Usages linguistiques des éléments supplémentaires dans l'Analyse factorielle des correspondances

Damon Mayaffre, Laurent Vanni

Usages linguistiques des éléments supplémentaires dans l'Analyse factorielle des correspondances

Damon Mayaffre[1], Laurent Vanni[2]

[1]CNRS – Université Côte d'Azur – damon.mayaffre@unice.fr

[2]CNRS – Université Côte d'Azur – laurent.mayaffre@unice.fr

Abstract

This contribution shows the interest of supplementary variables with Correspondence Analysis (CA). From a CA vector space crossing the main morpho-syntactic categories and the French Presidents of the fifth Republic, we project the lemma "indiquer" (to indicate) as a supplementary variable. Will it be located in the "verb subspace" of the graph? And if not, what should we conclude from this counterintuitive positioning? Beyond this example, it is the linguistic homogeneity of the rows of the contingency table (words, lemmas, grammatical categories, etc.) that we question, by projecting, in more or less interpretable ways, other linguistic elements into additional elements. The contribution ends with an open discussion on the complementarity of CA and deep neural networks.

Keywords: CA, supplementary variables, illustrative elements, mopho-syntax, political discourse, Hyperbase

Résumé

Cette contribution montre l'intérêt des éléments supplémentaires (ou éléments illustratifs) dans l'AFC. Sur un plan factoriel croisant les principales catégories morpho-syntaxiques et les présidents de la République, nous projetons en élément supplémentaire le lemme « indiquer ». Celui-ci se situera-t-il sur le graphique dans le sous-espace de la catégorie verbale ? Et si non, que faut-il conclure de ce positionnement contre-intuitif ? Au-delà de cet exemple, c'est l'homogénéité linguistique des éléments actifs du tableau de contingence, convoqués en ligne (mots, lemmes, catégories grammaticales, etc.), que nous interrogeons, en projetant, de manière plus ou moins interprétable, d'autres éléments linguistes en éléments supplémentaires. La contribution se termine par une discussion ouverte sur la complémentarité de l'AFC et les réseaux de neurones.

Mots clés : AFC, éléments supplémentaires, éléments illustratifs, morphosyntaxe, discours politique, Hyperbase

1. Introduction

Le crédit scientifique de l'ADT s'est pour beaucoup construit autour de la révolution benzécrienne (Benzécri 1973, 1981). Devant le potentiel heuristique de l'Analyse Factorielle des Correspondances (AFC), jamais démenti en matière d'analyse des textes, la communauté a traité des milliers de tableaux de contingence avec fruit, le plus souvent, reconnaissons-le, en ne croisant que les deux premiers facteurs (l'axe 1 et 2) et rarement en utilisant des outils de contrôle comme le *bootstrapping* et les ellipses de confiance. (Sur la révolution benzécrienne voir la conférence d'ouverture des JADTs 2016 : Beaudouin 2016).

Cette contribution entend revenir sur un des raffinements possibles de la méthode AFC, entendons l'usage essentiel et pourtant insuffisamment répandu des « éléments supplémentaires » (aussi nommés « éléments illustratifs »).

Qu'est-ce que le linguiste peut espérer du traitement des éléments supplémentaires ? Sont-ils interprétables linguistiquement ? Quelles sont leurs plus-values dans l'analyse et l'interprétation des textes que nous lui soumettons ? L'implémentation des éléments supplémentaires dans le logiciel *Hyperbase*, appliqués à un gros corpus des discours

présidentiel sous la Vème République permet d'apporter des éléments de réponses qui viennent s'additionner à ceux que présentent par exemple (Lebart, Piron et Steiner 2003) ou plus récemment (Lebart, Pincemin et Poudat 2019). En fin de contribution, les résultats AFC obtenus seront confrontés à des sorties-machines issues de réseaux de neurones artificiels, d'essence mathématique complémentaire.

2. AFC et éléments supplémentaires : bref rappel de l'usage commun

Dans leur ouvrage de synthèse sur l'ADT, (Lebart, Pincemin et Poudat 2019 : 155) résument plusieurs décennies de recherche, l'AFC et les éléments supplémentaires :

> *L'analyse [l'AFC]… permet de trouver des sous-espaces de visualisation des proximités entre variables et observations. Elle s'appuie, pour cela, sur des éléments (variables et observations) appelés* éléments actifs*. Elle permet aussi de positionner, dans ce sous-espace, des éléments (points-lignes ou points colonnes du tableau de données) n'ayant pas participé à l'analyse qui sont des* éléments supplémentaires *ou* illustratifs *(Gower, 1968).*

Ainsi dans un tableau de contingence qui croise, traditionnellement, les mots en ligne et les locuteurs en colonne, c'est-à-dire les éléments linguistiques observés en ligne (les mots donc, mais possiblement les lemmes, les catégories morpho-syntaxiques, etc.) et les variables ou métadonnées en colonne (les auteurs donc, mais possiblement les dates d'écriture, le genre de textes, etc.) l'usage le plus connu des éléments supplémentaires consiste à projeter *a posteriori*, sur le plan factoriel, une métadonnée ou variable (une colonne) « supplémentaire », étrangère au calcul qui a présidé à l'établissement du plan.

De la même manière, la projection d'un mot ou individu linguistique (une ligne) supplémentaire est possible (cf. section 3)

Ainsi une AFC du vocabulaire (en ligne) d'un corpus textuel littéraire organisé selon une première métadonnée « genre des textes » (en colonne) montrera sur le plan et dans les quadrants la proximité ou l'éloignement des mots considérés et des différents genres de discours analysés (roman, nouvelle, essai, etc.). L'ajout en éléments supplémentaires d'une autre métadonnée comme un « auteur », par exemple, permettra de projeter, *a posteriori*, dans l'espace factoriel précédemment créé, la variable « Hugo », « Proust » ou « Flaubert ».

3. Usages linguistiques

Cette contribution entend projeter en élément supplémentaire non pas une métadonnée supplémentaire, c'est-à-dire une colonne additive et inactive du tableau de contingence initial (dates des textes, genres des textes, locuteurs des textes, etc.) mais elle entend projeter un individu linguistique supplémentaire c'est-à-dire une ligne additive et inactive (un mot supplémentaire donc ou un lemme nouveau, une étiquette morphosyntaxique ajoutée, etc.).

Le principe mathématique de la projection comme éléments supplémentaires d'une colonne et d'une ligne est identique, comme le décrit (Lebart, Morineau et Piron 2000, 42-45, 99-100). Mais notre propos concerne l'usage et l'interprétation linguistique espérée en cas de projection en élément supplémentaire d'un trait linguistique ajouté sur un plan factoriel préalablement calculé sur d'autres traits linguistiques initiaux.

3.1. Morpho-syntaxe et lexique

En ADT, établir le plan factoriel des catégories morphosyntaxiques ou *parts of speech* d'un corpus de textes donne des indications descriptives essentielles au linguiste, comme ici à propos du corpus présidentiel français de 1958 à 2024 (figure 1).

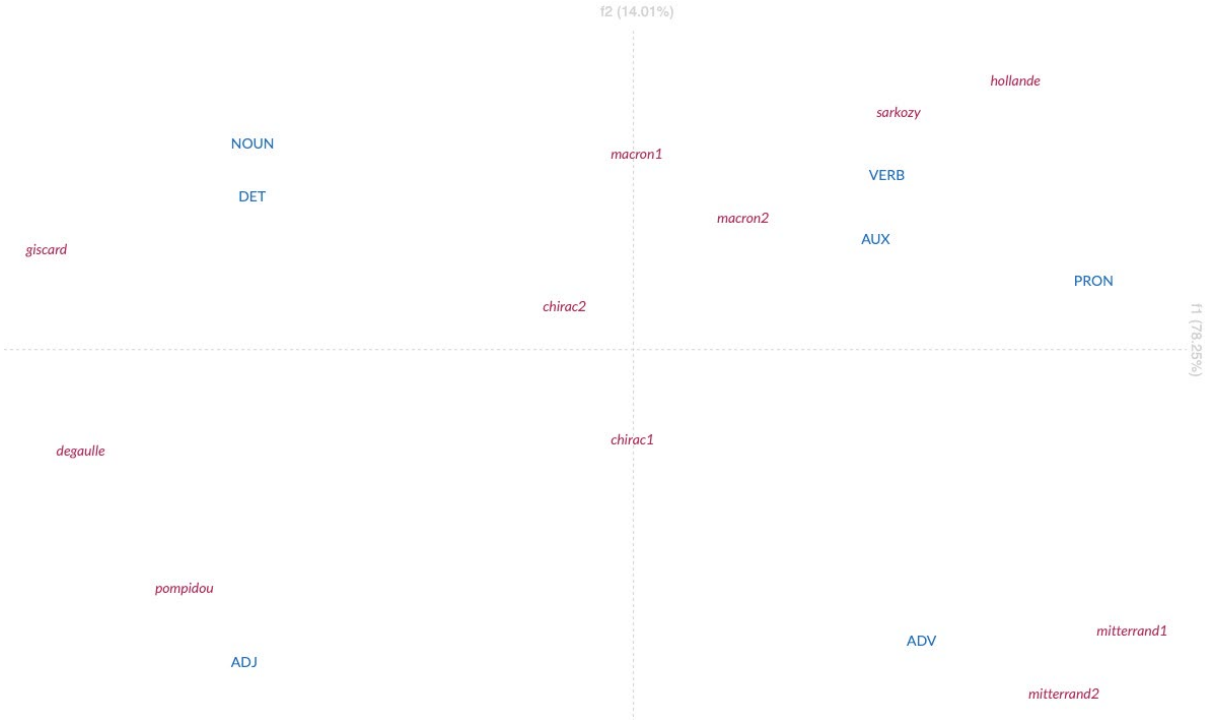

Figure 1. AFC des parts of speech *X présidents (noun=noms ; det=déterminants ; adj=adjectifs ; verb=verbes ; aux=auxiliaires ; pron = pronoms ; adv=adverbes)*

Sur divers corpus en effet, [Brunet : 2016] a montré que les textes s'inscrivaient le plus souvent entre deux pôles morphosyntaxiques opposés : les discours nominaux *versus* les discours verbaux. Ces discours opposés répondent à des caractéristiques statistiques que l'AFC permet d'explorer facilement. Les locuteurs prononçant/écrivant des discours nominaux sur-utilisent non seulement les noms, mais les catégories associées comme les déterminants et les adjectifs. Les locuteurs produisant des discours verbaux sur-utilisent quant à eux les verbes, les pronoms et les adverbes. Cette dichotomie linguistique tranchée se visualise en général comme une évidence sur l'axe 1 (comme sur notre figure 1), quand l'axe 2 mentionne pour sa part des oppositions secondaires plus subtiles ou plus particulières comme celle des adjectifs ou des adverbes avec les autres parties du discours (sur la figure 1, l'adjectif semble bien être statistiquement au discours nominal ce que l'adverbe est au discours verbal).

Quoi qu'il en soit nous avons montré aux JADTs 2006 que cette tonalité « rhétorico-grammaticale » fondamentale des discours (discours nominaux *versus* discours verbaux) permettait d'apprécier globalement le corpus politique des présidents sous la Vème République. Plus loin, nous avons proposé qu'une fois établie, elle devait être prise en compte dans l'appréciation statistique de la distribution des autres éléments linguistiques du corpus. Par exemple, dans le calcul des spécificités lexicales, la probabilité d'utilisation des verbes « aimer », « partir », « aller », etc. gagne à être calculée en fonction du nombre de verbes que le locuteur emploie et non en fonction du nombre de mots en général qu'il utilise (Mayaffre 2006).

C'est dans cette logique que nous nous proposons ici de projeter sur le plan factoriel morpho-syntaxique précédemment calculé (figure 1), un lemme en élément supplémentaire : le verbe « indiquer » (exemple lexical que nous avions pris aux JADTs 2006) (figure 1-bis)

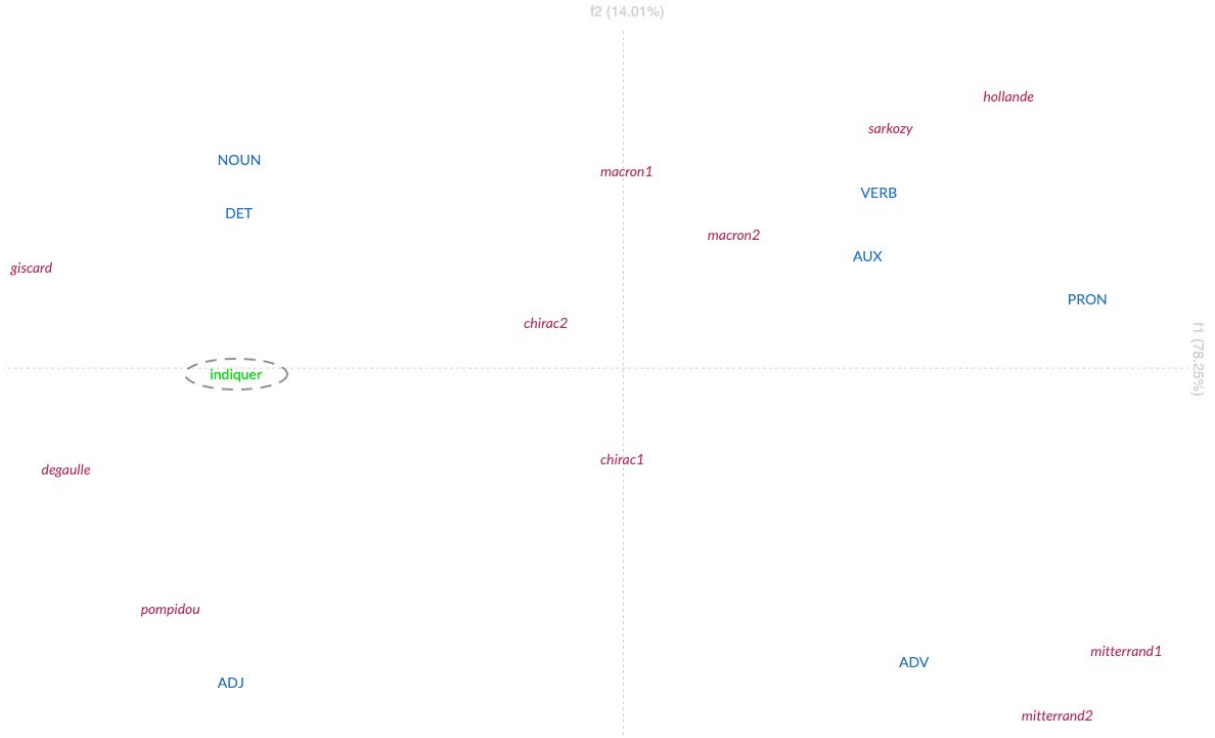

Figure 1-bis. AFC des parts of speech *X présidents – le lemme « indiquer » est projeté en élément supplémentaire*

Dans l'espace bien établi des verbes (à droite de l'axe 1) *versus* des noms (à gauche de l'axe 1), le verbe « indiquer » se situe à l'opposé de ce que l'on était en droit d'attendre (dans le détail, remarquons qu'il est également, légèrement, en contradiction avec sa position attendue sur l'axe 2).

L'interprétation devient puissante linguistiquement : dans le corpus « indiquer » ne répond pas au même fonctionnement statistique – donc sans doute au même usage linguistique – que sa catégorie morpho-syntaxique.

Les discours nominaux de Giscard, Pompidou et de Gaulle (à gauche de l'axe, avec les noms, les adjectifs, les déterminants, mais aussi les conjonctions de coordination) peuvent être qualifiés de *discours didactiques* énonciativement distendus (*versus* les *discours polémiques* tendus) selon la dichotomie de (Guespin 1976). Le président, investi des plus hautes fonctions, expose, explique, démontre, déclame magistralement son discours en se privant par exemple des verbes modaux éristiques très fréquents dans le discours politique (je *veux*, nous *devons*, la France *peut*, etc.) ou des verbes énonciatifs sur-utilisés (je vous *dis*, il faut *répéter*, nous *pensons*, etc.). C'est un discours de type « récit » selon l'historique typologie de (Benveniste 1970). Plus loin nous pouvons parler d'une expression politique réifiée : celle du constat, du concept, de l'idée (la France, la République, l'Etat, etc.) et non celle du combat, du leader et de l'engagement de l'orateur (je vous promets, nous réussirons, nous allons gagner, etc.).

Seulement, force est de constater que dans le cadre de ce discours nominal et didactique, le verbe « indiquer » (et quelques autres sans doute) fait exception. En retournant au texte, on peut constater que Giscard en particulier ne *pense* pas, ne *dit* pas, n'*affirme* pas, ne *veut* pas, etc. : il *indique*. Ainsi la description du discours s'affine et gagne en cohérence : le président pédagogue

enseigne des faits et *indique* aux Français des vérités qui semblent s'imposer à tous, sans autre tension verbale énonciative.

3.1. Du mot au syntagme

L'ADT discute depuis les origines des observables linguistiques pertinents des textes (ceux que l'on pourrait convoquer en ligne dans le tableau de contingence) ; en cela l'ADT ne fait que pointer la difficulté de la linguistique textuelle qui ne dispose pas de « textème » aussi naturel que le « phonème » peut l'être pour la phonologie.

Les mots se sont néanmoins empiriquement imposés, et la majorité des études en ADT entrent dans le corpus par les formes graphiques. Nous produisons ainsi ci-dessous une AFC croisant les 100 formes graphiques les plus fréquentes du corpus en ligne et les présidents de la république en colonnes.

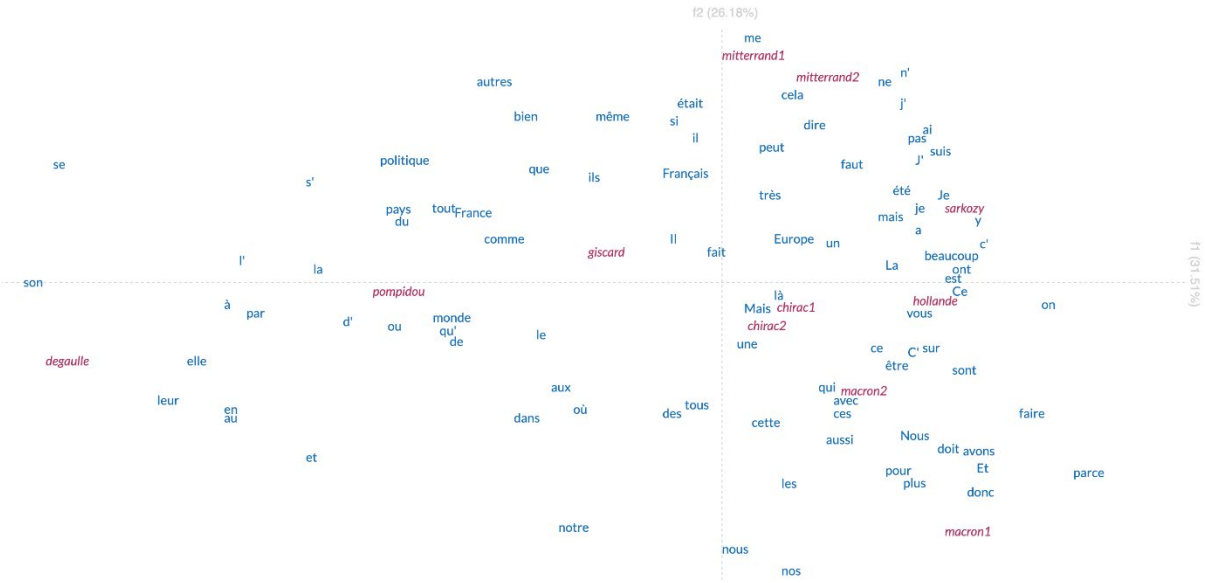

Figure 2. AFC des 100 mots les plus fréquents X présidents

L'espace vectoriel construit est aisément interprétable pour l'analyste du discours. Il oppose *grosso modo* les présidents du passé et les présidents plus récents autour de quelques mots comme « Etat », « France », « situation » etc. d'un côté et « travail » « vouloir », « aller » etc. de l'autre. Dans le cadre de la visualisation de cette opposition chronologique allant de de Gaulle à Macron et marquée par le retournement des années 1980, le poids des pronoms premières personnes singulier (« je », « Je », « J' », « j' ») et pluriel (« nous » voire « on », « nos ») parait non négligeable : les présidents récents depuis Mitterrand n'hésitent pas à utiliser le « je » ou le « nous » pour porter l'énonciation, jouer de l'argument d'autorité ou célébrer le tous ensemble, lorsque de Gaulle, Pompidou ou Giscard faisaient l'économie de ces marques énonciatives.

Seulement, les mots isolés – même rassemblés dans un tableau de contingence et *mis en correspondance ou en corrélation* – peuvent être trompeurs. Et si le mot peut être considéré comme le « textème » le plus commode pour entrer dans le corpus, les mots en cotexte c'est-à-dire les n-grams, segments répétés, collocations ou simplement syntagmes doivent souvent compléter l'analyse.

Projetons donc dans l'espace [100 mots les plus fréquents X présidents] précédemment créé, le syntagme verbal « je/nous croire » pour vérifier son positionnement (figure 2-bis)

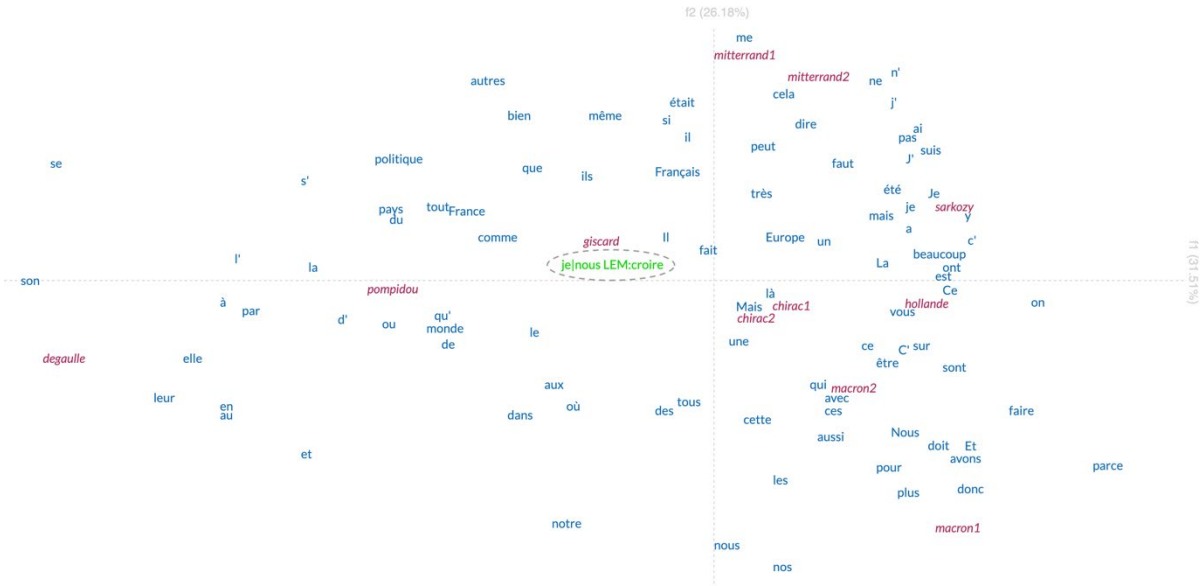

Figure 2-bis. AFC des 100 mots les plus fréquents X présidents – le syntagme « je/nous croire » est projeté en élément supplémentaire

A l'image de « indiquer » dans l'AFC précédente, le syntagme ne se situe pas dans le sous-espace auquel il semblait être dévolu, celui des « je », des « j' » ou des « nous ». Grossièrement : un de Gaulle, un Giscard ou un Pompidou sous-utilisaient massivement le « je » et le « nous », mais - observation d'autant plus remarquable - ils sur-utilisaient le « je crois », « je croyais » ou le « nous croyons » dans une énonciation à la fois plus modeste (*croire* est plus modeste que *savoir* ou *vouloir*) et plus optimiste durant les 30 Glorieuses (*croire* c'est aussi *espérer*).

4. Discussion

L'AFC benzécrienne reste encore aujourd'hui sans doute un des outils les plus heuristiques de l'ADT. Avec la projection d'éléments supplémentaires en ligne, l'analyse des correspondances gagne en puissance en permettant l'examen du comportement d'observables linguistiques sélectionnés. De la même façon que les variables illustratives en colonne permettent d'expliquer certaines métadonnées du tableau de contingence, l'ajout de mots ou syntagmes illustratifs en ligne permet de questionner un espace linguistique figé *a priori* (ex : Noms *versus* Verbes) dans lequel l'élément pourra être (ré)interprété.

Cette technique reste néanmoins contrainte par l'intuition du chercheur. La sélection puis la projection de mots illustratifs répondent à des hypothèses fortes, même si potentiellement suggérées par d'autres outils statistiques tel que le calcul des spécificités lexicales. La méthode demande d'une part d'établir un plan factoriel particulier (choix du tableau de contingence) et d'autre part de définir les observables qui seraient pertinents à positionner sur le graphique.

Pour repousser les limites des outils d'ADT classiques, il apparait aujourd'hui nécessaire de s'affranchir au maximum des *a priori* linguistiques visant à discriminer l'information. Le texte gagnerait à être traité dans son intégralité et les algorithmes seraient alors capables d'interroger l'ensemble des phénomènes en œuvre dans le corpus. Cela reviendrait à produire une AFC sur un tableau de contingence vertigineux, croisant une métadonnée avec non seulement le Tableau

Lexical Entier (TLE), mais aussi Tableau Grammatical Entier (TGE), mais encore l'ensemble des associations entre les mots qu'elles soient locales (syntagmes) ou distantes (champs sémantiques, paradigmes, …). Une AFC donc, où l'interrogation d'éléments supplémentaires correspondrait à projeter un texte nouveau (non inclus dans le tableau entier de contingence) en permettant d'une part de prédire l'appartenance ou proximité du nouveau texte avec un ensemble (sous-espace ou regroupement de textes) et d'autre part de projeter l'ensemble des observables linguistiques du texte sur le plan factoriel constitué.

En l'état, les analyses classiques ne permettent pas d'établir un tableau de contingence suffisamment large pour représenter tous les phénomènes linguistiques – si tant est que nous les connaissions – en œuvre dans les textes. De plus, la visualisation de l'information, même compressée sur deux axes principaux, reste difficile.

C'est pourquoi, nous nous proposons de contourner la difficulté en extrapolant l'usage de l'AFC aux réseaux de neurones profonds.

Depuis (Lebart 1997) nous savons en effet que l'AFC peut être considérée comme un réseau de neurones particulier où les lignes et les colonnes du tableau de contingence correspondent aux entrées/sorties du réseau, et où la compression de l'information se situe dans les couches intermédiaires chargées d'encoder la transformation entre les lignes (les entrées) et les colonnes (les sorties). En augmentant le nombre de neurones et de couches cachées et en utilisant certaines stratégies d'apprentissages issues des réseaux neuronaux prédictifs et génératifs, notamment la Convolution (Kim 2014) et la Self-Attention (Vaswani *et al*. 2017), il semble possible d'obtenir un système où la prédiction d'un nouveau texte revient à le placer, en supplément, sur un plan factoriel complexe à *N* dimensions, et l'analyse des couches cachées permet de projeter la position de variables illustratives (nouveaux observables) linguistiques automatiquement détectés par le réseau sur le même plan. Ici le tableau de contingence représente les données d'entrainement[1] nécessaires à la construction de l'espace vectoriel (le poids des neurones intermédiaires), et les éléments supplémentaires sont les textes à prédire (à positionner dans l'espace donc).

À titre d'exemple, la figure 3 montre un usage des réseaux de neurones profond comparable à la projection d'éléments supplémentaires sur une AFC. Le modèle de type *Multichannel Convolutional Transformer* (Vanni *et al*., 2024-*in press*) est entrainé sur un corpus des présidents français (1958-2024) et interrogé sur un discours inconnu (absent du corpus d'apprentissage comme il le serait du tableau actif de contingence), les vœux du président Macron le 31 décembre 2022.

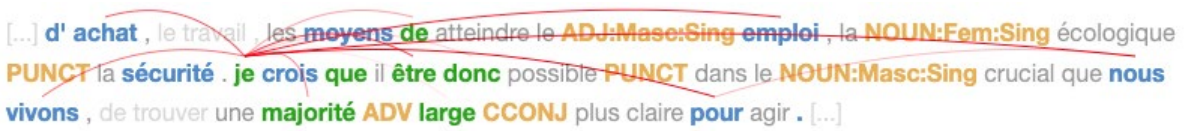

Figure 3 : Passage des vœux du président Macron aux Français (le 31 décembre 2022), attribué à Giscard par le réseau de neurone profond.

Si la grande majorité des segments du texte anonymisé de Macron sont bien attribués au locuteur Macron, la figure 3 représente un segment attribué, par erreur, mais de manière

[1] En réalité le corpus est découpé en segments de taille fixe pour pouvoir être manipulé par le réseau de neurones. La taille des segments fait partie des hyperparamètres qui définissent les bornes de l'analyse permise par le réseau.

appropriée linguistiquement au regard des emprunts intertextuels, au locuteur Giscard. Les marqueurs giscardiens détectés dans l'extrait de Macron nous renvoient en effet pertinemment à l'exemple 2 (section 3.1). Le syntagme « je crois » est fortement activé par la Self-attention (liens visualisés sur la figure 3), dans un passage plutôt nominal comme le détecte la Convolution (éléments surlignés sur la figure 3). De fait, ces éléments morphosyntaxiques et relations syntagmatiques représentent, dans le corpus, et pour le modèle, les traits caractéristiques non du président Macron mais ceux du président Giscard. La plus-value heuristique apparait importante, et le parallèle avec l'AFC, ici seulement esquissé à la suite de (Lebart 1997) mérite attention. Les marqueurs semblent comparables à des variables illustratives qui n'ont pas contribué à la construction du modèle, et leurs poids (score d'activation convolution/self-attention) les mettent en correspondance avec un président (ici Giscard) dans l'espace vectoriel figé par le modèle (résultat de l'apprentissage).

5. Conclusion

Depuis 50 ans la communauté ADT mesure toute la puissance de l'AFC. Elle doit néanmoins rester attentive au danger qui menace sa mise en pratique : la sur-interprétation.

L'interprétation des *correspondances* entre les lignes et les colonnes, et des corrélations des lignes ou des colonnes entre elles, nécessite d'abord, rappelons-le, que les tableaux de contingence répondent à une homogénéité raisonnable. Rassembler et croiser des observations et des variables trop hétéroclites permet certes de produire un graphique, mais celui-ci devient ininterprétable.

En ce qui concerne les éléments supplémentaires situés en ligne – c'est-à-dire les observations linguistiques – ils doivent nous interroger sur la pertinence de les exclure du calcul factoriel en les considérant comme éléments inactifs. Cette pertinence se mesure in fine à l'aune de considérations linguistiques. Quelle justification linguistique avons-nous de traiter, par exemple, 100 mots en ligne et de considérer le 101ème mot comme un élément supplémentaire, au lieu de le laisser participer au calcul ?[2] La philosophie même de la statistique multidimensionnelle de Benzécri milite plutôt en faveur de traiter ensemble toutes les données lexicales du tableau pour en extraire/compresser l'information la plus synthétique.

Les deux exemples que nous avons donnés montrent donc l'intérêt de l'usage contrôlé linguistiquement des éléments supplémentaires pour approfondir l'analyse d'un corpus textuel.

Dans le premier exemple par lequel nous concluons cette contribution, il nous aurait paru contestable de rassembler en ligne, comme éléments actifs, les grandes *parts of speech* (noms, verbes, adjectifs, etc.) et un certain élément lexical particulier (le verbe « indiquer »). Linguistiquement, l'établissement du plan factoriel morphosyntaxique du corpus n'a que faire d'être formé et possiblement déformé par un élément linguistique d'une autre nature (en l'occurrence lexicale). Du reste, des problèmes logiques et quantitatifs seraient survenus : les occurrences de « indiquer » sont déjà comptabilisées dans la catégorie « Verbe », et compter deux fois, dans un même tableau un élément n'est jamais sans danger, au regard notamment des totaux marginaux produits. De plus, quantitativement, la disproportion des effectifs d'une catégorie grammaticale en général, souvent de grande masse (les Verbes en l'occurrence), et

[2] A moins qu'il ne s'agisse d'exclure ce mot de la table active de contingence, au regard d'un profil trop discriminant. Mais il s'agirait alors d'une raison négative (exclure un *outlier* encombrant des éléments actifs) et non d'une plus-value interprétative positive de la projection d'un élément supplémentaire.

d'un sous-élément de cette catégorie, souvent de petite masse (le verbe « indiquer », en l'occurrence) est si importante qu'elle rend dérisoire leur traitement commun[3].

Ainsi pour reprendre la terminologie de l'AFC, le verbe « indiquer » semble n'avoir pas d'autre droit que d'« illustrer » les variables morphosyntaxiques actives ou constitutives du plan factoriel. En l'occurrence, avec le verbe « indiquer », nous avons choisi non pas une illustration mais une contre-illustration, pour une interprétation discursive du corpus présidentiel français, plus riche et moins attendue.

Bibliographie sélective

Dans leur ouvrage, Lebart L. Morineau A. et Piron M. recensent en bibliographie près de 1000 références sur la Statistique exploratoire multidimentionnelle *(Paris, Dunod, 2000 : 405-428) Aussi, les renvois bibliographiques ci-dessous devront être considérés comme illustratifs.*

Beaudouin V. (2016). Retour aux origines de la statistique textuelle : Benzécri et l'école française d'analyse des données. *JADT 2016 - Proceeding of 13th International Conference on Statistical Analysis of Textual Data*, Vol 1., 17-36.

Benveniste E. (1970). L'appareil formel de l'énonciation, *Langages*, 17, p. 12-18.

Benzécri J.-P. (1973). L'analyse des données. T2 : L'analyse des correspondances. Paris : Dunod.

Benzécri J.-P. (1981). Pratique de l'analyse des données. T3 : Linguistique et lexicologie. Paris : Dunod.

Brunet E. (2016). Tous comptes faits. Questions linguistiques. Paris : Champion, 2016

Gower J.C. (1968). Adding a point to vector diagram in multivariate analysis, *Bimetrika*, 55, 582-585.

Guespin L. (1976). Types de discours ou fonctionnements discursifs ? *Langages*, n°41.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014),* pages 1746–1751, Doha, Qatar. ACL.

Lebart L (1997). Réseaux de neurones et analyse des correspondances. *Modulad*, 18, p. 21-38.

Lebart L. Morineau A. et Piron M. (2000 – 3ème ed), Statistique exploratoire multidimensionnelle. Paris : Dunod.

Lebart L., Piron M. et Steiner J.-F. (2003). La Sémiométrie. Paris : Dunod

Lebart L. Pincemin B. et Poudat C. (2019). Analyse des données textuelles. Québec : Presse de l'Université du Québec.

Mayaffre D. (2006). Faut-il pondérer les spécificités lexicales par la composition grammaticale des textes ? Tests logométriques appliqués au discours présidentiel sous la Ve République, in J-M. Viprey (éd.), *JADT 2006*, Besançon, Presses univ. de Franche-Comté, pp. 677-685. [hal-00554681]

Mayaffre D et Vanni L (éds) (2021), *L'intelligence artificielle des textes. Des algorithmes à l'interprétation*. Paris : Champion.

Vanni L., Mahmoudi H., Longrée D. and Mayaffre D. (2024-in press). Multi-channel Convolutional Transformer and intertextuality: a Latin case study. In Giordano G. and Misuraca M. (Eds.) New Frontiers in Textual Data Analysis, Springer.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., and Polosukhin I. (2017). Attention is all you need. *In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*. Curran Associates Inc., Red Hook, USA, 6000–6010.

[3] Nous avons produit l'AFC du tableau de contingence Présidents X Catégories morphosyntaxiques (figure 1) en maintenant « indiquer » en élément actif. Comme attendu, le poids dérisoire de « indiquer » (*versus* le poids important des grandes catégories grammaticales traitées) l'empêche de participer à l'établissement du plan factoriel : il est situé près du point de croisement des deux premiers axes, sans plus-value interprétative.