Du texte profond. Textualité et deep learning

Damon Mayaffre, Laurent Vanni

Du texte profond. Textualité et deep learning

Le Français moderne, 90ème année, n°1, 2022, pp. 135-153

Damon Mayaffre et Laurent Vanni (CNRS – Université Côte d'Azur, UMR 7320, Bases, Corpus, Langage)

1. Liminaire

Faire texte, selon le titre de l'ouvrage de [Adam (éd.) 2015]. Qu'est-ce qui fait texte ? Une écriture, mais plus encore, peut-être, une lecture.

Dès lors, en quoi la lecture numérique contemporaine – *ie*. la lecture des textes que proposent les ordinateurs *via* l'Intelligence artificielle – contribue-t-elle aujourd'hui à fabriquer des textes inédits ?

La question mérite d'être posée car si le syntagme « intelligence artificielle » reste sujet à caution[1], « lecture numérique » peut être pris au sérieux. La machine lit aujourd'hui les textes avec une acuité qui n'a pas la pertinence de la lecture humaine bien sûr, mais possède sa logique mécanique. La lecture de l'ordinateur permet ainsi de repérer les saillances, les creux, les contrastes matériels ou formels d'un texte, ses régularités lexicales par exemple ou ses hapax. Les logiciels lisent et classent – parfois de manière plus systématique que l'esprit humain – le vocabulaire exhaustif de Shakespeare, les segments répétés de de Gaulle ou les paragraphes de Proust. Sans plus d'erreurs qu'un étudiant en Sciences du langage, ils lemmatisent et étiquettent morpho-syntaxiquement de gros corpus en très peu de temps. Avec une efficacité spectaculaire sur laquelle cette contribution s'appuiera, ils attribuent automatiquement, sans erreur, un texte anonyme à son auteur-père, là où le jury du prix Goncourt a pu se laisser duper par un pseudonyme de génie.

Évidemment, la beauté d'un texte reste inaccessible à la machine, ainsi que son sens profond. Mais un ordinateur lit et repère un certain nombre de signes (disons pour aller vite des tokens) à qui il donne sinon une *interprétation*, toutefois une *représentation* pratique (représentation numérique, le plus souvent sous forme vectorielle) et un sens machinal minimum. Ainsi, par exemple, sur des corpus standards, testés par la communauté scientifique internationale, les algorithmes identifient certes de manière grossière les critiques positives de film *vs*. les critiques négatives de film avec un taux marginal d'erreur.

Admettons cependant qu'un texte, sans rien dire d'une œuvre, est fait par la richesse d'un parcours interprétatif de lecture que seul l'humain peut aboutir [voir nécessairement Rastier 2001] ; l'ordinateur le plus puissant du monde ne saurait ainsi épuiser automatiquement le sens d'un objet aussi « complexe » [voir Adam dans Adam (éd.) (*2015*.), *1. Complexité et problèmes du texte. 1.1. Penser un objet complexe*, p. 11 et ss.].

Mais ce n'est point épuiser le sens du texte que nous demandons à la machine !

Nous postulons seulement que l'ordinateur doit pouvoir encadrer le parcours de lecture avec une plus-value heuristique intéressante. Il ne s'agit pas d'objectiver le sens puisqu'il n'est pas objectif, puisqu'il n'est pas immanent, puisqu'il n'est pas *déjà-là* dans le texte ; puisque, à l'image du texte, le sens est toujours un construit et non une donnée positive. Il ne s'agit donc pas d'objectiver le sens mais d'encadrer le parcours interprétatif, d'objectiver ses stations et ses méandres, ses traverses ou ses impasses : en d'autres termes, notre problème n'est pas l'objet (texte) mais l'objectivation (de la lecture)[2]. Nous postulons en effet que la machine peut sur des critères numériques mettre en lumière certaines *artefactures textuelles* qui sont possiblement sous-estimées par l'œil humain, aussi bien à un niveau *micro* ou *meso*

[1] Pour la plupart des auteurs, il s'agit de donner un sens métaphorique à *intelligence* lorsqu'on la qualifie *d'artificielle*, non pas un sens littéral. Cf. en France, par ex. [Bachimont 1994] et [Doueihi 2011].

[2] Nous reprenons ici naïvement la posture épistémologique de Rastier ou Saussure : « …nous n'avons plus affaire à des objets, mais à des objectivations, comme nous l'avons souligné à propos de la sémiosis chez Saussure, dès lors qu'il s'écarte décisivement de la conception ordinaire qui réduit la langue à des mots et à des règles ». [Rastier 2020, pp. 146-147]. *Cf*. aussi *infra* la question des « unités » pertinentes, à défaut d'être objectives, du texte.

que *macro*, et qui sont autant de ressorts vigoureux pour l'interprétation. Si le texte n'est pas un objet naturel mais bien un artefact[3], alors l'intelligence artefactuelle – on dit habituellement « artificielle » – et les programmes deep learning doivent permettre d'en envisager un sondage, une description, une manipulation (en l'occurrence *digitalisation*) pour initier de nouvelles interprétations.

2. Révolution

Qu'ils soient issus de documents nativement numériques [Paveau 2017] ou simplement scannés à partir d'un livre, les textes n'épousent plus aujourd'hui la feuille de papier, à laquelle ils ont parfois été assimilés abusivement. Cependant, la révolution trans-Gutenberg qui a vu en quelques décennies le transfert de l'essentiel de la culture humaine du *folio* à l'*e-book* et des bibliothèques à internet ne peut être envisagée comme révolution si elle est pensée seulement comme une modification technique du support physique.

Pour la linguistique textuelle, passer du livre à l'écran n'est pas simplement un changement matériel de support (la tablette, le papyrus, le parchemin, la feuille, l'écran) ni un changement seulement médiatique, ou médiologique au sens de Debray [Debray 1991]. Passer du livre à l'écran, aussi radicalement, provoque une rupture épistémologique encore négligée par les sciences du texte ; celle-ci ne doit certes pas faire table rase de la philologie traditionnelle ou de l'analyse littéraire classique, mais il convient d'en mesurer correctement les plus-values heuristiques et interprétatives.

En des termes simples, traiter un corpus de textes numériques aujourd'hui n'a un surplus d'intérêt scientifique que si les parcours de lecture proposés – c'est-à-dire l'établissement du sens ; c'est-à-dire la fabrication du texte – interrogent l'analyste réellement de manière nouvelle ; que si la représentation du texte proposée présente l'objet (le texte donc, la textualité ou la texture, les grandeurs textuelles pertinentes) sous une lumière effectivement inédite ; que si l'herméneute, fort de deux millénaires de pratiques auxquelles il ne doit pas renoncer, augmente d'un point de vue théorique et pratique la fertilité de son *cercle*.

A l'international, la communauté *text mining* et en France la communauté ADT (Analyse de Données Textuelles) ont proposé depuis plusieurs années des traitements de corpus numériques à même de renouveler la philologie classique et les points de vue des chercheurs sur leurs textes. Appuyées sur un appareillage statistique perfectionné, elles ont proposé des descriptions et des modélisations des corpus textuels utiles tant pour l'attribution d'auteur et la génétique des textes [ex. Brunet 2016] que pour la description des genres [ex. Magri 2009, pour les récits de voyage], la stylistique [ex. Viprey 1997], le commentaire littéraire [ex. Bernard 2003], l'analyse de discours [ex. Mayaffre 2012], etc. A titre d'illustration majeure – majeure et pourtant encore parfois ignorée – le traitement statistique des cooccurrences a permis dans de nombreuses études de souligner des structures réticulaires des textes jusqu'ici peu connues, d'en approcher leur *texture* ou simplement d'en objectiver, autour d'un mot pôle, les thèmes et les isotopies[4].

Avec l'avènement des *big data* textuelles (*google books* par exemple) et l'explosion de la puissance de calcul (processeurs centraux CPU, cartes graphiques GPU…), c'est l'Intelligence artificielle et les réseaux de neurones bio inspirés, autant que la statistique proprement dite, qui seront mobilisés dans cette contribution.

Si l'on admet que l'Intelligence artificielle (plus précisément le deep learning, avec des modèles dits convolutionnels CNN) est susceptible de révéler des structures profondes (*deep*) des textes et de mettre en lumière des observables jusqu'ici seulement pressentis, le programme de recherche apparait majeur.

[3] Voir nécessairement Adam et Heidmann (éds.), 2005 qui parlent du « leurre de l'évidence naturelle du texte » (p. 70) ou Adam 2018-a qui parle du « leurre de son évidence naturelle » (p. 40) ; voir plus explicitement Heidmann et Adam (2010) qui titrent « Les textes ne sont pas des objets naturels » (pp. 26-28). Voir également le postulat général d'*Arts et sciences du texte* de Rastier 2001.

[4] Il est impossible ici de résumer la bibliographie de 50 ans de recherche en la matière. On citera à l'échelle internationale les travaux de [Halliday et Hasan 1976] et en France les travaux de Viprey [ex. Viprey 2006]. Nous avons nous-mêmes fait le « plaidoyer » de la cooccurrence, du point de vue linguistique et statistique dans [Mayaffre 2014].

3. L'intelligence artificielle des textes. Principes généraux à l'usage de la linguistique textuelle

Le deep learning (ici le deep learning appliqué aux textes) est un univers informatique que cette contribution peut seulement résumer dans ses principes généraux. Pour un appareil critique et des considérations méthodologiques, nous renvoyons à nos écrits spécialisés [Vanni *et al*. 2018, Vanni *et al*. 2020, Vanni et Precioso 2021]. Nous renvoyons également aux contributions qui font autorité dans la littérature informatique internationale [Collobert and Weston 2008 ; Kim 2014].

L'idée princeps du deep learning est de demander à la machine d'apprendre (*learning*) : apprendre à discriminer des objets simples ou complexes comme une image ou comme un texte. Ainsi, après apprentissage, une machine sait reconnaitre avec un taux de réussite proche de 100% une image de chat *vs.* une image de chien. De la même manière[5], après entrainement, nous arrivons, avec un taux de réussite presque aussi important, à reconnaitre un texte de de Gaulle *vs.* un texte de Macron ou un roman de Flaubert *vs.* un roman de Dumas [pour l'apprentissage du discours littéraire voir l'étude monumentale de Brunet, Lebart et Vanni 2021 ; pour l'apprentissage du discours politique voir Guaresi et Mayaffre 2021].

La plupart des modèles d'apprentissage sont dits *supervisés* (nous ne pouvons ici considérer certains autres modèles révolutionnaires dits *non-supervisés*). En cas de supervision, il s'agit d'entrainer la machine sur la base d'un corpus d'apprentissage confectionné à cet effet par l'analyste : le chercheur donne sciemment à l'ordinateur un nombre important d'images de chat *vs.* un nombre important d'images de chien et lui demande d'y repérer les traits visuels saillants des deux mammifères (peut-être un type de pelage, peut-être une moustache, ou la taille de la queue, etc.) ; de la même manière, nous donnons à l'ordinateur un grand nombre de discours que nous savons être (*supervision*) de Macron, de de Gaulle ou de Mitterrand, et nous lui demandons d'y repérer les traits linguistiques caractéristiques des présidents (peut-être une prédilection pour le suffixe *tion chez Macron, peut-être son goût pour le mot « projet » ou son sur-usage du subjonctif, peut-être l'enchaînement syntaxique élémentaire « nom + adjectif », etc.).

Le reste relève de la puissance de l'algorithmique et du génie mathématique. Dans une logique connexionniste, l'ordinateur prend en compte en effet les millions de pixels de l'image (chaque pixel est alors considéré comme un neurone) et les millions de relations ou combinaisons de ces pixels entre eux (chaque relation est considérée comme une synapse) afin de reconnaitre les traits distinctifs du chat. De la même manière, entre connexionnisme et distributionalisme, la machine considère les milliers de mots, de lemmes, de catégories grammaticales, et leur combinaison ou distribution dans le corpus pour reconnaitre un discours de Macron.

Pour arriver à ses fins (bien reconnaitre l'image du chat ; bien reconnaitre un texte de Macron), l'ordinateur opère par *abstraction* successive : l'algorithmique réduit, augmente, combine, filtre, additionne, soustrait, synthétise, vectorise les informations dont il dispose (les données de départ, c'est-à-dire les pixels pour les images, les mots, les lemmes ou les catégories pour les discours) afin de réussir sa tâche de classification.

Par exemple, sur un corpus de roman français, la machine pourra essayer la pertinence de réduire ou filtrer un texte à ses seuls noms propres. Elle s'emparera alors de manière obvie de cette réduction pour reconnaitre un texte de Flaubert si le prénom « Emma » y figure *vs.* un texte de Dumas s'il y trouve « Dartagnan ». Plus subtilement, pour affermir sa conclusion, il repèrera que dans le premier texte le prénom « Emma » évolue souvent avec un imparfait mélancolique ; là où « Dartagnan » est suivi d'un passé simple plus dynamique. Ainsi, dans la fenêtre de convolution (*cf*. immédiatement *infra*, disons pour l'instant le « contexte »), le motif lexico-grammatical « nom propre [Emma] + temps verbal [imparfait] » devient pour la machine un chemin efficace pour discriminer Flaubert des autres romanciers français.

Le filtre lexical « Emma » combiné au filtre grammatical « indicatif imparfait » doivent ici être considérés comme des exemples parmi des milliers d'autres possibles dans un texte ; filtres pertinents pour reconnaitre l'auteur de Madame Bovary, mais

[5] Dans cette vulgarisation, nous nous appliquerons ci-dessous à mettre en parallèle le deep learning pour les images (chat/chien) et pour les textes (Macron/de Gaulle). En effet les implémentations que nous proposons pour le texte sont directement inspirées du cortex visuel et initialement appliquées au traitement des images par le cerveau et par la machine.

sans doute inutile pour reconnaitre de Gaulle ou Macron. Précisément, le deep learning explore (*apprend*) dans une compilation systématique des données et un jeu d'essais-erreurs surpuissant, tous les observables textuels dont il dispose et calcule toutes les combinaisons pour trouver et établir celles, discriminantes, qui *font texte* pour tel auteur, celles qui *font texte* pour telle œuvre, celles qui *font texte* pour tel genre.

4. *Description* autant que *prédiction*

Au plus haut niveau de notre proposition, il s'agit d'éclairer, du côté des SHS et de la linguistique textuelle, les modèles deep learning qui s'imposent partout à la ville comme à l'université mais qui restent encore obscurs aujourd'hui. L'Intelligence artificielle n'aurait d'intelligence que le nom si les résultats qu'elle obtient demeurent inintelligibles à l'esprit humain (celui des informaticiens eux–mêmes)[6] ; précisément, pour nous, les « couches cachées » du système (*grosso modo* les différents filtres, les différents paramétrages, les différentes pondérations), qui s'empilent lors du traitement pour obtenir un résultat certes pertinent, ne sauraient rester plus longtemps cachées. En effet, les amoureux des textes – linguistes, littéraires, historiens de la culture, etc. – n'ont que faire d'une « boite noire » susceptible de reconnaitre automatiquement sans faute un auteur, et de classer sans erreur les textes d'Emile Ajar avec ceux de Romain Garry, si elle reste incapable d'expliquer le protocole et d'exhiber les unités textuelles, le régime énonciatif, les mots, les thèmes, le style, les motifs, les passages responsables de cette classification réussie.

Fondamentalement, nous avons plaidé ailleurs pour une posture herméneutique face au texte numérique et face à la machine [Guaresi et Mayaffre 2021]. L'Intelligence artificielle donne des représentations originales du texte : représentations réticulaires, nous l'avons dit, ou rhizomatiques, représentations numériques, tabulaires, modulaires, vectorielles, hypertextuelles. Bref ! tout ce qui se passe lorsque l'on met un texte en machine du simple tableau hiérarchique des formes (les mots les plus utilisés) à une matrice de cooccurrences (les mots combinés ou associés), d'un simple concordancier (approximativement les phrases) à une analyse factorielle des correspondances (une nébuleuse), d'un simple dictionnaire exhaustif des lemmes (le vocabulaire) à un graphe de clustering morpho-syntaxique (un réseau). Mais ces représentations du texte sont heuristiques en SHS sous la condition non négociable d'en expliciter l'expression et le contenu, et tout simplement la raison d'être ; explicitation sans laquelle aucune interprétation sémantique n'est possible. La machine aujourd'hui sait *prédire* (*prediction* dans l'anglais du deep learning ; nous dirons plus simplement « classer » ou « reconnaitre »). Mais nous lui demandons désormais de *décrire* dans une perspective herméneutique [voir le collectif d'auteur dans Mayaffre et Vanni (éds) 2021]. Cette description doit nous révéler de nouveaux observables linguistiques constituants ou constitutifs du texte, *qui font texte*, que l'on pourrait appeler selon les modèles nombreux (nous y reviendrons) : « grandeurs textuelles », « marqueurs », « unités textuelles », « patterns », « n-grams », « expression régulière », « segments textuels », « unités phraséologiques », « lexies simples », « séquences », « paragraphes », « cooccurrences », « passages », « isotopie », « routine », « formule » « collocation », « texème », « herménème », etc. (*cf. infra*) ; toutes choses qui, précisément, participent, selon les différents auteurs, à la composition, à la structuration, à la caractérisation et à l'interprétation d'un texte.

Locaux dans le texte, ces observables linguistiques émergent grâce à un traitement global du corpus – le global détermine le local – selon les principes généraux de l'herméneutique ; selon les principes théoriques d'une sémantique différentielle ou sémantique de corpus (*cf.* le sous-titre de l'ouvrage de référence de Rastier 2011] ; et selon les principes techniques et indépassables du traitement informatique. En effet, l'IA (au même titre que la linguistique textuelle, par une heureuse coïncidence) a besoin de corpus de référence ou de corpus d'apprentissage qui constituent le global, l'architexte, la norme endogène ou l'horizon d'attente par rapport auxquels les textes et leurs unités se discriminent et prennent sens ; corpus globaux, homogènes en genre, équilibrés et contrastifs, à l'instar des corpus *réflexifs* en logométrie que nous avons théorisés ailleurs [Mayaffre 2002 et 2007], et finalement à l'instar de ceux de toute la linguistique des textes[7].

[6] En contrepoint de l'inintelligibilité du numérique notons le projet scientifique d'une nouvelle revue (premier numéro en 2020) au titre-profession de foi : « L'intelligibilité du numérique ».

[7] Nous ne pouvons ici revenir sur l'objet *corpus* – ce global qui détermine le local des textes. La linguistique textuelle contemporaine semble avoir fait de la réflexion sur le corpus un

5. Aborder le texte profond : entre fréquence et séquence

À un niveau plus technique mais qui parait tout aussi fondamental, l'hypothèse forte de cette contribution est la complémentarité de la statistique textuelle et du deep learning pour l'exploration des corpus de textes ; particulièrement la complémentarité de la statistique textuelle et du deep learning lorsqu'il s'agit de saisir ensemble la dimension paradigmatique et la dimension syntagmatique du texte. En effet, la statistique textuelle est d'essence paradigmatique : grossièrement, la lexicométrie considère le texte comme une *urne* composée d'unités discrètes (nous allons jusqu'à parler de *jetons* trad. *tokens*) ; et par calculs et tirages plus ou moins savants (schéma d'urne, loi normale, modèle hypergéométrique, etc.) les *choix* lexicaux ou les *choix* grammaticaux (des « paradigmes ») du locuteur sont ainsi évalués et décrits (Macron utilise beaucoup les mots suffixés en *tion ; de Gaulle aimait les temps du passé particulièrement les verbes au passé simple ; etc.). Le deep learning, lui, est d'essence syntagmatique : les *modèles convolutionnels* (CNN) opèrent par fenêtre coulissante (dans cette contribution, cette fenêtre glissante est de 6 mots) pour saisir le mot dans sa fenêtre cotextuelle : par convolution donc, ce sont les *combinaisons* ou l'enchainement des mots (ou « syntagme ») qui sont ainsi considérés. Avec l'IA, c'est en effet sur la chaine que le token prend une représentation unique – disons une *valeur* – qui apparait donc variable selon ses contextes ; nous pourrions dire *valeur différentielle*. Si les mots choisis et leur distribution fréquentielle sont importants pour le statisticien, la contextualisation et la distribution séquentielle des mots, essentielles dans la linguistique textuelle, sont satisfaites par l'IA. Concrètement, nous mettons à jour un schéma paru en 2021 dans *L'intelligence artificielle des textes. Des algorithmes à l'interprétation* pour illustrer la différence de traitement :

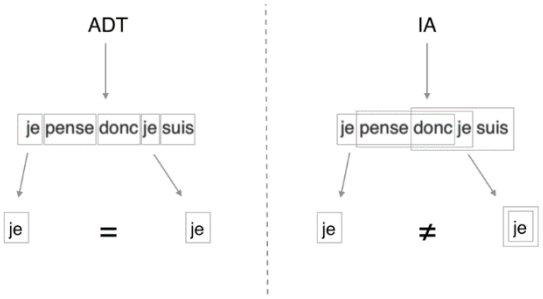

Illustration 1. *Modèle comparé de la représentation numérique des mots en ADT vs IA.*

<u>Commentaire du schéma</u> : dans le premier cas, pour la statistique, les unités sont discrètes et atomisées (tokenisées), et les deux « je » de la phrase ont une représentation identique ou isonomique ; les différentes occurrences d'une même forme, quels que soient leurs contextes, ont la même valeur (comme si les mots avaient toujours le même sens ?). Dans le deuxième cas, pour l'IA, l'approche est continue, coulissante ou convolutionnelle. En conséquence, les deux « je » prennent une valeur différente en fonction des mots précédents et suivants l'occurrence. Le lecteur notera que lorsque la fenêtre de convolution est assez importante (6, 7…15 mots), chaque occurrence du corpus devient un hapax puisqu'il est peu probable de retrouver des séquences identiques de 6, 7 ou 15 mots[8].

---

préalable scientifique (voir par exemple la revue dédiée éponyme *Corpus*). Un corpus problématisé (pléonasme) apparait comme la condition de toute recherche et de la sémantique différentielle. Cependant, la construction du corpus, nécessaire partout, apparait plus impérieuse dans les analyses numérisées. Car si le corpus peut garder la force de l'idéal ou le flou de l'idéel dans les études classiques, il prend un tour matériel concret indépassable dans les études informatisées : ce que l'on a réellement *saisi* (ou pas) et réellement *soumis* (ou pas) au traitement. Un texte envisagé ou pressenti mais non saisi se trouve projeté hors de l'analyse ; son vocabulaire est exclu de l'index alphabétique, ses phrases sont exclues du concordancier, ses unités exclues du calcul des fréquences ou des cooccurrences, etc. On ne badine pas avec la machine.

[8] Cette unicité de l'occurrence, implémentée dans l'IA, ne manque pas de faire penser à l'historique mise en garde du *Cours* à propos de la répétition de « messieurs » dans un discours. Les transcripteurs de Saussure précisent : « Lorsque, dans une conférence, on entend répéter à plusieurs reprises le mot Messieurs !, on a le sentiment qu'il s'agit chaque fois de la même expression, et pourtant les variations de débit et l'intonation la présentent, dans les divers passages, avec des différences phoniques très appréciables […] ; en outre, ce sentiment de l'identité persiste, bien qu'au point de vue sémantique non plus il n'y ait pas identité absolue d'un Messieurs ! à l'autre. » [*Cours de linguistique générale*, Deuxième partie. *Linguistique synchronique*, chapitre III. *Identités, réalités, valeurs*].

Syntagmatique complémentaire de paradigmatique, séquentiel complémentaire de fréquentiel, disons aussi simplement, à l'instar de la démonstration récente de la thèse de Camille Bouzereau [Bouzereau 2020], qualitatif (car contextualisant) complémentaire de quantitatif (possiblement décontextualisant) : le dialogue entre deep learning et statistique textuelle doit s'avérer productif.

6. Cercle herméneutique et *backpropagation*

(i) Posons d'abord la matérialité textuelle comme condition de l'interprétation ; si herméneutique il y a, elle devra être matérielle[9]. À la suite d'Umberto Eco, que l'on ne peut accuser ni de matérialisme ni de positivisme, concédons en effet que pour qu'il y ait interprétation il faut qu'il y ait, *d'abord*, quelque chose à interpréter [Eco, Les limites de l'interprétation – *Introduction*]. Ce *quelque chose à interpréter*, c'est le texte matériel ou l'objet empirique : ces mots qui sont écrits ou prononcés et que l'on soumet à la lecture puis à l'analyse. Or l'approche informatique, aveugle sur l'idée, traite obligeamment de la matière (on pourrait dire aussi du signifiant) : octets, bits, tokens, mots, lemmes, étiquettes morphosyntaxiques etc. Elle pourrait donc nous intéresser dans notre mouvement premier, empirique, philologique ou matérialiste, face au texte.

(ii) Posons ensuite l'indécidabilité problématique des unités textuelles et des « paliers de pertinence textuelle » [Monte *et al*. (éds) 2018] ; « énigme insoluble » écrit Dominique Legallois dans sa présentation de *Langages* consacré au(x) « Unité(s) du texte » [Legallois 2006 : 3]. Par exemple, nous savons que l'unité-mot n'a pas beaucoup de pertinence. Autre exemple problématique, l'unité-phrase est considérée avec méfiance par la linguistique textuelle : un texte est plus qu'une addition de propositions grammaticalement correctes[10] ; et les grammaires de texte, qui imaginaient le texte comme une simple extension de la phrase, se sont révélées des impasses scientifiques. À l'examen, si la question des unités du texte est donc insoluble, c'est qu'elle renvoie à des postures différentes entre approche réductionniste et approche holistique ; et conséquemment entre des méthodes compositionnelles *vs.* des méthodes relationnelles. François Rastier a présenté de manière peu contestée le débat sous la forme d'une opposition historique entre la tradition logico-grammaticale et la tradition rhétorico-herméneutique. En peu de mots, l'idée serait dans le premier cas (tradition logico-grammaticale) de définir les unités minimales ainsi que les règles d'assemblage pour produire/décrire un texte ; le texte serait une composition décomposable d'unités stables (disons les mots) sur la base d'une grammaire (qu'il reste cependant à établir). Dans le deuxième cas (tradition rhétorico-herméneutique), il s'agirait de poser le texte comme un tout, dans lequel il convient dès lors de repérer des fragments pertinents, ainsi que les régimes de relation entre ces fragments ; le texte devient un ensemble sémantique, dynamique, intégratif, différentiel, relationnel. Et à nos yeux, nous ne sommes plus très loin de l'idée de *réseau*, que l'intelligence artificielle néo-connexionniste ou que les réseaux de neurones artificiels modélisent et mécanisent efficacement aujourd'hui. Quoi qu'il en soit, unités constituantes ou fragments constitutifs, la définition des « éléments » ou des « grandeurs » du texte reste fragile. Et l'approche numérique - ce serait alors son intérêt - doit nourrir ce débat fondamental : au-delà du simple *token*, sur quelles « unités », « zones de localité » ou « observables » s'appuie la machine pour reconnaitre pertinemment un texte de Macron ou un texte de Flaubert ?

(iii) Revendiquons enfin notre posture herméneutique : faire texte comme faire sens, c'est admettre que le global (le monde, le point de vue, l'intention, l'œuvre, le corpus, le texte dans sa totalité) détermine le local (par exemple le mot, le syntagme, le paragraphe). Seulement, si le texte dans son entier donne le sens aux parties (on dira au sens fort et étymologique : *informe* ses parties), les parties ainsi informées nourrissent à leur tour et précisent l'interprétation : il s'agit, dans l'épistémologie moderne, depuis Friedrich Schleiermacher et Wilhelm Dilthey, du *cercle herméneutique*, de ses vertus et de ses problématiques. Du tout aux parties, des parties au tout (c'est à dire pour nous du texte aux mots, des mots au texte), le cheminement n'est ni linéaire ni en sens unique.

---

[9] François Rastier à la suite de Spitzer, Szondi, Bollack propose le programme ambitieux de refondre le rapport entre philologie et herméneutique [Rastier 2001-a ; Rastier 2001-b] et Jean-Michel Adam insiste souvent sur « le moment philologique » préalable à toute analyse textuelle et toute herméneutique digne de ce nom [par ex. Adam, ed (2005), p. 83]. Nous avons pris quant à nous le risque de titrer notre thèse HDR, « Vers une herméneutique matérielle numérique » [Mayaffre 2010].

[10] Il faut sans doute aller plus loin que cette formulation : le texte est, précisément, ce qui est *en plus* ou *au-delà* de l'addition de phrases grammaticalement correctes.

Or il se trouve que le point évoqué en iii (le cercle herméneutique), appuyé sur le point i (la matérialité textuelle comme point de départ) et sur le point ii (la réflexion sur les unités pertinentes du texte), trouve une implémentation technique mais centrale dans l'édifice du deep learning : la *backpropagation*.

La *backpropagation* est en effet la clef des modèles deep learning et peut se résumer en peu de phrases et en deux schémas (illustration 2 et 3) qui reproduisent l'idée du cercle herméneutique.

Propagation. — Dans un premier mouvement, nous semblons donner à la machine, en entrée du système, les mots du texte comme unités élémentaires et nous trouvons en sortie du traitement un résultat global ou unifié : un texte ; un texte classé ou identifié. La démarche s'annonce donc analytique ou réductionniste (des unités vers la totalité, c'est-à-dire, des mots en entrée vers le texte en sortie). (illustration 2).

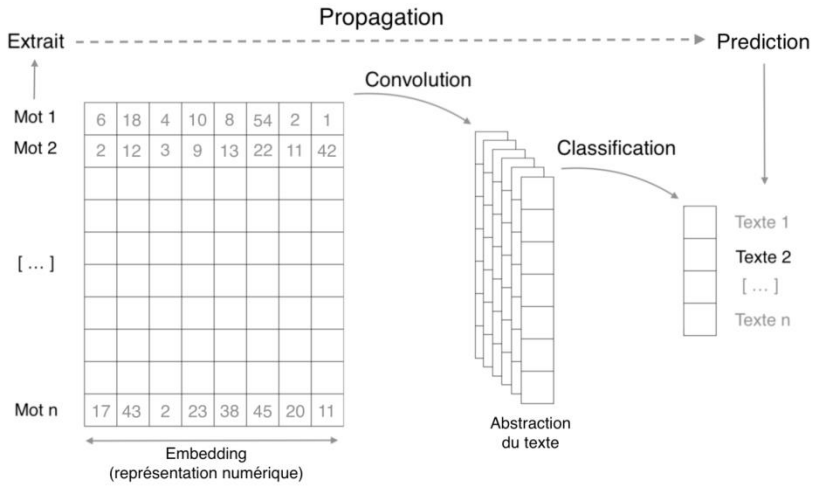

Illustration 2. *Propagation – des mots au texte dans le traitement deep learning*

Bien sûr, les modèles que nous utilisons essayent d'être intelligents linguistiquement et l'information est en vérité triple pour être riche : en entrée du système nous donnons non seulement les mots (unités contestables), mais les lemmes associés et l'étiquette morphosyntaxique ; le traitement est dit *multi-channel* (3 niveaux ; ici trois caractérisations linguistiques). Mais la démarche semble bien réductionniste car c'est de ces mots, lemmes et étiquettes, mis bout à bout, que nous concluons le texte dans son ensemble.

Bien sûr encore, nous l'avons dit, ces unités élémentaires sont traitées de manière convolutionnelle, c'est à dire en fonction de leur contexte ; l'approche des éléments du texte n'est plus discrète mais continue ; et le mot n'est pas seul puisqu'il est pris dans son syntagme, sa phrase, son paragraphe et plus généralement dans sa distribution dans le corpus. Cependant, admettons toujours que la démarche, nous en déplaise, reste analytique puisque nous avons tokenisé *a priori* le texte pour définir et traiter des unités minimales, desquelles sera déduit ou (re)composé le texte dans son entier.

Backpropagation. — Seulement, partant de la représentation des mots-unités pour en conclure une classe-texte, le système deep learning s'autorise, en cours de traitement, et autant de fois qu'il lui paraitra nécessaire, de *corriger* la représentation et le poids des mots-unités en question, pour optimiser le résultat souhaité. Ainsi, c'est en sachant, par supervision, que ce texte est un texte de Macron, que le deep learning modifie la représentation des mots présents dans le texte. La démarche s'inverse donc tout à coup. Nous partons de ce qui avait été présenté comme la sortie (le texte) pour en informer (donner une forme nouvelle, donner un nouveau « sens ») les unités de départ (les mots). Nous ne concluons plus le texte des mots, mais nous concluons les mots (leur représentation corrigée, leur valeur corrigée) à partir du texte et même du corpus dans lesquels ils se trouvent. (illustration 3)

Cas d'école élémentaire. — Caricaturons pour nous faire comprendre du lecteur non expert et renvoyons le lecteur expert à la section suivante. Dans un genre donné, celui des récits de voyage, les mots « navire » et « capitaine » sont respectivement représenté par une suite de nombres : on parle alors de représentation numérique ou

d'*embbeding*[11]. Ainsi, dans ce corpus de récit de voyage, « navire » et « capitaine » sont respectivement identifiés par [2, 6, 4, 9, 12…] et [2, 4, 8, 3, 10]. Le lecteur remarque immédiatement que la représentation numérique des deux mots est assez proche puisqu'il y a, dans les deux cas, presqu'uniquement des petits nombres pairs. Dans cet exemple, la représentation numérique commune aux deux mots pourra sans doute être imputée au fait qu'ils appartiennent tous les deux au même champ lexical de la mer, que tous les deux sont notamment cooccurrents dans le corpus du verbe « naviguer », que tous les deux sont utilisés par les mêmes auteurs-baroudeurs, etc ; les petits nombres pairs semblent ainsi un marqueur numérique positif du vocabulaire maritime des récits de voyage, comme le nombre 1 au début d'une carte vitale est un marqueur positif du sexe masculin d'un individu considéré.

Seulement, dans un autre genre et un autre corpus, ceux du journalisme politique « capitaine » se retrouve seulement dans le cadre de considérations guerrières ; les deux mots « navire » et « capitaine » n'appartiennent plus, pour ce nouveau genre, à la même isotopie ou à la même thématique ; ils ne cooccurrent plus ensemble avec le verbe « naviguer » ; ne sont plus employés par les mêmes auteurs ; n'ont plus la même distribution, etc.. C'est alors que la machine ayant considéré ce genre nouveau s'autorise à changer les valeurs numériques de « capitaine » en lui attribuant une autre représentation numérique (de grands nombres impairs par exemple) afin que cette représentation corrigée (*backprogation*) performe mieux la classification recherchée (décrire les récits politiques et non plus les récits de voyage). Le global (le corpus, les textes de journalisme politique) a ainsi réidentifié ou requalifié le local (le mot « capitaine » qui change de signification ou de représentation) afin d'optimiser le système et améliorer la tâche.

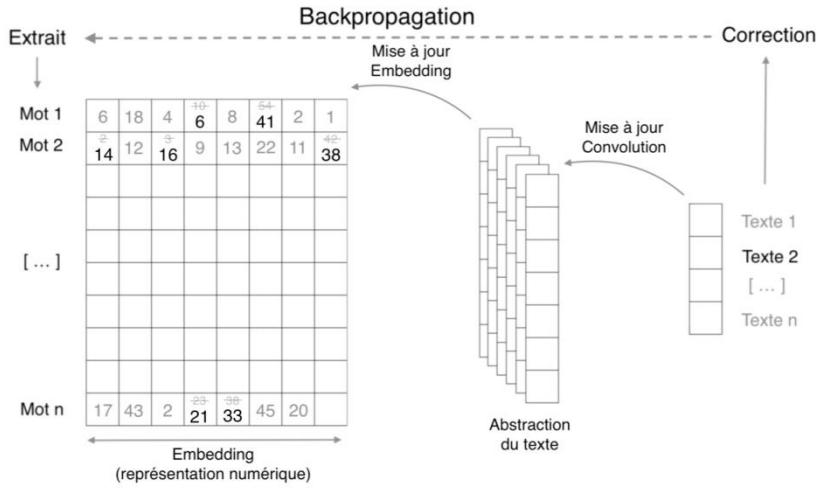

Illustration 3. Backpropagation – du texte aux mots dans le traitement deep learning

C'est en cela que la propagation/backprogagation nous parait une implémentation concrète du cercle herméneutique : le système, en rond, s'appuie sur les unités matérielles locales initiales (les mots) pour identifier le texte au sein d'un corpus global, en même temps que le système assure, dans un jeu d'aller-retour ou un cercle, la requalification des données initiales pour obtenir le résultat général souhaité. Fort de cette requalification des données et de l'amélioration des performances du traitement, le corpus sera à nouveau analysé pour être encore mieux reconnu ; et ainsi de suite va l'apprentissage jusqu'à satisfaction[12].

L'essentiel de la puissance de calcul de la machine se trouve investi à cet endroit : par des millions d'essais/erreurs, de qualifications/requalifications, propagation/backpropagation l'ordinateur optimise la performance de sa classification et découvre, et pour tout dire « invente », de nouvelles représentations : celles qui

[11] Qu'il nous soit permis de vulgariser plus encore. Un mot est représenté dans le corpus par une suite de nombres, un peu comme un individu peut être représenté sur sa carte vitale par la suite 1800675… ou 2400201… selon qu'il s'agisse, dans le premier cas, d' un homme (1) né en 1980 (80) au mois de juin (06) à Paris (75)… et dans le deuxième cas d'une femme (2) née en 1940 (40) au mois de février (02) dans l'Ain (01). On attribue ainsi aux mots des nombres (ou *embbeding*) qui vont permettre à la machine de les caractériser, trier, manipuler.

[12] Ce *ainsi de suite* se chiffre et est formalisé en IA. Il s'agit techniquement l'*EPOCH* (le nombre de lectures de l'ensemble des textes permis au modèle pour s'entrainer), que nous effectuons 100 fois dans cette contribution [Vanni et Precioso 2021, pp. 25, 31-32].

s'avèrent efficaces pour la tâche décisionnelle et qui pointent en conséquence de nouveaux observables pertinents (mots simples contextualisés ou motifs plus complexes qui *font texte* chez Macron *vs.* qui *font texte* chez de Gaulle). Ces nouveaux observables, interprétés par la machine, interprétable pour l'analyste, semblent s'approcher, techniquement, de la définition herméneutique des *passages* de [Rastier 2007 et 2011][13].

6. Résultats : les profondeurs des textes d'Emmanuel Macron

Soit le corpus des discours présidentiels français que nous maitrisons bien [Mayaffre 2021] ; corpus homogène en genre (les discours officiels élyséens par lesquels les présidents s'adressent au pays) ; corpus homogène chronologiquement (le temps de la Vème République, 1958-2021) ; corpus contrastif (de Gaulle, Pompidou, Giscard, Mitterrand, Chirac, Sarkozy, Hollande et Macron) ; corpus volumineux et représentatif (>3m. de mots, environ 400.000 mots par présidents) ; corpus lemmatisé et étiqueté morpho-syntaxiquement, normalisé typographiquement, etc.

Nous entrainons la machine à reconnaitre les textes respectifs des 8 présidents successifs. Après quelques minutes de traitement, le réseau *deep* affiche un taux de reconnaissance satisfaisant de 91 % : sur la base de cet apprentissage profond, n'importe quel texte élyséen inconnu ou anonymisé d'un des présidents sera attribué justement à son auteur avec ce taux marginal d'erreur[14].

Plus précisément, le corpus a été découpé en paragraphes qui deviennent la grandeur d'observation centrale et *meso* du texte, et le taux affiché de 91% signifie que 91% des paragraphes du corpus ont pu être attribués fidèlement à leur président-père.[15] Pour des raisons techniques ces paragraphes s'approchent de la définition linguistique que [Adam 2018-b] théorise dans *Le paragraphe : entre phrases et texte*, mais sans toutefois l'épouser complètement : le traitement exige en effet une fenêtre fixe ; le découpage est donc grossier et mécanique, même si l'empan de 50 mots que nous avons choisi ici représente la taille moyenne d'un paragraphe typographique dans le corpus.

Sur la base de cet apprentissage, la démarche qui nous intéresse est de comprendre les phénomènes de textualisation propres aux auteurs, à Macron ou à de Gaulle, à Chirac ou à Mitterrand ; ce qui *fait texte* pour les uns, ce qui *fait texte* pour les autres. Dit autrement, le logiciel Hyperbase, dans sa version *deep*, révèle les passages les plus significatifs – que nous avons appelé *passages-clefs* [Vanni *et al.* 2020] – et au sein de ces passages les motifs textuels qui ont permis d'interpréter un texte comme macronien ou gaullien, chiraquien ou mitterrandien.

Examinons une sortie machine qui sélectionne un fragment (50 mots donc) comme extrêmement typique du texte de Macron. Le réseau de neurones artificiels active les zones du texte saillantes, aux trois niveaux linguistiques imbriqués : formes, lemmes, code grammaticaux (surlignés dans l'illustration). Tous les ressorts de la prose de Macron semblent alors être mis au jour.

> ⋯ **permettre** aux **acteur européen** d'émerger dans un marché loyal et qui VER:FUTUR aussi de compenser les *profondes désorganisations* sur l'économie traditionnelle que DET:DEM *transformation* parfois **créer**. Les *grandes plateformes numériques*, la protection des *données* sont au cœur de **notre** *souveraineté* à cet égard. Et⋯

[13] Rappelons, avant d'y revenir, la définition élémentaire dans Rastier, 2007, §6 note 4 : « Le passage est une grandeur [du texte] établie ou reconnue par l'analyse… » ; établissement et reconnaissance par contextualisation au sein du corpus et *via* un parcours interprétatif, s'entend.

[14] Les protocoles méthodologiques deep learning sont très rigoureusement définis par les informaticiens. Le corpus d'apprentissage est composé d'un set d'entrainement proprement dit et d'un set de vérification [Vanni et Precioso 2021] c'est-à-dire de textes anonymisés, exclus de l'entrainement, sur lesquels la machine doit valider en aveugle ses performances.

[15] Que signifie, dès lors, les 9% mal attribués ? Soit la machine touche à ses limites, soit un président (Macron par exemple) a « emprunté » ou « empreinté » quelques paragraphes aux autres présidents (à de Gaulle ou à Mitterrand par exemple) selon le grand paradigme de la polyphonie, de l'intertextualité ou du discours rapporté (cf. session suivante). Ces paragraphes « empreintés » sont donc attribués par la machine, fautivement mais non sans pertinence, au président-inspirateur (cf. *infra*).

Illustration 4. Un passage-clef de Macron – les items surlignés sont les zones du texte activées par le réseau qui ont permis à la machine de reconnaitre la prose du président (en majuscule les codes, en gras les lemmes, en italique les formes)

Les motifs lexico-grammaticaux d'une rhétorique du processus, du mouvement, de la promesse sont repérés chez le leader *En Marche !* À l'analyse, c'est elle qui effectivement signe le macronisme en discours [Mayaffre 2021]. Par exemple, les verbes (lemmes) encourageants « permettre » ou « créer » sont typiques, dans un passage où le deep learning repère un verbe au futur promissif qui renforce la projection vers l'avant. Le mot « transformation » (et à travers lui beaucoup de mots en *.tion) est, comme on le sait, la clef de voûte statistique de ce discours en mouvement (illustration 5), et nous remarquons que le réseau repère ici l'enchainement « déterminant démonstratif + 'transformation' » qui permet à Macron de surligner (la monstration, sinon la démonstration) des innovations souhaitées[16]. Les « plateformes numériques », renforcées par l'adjectif « grandes », sont également typiques d'un président *high tech* qui promet le progrès et la modernité. À propos de renforcement, la convolution remarque encore « profonde » qui qualifie « désorganisation » : de fait, l'adjectif « profond » est l'adjectif préféré de Macron surtout quand il qualifie le mouvement (*transformation profonde*, *innovation profonde*…, ou *refondation en profondeur*, etc.) (illustration 5). Et toujours, car le passage repéré par la machine est magistral, la « souveraineté » (*souveraineté numérique* s'entend) signe la fin de l'extrait et apparait à l'analyse comme une préoccupation majeure des discours de Macron ; là où les « acteurs européens » en début de passage renvoient de manière caricaturale, en un syntagme unique, à la double posture fondamentale du macronisme : le pragmatisme affiché ou l'action entrepreneuriale (*vs.* l'idéologie ou la politique) symbolisés ici par le mot-clef « acteurs » (*vs.* les « citoyens » ou le « peuple » ou les « ouvriers »…), et sa foi dans l'Europe, de sa campagne électorale de 2017 jusqu'à sa politique vaccinale de 2021, ici verbalisée par « européenne ».

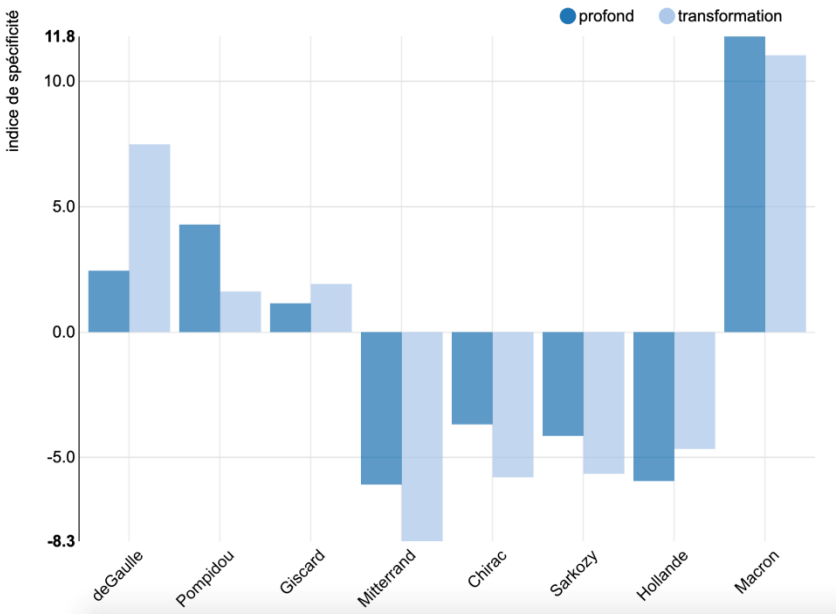

Illustration 5. Sur-utilisation statistique de « profond » et « transformation » par Macron

Les spécialistes d'Emmanuel Macron ainsi que ceux de logométrie remarqueront, au fond, que la plupart des mots ou codes grammaticaux activés par le réseau IA sont des mots statistiques de Macron, à l'image de l'illustration que nous avons donnée de « transformation » ou de « profond » (illustration 5). Ce passage clef serait ainsi, peut-

[16] Nous ne pouvons développer ici. Dans un discours dynamique mais non abouti, Macron célèbre « la transformation » (sans complément de nom) ou utilise de manière intransitive le verbe « transformer » (sans complément d'objet : *je veux transformer*.). Dans ce cadre discursif qui fait de la transformation une finalité sans fin, le président semble souvent utiliser de manière rhétorique ou circulaire le démonstratif (*cette transformation*) : il feint ainsi de renvoyer à une transformation qui, précisément, n'aura jamais été définie auparavant dans le discours autrement que par un absolu. (Pour un développement approfondi, nt. la monovalence du verbe « transformer », voir *Macron ou le mystère du verbe. Ses discours décryptés par la machine*. Mayaffre 2021).

être, seulement une somme de *spécificités*[17]. Pourtant, c'est bien la combinaison de ces mots (la convolution) c'est-à-dire leur contextualisation (on pourrait aussi dire leurs « cooccurrences » ou leurs « collocations » si l'on veut rester dans un cadre statistique) qui fait sens. Ces mots *font texte* chez Macron par leur *engrenage* sur la chaine, dans ce passage précis, et pas nécessairement ailleurs. Par exemple, le futur surligné par la machine n'est pas statistiquement très marqué, mais devient pertinent, chez Macron et dans ce passage, lorsqu'il s'agit de promouvoir la « transformation » ou de compenser une « désorganisation ». De même, « grandes » n'est aucunement un adjectif macronien (il est plus utilisé par Giscard, par Mitterrand ou par Chirac), mais devient pertinent, chez Macron, pour qualifier la « révolution numérique » que le président de la *start up nation* appelle de ses vœux. C'est en cela que les mots et motifs remarqués semblent répondre à la définition de « passage » proposé par Rastier : ils font sens non pas en eux-mêmes pour eux-mêmes, mais à point nommé, dans le cadre d'un contexte d'utilisation – « parce que seuls les contextes sont constituants » [Rastier 2020 : 147] – , c'est-à-dire, plus généralement, dans le cadre d'un parcours de lecture, que la machine objective non seulement localement (les « unités » dans leurs environnements immédiats) mais globalement (puisque toute sortie-machine se trouve informée par le corpus d'apprentissage en général).

7. Du texte à l'intertexte

Déplaçons pour finir la question vers le haut avec Lotman ou Voloshinov, Kristeva ou Barthes : le texte profond, c'est l'intertexte.

La profondeur (*deep*) d'un texte, ou son épaisseur, semble en effet devoir se mesurer par ses échos intratextuels (et la statistique cooccurrentielle et la convolution sont alors de précieux outils) mais plus encore par ses résonnances intertextuelles. Un texte apparent renvoie dans son substrat à d'autres : c'est une échographie du texte que le deep learning peut réaliser, en outillant ainsi un concept au cœur d'une littérature scientifique pluri décennale[18].

Nous avons plaidé en effet ailleurs [Mayaffre *et al* 2020-a et Mayaffre *et al* 2020-b] que le deep learning permet de découdre le texte-patchwork ou révéler les encres successives du palimpseste que décrit Genette en 1982, et que Ute Heidmann reprend dans plusieurs tête de chapitre [Heidmann et Adam, 2010, *Le petit chaperon rouge palimpseste* (pp. 81-112) ; *La barbe bleue palimpseste* pp. 113-152 ; voir également, Adam 2018-a, Deuxième partie - *Variations intertextuelles*, pp. 207-318].

Illustrons par une sortie-machine inédite, le parcours théorique et méthodologique.

L'algorithme est entrainé à bien reconnaitre nos présidents et est susceptible d'extraire les zones textuelles typiques des uns et des autres (*supra*). Nous lui soumettons maintenant un texte jusqu'ici inconnu : le dernier discours de vœux d'Emmanuel Macron pour l'année 2021.

Sans erreur, HyperDeep reconnait ce texte inédit comme un texte de Macron ; nous voici rassurés. Seulement, il attribue fautivement certains passages à d'autres présidents qu'au président Macron. Ces passages, minoritaires, mal attribués, ont donc pour la machine une empreinte gaullienne ou pompidolienne, mitterrandienne ou hollandaise : c'est, selon nous, l'intertexte révélé (au sens des révélateurs chimiques) des discours. Pour illustration, l'extrait suivant du discours de Macron est attribué à Pompidou en fonction des observables surlignés (illustration 6).

> … d'une NOM *européenne*. Restons **ce** *peuple uni* VIRGULE ADJ VIRGULE fier de son *histoire*, de ses NOM VIRGULE, de sa **culture**, confiant *dans* l'*avenir* et le **progrès**, sûr de son *talent* et de son énergie et *ambitieux* pour lui-même…

Illustration 6. Extrait de Macron attribué à Pompidou dans le discours des vœux du 31 décembre 2020 (en majuscule les codes, en gras les lemmes, en italique les formes qui ont été activés par le réseau de neurones artificiels)

À l'analyse, nous pouvons facilement argumenter que la prose macronnienne imite subtilement ici la prose pompidolienne. D'abord la phrase est très nominale

[17] Spécificité dans le sens technique de la statistique textuelle depuis les années 1980 (indice de sur-utilisation d'un terme par un locuteur par rapport à l'usage moyen du corpus).

[18] Le concept d'intertextualité est traité dans des centaines de références. Son histoire a pu faire l'objet de tentative de synthèse [par ex. N. Limat-Letellier, « Historique du concept d'intertextualité », *in* Marie Miguet-Ollagnier et Nathalie Limat-Letellier (dir.), *L'intertextualité*, PU de Franche Comté, 2002, pp. 17-64.]

comme le discours de Pompidou était le plus souvent nominal. Ensuite, Macron convoque la « culture », le « progrès », « l'histoire », le « peuple » « l'avenir » : autant de mots que l'anthologiste de la poésie française, attaché à Malraux comme à de Gaulle, aimait convoquer dans des discours souvent épiques. La structure « NOM + européenne » (chez Macron sous la forme de « renaissance européenne ») est également une structure forte du discours de Pompidou qui parle plus que les autres dans le corpus élyséen de la « foi européenne », « politique européenne », « construction européenne », « agriculture européenne, « union européenne », « identité européenne », « défense européenne ». L'énumération d'adjectifs, encore, avec la mise entre virgules (« peuple uni, ADJ, fier ») se trouve être typiquement pompidolienne : ce type d'énumération participe à la richesse du discours qui caractérise dans le corpus élyséen le discours de l'agrégé de Lettres ; et que Macron cherche parfois à atteindre.

Dans les deux derniers exemples donnés, on remarque que Pompidou n'utilise pas le nom « renaissance » ni le syntagme « renaissance européenne », et il n'utilise pas l'adjectif « solidaire » dans ses énumérations adjectivables, mais l'Intelligence artificielle multi-critères (mot, lemme, code) fait abstraction de ces non-réalisations lexicales pour souligner pertinemment le motif grammatical « NOM+européenne » ou « peuple uni, ADJ, fier ».

Ces échos intertextuels au sens élargi – c'est-à-dire au sens général de *variations* ou *modulations* d'un même motif ou d'une même structure indentifiables, et non pas au sens étroit de citations explicites ou de plagiats – ne surprendront pas : le genre ritualisé des « discours de vœux » et la charge institutionnelle et historique qui pèse sur les présidents, font que la prose de Macron, le 31 décembre, tend à s'inscrire dans des « déjà-dits » que la machine repère, et qui sont attribués – en fonction du corpus de référence analysé – à leur étymon textuel.

7. Au-delà des textes…

Les humanités numériques ne consistent pas seulement dans la numérisation des humanités mais dans l'humanisation du numérique. L'historien, le géographe, le linguiste du texte ne saurait se contenter aujourd'hui de scanner ses archives, ses cartes ou ses corpus ; il aspire à revoir fondamentalement, à la lumière du numérique ses objets, ses protocoles et ses conclusions.

Cependant, humaniser le numérique ne signifie nullement anthropomorphiser la machine en lui prêtant on ne sait quelle sensibilité ou intelligence, ni quelques forces probatoires. Humaniser le numérique, c'est le dé-positiver et adopter face à lui, plus que jamais, une posture herméneutique.

Les sorties-machines, au fond, comme le sens, comme le texte, ne sont pas positives, mais sont elles-mêmes interprétation des données, et restent pour l'essentiel *à interpréter*. La machine ne conclut rien mais re-présente à l'esprit humain les objets éternels des humanités ; ces *représentations numériques* entrent avec dynamisme dans le cercle de nos pratiques interprétatives multi-séculaires.

En la matière, si l'IA peut constituer une révolution, c'est précisément parce qu'elle renonce à réifier, naturaliser, positiver l'objet pour nous interroger sur le processus, c'est-à-dire l'objectivation ; elle nous donne moins à voir le texte pour lui-même, qu'elle participe à la textualisation, *via* le numérique ; c'est-à-dire qu'elle constitue moins pour nous un instrument extérieur à l'objet déjà-là qu'une instrumentation de nos parcours de lecture qui échafaudent le sens. Le deep learning en effet apprend des données textuelles jusqu'à corriger (*backpropagation*) le point de vue initial – or c'est le point de vue qui crée l'objet. Reste alors à l'analyste l'essentiel et le dernier mot : apprendre de cet apprentissage numérique, et comprendre – prendre par devers soi – les nouveaux observables que la machine performe pour faire texte.

Références

ADAM, Jean-Michel (2018-a), *Souvent textes varient*, Paris, Classiques Garnier.

ADAM, Jean-Michel (2018-b), *Le paragraphe : entre phrases et texte*, Paris, Colin, 2018.

ADAM, Jean-Michel (2020), « Postface. Le texte est-il soluble dans le textiel ? », *Corela*, HS-33 (https://journals.openedition.org/corela/11938)

ADAM, Jean-Michel (éd.) (2015). *Faire texte. Frontières textuelles et opérations de textualisation*, Besançon, Presses universitaires de Franche-Comté.

ADAM, Jean-Michel et HEIDMANN, Ute (éds.) (2005), *Sciences du texte et analyse de discours. Enjeux d'une interdisciplinarité*, Genève, Slatkine Érudition.

BACHIMONT, Bruno (1994), *Le contrôle dans les systèmes à base de connaissances. Contribution à l'épistémologie de l'intelligence artificielle*, Paris, Hermès, 1994.

BERNARD, Michel (2003), « "Mes mots à moi" : aperçus lexicométriques sur l'œuvre de Nathalie Sarraute », dans Agnès Fontvielle et Philippe Wahl (dir.), Nathalie Sarraute : du tropisme à la phrase, Lyon, Presses Universitaires de Lyon, pp. 59-69.

BOUZEREAU, Camille (2020), *Doxa et contredoxa dans la construction du territoire discursif du front national (2000-2017)* (thèse de doctorat en Sciences du langage soutenue à Nice le 27 novembre 2020).

BRUNET, Etienne (2016), *Tous comptes faits. Questions linguistiques*, Paris, Champion.

BRUNET, Etienne, LEBART, Ludovic et VANNI, Laurent (2021), « Littérature et intelligence artificielle » *in* Mayaffre, D. et Vanni, L. (éds), *L'intelligence artificielle des textes. Des algorithmes à l'interprétation*, Paris, Champion, pp. 73-130.

COLLOBERT, Ronan and WESTON, Jason (2008), « A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning. », *Proccedings of the 25th International Conference on Machine Learning*, 160–67. ICML'08. New York, USA:ACM.

DEBRAY, Régis (1991), *Cours de médiologie générale*, Paris, Gallimard.

DOUEIHI, Milad (2011), *Pour un humanisme numérique*, Paris, Le Seuil.

ECO, Umberto (1987), *Les limites de l'interprétation*, Paris, Grasset.

GENETTE, Gérard (1982), *Palimpsestes. La Littérature au second degré*, Paris, éditions Le Seuil.

GUARESI, Magali et MAYAFFRE, Damon (2021), « Intelligence artificielle et discours politique », *in* Mayaffre D. et Vanni L. (éds), *L'intelligence artificielle des textes. Des algorithmes à l'interprétation*, Champion, pp. 131-182.

HALLIDAY, Michael and HASAN Ruqaiya (1976), *Cohesion in English*. London, Longman

HEIDMANN, Ute et ADAM, Jean-Michel (2010), *Textualité et intertextualité des contes*. *Perrault, Apulée, La Fontaine, Lhéritier*..., Paris, Garnier.

KIM, Yoon (2014), « Convolutional neural networks for sentence classification », *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746-1751.

LEGALLOIS, Dominique (2006), « Présentation générale. Le texte et le problème de son et ses unités : propositions pour une déclinaison », *Langages*, 163, pp.3-9.

MAGRI, Véronique (2009), *Le Voyage à pas comptés. Pour une poétique du voyage au xixe siècle*, Paris, Champion.

MAYAFFRE, Damon (2002), « Les corpus réflexifs : entre architextualité et hypertextualité », *Corpus*, n°1, pp. 51-69.

MAYAFFRE, Damon (2007), « Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques », *in* Rastier F. et Ballabriga M. (éds), *Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation*, Toulouse, Put, pp. 15-26.

MAYAFFRE Damon, *Vers une herméneutique matérielle numérique. Corpus textuels, Logométrie et Langage politique*, 3 vol. 107, 232, 414 p. Soutenue à Nice, le 30 avril 2010.

MAYAFFRE D. (2012), *Le discours présidentiel*, Paris, Presses de ScPo.

MAYAFFRE Damon (2014), « Plaidoyer en faveur de l'Analyse de Données co(n)Textuelles. Parcours cooccurrentiels dans le discours présidentiel français (1958-2014) », *JADT 2014*, édité par E. Néé, M. Valette, J.-M. Daube et S. Fleury, Paris, Inalco-Sorbonne nouvelle, pp. 15-32. [hal-01181337].

MAYAFFRE, Damon (2021), *Emmanuel Macron ou le mystère du verbe. Ses discours décryptés par la machine*, La Tour d'Aigues, L'Aube.

MAYAFFRE, Damon *et al*. (2020-a), « Du texte à l'intertexte. Le palimpseste Macron au révélateur de l'intelligence artificielle » (avec C. Bouzereau *et al*.), *7ème Congrès Mondial de Linguistique Française*, 2020. [hal-02520224]

MAYAFFRE, Damon *et al*. (2020-b), « Objectiver l'intertexte ? Emmanuel Macron, deep learning et statistique textuelle », *JADT 2020*. [hal-02894990]

MAYAFFRE, Damon et VANNI, Laurent (éds.) (2021), *L'intelligence artificielle des textes. Des algorithmes à l'interprétation*, Paris, Champion.

MONTE, Michèle *et al*. (2018), *Stylistique & Méthode. Quels paliers de pertinence textuelle*, Lyon, PUL.

PAVEAU, Marie-Anne (2017), *L'analyse du discours numérique*, Paris, Hermann.

RASTIER, François (2001-a), *Arts et sciences du texte*, Paris, PUF.

RASTIER, François (2001-b), Sémiotique et sciences de la culture, *Linx*, n° 44-45, pp. 149-168.

RASTIER, François (2007), « Passages », *Corpus*, n°6, pp. 25-54.

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

RASTIER, François (2020). « Sémiosis et métamorphoses », *Semiotica*, vol. 2020, pp. 145-162.

VANNI, Laurent, DUCOFFE, Mélanie, MAYAFFRE, Damon, PRECIOSO Frédéric *et al*. (2018), « Text Deconvolution Saliency (TDS) : a deep tool box for linguistic analysis », *56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne [hal-01804310].

VANNI, Laurent, CORNELI, Marco, LONGREE, Dominique, MAYAFFRE, Damon and PRECIOSO, Frédéric (2020), « Key Passages : From Statistics to Deep Learning », *Text Analytics. Advances and Challenges*, Cham, Springer, pp. 41-55.

VANNI, Laurent et PRECIOSO, Frédéric (2021), « Deep learning et description des textes. Architecture méthodologique », *in* Mayaffre D. et Vanni V, *L'intelligence artificielle des textes*, Paris, Champion, pp. 15-72.

VIPREY, Jean-Marie (1997), *Dynamique du vocabulaire des Fleurs du mal*, Paris, Champion.

VIPREY, Jean-Marie (2006), « Structure non-séquentielle des textes », *Langages*, n° 163, pp. 71-85.

Damon Mayaffre – CNRS – Université Côte d'Azur
damon.mayaffre@unice.fr
Laurent Vanni – CNRS – Université Côte d'Azur
laurent.vanni@unice.fr

Résumé : Cette contribution revient sur la définition d'un texte et de ses unités en supposant que l'Intelligence artificielle est susceptible de modifier nos représentations et nos parcours de lecture. Elle propose une vulgarisation du deep learning du côté de la linguistique textuelle. Ce faisant, elle revient, à la lumière du numérique, sur quelques notions fondamentales comme la textualité, l'intertextualité ou le cercle herméneutique dans la sémantique de corpus.

Mots-clefs : texte, textualité, intertextualité, deep learning, logométrie, intelligence artificielle, lecture numérique.

Abstract : This contribution returns to the definition of a text and its units by assuming that artificial intelligence is likely to modify our representations and our reading paths. It proposes a popularization of deep learning from the perspective of textual linguistics. In doing so, it returns, in the light of digital technology, to some fundamental notions such as textuality, intertextuality or the hermeneutic circle in corpus semantics.

Keywords: text, textuality, intertextuality, deep learning, logometry, artificial intelligence, digital reading