Objectiver l'intertexte ? Emmanuel Macron, *deep learning* et statistique textuelle

Damon Mayaffre[1] et Laurent Vanni[2]

[1] Université Côte d'Azur, CNRS, BCL, France - damon.mayaffre@univ-cotedazur.fr

[2] Université Côte d'Azur, CNRS, BCL, France - laurent.vanni@univ.cotedazur.fr

Abstract

The present paper suggests that intertextuality can be brought out objectively by resorting to specific methodological tools. The case in point is political intertextuality in the speeches of the French president Emmanuel Macron.

Deep learning (convolutional model) is first used to "learn" (satisfactory accuracy rate of 92.3%) the French presidential speeches since 1958: the speeches of De Gaulle, Pompidou, Giscard, Mitterrand, Chirac, Sarkozy and Hollande are then considered as the potential intertext of Macron's own speeches.

Next, Macron's texts - hitherto unknown to the machine - are included in the model and the machine is instructed to assign Macron's quotations to one of his predecessors based on their linguistic content.

Finally, the algorithm extracts and describes Macron's quotations and linguistic units (wTDS, lexical specificities, co-occurrences, morpho-syntactic labels) as they were interpreted by the machine in comparison to those of De Gaulle or Sarkozy, of Mitterrand or Holland.

Macron's discourse is permeated with, sometimes explicitly but more often than not implicitly, by the discourse of former French presidents - a phenomenon that we shall refer to as "intertextuality" - and it turns out that Artificial Intelligence and textual statistics are able to identify such phenomena of borrowing, imitation and even plagiarism.

Keywords: text mining, deep learning, text, intertextuality, political discourse, Macron

Résumé

Cette contribution propose un parcours méthodologique susceptible d'objectiver l'intertexte ; l'intertexte politique des discours du président français Emmanuel Macron en l'occurrence.

Le *deep learning* (modèle convolutionnel) est d'abord utilisé pour « apprendre » (taux d'accuracy satisfaisant de 92,3%) le discours présidentiel français depuis 1958 : les discours de de Gaulle, Pompidou, Giscard, Mitterrand, Chirac, Sarkozy et Hollande sont alors considérés comme l'intertexte potentiel des discours de Macron.

Ensuite, les textes de Macron – inconnus jusqu'ici du système – sont versés dans le modèle et nous forçons la machine à attribuer les passages de Macron à l'un de ses prédécesseurs en fonction de leur composition linguistique.

Enfin, l'algorithme extrait et décrit les passages et les unités linguistiques (wTDS, spécificités lexicales, cooccurrences, étiquettes morpho-syntaxiques) de Macron interprétées par la machine comme ressemblant à celles de de Gaulle ou Sarkozy, à celles de Mitterrand ou de Hollande.

Le discours de Macron est traversé, de manière explicite parfois, de manière implicite le plus souvent, par les discours de ses prédécesseurs – phénomène que l'on appellera « intertextualité » – et l'Intelligence artificielle et la statistique textuelle peuvent repérer les phénomènes d'emprunt, d'imitation voire de plagiat.

Mots clés : intertexte, intertextualité, statistique textuelle, Macron, logométrie, *deep learning*, convolution, déconvolution, discours politique

1. Introduction

Malgré des performances pluri-décennales spectaculaires, l'ADT (Lebart, Pincemin, Poudat 2019) bute encore sur quelques tâches importantes de l'analyse linguistique des textes : les figures du discours par exemple dont l'esthétique reste pour l'essentiel inaccessible au traitement statistique, ou encore, à un tout autre niveau, l'intertextualité[1].

Il est vrai que l'intertexte et l'intertextualité sont des concepts linguistiques majeurs mais fuyants depuis Kristeva, Barthes, Riffaterre ou Genette sinon depuis Bakhtine[2]. Ils apparaissent rétifs à toute forme d'objectivation, de formalisation, de mesure, c'est-à-dire, pour nous, à toute forme d'implémentation informatique définitive. Si tous les linguistes conviennent que l'intertextualité est une condition de l'interprétation d'un texte-cible, aucun ne prétend pouvoir clairement l'expliciter et la circonscrire au-delà de la convocation discrétionnaire de telle référence supposément éclairante ici, de tel discours prétendument inspirateur là, de telle reprise textuelle soupçonnée ailleurs. Sur les bases d'une définition aussi fragile[3], l'analyse statistique des données textuelles semble condamnée à l'impuissance méthodologique.

Nous considérons ici les *corpus réflexifs* numériques (Mayaffre 2002, Rastier 2004)[4] comme la matérialisation d'un certain intertexte du texte-cible et proposons un protocole méthodologique pour objectiver cet intertexte désormais matérialisé. Dans cette contribution, le texte-cible est le discours de Macron depuis son élection en 2017, et l'intertexte pressenti de ce texte-cible est l'ensemble du discours élyséen depuis 1958 (de Gaulle, Pompidou, Giscard, Mitterrand, Chirac, Sarkozy et Hollande) dans lequel le nouveau président peut puiser des références conscientes ou inconscientes, explicites ou implicites. On notera que cet intertexte pressenti – un intertexte parmi d'autres possibles – se justifie par des raisons génériques (discours officiel présidentiel), énonciatives (locuteurs dans une même situation d'énonciation), historiques (l'unité chronologique que constitue la V^{ème} République) ou encore politiques (prétention centriste de Macron à être « en même temps » gaulliste et mitterrandien, sarkozyste et hollandais, giscardien et chiraquien).

Le *deep learning* – ici un modèle convolutionnel (CNN) implémenté dans le logiciel Hyperbase (cf. d'abord l'étude très citée de Kim 2014, puis Ducoffe et al. 2016, Montavon et al. 2018, Vanni et al. 2018a, Vanni et al. 2020-*submitted*) – est d'abord convoqué pour extraire du texte-cible les passages empruntés aux corpus de référence (ou corpus intertextuel). Puis l'ADT permet d'affermir les résultats de l'IA, pour mesurer les spécificités,

[1] Ce travail 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

[2] Au-dela des auteurs majeurs et fondateurs évoqués ici, cette contribution s'appuie spécifiquement, du point de vue conceptuel, sur un numéro de revue (*Cahiers de praxématique*, 33, 1999) et un ouvrage collectif (Bres et al. (dir), 2005). Pour une discussion théorique sur la notion d'intertexte cf. notre contribution récente (Mayaffre et al. 2020).

[3] On se rappelle par exemple de la définition fondatrice mais évanescente de Barthes : « L'intertexte est un champ général de formules anonymes, dont l'origine est rarement repérable, de citations inconscientes ou automatiques, données sans guillemets ». (*Encyclopaedia universalis*, ed. 1995, Tome 22, p. 372).

[4] Rappelons : un *corpus réflexif* est un corpus qui contient ses propres ressources interprétatives, c'est-à-dire dans lequel chaque texte constitue le contexte interprétatif de tous les autres, et l'ensemble des textes constitue le contexte interprétatif de chacun ; dans le but d'une interprétation « endogène » mieux contrôlée.

les segments répétés, les cooccurrences ou les motifs de Macron qui portent, de fait, statistiquement, l'empreinte de ses prédécesseurs.

2. Cadrage méthodologique : ADT et *deep learning*

2.1. Deux approches complémentaires

Le dialogue que nous avons engagé entre ADT et IA (*deep learning*, ici avec un modèle CNN) permet d'espérer une complémentarité réciproquement éclairante entre la statistique textuelle et les réseaux de neurones artificiels (Vanni et al. 2018b ; Brunet et al. 2019 ; Mayaffre et al. (dir.) 2020-sous presse). A terme, il vise à « intégrer » ces méthodes.

L'ADT procède par tokenisation (i), et se trouve fondamentalement sensible au fréquentiel (ii). Le *deep learning* (modèle CNN), lui, procède par contextualisation (i-bis), et se trouve sensible au séquentiel (ii-bis).

Quels que soient les efforts consentis ultérieurement (par exemple dans l'étude des cooccurrences ou des segments répétés, ou simplement par l'usage d'un concordancier), l'ADT commence en effet par la tokenisation, c'est-à-dire par la segmentation ou atomisation du texte en unités discrètes et indépendantes (i), et se prolonge par le dénombrement puis la distribution quantitative de ces *tokens* dans le corpus (ii). Le schéma d'urne par lequel on procède au tirage de boules indépendantes ou jetons (*tokens*) reste fondateur de la discipline. Les tirages avec ou sans remise apparaissent comme une déclinaison experte de ces principes fondamentaux notés ici par (i) et (ii).

L'IA a certes également besoin de *tokens*. Mais le modèle convolutionnel les considère aussitôt dans leur cotexte immédiat (i-bis). Pour le modèle, le mot n'est jamais pris seul, isolément. Le mot n'a pas une représentation pour lui-même. Il a la représentation de sa fenêtre cotextuelle. La conséquence la plus spectaculaire de cette remise en cotexte est que les différentes occurrences d'un mot ne sont pas considérées comme identiques entre elles, comme c'est le cas en ADT ; elles varient selon leur cotexte. L'ensemble du texte est balayé par une fenêtre coulissante de 3, 5 ou 10 mots, dans laquelle l'unité est prise en considération. Ainsi dans la phrase "je pense donc je suis" le mot "donc" sera envisagé selon la figure 1.

Figure 1. Représentations de « donc » en ADT et en IA (modèle convolutionnel, fenêtre de 3)[5]

[5] Le schéma figure la représentation différente (ici la couleur) du mot « donc » dans la tokenisation propre à l'ADT et dans la convolution propre au *deep learning*. En creusant le schéma nous aurions pu également remarquer que les deux « je » de la phrase ont la même représentation (la même couleur) dans le cas de la tokenisation, alors qu'ils n'ont pas la même représentation (deux couleurs différentes) en cas de convolution.

Avec ce balayage contextualisant, l'approche cesse ainsi d'être discrète et devient continue. L'urne devient réseau ou suite. Et dans ces conditions, le fréquentiel (ii) cède le pas au séquentiel (ii-bis) car le décompte, la distribution ou la probabilité de voir apparaitre plusieurs fois des séquences de 3, 5, 10 mots dans le corpus sont dérisoires (ie. sur une séquence longue, nous avons presque toujours à faire à des hapax).

Dit autrement enfin, au plus haut niveau linguistique cette fois-ci, l'ADT est d'essence paradigmatique (tokenisation, puis *sélection* des *tokens* en fonction du dictionnaire des fréquences, en fonction de l'étiquette morpho-syntaxique, en fonction de la lettre à l'initiale, etc.). L'IA, elle, est d'essence syntagmatique (cotextualisation, c'est à dire *combinaison* ou séquentialisation ou convolution du *token* pris dans la linéarité du texte).

Ainsi, la complémentarité de l'IA et de l'ADT est grosse d'espoir pour l'analyste des corpus textuels. Jusqu'ici séparées (l'IA appartenait pour l'instant surtout au TAL), les deux visions pourraient être combinées dans une démarche intégrative prometteuse.

2.2. Protocole

L'efficacité de l'apprentissage (*learning*) puis de la classification (*prediction*) des textes par le *machine learning* ou le *deep learning* ne sont plus en débat aujourd'hui. Depuis plusieurs années désormais, elles se chiffrent par des taux d'exactitude (*accuracy rate*) spectaculaires. En un mot, il n'y a pas plus de doute que la machine apprenne et reconnaisse un discours (une série de mots) de Macron *versus* un discours de de Gaulle, qu'il n'y a de doute que la machine apprennent et reconnaisse une image de chat (une série de pixels) *versus* une image de chien [cf. par exemple Ducoffe et al. 2016 ou Brunet et al. 2019].

2.2.1. Sur un jeu de données de 1000 discours présidentiels sous la Vème République, avant la présidence Macron, équivalant 3 millions d'occurrences, l'algorithme, implémenté dans le logiciel Hyperbase, apprend (*learning*) à reconnaître les discours de de Gaulle (phrases longues, plutôt nominales ou adjectivales, avec par exemple la « France » ou « l'État » comme premiers noms) comme ceux de Mitterrand (phrases courtes, plutôt verbales, avec le « je » et le « moi » comme centre d'intérêt personnel et « l'Europe » comme horizon). Il apprend à reconnaître les discours de Sarkozy (un lexique fort et une syntaxe simple) et ceux de Hollande (une syntaxe compliquée et un lexique affadi). Il apprend à reconnaître les textes de Giscard (phraséologie technocratique ou didactique), ceux de Chirac (idéologiquement pleins de vide) ou ceux de Pompidou (style littéraire, riche voire ampoulé)[6] ; et le corpus de validation (corpus-test composé de discours présidentiels inconnus et anonymisés) permet de chiffrer que le programme retrouve automatiquement au-dessus de 92,3 % des fois le bon auteur-président des discours. Sur un jeu de textes littéraires composé de l'essentiel de la littérature française aux 18ème, 19ème et 20ème siècles, des résultats supérieurs, proches de 100% ont même été établis par Etienne Brunet, Laurent Vanni et Ludovic Lebart dans une étude monumentale qui brasse 50 auteurs de Racine à Giono, de Hugo à Yourcenar, de Proust à Le Clézio [Brunet et al. 2019 et Brunet, Lebart et Vanni 2020-sous presse].

Ainsi, en apprenant à la machine le discours élyséen de de Gaulle à Hollande, entre 1958 et 2017, nous avons construit un certain « horizon d'attente » (selon la terminologie de la sémiologie des décennies précédentes), ou au contraire un certain point de départ des discours de Macron : selon nous, Hyperbase a appris un certain *intertexte* dans lequel Macron pourra

[6] Pour une analyse des discours et du style des présidents sous la Vème République voir nos ouvrages Mayaffre 2012-a et Mayaffre 2020-sous presse.

emprunter pour construire ses propres discours ou duquel Macron pourra s'inspirer pour parler ; ou encore, un certain *intertexte* dans lequel l'auditeur puis l'analyste pourront faire résonner les discours de Macron pour les comprendre et les interpréter.

2.2.2. Puis, nous versons les discours de Macron, inconnus du système, dans ce corpus élyséen de référence (ou intertexte élyséen), en demandant à l'algorithme de rapprocher (classer) chaque paragraphe (ie. des fenêtres de 50 mots) de Macron d'un des présidents précédents qu'elle connait ; non sans artificialité donc, nous forçons le programme à attribuer (*prediction*) les paragraphes de Macron à un de ses prédécesseurs en raison de ressemblances linguistiques détectées. Ainsi, si Macron devait s'écrier à la tribune « Vive la Syrie libre ! », la machine attribuera ce passage à de Gaulle en référence sans doute au « Vive le Quebec libre ! » de Montréal en 1967. Si Macron devait prononcer « vous n'avez pas le monopole des sentiments », ou peut-être « vous n'avez pas l'exclusivité des sentiments », la machine attribuera le passage à Giscard d'Estaing en référence au débat télévisé avec Mitterrand en 1974 durant lequel le candidat conservateur avait répliqué au candidat socialiste « vous n'avez pas le monopole du cœur ». Pour donner un premier résultat réel du travail, lorsque Macron déclare « on ne peut pas travailler moins et gagner plus »[7], Hyperbase attribue automatiquement la phrase à Sarkozy en référence sans doute au « il faut travailler plus pour gagner plus » que le président de droite avait souvent répété durant son mandat.

2.2.3. Enfin, dernière étape décisive pour l'étude linguistique, l'algorithme de déconvolution présenté à la communauté internationale par (Vanni et al. 2018) et affiné aujourd'hui (Vanni et al. 2020-*submitted*) permet de décrire le corpus : il s'agit non seulement d'extraire les phrases de Macron attribuées à de Gaulle ou Pompidou, Sarkozy ou Hollande, mais de surligner les éléments linguistiques qui ont participé, via les couches cachées du système, à la décision. Ici, la méthode d'intelligence artificielle qui sert de guide (convolution puis déconvolution, et indice de reconnaissance des unités saillantes du réseau : *Text Deconvolution Saliency* –TDS et wTDS) est doublée par les procédés logométriques traditionnels (l'historique calcul des *spécificités* (Lafon 1980) ou le calcul des cooccurrences), afin d'affermir les résultats de l'IA par la statistique textuelle.

Dans tous les cas, en matière de classification ou de description, précisons que l'algorithme utilisé est multiniveaux (*multichannel*) c'est-à-dire que l'analyse se fait sur trois niveaux linguistiques empilés : le texte sous sa forme graphique, le texte lemmatisé, le texte étiqueté morpho-syntaxiquement. Tant est si bien que tel extrait de Macron pourra être rapproché du discours de Giscard par exemple à cause de sa nature fortement nominale (combinaisons marquées de noms, d'adjectifs et de déterminants) ; tel extrait de Macron sera rapproché de Pompidou par exemple pour la variété des lemmes utilisés (concentration de plusieurs lemmes rares et précieux) ; tel extrait sera rapproché de Hollande pour sa combinaison de mots graphiques significatifs comme « territoires » (au pluriel) ou « investissements » (également au pluriel). Tant est si bien surtout que les saillances linguistiques repérées par l'IA, à l'image des *motifs* théorisés en ADT par (Longrée et Mellet 2013) sont potentiellement multiniveaux c'est-à-dire grammatico-lexicaux : c'est parfois le mariage de tel mot (« transformation » par exemple) avec telle étiquette (verbe au futur par exemple) qui explique la décision prise par la machine.

[7] E. Macron, Vœux aux Français, 31 décembre 2018.

3. Résultats

3.1. Classification

Les emprunts ou empreintes du discours macronien sont riches et variés ou, autrement dit, l'intertexte des discours de Macron est ample. Sur les 10.000 paragraphes (fenêtres de 50 mots) analysés du corpus Macron[8], les taux d'inspiration ou d'intertextualité – c'est-à-dire ces passages de Macron que l'algorithme attribue à de Gaulle, Mitterrand ou Hollande… – se hiérarchisent selon la figure 1.

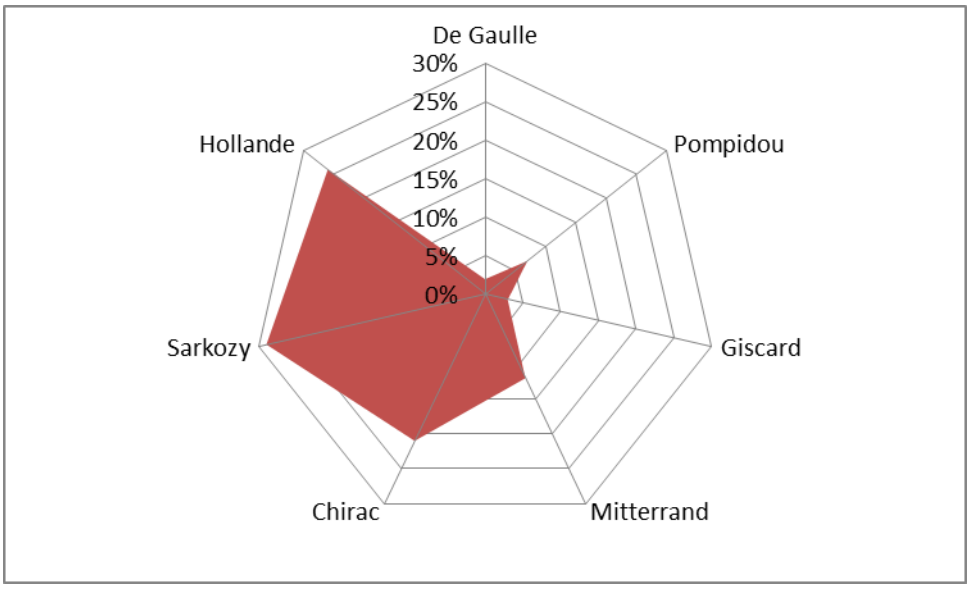

Figure 2. Périmètre de l'intertexte de Macron

Conformément à la contrainte du temps qui pèse sur les séries textuelles chronologiques (Salem 1988), Macron emprunte plus à ses prédécesseurs immédiats des années 1990-2010 (Hollande, Sarkozy, Chirac) qu'il n'emprunte à ses prédécesseurs plus lointains des années 1950-1970 (de Gaulle, Pompidou ou Giscard) ; la mémoire discursive est une mémoire de plus ou moins courte durée.

Dès lors, sur la base de ce constat chronologique, les hiatus attirent l'attention de l'analyste (Mayaffre 2020-sous presse).

Ainsi, à rebours de la chronologie, les empreintes pompidoliennes dans le discours de Macron sont plus nombreuses (7%) que les empreintes giscardiennes (3%) : par sa tenue littéraire (richesse et variété du vocabulaire), Macron imite souvent l'Anthologiste de la poésie française ; un souffle pompidolien traverse souvent les discours du jeune président.

De même, en dépit de l'impératif chronologique, l'intertexte sarkozyste (29%) du discours de Macron est plus important que l'intertexte hollandais (26%) : ici ce chiffrage peut éclairer le débat sur l'identité discursive d'Emmanuel Macron et son positionnement politique fondamental peut-être plus à droite de l'échiquier qu'à gauche.

[8] Rappelons le corpus : composé de plus de 1100 discours équivalant 3,5 millions de mots, le corpus recueille les principaux discours de De Gaulle, Pompidou, Giscard, Mitterrand, Chirac, Sarkozy, Hollande et Macron (1958-2020). Afin de donner une bonne représentativité au corpus, nous avons saisi les allocutions solennelles, les grands discours d'estrade, les interviews télévisées, les vœux, les tribunes dans la presse, etc.

Cependant, si la classification est intéressante – il s'agit d'une force reconnue du *deep learning* ici exploitée de manière originale pour mesurer l'intertexte – c'est la description linguistique du corpus qui constitue la plus-value méthodologique que nous cherchons à souligner dans cette contribution.

*3.2. Description triple niveau (*channel*)*

La force de l'algorithme implémenté dans le logiciel est de faire remonter, par déconvolution, les zones d'activation du réseau (indice intitulé en 2018 « TDS » (*Text Deconvolution Saliency*) et amélioré en 2020 sous le nom de « wTDS ») à un triple niveau : les formes graphiques, les lemmes et les étiquettes morpho-syntaxiques. Rappelons surtout que les passages sélectionnés et les zones d'activation dans le passage doivent être considérés en contexte, c'est-à-dire non plus dans une logique occurrentielle (tokenisation et distribution quantitative de chaque *token* ou de chaque étiquette dans le passage) mais dans une logique co-occurrencielle ou convolutionnelle (combinaison ou co-présence des éléments surlignés dans le passage).

3.2.1. Exemple d'intertexte sarkozyste dans le discours de Macron

A titre illustratif, les sorties machines de cette contribution sont issues du discours des voeux aux Français pour l'année 2020 d'Emmanuel Macron. Dans ce discours donc, comme dans le corpus en général, certains passages sont jugés inspirés par Sarkozy, qui est, selon la classification globale (*supra* figure 2), le premier inspirateur des discours de Macron :

" [...] travailler mieux , de partager la richesse créer dans toutes les NOM , à aider notre agriculteur et notre pêcheur à vivre dignement de leur travail comme à tout le entrepreneur et salarié . C' VER:pres DET:ART NOM PRP une Nation forte et indépendante . Si PRO:PER voulons lutter efficacement [...] "

Figure 3. Intertexte sarkozyste dans le discours de Macron. (Passage attribué à Sarkozy par l'algorithme qui souligne (wTDS) les lemmes (en vert), les formes (en bleu) et les étiquettes (en orange) responsables de la prédiction.)

De fait, tout dans ce passage de Macron peut renvoyer au discours de Sarkozy. Au niveau lexical, l'algorithme souligne le lemme « travailler » (verbe) et le lemme « travail » (nom). De fait, statistiquement, il s'agit de deux grandes *spécificités* lexicales sarkozystes (et aussi macronistes) comme l'atteste la distribution quantitative des termes dans le corpus.

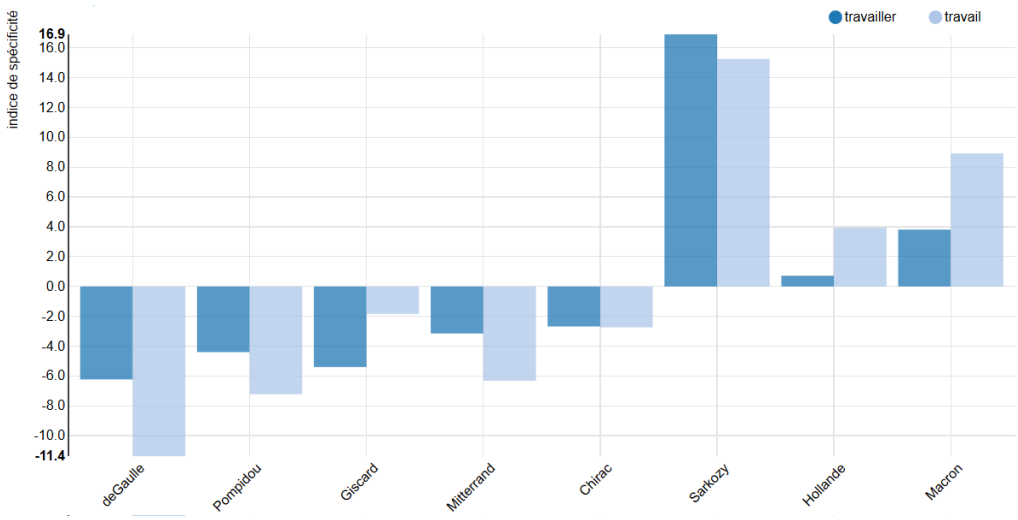

Figure 4. Les lemmes « travailler » et « travail » dans le corpus présidentiel 1958-2020

Dans la logique convolutionnelle (i.e. contextuelle) dont on a parlé, on remarquera aussi dans le passage la combinaison de « travailler + mieux » : les adverbes mélioratifs sont ici aussi, à l'analyse, très fortement sarkozystes dans le corpus (en l'occurrence sous la forme « travailler plus », « travailler davantage », « travailler mieux »).

Mais à ce socle lexical autour du « travail », s'ajoutent d'autres traits lexicaux sarkozystes qu'Hyperbase remarque, comme l'énumération de catégories socio-professionnelles que le président de droite aimait en son temps caresser : les « pêcheurs » ou les « agriculteurs » par exemple qui constituent électoralement la catégorie socio-professionnelle la plus fidèle à la droite avec les médecins, et encore les « entrepreneurs » et les « salariés ».

Ajoutons encore que dans ce passage, deux formes graphiques (en bleu) sur-lignées par le wTDS ne surprendront pas : « voulons » et « efficacement ». Macron emprunte en effet souvent à Sarkozy son volontarisme discursif, et son affirmation d'actions et d'efficacité. Du reste, c'est ce volontarisme verbal et cette rhétorique de la conviction qui expliquent sans doute le repérage du motif grammatico-lexical [C'+verbPrésent+Det+Nom]. Le pronom démonstratif neutre est la grande signature du sarkozysme en discours (Mayaffre 2012-b). Ici, suivi d'un verbe conjugué au présent et suivi d'un nom (« c'est la France… », « c'est un pays… », « c'est le travail… », « c'est la condition… », etc.), le pronom bien nommé *démonstratif* donne une force de conviction qui n'est pas sans rappeler Sarkozy dans la bouche d'Emmanuel Macron.

3.2.2. Exemple d'intertexte hollandais dans le discours de Macron

Dans son discours de vœux aux Français pour l'année 2020, Macron reprend, à côté du discours de Sarkozy, le discours de Hollande, afin d'équilibrer politiquement son propos :

" [...] d' un projet de justice et de progrès social . Un NOM de justice et de progrès social parce qu' PRO:PER assure l' universalité : il s' agit de faire en sorte KON un euro de cotisation versé ouvre les mêmes droits pour tous dès la première heure de travail [...] "

Figure 5. Intertexte hollandais dans le discours de Macron. (Passage attribué à Hollande par l'algorithme qui souligne (wTDS) les lemmes (en vert), les formes (en bleu) et les étiquettes (en orange) responsables de la prédiction)

Sans surprise, le passage repéré (figure 5), qui vise à justifier une réforme des retraites contestée par les syndicats, cultive un lexique de gauche : « progrès social », « justice », « universalité », « les mêmes droits pour tous ». Par ces mots, l'écho intertextuel à Hollande et, à travers lui, au parti du mouvement sonne éloquemment. Par exemple, dans le corpus élyséen (1958-2020), « universalité » est bien une spécificité statistique de Hollande.

Et de droite et de gauche, le discours de Macron emprunte ainsi des deux côtés de l'échiquier. Dans un contexte politique difficile qui vise à satisfaire et l'électorat de droite et l'électorat socialiste (ou CFDétiste), Macron multiplie et diversifie ainsi les reprises phraséologiques, les emprunts lexicaux ou les clins d'œil discursifs (Mayaffre 2020 – sous presse).

3.2.3. Prolongement : les marques du macronisme

La puissance classificatoire puis descriptive (wTDS) du *deep learning* nous permet enfin de faire chemin inverse et, du texte à l'intertexte, nous nous proposons de conclure en passant de l'intertexte au texte. Si la machine est sensible aux discours autres qui traversent les discours de Macron (intertexte), elle peut repérer les zones du discours qui sont jugées d'essence pure – purement macroniennes pourrait-on dire.

L'extrait suivant est sélectionné par la Hyperbase à ce titre :

" [...] de DET:POS société . Nous VER:futu DET:ART revalorisation et le transformation des carrières des enseignants , des professeurs , du soignant . Nous mener un politique ambitieux pour l' hôpital auquel je tenir tant et pour une médecine plus humaine centrée sur DET:ART patient . Nous aurons aussi ce année [...] "

Figure 6. Passage jugé purement macronien. (L'algorithme souligne (wTDS) les lemmes (en vert), les formes (en bleu) et les étiquettes (en orange) responsables de la prédiction)

L'examen statistique des occurrences et des cooccurrences de ce passage attestent de la pertinence du wTDS. Par exemple, la distribution de « société », « transformation » ou des verbes conjugués au futur, présents dans ce passage, montre que ces éléments sont, de fait, typiques (*spécificités*) de Macron (figure 7).

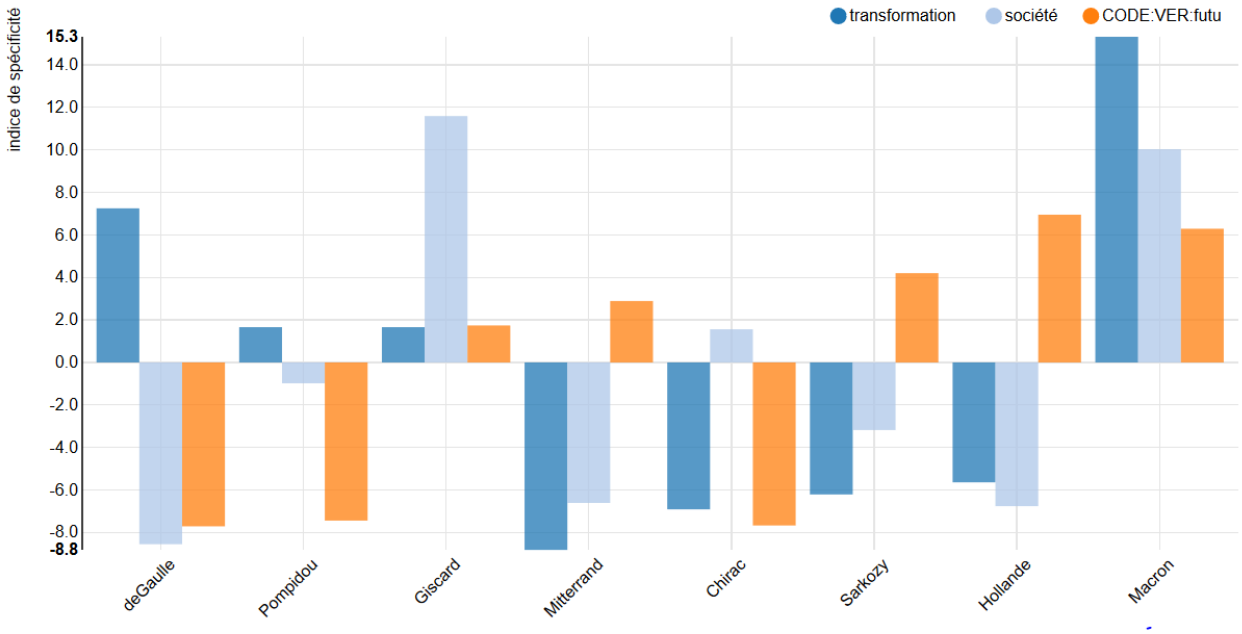

Figure 7. Spécificités de Macron (repérées précédemment dans le passage par le wTDS)

Et non seulement Macron sur-utilise « société » et « transformation » selon le calcul habituel des spécificités, mais le traitement statistique-roi de l'ADT (les cooccurences) montre qu'il aime les associer dans un même paragraphe : le cooccurrent statistique le plus indicés de « transformation » dans le discours de Macron est ainsi « société » (+15), lorsque, symétriquement, « société » est associé préférentiellement à « transformation » à hauteur de +8. Evidemment au regard de ces premières données, lorsqu'on calcule la distribution du motif cooccurrentiel [« transformation » + « société » + verbe au futur »] » la sortie-machine est spectaculaire (figure 8) : IA (le wTDS des mots de l'extrait ci-dessus) et ADT (le calcul des spécificités et des cooccurrences) conspirent ainsi au même résultat.

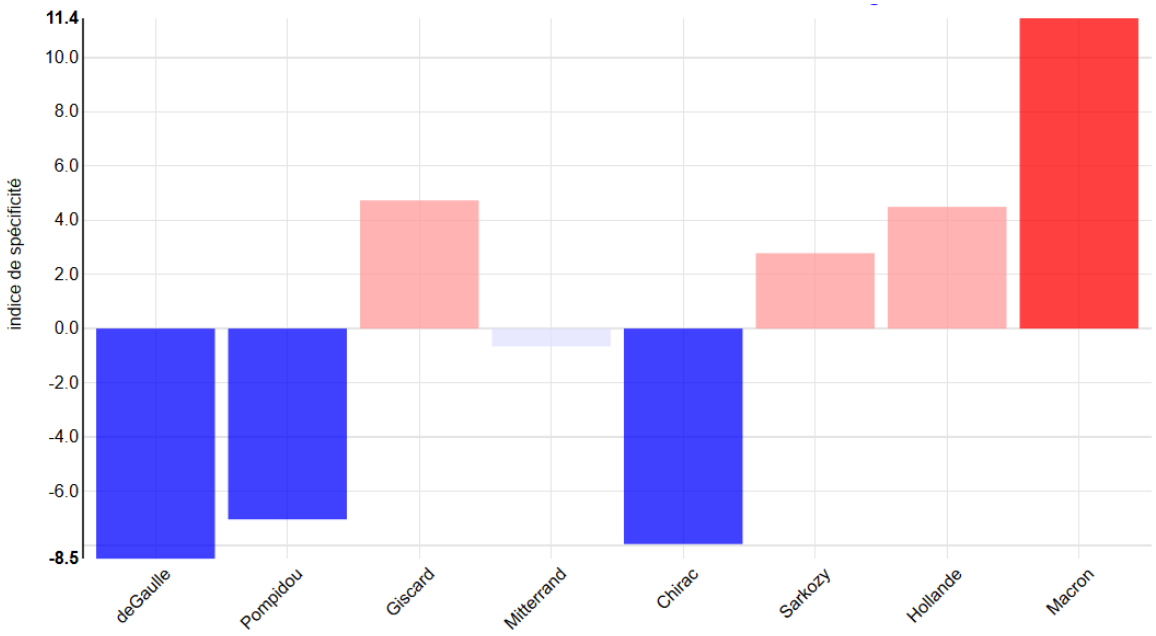

Figure 8. Distribution (spécificités) du motif cooccurrentiel [« transformation » + « société » + verbe au futur] dans le corpus présidentiel (1958-2020)

4. Conclusion

En articulant *deep learning* et ADT, cette contribution consacrée à l'objectivation de l'intertexte des discours de Macron a cherché à mettre en avant 3 tensions méthodologiques fondamentales dont la communauté ADT peut tirer profit.

1. Croiser l'urne (statistique) et le réseau (de neurones), croiser la tokenisation et la convolution, et, d'une autre manière, croiser l'occurrence et la co-occurrence. Le schéma d'urne reste d'une efficacité incontestable pour traiter un corpus textuel contrastif. Cependant, la convolution permet d'ajouter une approche co-textualisée, prenant le mot en contexte, prenant en compte la linérarité du texte ou sa réticularité. La clef de voûte de ces deux approches croisées est sans doute la cooccurrence (les unités du texte dans leur éco-système) comme nous l'avions affirmé, après d'autres, aux JADTs 2014 (Mayaffre 2014).

2. Croiser la classification et la description. L'IA est performante pour classer des textes. L'ADT reste essentiel pour décrire (chiffrer) les régularités et les écarts linguistiques. Le wTDS permet d'expliquer la prédiction du réseau de neurones. En d'autres termes, il permet d'ouvrir la boite noire du *deep learning* et, dans cette ouverture, l'ADT (la statistique) permet de vérifier la pertinence des unités linguistiques exhumées des couches cachées du système.

3 Croiser enfin l'axe paradigmatique et l'axe syntagmatique. La langue et les textes sont faits de *sélection* et de *combinaison*. Les traitements informatiques et statistiques ADT se sont depuis plusieurs décennies surtout consacrés à repérer les sélections des locuteurs (les mots les plus fréquents, les mots statistiquement préférés, la distance intertextuelle au regard des mots partagés). L'IA (modèle convolutionnel) est sensible à l'axe syntagmatique (les mots dans la fenêtre, dans leurs enchainements ou leurs combinaisons, c'est-à-dire dans leur singularité cotextuelle).

Gageons qu'articulés ensemble les deux modèles rendent compte de l'objet textuel dans toute sa complexité.

Références

Bres J., Haillet P.-P., Mellet S., Nølke H. et Laurence Rosier L. (2005). *Dialogisme, polyphonie : approches linguistiques*. De Boeck.

Brunet E. et Vanni L. (2019). *Deep learning* et authentification des textes. *Texto!*, vol. XXIV - n°1 [http://www.revue-texto.net/docannexe/file/4194/texto_brunetvanni_deep_final.pdf], consulté le 06/01/2020].

Brunet E., Lebart L. et Vanni L. (2020 – sous presse). Littérature et intelligence artificielle. In Mayaffre D. et al. (sous la dir.), *L'intelligence artificielle des textes. Points de vue critique, points de vue pratique*. Honoré Champion.

Cahiers de praxématique (1999). "Sémantique de l'intertexte", n°33 [en ligne : https://journals.openedition.org/praxematique/1965].

Ducoffe M., Mayaffre D., Precioso F., Lavigne F., Vanni L. et Tre-Hardy A. (2016). Machine Learning under the light of Phraseology expertise: use case of presidential speeches, De Gaulle - Hollande (1958-2016). In Mayaffre D. et al. (sous la dir), *JADT 2016 - Statistical Analysis of Textual Data*, Nice, Jun 2016, France, p.157-168. [hal-01343209V2].

Kim Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).*

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *Mots, Ordinateur, Textes, Société MOTS*, n°1.

Lebart L., Pincemin B. et Poudat C. (2019). *Analyse des données textuelles*. P.U n°Québec.

Lebart L. et Salem A. (1994). Statistique textuelle. Dunod.

Longrée D. et Mellet S. (2013). Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours. *Langages*, 189, p. 65-79.

Mayaffre D. (2002). Les corpus *réflexifs* : entre architextualité et hypertextualité. *Corpus*, 1 [En ligne : http://journals.openedition.org/corpus/11] consulté le 02 janvier 2020].

Mayaffre D. (2012-a). *Le discours présidentiel sous la Vème République. Chirac, Mitterrand, Giscard, Pompidou, de Gaulle*, Presses de Sciences Po.

Mayaffre D. (2012-b). *Mesure et démesure du discours. Nicolas Sarkozy (2007-2012).* Presses de Sciences 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 Née E. et al. (dir), *JADT 2014, Proceedings of the 12th International Conference on Textual Data Statistical Analysis*, Inalco-Sorbonne nouvelle, p. 15-32. [hal-01181337].

Mayaffre D. (2020-sous presse). *Macron par l'intelligence artificielle. Ses discours décryptés par la machine*. Les éditions de l'Aube.

Mayaffre D., Bouzereau C, Guaresi M., Precioso F. et Vanni L. (2020). Du texte à l'intertexte. Le palimpseste Macron au révélateur de l'intelligence artificielle, *7ème Congrès Mondial de Linguistique Française*.

Mayaffre D (sous la dir.) (2020-sous presse). *L'intelligence artificielle des textes. Points de vue pratique, point de vue critique*. Honoré champion.

Montavon G., Samek W. and Müller K. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73.

Rastier F. (2004). Enjeux épistémologiques de la linguistique de corpus. *Texto !* [en ligne, http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html], consulté le 06 janvier 2020).

Salem, A. (1988). Approches du temps lexical. Statistique textuelle et séries chronologiques. *Mots* 17, pp. 105-143.

Vanni L., Ducoffe M., Mayaffre D., Precioso F., Longrée D. (2018a). Text Deconvolution Saliency (TDS) : a deep tool box for linguistic analysis. *56th Annual Meeting of the Association for Computational Linguistics* [hal-01804310].

Vanni L., Mayaffre D., Longrée D. (2018b). ADT et *deep learning*, regards croisés. Phrases-clefs, motifs et nouveaux observables. In D. Iezzi et al. (dir.) *JADT' 2018*, UniverItalia, p. 459-466. [hal-01823560].

Vanni L., Corneli M., Mayaffre D., Precioso D. (2020-*submitted*). From text saliency to linguistic objects:learning linguistic interpretable markers with a multi-channels convolutional architecture. *57th Annual Meeting of the Association for Computational Linguistics*.