Texto! Textes et cultures, Vol. XXIX, nº 3–4 (2024).

L'ADN de l'ADT.
Aux limites de l'interdisciplinarité en analyse des données textuelles[1].

Damon Mayaffre

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

Résumé. — L'ADN de l'ADT tient dans son interdisciplinarité. La description statistique nourrit l'interprétation linguistique. Le quantitatif féconde le qualitatif. Les chiffres dialoguent avec les mots. Cependant, le rendez-vous interdisciplinaire implique une double conscience, celle linguistique de l'objet texte, celle mathématique de l'indice statistique. Dans cette contribution c'est la dimension herméneutique du texte qui est posée, pour mieux convoquer une statistique non pas probatoire mais exploratoire. Les textes, réunis en corpus, sont des parcours interprétatifs : l'ADT est une heuristique qui intègre la mesure, le coefficient, l'arbre ou le vecteur à ces parcours de lecture.

Mots clés. — texte, statistique textuelle, humanités numériques, AFC, textométrie, logometry, sémantique de corpus, herméneutique numérique.

Abstract. — The DNA of ADT lies in its interdisciplinary nature. Statistical description informs linguistic interpretation. The quantitative feeds the qualitative. Numbers interact with words. However, the interdisciplinary approach requires a dual awareness: the linguistic awareness of the text object and the mathematical awareness of the statistical index. In this contribution, the hermeneutical dimension of the text is posited, the better to summon a statistic that is not probative but exploratory. The texts, brought together in a corpus, are interpretative paths: the ADT integrates the measure, the coefficient, the tree or the vector into these reading paths.

Keywords: text, textual statistics, digital humanities, CA, textometry, logometry, corpus semantics, digital hermeneutics.

[1] Conférence inaugurale, JADT 2023 : 17th International Conference on Statistical Analysis of Textual Data.

1. Introduction

L'ADT, sa définition, sa valeur scientifique et ses limites tiennent en un mot principal, hélas aujourd'hui galvaudé : *interdisciplinarité*. Le terme est usé de nos jours à force d'avoir servi ces dernières décennies (figure 1), mais, au détour des années 1950, la puissance heuristique nouvelle du croisement de la linguistique philologique ou textuelle d'un côté et des statistiques ou des mathématiques de l'autre a bel et bien produit un trésor de productions, de thèses, d'ouvrages, d'articles, de colloques dont les JADTs sont depuis un demi-siècle une jolie illustration.

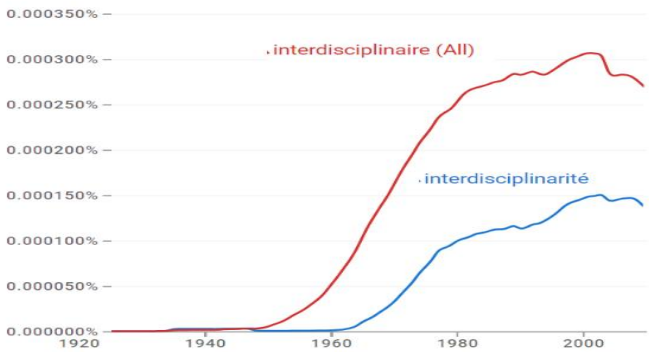

Figure 1. Fréquences relatives de "interdisciplinaire" et "interdisciplinarité" sur N-gramViewer

Elargissons la focale pour mesurer l'ambition. L'ADT fut au milieu du siècle précédent la rencontre interdisciplinaire d'un objet millénaire – le texte – avec un outil prodigieux en devenir – l'ordinateur. C'est-à-dire, la rencontre entre l'histoire ou la culture (puisque l'archive textuelle est la définition de l'histoire, et que la culture humaine est pour l'essentiel portée par le texte) et l'avenir (puisque le numérique est la révolution technologique, la révolution épistémologique et peut-être même la révolution anthropologique en cours).

Plus prosaïquement, selon une expression facile, plusieurs fois empruntée, l'ADT est le rendez-vous interdisciplinaire des chiffres et des lettres, ou la fécondation par le quantitatif du qualitatif ; « Mots chiffrés et déchiffrés » titrent facétieusement des mélanges offerts à Etienne Brunet en 1998.

À ces égards, l'ADT ne saurait être ramenée à une simple pratique, à une simple technè ou à des logiciels qui se multiplient et se perfectionnent. Elle apparait au plus haut niveau comme une *heuristique* qui questionne fondamentalement la production du savoir – particulièrement nos savoirs en SHS ou sciences de la culture – et interroge les disciplines.

Le croisement entre la description statistique et l'interprétation linguistique produit en effet un surplus de connaissance et une plus-value scientifique dont l'ADT peut se prévaloir. Le texte est appréhendé, représenté, interprété dans un parcours de lecture qui allie et entremêle, dans un seul élan, approche numérique et approche oculaire ; approche « digitale » pourrait-on dire puisque le mot, finalement, pointe à la fois par son étymologie la dimension nécessairement humaine ou manuelle de la lecture et, par son usage contemporain, la dimension numérique et quantitative des *digitals humanities*.

Cette contribution ne prétend pas faire l'historique de l'ADT ni celui des JADTs. Elle entend simplement rappeler, dans un discours amoureux de l'Analyse statistique des données textuelles,

la puissance et les contraintes de cette interdisplinarité définitionnelle, gravée dans l'ADN de notre communauté. C'est en cultivant l'interdisplinarité d'un point de vue théorique et pratique que notre communauté scientifique a construit, et préserve aujourd'hui, son identité face à des continents scientifiques anciens, côté SHS, comme la Philologie classique, la Littérature ou l'Analyse du discours traditionnelle et des continents majeurs plus récents dans l'histoire des disciplines comme la Computational linguistics, le NLP ou TAL ou l'Intelligence artificielle. Apologie d'un entre-deux. Hymne à l'intersection. Fertilité interdisciplinaire.

2. Au-delà du langage, le texte

Il ne saurait y avoir d'interdisciplinarité sans disciplines.

Il ne peut donc y avoir d'ADT sans une discipline statistique mature d'un côté, nous y reviendrons, croisée à une discipline linguistique consciente d'elle-même de l'autre : la linguistique textuelle.

Dans les termes de la dichotomie saussurienne dont on connait les limites, l'objet et la fin de l'ADT sont non la Langue mais la Parole, et plus précisément les *textes* que l'on définira avec Jean-Michel Adam comme les formes empiriques des discours (Adam 2020 chap. 1 Introduction l'analyse textuelle des discours : 23-67) ; ou avec François Rastier comme « les objets empiriques de la linguistique », ou les « unités minimales (mais non élémentaires) de la description linguistique » (Rastier 2020 : 11-12). Ces textes, précisons-le d'emblée, seront rassemblés en corpus textuels dans lesquels ils prendront sens ; sans lesquels aucune étude ADT aboutie ne peut se déployer.

Or, si la Langue répond à des règles, les textes relèvent des usages ; si la langue relève, dit-on, de la logique, le texte relève de l'interprétation. L'identité profonde de l'ADT se trouve là, face à notre objet-texte, dans une posture épistémologique fondamentale que François Rastier qualifie de « rhétorico-herméneutique » (*versus* « logico-grammaticale » de la linguistique formelle qui étudie la Langue). Cette identité « rhétorico-herméneutique » doit être rappelée car elle est contre-intuitive pour le néophyte en ADT. Elle semble en effet contrevenir, de prime abord, à l'idée naïve et à l'usage pratique que l'on pourrait (se) faire de l'informatique ou des mathématiques. Elle constitue, pourtant, à nos yeux, le cœur battant et original de nos pratiques.

Objet et finalité, le texte nous oblige en effet par sa dimension herméneutique ; il oblige la communauté ADT techniquement et méthodologiquement, nous le rappellerons, avec une statistique générale adaptée à quelques phénomènes textuels remarquables comme les phénomènes de co-occurrence nécessaires à appréhender pour co(n)textualiser c'est-à-dire sémantiser (cf. *infra* la conférence d'ouverture que nous avions pu faire aux JADTs de Paris : Mayaffre 2014) ; il nous oblige plus fondamentalement dans la posture et dans l'approche.

Il n'existe pas de grammaire de texte comme il y en a une de la phrase. Un texte ne se programme pas, ne s'échantillonne pas, ne se modélise pas, ne se génère pas, au sens de la Grammaire *générative* ou de l'IA *générative* en dépit des tentatives d'apparence réussies de Chat GPT. Un texte se lit – mieux, c'est la lecture qui co-construit le texte –, il se décrit, s'explore et

s'interprète[2]. C'est pourquoi nous avons plaidé modestement ailleurs, en référence à des contributions majeures comme *Introduction à l'herméneutique littéraire* (Szondi 1989) ou *Arts et Sciences du texte* déjà cité (Rastier 2001), en faveur d'une réconciliation de la philologie (numérique) et de l'herméneutique (numérique) : c'est à nos yeux le projet fort, quoique sous-jacent et parfois oublié, de l'ADT (Viprey 2005 ; Mayaffre 2007 et 2010 ; Rastier 2011, De Angelis 2020).

L'ADT en effet établit (philologie numérique) et interprète (herméneutique numérique), dans un même mouvement, des textes numérisés. Cet établissement *et/ou* interprétation du texte se fait grâce à des moyens informatiques *ad hoc* qui combinent traitements statistiques et retour contrôlé au texte.

Seulement, en disant qu'il ne saurait y avoir d'établissement du texte sans son interprétation (affirmation herméneutique) pas plus qu'il ne peut y avoir d'interprétation sans établissement de ce qu'il y a à interpréter (affirmation philologique), nous bornons aux deux extrémités la démarche, et décryptons l'ADN de l'ADT.

Objectiver le parcours interprétatif – D'un côté, du côté de l'herméneutique numérique, l'Analyse de textes assistée par ordinateur, comme on disait autrefois, ne saurait participer au néo-positivisme ambiant qui peut affecter le Traitement automatique des langues (Valette 2016) ou l'Intelligence artificielle lorsqu'elle est mal maitrisée. Loin de toute posture scientiste ou technicisante, l'ADT est étrangère à la prétention prescriptive et probatoire : la statistique est exploratoire, les corrélations ne sont pas causalité, l'effort d'objectivation concerne le parcours interprétatif et non le sens lui-même. En matière de sémantique de texte, la statistique est en effet au mieux confirmatoire (en réalité descriptive et exploratoire) mais jamais probatoire, pour la simple et définitive raison que le sens des textes ne se prouve pas, ne se re-trouve pas (comme s'il était déjà là), n'est pas ontologique. Le sens se construit et s'interprète grâce à un protocole de lecture que l'ADT, précisément, encadre et explicite. L'ambition de l'ADT est donc humaniste (elle s'adresse en tout cas aux Humanités) : il ne s'agit pas d'objectiver le sens mais d'objectiver le parcours interprétatif ; il ne s'agit pas de réifier la vérité supposée des textes, mais d'outiller, de contrôler, d'augmenter une lecture exploratoire ou interprétative.

Entre description à visée interprétative et probation scientiste, la frontière est certes ténue mais elle doit être bien défendue car elle distingue une démarche scientifique d'une démarche qui prétend l'être. On connait des usages, dans notre communauté même, qui ont pu donner à des indices statistiques valeur de preuve, là où il n'y avait matière qu'à interprétation. Or un indice, fût-il statistique, est un « signe qui met sur la trace de quelque chose » non la chose elle-même (Larousse en ligne).

Dès 1984, Etienne Brunet avait prévenu :

[2] En posant le texte comme objet et finalité de l'ADT, et en précisant qu'un texte nécessairement se lit (lecture assistée par ordinateur évidemment) nous réservons à d'autres communautés parentes comme le TAL les traitements automatisés et le *benchmarking* d'algorithmes qui opèrent sans égard aux SHS pour les corpus textuels. Pour l'ADT, les textes ne sont pas prétexte.

> *En matière lexicale la règle statistique ne permet, elle aussi, que la mesure. Il ne s'agit que de décrire, nullement d'expliquer, moins encore de prévoir. (Brunet 1984 : 6)*

Et dans une formulation sans doute trop définitive, François Rastier aujourd'hui va plus loin :

> Taille du corpus et finitude herméneutique. — *Les sorties d'un logiciel ne sont pas des résultats scientifiques, mais des indices qui peuvent revêtir un intérêt dans un processus interprétatif. Elles ne peuvent même pas prétendre à la dignité de fait scientifique, tant qu'elles n'ont pas pu être interprétées. Ce sont donc des questions, plutôt que des réponses ; toutefois, les questions imprévues recèlent par là un potentiel heuristique (Rastier 2020 : 14).*

Dit autrement : peut-on faire confiance à un indice de confiance ? Certes ! Mais poser le texte comme un objet herméneutique doit nous protéger, au plus haut niveau, contre tous les mésusages et éblouissements méthodologiques qui mèneraient à des conclusions hâtives et définitives. Particulièrement, aujourd'hui, poser le texte avec l'ADT dans sa dimension herméneutique doit nous protéger contre l'usage incontrôlé de l'Intelligence artificielle qui délivre, en boite noire[3], des pastiches d'analyses, prétend à des synthèses ou à des abstracts, qui ne sont pourtant que des *avatextes* numériques, probabilistes, non explicités. Nous avons porté avec Laurent Vanni et Dominique Longrée dès les JADTs 2018 à Rome « un regard croisé » sur l'ADT et le deep learning (Vanni, Mayaffre et Longrée 2018) et il est heureux de voir, dans ce millésime 2024, plusieurs contributions alimenter un débat nécessaire et d'avenir.

Finalement, c'est d'un point de vue statistique que Ludovic Lebart, Bénédicte Pincemin et Céline Poudat dans la conclusion du dernier chapitre de *L'Analyse des données textuelles* réfléchissent aux liens et relations existant

> *entre exploration et prévision, entre description et décision, entre observation et modélisation, entre approches supervisées et non supervisées, et ce, en résumant toutes les oppositions (pas exactement équivalentes) par les termes* exploratoire *et* confirmatoire.(Lebart, Pincemin, Poudat 2019 : 412)

C'est d'un point de vue linguistique que nous renchérissons : les textes s'explorent et s'interprètent avant tout, puisque c'est de leur exploration – parcours de lecture interprétatifs dans lesquels l'outil informatique et le traitement statistique sont intégrés – que nait le sens.

Matérialité textuelle – De l'autre côté, du côté de la philologie numérique, le texte doit être établi. Le texte en ADT n'est pas une idée ou un idéal, mais une matérialité. Le texte est la forme empirique de la langue disions-nous. C'est la forme matérielle du discours pourrait-on ici préciser. L'ADT ne saurait en effet traiter autre chose qu'un matériau textuel dûment saisi et constitué en

[3] L'image de la boite noire peut être explicitée avec ChatGPT: avec la boite noire, nous ignorons (i) les corpus d'apprentissage qui s'appliquent à être non référencés ; nous ignorons (ii) le paramétrage des algorithmes et la profondeur (*deep*) des couches cachées qui rendent invérifiables, infalsifiables, ininterprétables les sorties-machines.

corpus numérique ; matériau au sein duquel s'épanouit de manière contrainte le parcours interprétatif outillé.

La matérialité textuelle se retrouve par exemple chez Szondi dans l'affirmation première de la « nécessité du texte » (de Launay (éd). 2013) afin de revendiquer une herméneutique critique. Elle se retrouve pour finir, dans une forme d'ultime concession, chez Eco qui écrit *aux limites de l'interprétation*, que pour qu'il y ait interprétation, il faut qu'il y ait, *d'abord*, quelque chose à interpréter (Eco 1992, *Introduction*).

Et cette herméneutique matérielle du texte, qui rompt avec l'herméneutique religieuse ou avec l'approche romantique des œuvres et des auteurs, qui est valable partout, est rendue impérieuse par l'ADT, c'est-à-dire par l'appariement informatique de notre heuristique[4]. Un ordinateur, même extrapolé en Intelligence artificielle, ne traite que des bits et des octets, c'est-à-dire pour nous des caractères, des lettres, des mots, des signifiants ou signaux numériques ; il ne traite pas de l'idée, du concept, de l'intuition, du sens ou du signifié. Le terme de *saisie* est à ce titre saisissant : loin d'être virtuels, comme pouvaient l'être certains corpus traditionnels désirés ou envisagés – insaisissables –, les corpus textuels numériques, eux, sont tangibles, constitués pratiquement par ce que le chercheur a pu ou non matériellement saisir ; c'est-à-dire établir. A ce titre, nos corpus effectivement saisis nous protègent à la fois contre les fantasmes des corpus imaginaires « en langue » et contre les fantômes des *big data* informes et inexploitables. A ce titre, en revanche, il n'est pas étonnant que l'ADT, en tant qu'herméneutique/philologie numérique, ait pu recevoir un écho fort auprès des littéraires bien sûr, toujours proches de leurs textes, mais aussi des historiens du discours ou linguistes-historiens comme ceux du laboratoire de Saint Cloud (Maurice Tournier, Jacques Guilhaumou, Annie Geffroy, Pierre Fiala, Benoit Habert, Denis Peschanski, etc.) : sensibles à la « matérialité de l'archive », à la « matérialité discursive », à la « matérialité propre des textes », aux « matériaux empiriques » des ressources textuelles ou à la « matérialité propre des énoncés », ils ont précocement compris la force de l'ADT pour décrire et co(n)textualiser la factualité empirique ou matérielle de leur objet, que cela soit sous forme d'un simple calcul de spécificités lexicales ou d'un simple concordancier (les passages entre guillemets sont de J. Guilhaumou 2004).

3. L'urne et ses *au-delà*

Au commencement de l'ADT est souvent considérée, en France, cette affirmation de Pierre Guiraud en 1959, plusieurs fois reproduite depuis, récemment mis en exergue par (Magri 2020 : 3) : « la linguistique est la science statistique type ; les statisticiens le savent bien ; la plupart des linguistes l'ignorent encore ». Et, de fait, 60 ans après, si l'ignorance a certes reculé, certains imaginent encore pouvoir caractériser sans mesurer et qualifier sans quantifier ; comme un océanographe prétendrait rendre compte de son objet – un océan de mots – sans connaitre le

[4] C'est en ce sens que nous avions titré notre *Habilitation à diriger les recherches* en trois mots intimement reliés : *herméneutique matérielle numérique* (Mayaffre 2010).

taux de salinité, la profondeur de ses fosses, la hauteur de ses vagues, le volume ou l'étendue de ses eaux.

Par « science statistique type », Pierre (Guiraud 1954 et 1959) puis Charles (Muller 1964a et 1964b) entendaient au fond que le texte et ses mots, dans une statistique standard, répondraient au schéma d'urne, à l'hypothèse nulle (c'est-à-dire au hasard), à la loi normale, aux distributions gaussiennes, etc. Les mots seraient aléatoirement distribués et uniformément répartis ; ils seraient ainsi dotés d'une identité mesurable et d'une fréquence théorique. Dès lors, le nombre constaté de leurs occurrences (ou fréquence observée) pourrait être évalué par une probabilité, en raffinant statistiquement la soustraction élémentaire mais implacable : *fréquence théorique – fréquence observée*. Au fond, l'espoir de la discipline reposait sur une espérance ; la plus-value interprétative en linguistique se nourrissait de l'espérance mathématique supposée des mots.

De l'urne au corpus – La critique du schéma d'urne est depuis longtemps consommée dans notre communauté : l'urne, comme le « sac de mots », est une représentation imparfaite du texte ; imparfaite mais nécessaire ; nécessaire et heureusement gérée (Brunet 1984).

Il n'existe pas en effet de fréquence théorique en Langue, car la langue théorique, que certains appellent « anglais standard », « italien standard » ou « français standard », n'existe pas ; et sans fréquence théorique, point d'écart par rapport à la moyenne ou au hasard ; point de statistique. Ainsi le nombre d'occurrences du mot « fourchette » est proche de zéro (0 occurrence) dans nos gigas corpus politiques français du XXème et XXIème siècles de Clemenceau à Macron, de Jaurès à Sarkozy. Là où ce nombre est conséquent dans les corpus littéraires de *Frantext* ; et serait plus conséquent encore, par spécialisation lexicale, si nous enregistrions nos conversations domestiques à l'heure du repas. Dès lors, comment espérer calculer pertinemment la fréquence théorique de « fourchette » en langue française ?

Il n'existe donc pas de fréquence théorique en Langue mais il en existe une *en corpus*. C'est l'affirmation fondatrice de l'ADT, nécessaire à l'acceptation de la statistique de Muller et de ses continuateurs. Notre statistique ne cherche pas à témoigner d'une Langue hypostasiée, pas plus qu'elle ne cherche à témoigner de *big data* incontrôlables. Elle consiste, pour l'essentiel et de manière consciente, à calculer les écarts d'un texte particulier par rapport au corpus textuel dans lequel il est plongé.

Il ressort ainsi avec force que l'ADT, contrairement à d'autres domaines connexes de la computationnal linguistics, est fondamentalement une *linguistique en corpus*, une *linguistique sur corpus*, une *linguistique de corpus* : si le texte est son objet, le corpus est sa condition[5].

[5] En soulignant linguistique *de/sur/en* corpus, nous cherchons seulement à imposer, par répétition, l'idée fondamentale : le corpus. Précisons que l'ADT adopte préférentiellement une démarche *corpus-driven* (versus *corpus-based*) : de manière heuristique, la statistique exploratoire fait émerger du corpus des hypothèses de travail ; le corpus est champ d'exploration et de découverte et non un recueil de tests ou d'expérimentation pour une thèse linguistique préalable (Tognini-Bonelli 2001). Ajoutons que les « corpus d'apprentissage » de l'IA générative contreviennent à la définition scientifique de corpus en ADT puisque les textes recueillis ne sont ni référencés, ni critiqués.

Le corpus est la norme ou la référence sans lesquelles nous ne pouvons rien pratiquer. Le corpus est le fond sur lequel les textes prennent forme. Nos parcours sémantiques sont endogènes au corpus : une « sémantique de corpus » selon le titre de l'ouvrage de (Rastier 2011). Notre stylistique est endogène au corpus. Notre statistique est – par définition – endogène au corpus.

Répétons : le corpus est, dans une implacable cohérence épistémologique, la norme herméneutique et la norme statistique pour les chercheurs en ADT. Norme herméneutique, nous ne pouvons y revenir que rapidement : les « corpus réflexifs » (Mayaffre 2002) mettent en dialogue, en *reflet* ou en *réflexion*, les textes entre eux, qui dès lors se contextualisent mutuellement ; les corpus réflexifs permettent d'internaliser les ressources interprétatives ; ils sont la condition de l'interprétation. Selon le principe d'architexte ou d'architextualité énoncé par François Rastier : « tout texte placé dans un corpus en reçoit des déterminations sémantiques, et modifie potentiellement le sens de chacun des textes qui le composent. » (Rastier, 2001 : 92). Norme herméneutique donc, mais norme statistique ou quantitative également : le corpus est l'étalon-mesure qui donne sens au fréquentiel, permet la moyenne et l'écart à la moyenne, autorise le calcul de la probabilité.

Les corpus sont donc la clef, et les meilleurs pionniers ou praticiens de l'ADT ont tous réfléchi à leur nature, leur structure et leur exigence. André Salem par exemple a labouré les particularités des corpus textuels chronologiques pour montrer les dynamiques lexicales caractéristiques qui s'y manifestaient : le *temps lexical*, qui se visualise par une parabole sur une AFC (effet Guttman) et se décrit par des « spécificités chronologiques » que le logiciel Lexico implémente (Salem 2021, 1991, 1988). Etienne Brunet a insisté sur les traits distinctifs des corpus génériques, particulièrement du point de vue de la distribution statistique des traits morphosyntaxiques (Brunet 2016). Bénédicte (Pincemin 2012 ou Rastier et Pincemin 1999) ou Céline (Poudat et Landragin 2017) ont typologisé, plus généralement, les corpus (de référence, de travail, étiquetés, etc.) dans la chaine du traitement. De ces réflexions essentielles, nous ne retiendrons ici que la plus importante, sans laquelle aucune pratique ADT ne semble adulte : la dialectique entre contrastivité et homogénéité. Contrastivité d'abord et avant tout : notre statistique contraste les textes ou les compare, pour décrire et explorer les caractéristiques des uns par rapport aux autres (ou par rapport à l'ensemble). Notons que cette réalité statistique fait directement écho ou implémente, en linguistique, les approches de «comparaison différentielle » pratiquées par exemple par Jean-Michel Adam et Ute (Heidmann 2005a et 2005b) ou le projet de « caractérisation contrastive » qui signe la sémiotique des cultures selon François (Rastier 2001 et 2020.) Homogénéité des corpus ensuite, néanmoins : les contrastes, premiers et nécessaires, n'auront de sens que si l'ensemble (c'est-à-dire le corpus) est cohérent et homogène, et que les textes rassemblés sont comparables. Par-là, l'ADT admet que la qualification des données (les textes dument sélectionnés pour contraster) préside au traitement quantitatif.

Quoi qu'il en soit, nous retombons sur l'idée d'une approche empirique de la langue – les corpus textuels – loin des dogmes et des formalismes. Et tout devient limpide : il n'y a pas de sémantique aboutie en Langue et il n'y a pas de statistique possible en Langue (au-delà de quelques régularités zipféennes). Restent seulement les textes, structurés en corpus, à décrire, à

mesurer, à comparer, à explorer, à interpréter : l'Analyse statistique des données textuelles tient son cahier des charges.

Du token au tableau de données textuelles – La critique de l'urne, en tant qu'approximation rapide du texte, a été poussée également dans une autre direction, et ceci précocement dans l'histoire de l'ADT.

Avec ou sans remise, les boules de l'urne sont considérées indépendamment. C'est une statistique atomique qui a été d'abord mise en place, dont on doit ici vanter – avant de la critiquer – les plus grandes réussites, comme le calcul des spécificités vulgarisé par Pierre Lafon dès 1980 dans le premier numéro de la revue *Mots* (Lafon 1980), et encore positivement discuté aujourd'hui, ici même, par Bénédicte Pincemin en tant qu'implémentation particulièrement pertinente du test exact de Fisher (Pincemin 2024).

Pourtant, le texte est un tissu aux fils reliés, entrecroisés, plus qu'une urne aux tokens indépendants. Entrelac, mailles, tissage : l'étymologie de *texte* a souvent été pertinemment rappelée.

Il y aurait d'abord matière à parler ici des cooccurrences que la plupart d'entre nous avons traitées depuis au moins un demi-siècle. C'est le sens de notre conférence d'ouverture à Paris en 2014 : « Plaidoyer en faveur de l'Analyse de Données co(n)Textuelles Parcours cooccurrentiels dans le discours présidentiel français (1958-2014) ». Les calculs des cooccurrences en effet mettent en équation fondamentalement une linguistique contextualisante (Firth 1957, Halliday and Hasan 1976), selon la double affirmation que le sens nait *en/du* contexte, et que la cooccurrence est la forme minimale (mais calculable) de celui-ci. Dans le texte, l'item (le mot) n'est jamais seul. Les paires ou triplets cooccurrenciels (['classe' – 'ouvrier' –'moyenne'] *versus* ['classe' – 'maitresse' – 'école']) sont les premières unités de sens (*qui font sens*), sinon les premiers *textèmes*, que la statistique permet de mettre au jour (Mayaffre 2014).

Mais nous voulons parler ici de l'expansion de la statistique de Pierre Guiraud ou de Charles Muller vers une *Statistique exploratoire multidimensionnelle* selon l'ouvrage de (Lebart, Morineau et Piron 2000-3ème éd) plusieurs fois rééditée ou *Analysi multidomentionale dei dati* selon celui de Sergio (Bolasco 2022)[6] : la révolution benzécrienne qui a irradié, en France, l'ADT naissante dès les années 1970. (Sur l'héritage laissé par Jean-Paul. Benzécri en ADT voir la conférence d'ouverture des JADTs 2016 : Beaudouin 2016).

En ADT, de l'urne au tableau, les tables de contingences, parfois immenses, que l'AFC ou *Correspondence Analysis* traitent, croisent les variables des textes en colonne (les auteurs des textes, les dates des textes, les genres des textes, etc.) et les individus linguistiques en ligne (les mots, les lemmes, les parts-of-speech, etc.). Dès lors qu'ils sont rassemblés en tableau, les individus linguistiques ne sont plus considérés isolément comme dans une urne aveugle. Ils contribuent ensemble à la construction puis à la visualisation d'un espace, sorte de représentation, certes

[6] Nous citons l'ouvrage de Sergio Bolasco aussi pour son sous-titre roboratif pour cette contribution : « strategie e criteri d'interpretazione ». Statistique exploratoire pour Ludovic Lebart, stratégie interprétative pour Sergio Bolasco : les meilleurs statisticiens semblent s'inscrire dans la posture « rhétorico-herméneutique » dont nous parlions.

partielle, du texte. Mieux : dans un usage pressenti par (Benzécri 1973, 1981) lui-même, puis appliqué de manière magistrale par Max (Reinert 1983 et 1986), disparu cette année et auquel les JADTs 2024 veulent rendre hommage, ou, d'une autre manière par Jean-Marie (Viprey 2004), les variables du texte peuvent être négligées pour se concentrer uniquement sur les contenus linguistiques du texte. Les tables de contingence dites carrées (ou triangulaires puisque, de part et d'autre de la diagonale, la deuxième moitié du tableau reproduit à l'identique la première moitié du tableau), qui croisent les mots en ligne avec les mots en colonne, permettent de rendre compte de la *texture* et du tissage thématique ; les cellules du tableau constituant les mailles cooccurrentielles (approximation des textèmes ?) ou la rencontre *d'unités de contexte* selon l'expression de Max Reinert (figure 2).

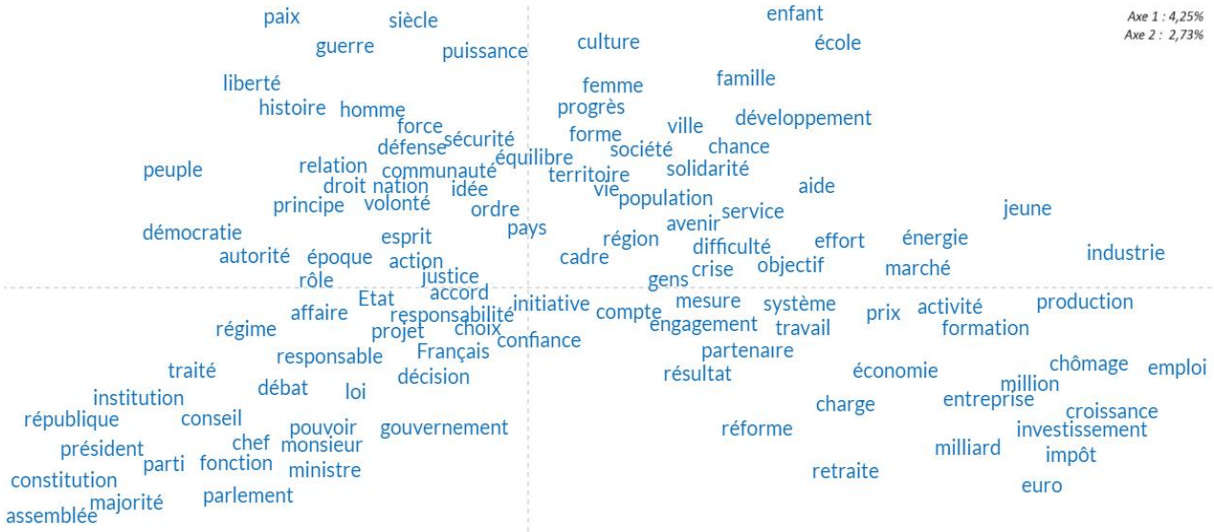

Figure 2. AFC des principaux lemmes nominaux du discours présidentiel (1958-2024) – Sur le plan, l'organisation lexicale du corpus semble mener vers la visualisation d'un discours régalien (quadrant haut-gauche), d'un discours institutionnel (quadrant bas-gauche), d'un discours sociétal (quadrant haut-droit) et d'un discours économique (bas-droit)

Évidemment, la révolution benzécrienne ne peut se résumer à cet usage très particulier, mais nous retenons à travers lui qu'elle permet, par la puissance mathématique, d'embrasser le texte comme objet complexe là où l'urne le considérait comme un objet élémentaire (ie. composé d'éléments traités séparément). Nous retenons surtout de lui la dimension exploratoire, non supervisée, non prescriptive de la statistique : outil typique de l'ADT, l'AFC vient implémenter la posture herméneutique modeste mais fondamentale qui fait notre identité scientifique face au texte.

4. Prospectives

L'ADT en posant le texte comme objet et la statistique comme méthode repose sur deux propositions interdisciplinaires simples à énoncer et complexes à articuler.

La première s'adresse aux SHS et au continent des textes : impossible de caractériser sans mesurer. Dans sa plus sommaire expression, caractériser consiste à passer de 0 à 1 ou de 1 à 0. Il s'agit d'un dénombrement primaire pour distinguer ce qui existe (1) de ce qui n'existe pas (0), et ce dénombrement élémentaire des choses prend très vite, et plus sérieusement, une dimension statistique lorsqu'il s'agit d'embrasser l'ensemble d'un vocabulaire, de comparer la distribution

d'un mot dans un corpus contrastif, de calculer et représenter le profil co-occurrentiel des lemmes, etc. Ainsi, les saillances, les différences, les spécificités nécessaires à toute caractérisation contrastive ou différentielle seront d'autant plus convaincantes qu'elles auront été mises en valeur par une lecture augmentée, par la puissance et la systématicité de l'ordinateur, par l'évaluation statistique de ce qui est significatif et de ce qui ne l'est pas.

La seconde proposition s'adresse aux statistiques et au continent des nombres : impossible de mesurer sans caractériser. Qu'est-ce qui est quantifiable ? Que peut-on compter ? Quelle valeur sémantique donner à un indice statistique ? Sous quelle condition rend-on le significatif signifiant ? La quantification exige ainsi une caractérisation du corpus dans son ensemble comme celle des unités du texte à compter. Un « moment philologique » préalable comme le disent Jean-Michel (Adam et Heidmann 2005 : 83), et un moment herméneutique, président toujours à l'analyse.

Loin d'être contradictoires ces deux propositions doivent être articulées de manière complémentaire et vertueuse par l'ADT : il s'agit à nos yeux d'une déclinaison concrète et réussie du cercle herméneutique souvent évoqué.

La dimension exploratoire et interprétative nous semble ainsi essentielle à tous les niveaux de l'ADT ; c'est sans doute elle qui distingue notre communauté de celle plus formelle du NLP. Au commencement du processus, l'objet texte exige une posture herméneutique et au bout du processus, nous ne l'avons pas assez souligné, l'exploration et l'interprétation se traduisent par un retour systématique au texte (notamment grâce à la navigation hypertextuelle). Au cœur du processus surtout, l'AFC fournit un exemple emblématique : avec l'AFC, l'interprétation apparait partout nécessaire, dans le choix des lignes et des colonnes, dans la mise à l'écart d'éléments perturbateurs, dans le croisement discrétionnaire des axes 1, 2, 3, 4…, dans la lecture même du graphique savamment généré. Le plan factoriel et les *correspondances* qu'il donne à voir ne se lisent pas intuitivement à livre ouvert, mais s'interprètent. Du reste, toutes les visualisations de données sont des représentations, c'est-à-dire des interprétations ; nos sorties-machines dans leurs ergonomies sont des points de vue, non des vérités ; « une herméneutique des sorties logicielles » (Rastier 2011 : 44) reste à construire.

Pour finir, la dimension exploratoire et interprétative de l'heuristique ADT interroge l'identification des observables du texte et, par-là, l'usage inductif ou déductif de nos pratiques. Une chose est de définir les unités jugées pertinentes du texte pour en mesurer la distribution quantitative ; c'est la pratique usuelle de l'ADT dont il ne s'agit pas de contester ni la fertilité, ni la validité. Seulement, cette pratique implique l'existence a priori de textèmes, unités discrètes élémentaires du texte, que l'analyste serait dès lors en droit d'étudier. Les théories textuelles ont hélas montré que ces unités élémentaires ou fondamentales sont mal identifiées : « lorsque la linguistique se donne le texte pour objet, force est d'admettre que la question des unités demeure une énigme insoluble » (Legallois 2006 : 3). Ainsi, faute de textèmes consensuels, préalables, immanents, Jean-Michel Adam traite des *séquences* ou des *paragraphes* (Adam 1987 et Adam 2018) et François Rastier théorise le *passage* (Rastier 2007). Les mots graphiques, sur lesquels certains d'entre nous posent des étiquettes morpho-syntaxiques, ne sont pertinents que lorsqu'ils sont interprétés, contextualisés, mis en contraste et en relation.

Autre chose est de laisser la statistique explorer le corpus pour identifier les unités qui, précisément, *font texte*. André Salem, tout en restant fidèle au matériau textuel brut, a précocement

implémenté le repérage puis le calcul de *segments répétés*, sans préconception de leur longueur ou de leur nature syntaxique (Salem 1987). Sylvie Mellet et Dominique Longrée ont établi dans plusieurs articles les *motifs multidimensionnels* comme unité textométrique essentielle (Mellet et Longrée 2012). Aujourd'hui l'usage d'une IA exploratoire – celle qui va explorer les paramétrages, les unités apprises et les couches cachées du modèle – permet de trouver des *passages profonds* (Vanni et. al. 2020). Ainsi, de manière cumulative, l'ADT n'a eu de cesse depuis plusieurs décennies d'explorer les textes. Son potentiel heuristique aujourd'hui tient moins, à nos yeux, dans le raffinement des indices – peut-être peut-on penser néanmoins à l'introduction d'une statistique bayésienne ou conditionnelle répandue en psycholinguistique ou en science cognitive ? – que dans la découverte statistique de nouveaux observables linguistiques – interprétés et interprétables – qui viennent s'additionner aux anciens.

Bibliographie

Adam J.-M. (1987). *Textualité et séquentialité. L'exemple de la description*. Langue française, 74, pp. 51-72.

Adam J.-M. (2018). *Le paragraphe : entre phrases et texte*. Paris, Armand Colin.

Adam J.-M. (2020). *La linguistique textuelle. Introduction à l'analyse textuelle des discours*. Paris, Armand Colin.

Adam J.-M. et Heidmann U. (éds). (2005). *Sciences du texte et analyse de discours. Enjeux d'une interdisciplinarité*, Genève, Slatkine Érudition.

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., pp. 17-36.

Benzécri J.-P. (1973). *L'analyse des données. Tome 2 : L'analyse des correspondances*. Paris : Dunod.

Benzécri J.-P. (1981). *Pratique de l'analyse des données. Tome 3 : Linguistique et lexicologie*. Paris : Dunod.

Bolasco S. (2022). *Analisi multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione* Rome : Carocci Editore.

Brunet E. (1984). *Le viol de l'urne*. In *La recherche française en langue et Littérature*, Paris, Champion, pp. 253-264.

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

Legallois D. (2006). *Le texte et le problème de son et ses unités : propositions pour une déclinaison*. Langages, 163, pp. 3-9.

De Launay M. (éd) (2013). *L'herméneutique littéraire et son histoire*. Peter Szondi. Revue germanique international,17.

De Angelis R. (2020). *De l'herméneutique matérielle à l'herméneutique digitale ou numérique*. Texto ! Textes et cultures, vol. XXV, n° 4 (http://www.revue-texto.net/index.php?id=4470). [hal-03926958].

Eco U. (1992). *Les limites de l'interprétation*. Paris : Grasset.

Firth R. (1957). *Papers in Linguistics 1934-1951*. London : Oxford University Press.

Guilhaumou J. (2004). *Où va l'analyse de discours ? Autour de la notion de formation discursive*. Texto ! Textes et cultures [http://www.revue-texto.net/Inedits/Guilhaumou_AD.html].

Guiraud P. (1954). *Les caractères statistiques du vocabulaire*. Paris : PUF.

Guiraud P. (1959). Problèmes et méthodes de la statistique linguistique, Dordrecht : Reidel, Publishing Company.

Halliday M.A.K. & Hasan R. (1976). *Cohesion in English*. London : Longman.

Heidmann U. (2005a). *Comparatisme et analyse de discours. La comparaison différentielle comme méthode*, in J.- M.

Adam & U. Heidmann éds., Sciences du texte et analyse de discours. Enjeux d'une interdisciplinarité. Genève : Slatkine, pp. 99-118.

Heidmann U. (2005b). *Epistémologie et pratique de la comparaison différentielle*, in M. Burger & C. Calame éds, *Comparer les comparatismes*. Lausanne : Etudes de Lettres, pp. 141-159.

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

Lebart L., Morineau A., Piron M. (2000, 3ème édition). *Statistique exploratoire multidimentionnelle*. Paris : Dunod.

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

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

Magri V. (2020). *La statistique et le nombre*. Le français moderne, 1-88ème année, pp. 3-11.

Mayaffre D. (2007). *Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques*, in François Rastier et Michel Ballabriga (éds), Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation, Toulouse, Put, 2007, pp. 15-26. [hal-00551477]

Mayaffre D. (2008). *L'entrelacement lexical des textes. Cooccurrences et lexicométrie*. Texte et corpus, n°3, pp. 91-102. [hal-00553808]

Mayaffre D. (2008). *De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie. Sémantique & Syntaxe,* 9, pp. 53-72. [hal-00551114]

Mayaffre D. (2010). *Vers une herméneutique matérielle numérique. Corpus textuels, Logométrie et Langage politique* – Thèse HDR soutenue à Nice le 30 avril 2010. [tel-00655380]

Mayaffre D. (2014). *Plaidoyer en faveur de l'Analyse de Données co(n)Textuelles Parcours cooccurrentiels dans le discours présidentiel français (1958-2014),* Proceedings of the 12th International Conference on Textual Data Statistical Analysis JADT 2014 ¬ conférence invitée, édité par E. Néé, M. Valette-M. Daube et S. Fleury, Paris, Inalco-Sorbonne nouvelle, p. 15-32. [hal-01181337].

Mayaffre D., Pincemin B., Poudat C (2019). *Explorer, mesurer, contextualiser. Quelques apports de la textométrie à l'analyse de discours*. Langue française, 21, pp. 101-115.

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

Mellet S. et Longrée D. (2012). *Légitimité d'une unité textométrique : le motif*. In, JADT 2012, 11èmes Journées internationales d'analyse statistique des données textuelles, Jun 2012, Liège, Belgique. pp.715-728. ⟨hal-01365002⟩

Mellet S. et Vuillaume M. (éds) (1998). *Mots chiffrés et déchiffrés. Mélanges offerts à Etienne Brunet*. Paris : Champion.

Muller Ch. (1964-a). *Essai de statistique lexicale. "L'Illusion comique" de Pierre Corneille*. Paris : Klincksieck.

Muller Ch. (1964-b). *Calcul des probabilités et calcul d'un vocabulaire*. Travaux de linguistique et de littérature, t. II-1.

Pincemin B. (2012). *Hétérogénéité des corpus et textométrie*. Langages 187, pp. 13-26.

Pincemin B. (2024). *Specificities and other applications of the Fisher's exact test to textual data: What's the matter with lexical frequencies ?, in* A. Dister et D. Longrée (eds), Mots comptés, textes déchiffrés. Louvain : Presses Universitaires de Louvain, vol. 2, pp. 703-713.

Poudat C. et Landragin F. (2017). *Explorer un corpus textuel. Méthodes - pratiques – outils*. Paris : De Boeck.

Rastier F. (2001). *Arts et sciences du texte*. Paris : PUF.

Rastier F. (2007). « Passages ». *Corpus*, n° 6 (http://journals.openedition.org/corpus/832).

Rastier F. (2011). *La mesure et le grain*. Paris : Honoré Champion.

Rastier F. (2020). « Mesure et démesure. Quantité et qualité en linguistique de corpus ». *Le français moderne*, 1-88ème année, pp. 11-26.

Rastier F. et Pincemin B. (1999). *Des genres à l'intertexte*. Cahiers de praxématique, 33 (http://journals.openedition.org/praxematique/1974)

Reinert M. (1983) *Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte*. Les Cahiers de l'Analyse des Données, 2, pp. 187-198.

Reinert M. (1986) *Un logiciel d'analyse lexicale*. Les cahiers de l'analyse des données, Tome 11, no. 4, pp. 471-481.

Salem A. (1987). *Pratique des segments répétés. Essai de statistique textuelle*. Paris : Klincksieck.

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

Salem A. (1991). *Les séries textuelles chronologiques*. Histoire & Mesure, vol. VI, n°1/2, pp. 149-175.

Salem A (2021). *Le temps lexical. Un bilan méthodologique sur l'analyse des séries textuelles chronologiques*. Histoire & mesure, vol. XXXVI, pp. 21-56

Szondi P. (1989 - trad M. Bollack). *Introduction à l'herméneutique littéraire*. Paris : Cerf

Tognini-Bonelli E. (2001). *Corpus Linguistics at Work*. London : John Benjamins Publishing Company.

Valette M. (2016). *Analyse statistique des données textuelles et traitement automatique des langues. Une étude comparée*. In *JADT 2016 - Proceeding of 13th International Conference on Statistical Analysis of Textual Data*, Vol 2., pp. 697-706.

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

Vanni L *et al*. (2020). *Key passages : from statistics to deep learning. Text Analytics. Advances and Challenges*. Stella Iezzi et al. (eds), Text Analytics. Advances and Challenges. Springer, pp. 41-54 [hal-03099658]

Viprey J.-M. (2004). *Analyses textuelles et hypertextuelles des Fleurs du mal*. Paris : Champion.

Viprey J.-M. (2005). *Philologie numérique et herméneutique integrative*, *in* J.-M. Adam et U. Heidmann (éds), Sciences du texte et analyse du discours, Genève, Slatkine, pp. 51-68.