenoble : Laboratoire d'Informatique de Grenoble.

Arnold Edward, Labbé Dominique (2015). Vote for me. Don't vote for the other one. *Journal of World Languages*. Routledge, p. 1-18.

Benveniste Emile (1956). La nature des pronoms. Reproduit dans Benveniste 1966, p.251-257.

Benveniste Emile (1958). De la subjectivité dans le langage. Reproduit dans Benveniste 1966, p.258-265.

Benveniste Emile (1959). Les relations de temps dans le verbe français. Reproduit dans Benveniste 1966, p. 237-250.

Benveniste, Émile. (1965). Structure des relations d'auxiliarité. Reproduit dans Benveniste 1970, p. 177-193.

Benveniste Emile (1966 & 1970). *Problèmes de linguistique générale*. Paris, Gallimard (rééd. 1980).

Bernoulli Jacques (1713). *Ars conjectandi*, Bâle : Thurneysen Frères.

Camerer Colin F. et Al. (2018). Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior*. August 2018.

Cotteret Jean-Marie, Moreau René (1969). *Le vocabulaire du général de Gaulle*. Paris, Presses de la Fondation des sciences politiques.

Desrosières Alain (1988). *La partie pour le tout : comment généraliser ? Cinq contributions à l'histoire de la statistique*. Paris : Economica.

Dodge Yadolah (1993). *Statistique. Dictionnaire encyclopédique*. Paris : Dunod.

Dubois Jean (1969). "Enoncé et énonciation". *Langages*. 13, p 100-110.

Gross Maurice (1999). Sur la définition d'auxiliaire du verbe. *Langages*, 135, p 8-21.

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

Labbé Cyril, Labbé Dominique (2010). Ce que disent leurs phrases. In Bolasco S., Chiari I., Giuliano L. (Eds). *Proceedings of 10th International Conference Statistical Analysis of Textual Data*. Rome : Edizioni Universitarie di Lettere Economia Diritto. Vol 1, p. 297-307.

Labbé Cyril & Labbé Dominique (2013). La modalité verbale en français contemporain. Les hommes politiques et les autres. Banks David (ed). *La modalité, le mode et le texte spécialisé*. Paris : L'Harmattan, p. 33-61.

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

Labbé Dominique (1997). Le *nous* du général de Gaulle. *Quaderni di studi linguistici*, 4/5, 1998, p 331-354.

Lafon Pierre (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, 1, p. 127-165.

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

Lebart Ludovic, Salem André (1992). *Statistique textuelle*. Paris : Dunod.

Mandelbrot Benoît (1957). Étude de la loi d'Estoup et de Zipf Fréquences des mots dans le discours. Apostel L et al. *Logique, langage et théorie de l'information*. Paris, PUF, p. 22-53.

Meusnier Norbert (1987). *Jacques Bernoulli et l'ars conjectandi. Documents pour l'étude de l'Emergence d'une Mathématisation de la Stochastique*. Mont-Saint-Aignan : Institut de Recherche sur l'Enseignement des Mathématiques.

Meusnier Norbert (1992). *Christian Huygens et Jacques Bernoulli : la première partie de l'Ars Conjectandi*. Paris : Centre d'Analyse et de Mathématique Sociale.

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

Monière Denis, Labbé Cyril et Labbé Dominique (2008). Les styles discursifs des premiers ministres québécois de Jean Lesage à Jean Charest. *Canadian Journal of Political Science / Revue canadienne de science politique*. 41:1, p. 43-69.

Monière Denis, Labbé Dominique (2012). Le vocabulaire caractéristique du Premier ministre du Québec J. Charest comparé à ses prédécesseurs. Dister A. et al. (éds). *Proceedings of the 11th International Conference on Textual Data Statistical Analysis*. Liège : LASLA - SESLA, p.737-751.

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

Muller Charles (1964). Calcul des probabilités et calcul d'un vocabulaire. Reproduit dans : *Langue française et linguistique quantitative*. Genève-Paris: Slatkine-Champion, 1979, p. 167-176.

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

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

Shafer Glenn (1996). The Significance of Jacob Bernoulli's *Ars Conjectandi* for the Philosophy of Probability Today. *Journal of Econometrics*, 75-1, p. 15-32

Zipf G. K. (1935). *La psychobiologie du langage*. Paris : CEPL, 1974.

Annexe 1

La bibliothèque électronique du français contemporain et la section "présidents de la République française" (mars 2019)

Corpus	Dates	N textes	Mots
Discours politique *	XIX –XXIe siècle	14 409	28 632 865
Littérature	XVII-XXe siècle	1 142	22 522 956
Presse	XXe siècle	1 513	2 039 631
Divers écrits	XXe siècle	182	1 323 194
Français oral	XXe siècle	438	3 637 009
Total		17 684	58 155 655

Les présidents français (1958-2018)*

Président	Dates	N textes	Mots	Vocables
De Gaulle	1958-1969	459	410 492	9 002
Pompidou	1969-1974	137	259 918	8 076
Giscard d'Estaing	1974-1981	191	660 560	9 535
Mitterrand	1981-1995	2 547	5 576 244	25 695
Chirac	1995-2007	2 478	4 081 697	24 054
Sarkozy	2007-2012	1 074	3 221 250	21 597
Hollande	2012-2017	1 544	3 182 939	20 403
Macron	2017-2018	318	1 037 988	14 819
Total	1958 - 2018	8 748	18 431 088	45 575

* En ligne sur le site du Centre de linguistique de corpus de l'Université de Neuchâtel.

Annexe 2. Densités des catégories grammaticales dans les interventions d'E. Macron comparées aux autres présidents (‰)

Catégories	A Présidents (‰)	B Macron (‰)	(B-A)/A (%)	Indice $\alpha = 0{,}001$
Verbes	155,3	152,5	-1,8	-
Futurs	6,3	7,5	+19,4	+
Conditionnels	3,1	1,6	-49,2	-
Présents	78,4	77,0	-1,8	-
Imparfaits	7,2	4,5	-37,2	-
Passés simple	0,6	0,5	-7,9	≈
Participes passés	22,8	20,0	-12,3	-
Participes présents	2,2	2,8	+28,9	+
Infinitifs	34,8	38,6	+11,0	+
Noms propres	24,6	20,5	-16,7	-
Noms communs	179,7	184,9	+2,9	+
Adjectifs	56,7	58,9	+4,0	+
Adj. participe passé	5,9	6,6	+12,1	+
Pronoms	122,6	121,1	-1,2	-
Pronoms personnels	62,8	58,4	-7,0	-
Déterminants	182,3	179,3	-1,6	-
Articles	129,9	123,6	-4,8	-
Nombres	17,8	13,6	-23,7	-
Possessifs	15,5	16,8	+8,9	+
Démonstratifs	9,3	15,3	+63,4	+
Indéfinis	9,8	10,0	+1,7	≈
Adverbes	68,3	63,6	-6,8	-
Prépositions	152,9	158,5	+3,7	+
Coordinations	32,5	40,6	+24,9	+
Subordination	23,0	19,0	-17,3	-
Mots étrangers	0,1	0,5	+499,2	+

La lecture du tableau est horizontale. Par exemple, chez l'ensemble des présidents, il y a en moyenne 155,3 verbes pour mille mots. Chez E. Macron, cette densité est de 152,5, soit 1,8% de moins que dans l'ensemble de référence. Cette différence négative est statistiquement significative avec un risque d'erreur inférieur à 1 pour mille (indice égal à 0,000). Dans cette catégorie du verbe, E. Macron préfère le futur, le participe présent et l'infinitif et évite surtout le conditionnel et les passés (imparfait et participe). Pour le passé simple et les adjectifs indéfinis, on ne peut écarter l'hypothèse selon laquelle la densité chez E. Macron ne s'écarte pas de l'usage commun (avec un risque d'erreur de 1 pour mille).

Annexe 3 Les trente verbes préférés d'E. Macron avec leur rang chez les présidents et leur indice de caractéristique

Rang Macron	Rang Présidents	Vocable	Effectif (F_{im})	Fréquence (‰)	Ecart % (F_{im}/F_{ip})	α = 0,001
1	1	être	32 460	31,27	-1,9	≈
2	2	avoir	21 274	20,50	-7,5	-
3	3	faire	5 455	5,26	2,5	+
4	4	pouvoir	4 601	4,43	-8,8	-
5	6	devoir	3 623	3,49	10,8	+
6	7	vouloir	3 188	3,07	10,4	+
7	5	dire	2 678	2,58	-28,9	-
8	9	aller	2 405	2,32	11,5	+
9	13	permettre	2 128	2,05	73,7	+
10	10	savoir	1 937	1,87	-0,5	≈
11	22	souhaiter	1 599	1,54	95,4	+
12	12	venir	1 585	1,53	21,4	+
13	8	falloir	1 569	1,51	-31,7	-
14	11	prendre	1 353	1,30	2,4	+
15	19	croire	1 164	1,12	23,1	+
16	31	porter	1 120	1,08	120,4	+
17	14	voir	1 008	0,97	-11,8	-
18	24	vivre	970	0,93	32,9	+
19	15	penser	964	0,93	-13,1	-
20	18	mettre	939	0,90	-3,2	≈
21	74	construire	919	0,89	268,3	+
22	16	parler	894	0,86	-14,9	-
23	17	donner	865	0,83	-12,0	-
24	46	évoquer	856	0,82	117,4	+
25	42	continuer	843	0,81	103,1	+
26	49	conduire	798	0,77	129,0	+
27	25	tenir	769	0,74	12,6	+
28	86	accompagner	722	0,70	224,1	+
29	29	engager	712	0,69	21,3	+
30	23	agir	674	0,65	-17,2	-

Lecture : le verbe "être" occupe le premier rang des verbes chez E. Macron comme chez les autres présidents. Chez E. Macron, on en rencontre 31,27 pour mille mots, soit -1,9% de moins que chez les autres. Ce vocable ne peut être considéré comme caractéristique d'E. Macron au seuil d'erreur de 1 pour mille.

Annexe 4. Les trente vocables caractéristiques les plus sur-employés (C+) par E. Macron comparé à ses prédécesseurs*

Rang	Vocable	Effectif total (F_{ip})	Effectif Macron (F_{im})	E_{iu}	F_{im}/F_{ip} (%)
1	ce (pro.)	242 736	14 415	13 512	5,9
2	qui (pro.)	243 992	14 969	13 581	6,1
3	mener (v.)	3 692	440	204	11,9
4	refonder (v.)	197	63	11	32,0
5	chancelière (n. f.)	672	117	39	17,4
6	et (conj.)	431 160	30 978	25 983	7,2
7	soie (n. f.)	84	56	5	66,7
8	évidemment (adv.)	1 941	366	102	18,9
9	souhaiter (v.)	14 521	1 599	804	11,0
10	certain (pro.)	3 812	383	212	10,0
11	accès (n. m.)	2 381	309	138	13,0
12	francophonie (n. f.)	1 662	216	96	13,0
13	regarder (v.)	3 848	399	213	10,4
14	remercier (v.)	7 606	655	421	8,6
15	axe (n. m.)	610	103	35	16,9
16	structurant (adj.)	138	81	8	58,7
17	préfet (n. m.)	752	143	44	19,0
18	partager (v.)	3 887	402	215	10,3
19	pilier (n. m.)	411	99	24	24,0
20	durant (pré.)	952	341	55	35,8
21	déstabilisation (n. f.)	156	42	9	26,9
22	start up (n. f.)	225	60	13	26,7
23	revenir (v.)	5425	474	300	8,7
24	méthode (n. f.)	1694	189	98	11,2
25	radicalisation (n. f.)	118	53	7	44,9
26	peur (n. f.)	1700	201	98	11,9
27	présent (adj.)	3951	393	231	10,0
28	accompagner (v.)	3952	722	219	18,2
29	zone (n. f.)	3957	361	229	9,1
30	langue (n. f.)	3967	509	229	12,8

* Sauf les noms propres (voir tableau 3). Pour tous ces vocables, l'indice est égal à 1 (moins de une chance sur un milliard de se tromper en affirmant qu'il y en a "trop" chez E. Macron.

** Lecture horizontale : le pronom "ce" est le vocable le plus caractéristique d'E. Macron comparé à ses prédécesseurs. Dans tout le discours présidentiel, il est employé 242 736 fois. E. Macron l'utilise 14 415 fois alors que l'on en attend 13 512. Dernière colonne : E. Macron pèse 5,9 % de l'effectif total (alors que l'hypothèse d'équi-répartition en laisse attendre seulement 5,3%).

Annexe 5. Les trente vocables caractéristiques les plus sous-employés (C-) par E. Macron comparé à ses prédécesseurs*

Rang	Vocable	Effectif total (F_{ip})	Effectif Macron ($_{Fim}$)	E_{iu}	F_{im}/F_{ip} (%)
45547	en (pro.)	43 170	1 662	2 403	3,9
45548	on (pro.)	88 906	3 877	4 949	4,4
45549	pays (n. m.)	59 139	2 507	3 421	4,2
45550	mon (dét.)	33 756	1 202	1 872	3,6
45551	son (dét.)	78 760	3 322	4 367	4,2
45552	croissance (n. f.)	6 786	110	393	1,6
45553	puisque (conj.)	8 651	185	521	2,1
45554	se (pro.)	118 878	5 298	6 617	4,5
45555	même (adv.)	21 032	582	1 108	2,8
45556	notamment (adv.)	12 775	266	673	2,1
45557	dire (v.)	66 878	2 678	3 703	4,0
45558	de (pré.)	1 451 646	80 928	84 572	5,5
45559	poser (v.)	7 791	111	431	1,4
45560	progrès (n. m.)	7 169	99	415	1,4
45561	ne (adv.)	198 570	8 757	10 460	4,4
45562	quatre (dét.)	19 065	493	1 057	2,6
45563	cela (pro.)	45 881	1 645	2 554	3,6
45564	façon (n. f.)	9 163	140	530	1,5
45565	si (conj.)	39 940	1 492	2 407	3,7
45566	je (pro.)	308 252	14 641	17 159	4,8
45567	question (n. f.)	17 104	371	989	2,2
45568	bien (adv.)	43 828	1 312	2 309	3,0
45569	cent (dét.)	35 313	1 032	1 958	2,9
45570	neuf (dét.)	20 811	455	1 154	2,2
45571	pas (adv.)	156 985	6 266	8 270	4,0
45572	naturellement (adv.)	8 603	11	453	0,1
45573	il (pro.)	234 088	10 008	13 031	4,3
45574	problème (n. m.)	18 493	238	1 070	1,3
45575	que (conj.)	268 604	11 517	16 187	4,3
45576	le (dét.)	2 046 164	109 557	113 454	5,3

* Sauf les noms propres (voir tableau 3). Pour tous ces vocables, l'indice est égal à 0 (moins de une chance sur un milliard de se tromper en affirmant qu'il y en a "pas assez" chez E. Macron.

** Lecture horizontale : l'article "le" est le vocable le plus sous-employé chez E. Macron comparé à ses prédécesseurs. Dans tout le discours présidentiel, il est employé 2 046 164 fois. E. Macron l'utilise 109 557 fois alors que l'on en attend 113 454. Dernière colonne : E. Macron pèse 5,3 % de l'effectif total de "le" (alors que l'hypothèse d'équi-répartition en laisse attendre 5,6%).

Annexe 6. Les trente vocables communs ($C_{\approx}$) les plus employés par E. Macron et ses prédécesseurs (avec $\alpha = 0,05$)

Rang	Vocable	Catégorie grammaticale	Présidents (F_{ip})	Macron (F_{im})	F_{theo}
511	choisir	v.	3 945	221	218
494	entier	adj.	3 899	229	228
598	suivre	v.	3 324	183	184
560	agricole	adj.	3 258	195	191
648	essayer	v.	3 078	163	170
611	réussite	n. f.	3 021	177	175
600	alliance	n. f.	3 005	181	174
623	conviction	n. f.	2 998	172	173
669	attacher	v.	2 976	157	164
679	passé	n. m.	2 769	156	160
657	éviter	v.	2 758	160	153
643	fond	n. m.	2 739	165	158
717	courage	n. m.	2 619	142	151
704	vite	adv.	2 617	146	138
685	contrat	n. m.	2 606	152	151
729	régler	v.	2 605	140	144
730	convaincre	v.	2 594	139	143
690	perspective	n. f.	2 590	151	150
699	troisième	dét.	2 568	149	142
678	ouverture	n. f.	2 552	156	148
726	historique	adj.	2 482	140	145
698	immense	adj.	2 437	149	143
780	offrir	v.	2 407	128	133
778	maintenir	v.	2 364	128	131
798	utiliser	v.	2 296	125	127
722	victime	n. f.	2 287	141	132
826	vôtre	pro.	2 180	119	121
791	déclaration	n. f.	2 166	126	125
771	logement	n. m.	2 143	129	124
847	fier	adj.	2 093	113	122