



**HAL**  
open science

## Littérature et intelligence artificielle

Étienne Brunet, Ludovic Lebart, Laurent Vanni

► **To cite this version:**

Étienne Brunet, Ludovic Lebart, Laurent Vanni. Littérature et intelligence artificielle. L'intelligence artificielle des textes, 15, Honoré Champion, pp.73-130, 2021, Lettres numériques, 9782745356406. hal-04436066

**HAL Id: hal-04436066**

**<https://hal.science/hal-04436066>**

Submitted on 3 Feb 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# *Littérature et intelligence artificielle*

Étienne Brunet, Ludovic Lebart, Laurent Vanni

BCL(CNRS)

Université de Nice Côte d'Azur

**Abstract.** *Using Deep Learning to attribute authorship of French literary texts*

While problems of attributing authorship or dating a text can be tackled using the usual methods of literary historians, it is equally possible to turn to statistical and computing tools. A range of intertextual measures have been proposed to describe variation within and across authors. To date no single method can claim an uncontested superiority comparable to the use of DNA in paternity suits or criminal investigations. The present study asks whether artificial intelligence may be able to play this role, and seeks the answer in research involving two corpora. The first concerns 20th century French literature: a deep learning algorithm is used on 50 texts by 25 authors (e.g., Roman Gary, Émile Ajar) with the goal of matching the two texts by the same author. Deep learning is perfectly accurate. The second corpus is drawn from French classical drama and here the algorithm also categorically distinguishes and matches plays by Racine, Corneille, and Molière. The only errors concern two plays (the French texts of Molière's *Don Garcia of Navarre* and Racine's *The Litigants*) where the comic genre takes precedence over authorial voice. This paper investigates the mechanisms of deep learning.

**Key words:** *artificial intelligence, deep learning, text authentication, literary authorship, intertextual distance, lexicometry, 20th century French novel, French classical drama, Molière authorship question.*

**Résumé.** *Littérature et intelligence artificielle*

Les problèmes de paternité ou de datation peuvent être abordés avec les moyens habituels de l'histoire littéraire, mais aussi en recourant aux ressources de la statistique et de l'informatique. Diverses mesures intertextuelles ont été proposées pour tenter de distinguer les distances intra (entre les textes d'un même auteur) et les distances inter (entre les auteurs). Malheureusement aucune jusqu'ici n'a pu prétendre au rang de juge suprême, comparable à l'ADN dans les recherches de paternité ou de criminalité. L'Intelligence artificielle peut-elle jouer ce rôle? C'est l'objet de la présente étude, menée conjointement dans deux corpus. Dans le premier, on aborde le roman au XX<sup>ème</sup> siècle en proposant à l'algorithme du Deep Learning un panel de 50 textes et de 25 écrivains (parmi lesquels Roman Gary et Émile Ajar). Il s'agit de reconnaître les textes qui ont le même auteur. Le Deep Learning réussit l'épreuve sans faillir. Fort de cette réussite, le même algorithme est appliqué au théâtre classique. La conclusion est catégorique : Racine, Corneille et Molière se distinguent parfaitement sauf dans deux cas (*Don Garvie* et *Les Plaideurs*) où le genre vient brouiller la signature. Le présent article s'interroge sur les mécanismes mis en œuvre dans le Deep Learning.

**Mots-clés :** *Intelligence artificielle, Deep Learning, authentification des textes, paternité littéraire, distance intertextuelle, lexicométrie, roman au XX<sup>ème</sup> siècle, théâtre classique, affaire Corneille-Molière.*

1 – L’authentification des textes c’est l’activité préférée de la recherche littéraire car elle peut mener au plaisir, et parfois à la gloire, de la découverte. Quand le critique s’emploie seulement à exprimer ses goûts, à approfondir son sentiment, à multiplier les rapprochements, il n’atteint jamais le niveau de l’absolu, si brillantes que soient ses analyses. Mais quand le chercheur s’attache aux sources, qu’il trouve la preuve matérielle d’une erreur à redresser, d’une idée reçue à corriger, d’une incertitude à combler, d’une signature à authentifier, il accède au niveau suprême de la vérité, même s’il ne s’agit que d’une virgule. On observe aussi dans les sciences dures un écart entre la théorie et l’observation de terrain, mais dans bien des cas l’expérimentation permet de rapprocher les deux approches et d’accorder l’une à l’autre par d’incessants recoupements, le fin du fin étant la découverte expérimentale (comme celle de la planète Neptune par Le Verrier) vérifiant une prédiction théorique antérieure. Or l’expérimentation n’est guère réalisable dans le domaine littéraire : aucun écrivain n’accepterait de se soumettre à des conditions imposées par quelque théorie et d’écrire un texte qui aurait été prévu et conditionné. Les tentatives des surréalistes et de l’*OuLiPo* ont pu amuser sans parvenir à convaincre. Et c’est aussi le cas des premières expériences où la plume a été confiée à l’ordinateur, après qu’on l’a doté d’un dictionnaire, d’une grammaire sommaire et de quelques relevés statistiques destinés à guider le hasard.

2 – À défaut d’un automate poète et créateur, peut-on imaginer un algorithme imitateur et composant des textes proustiens et crédibles au point que les spécialistes puissent s’y méprendre. Jusqu’ici la machine n’a signé aucun pastiche qui puisse tromper un lecteur un peu attentif. Dans le domaine du son ou de l’image, l’imitation s’approche peut-être plus facilement de la création. Ainsi en analysant 300 tableaux de Rembrandt l’intelligence artificielle a pu proposer un faux portrait de jeune homme que les experts pourraient authentifier, abstraction faite de la matérialité du support et des ingrédients chimiques utilisés (voir le site <https://www.nextrembrandt.com/>). Les mêmes techniques de *deep learning* appliquées à la composition musicale peuvent proposer une orchestration imitée de Bach sans recourir à la simple copie, ni à la couture des extraits, ni même à l’expertise documentée des musicologues ([http://www.mlsalt.eng.cam.ac.uk/foswiki/pub/Main/ClassOf2016/Feynman\\_Liang\\_8224771\\_assignsubmission\\_file\\_LiangFeynmanThesis.pdf](http://www.mlsalt.eng.cam.ac.uk/foswiki/pub/Main/ClassOf2016/Feynman_Liang_8224771_assignsubmission_file_LiangFeynmanThesis.pdf)).

3 – Notre ambition est moindre. Il ne s’agit ici ni d’imitation, ni de création mais seulement de reconnaissance. On devrait préférer le terme plus exact de *prédiction*, qui est celui qu’utilise la discipline. En effet lorsque les techniques de l’intelligence artificielle sont mises en œuvre, la reconnaissance (ou identification) n’est pas absolue. Le processus de classification aboutit à un continuum de valeurs échelonnées de 0 à 100. Cette approximation en pourcentages est assez habituelle dans les conclusions

statistiques, et particulièrement dans les méthodes éprouvées de la lexicométrie. Mais il faut ici souligner une différence essentielle dans l'approche des faits. L'habitude s'est imposée dans les études lexicométriques de constituer un corpus en rassemblant des textes que l'on oppose les uns aux autres en s'aidant de la « norme » interne représentée par l'ensemble du corpus. Il y a certes des avantages pratiques à procéder de la sorte mais cela ne va pas sans bousculer la vraie démarche statistique qui exclut radicalement l'échantillon de la population. Pour savoir si un échantillon (ou un texte) peut ou non appartenir à la population (ou corpus), les tests autorisés n'admettent pas sa présence, même proportionnellement faible, dans les données de référence, dites aussi données d'apprentissage. La procédure que suit en particulier le *deep learning* respecte intégralement cette distinction, et dénonce immédiatement les forfaitures, en parlant de façon assez dérogatoire d'*apprentissage par cœur* (overfitting). Quand le texte à examiner figure par mégarde ou ignorance dans les données d'apprentissage, un seuil scandaleux de reconnaissance signale le non-respect du principe de séparation. Dans la même situation, la lexicométrie traditionnelle enregistre peu de différence dans les résultats, que le texte à examiner soit extérieur ou non au corpus. On peut en conclure a priori que l'approche du *deep learning*, étant plus exigeante et plus sévère dans les conditions de traitement, laisse espérer une sensibilité plus grande et plus exacte à la spécificité des textes.

## I – Le roman au XXe siècle

### Reconnaître que deux textes sont d'un même auteur

#### I.1 – Une expérience en double aveugle

Nous nous proposons d'en apporter la confirmation avec une expérience en double aveugle mettant en jeu deux chercheurs dont les compétences et les parcours sont si éloignés, que les travaux de l'un sont presque opaques à l'intelligence de l'autre. Le premier, qui est d'origine littéraire et qui tient momentanément la plume, a cinquante ans de statistique linguistique derrière lui. Le second, qui est né cinquante ans plus tard et qui a vécu dans l'informatique depuis le berceau, s'est spécialisé dans l'intelligence artificielle. L'idée leur est venue de s'associer et de résoudre les problèmes de l'un avec les ressources de l'autre<sup>1</sup>. En réalité le mot problème est impropre quand la solution de l'authentification est apportée par l'histoire. Il s'agit en effet, dans un premier temps, de vérifier par la seule analyse textuelle si **Romain Gary et Emile Ajar**, tous les deux lauréats du prix Goncourt, ne font qu'un seul et même auteur. Ce qui fut

---

<sup>1</sup> En cours de route le besoin s'est fait sentir d'une garantie supplémentaire pour couvrir le champ mathématique. Ludovic Lebart s'est alors associé à l'équipe en y apportant sa compétence statistique et le renfort précieux de son logiciel DtmVic, dont certaines fonctions ont été étendues pour le présent projet.

quelque temps un mystère, une rumeur, un soupçon, a fait l'objet d'une révélation explicite qui ne permet plus aucun doute<sup>2</sup>. Mettons-nous cependant à la place d'un membre du jury, plus méfiant que les autres et doté des moyens lexicométriques de vérification. À l'époque, c'est-à-dire en 1975, certaines méthodes existaient déjà pour mesurer la distance intertextuelle, ou, à l'inverse, la proximité, à travers la terminologie de Charles Muller (*connexion lexicale*) ou de Etienne Evrard (*affinités*)<sup>3</sup>. Mais on aurait eu du mal à cette date à trouver sur un support informatique le texte des romans de Gary et des autres écrivains invités à la comparaison. Les moyens actuels, même embarrassés par le copyright, nous ont permis d'incorporer 50 romans dans un corpus où 25 auteurs du XX<sup>ème</sup> siècle sont représentés, chacun fournissant deux titres. En réalité, le problème étant centré sur Romain Gary, la part de cet écrivain est plus avantageuse, puisque deux titres sont alloués à Ajar, deux autres à Gary jeune, et deux autres à Gary vieillissant. Le choix des auteurs n'a rien d'un palmarès rigoureux : l'expérience n'aurait guère changé avec une sélection différente, l'essentiel étant de s'en tenir au roman, pour écarter l'influence dangereuse du genre.

Le dispositif adopté ressemble un peu au protocole familial à la recherche médicale qui aime à étudier le cas des jumeaux. Il s'agit en effet de vérifier si les deux textes qui ont le même père ont des caractéristiques qui les distinguent des autres familles et qui en font des frères, sinon des jumeaux<sup>4</sup>. On s'attend que le sujet, l'époque de la rédaction ou le désir de renouvellement puissent brouiller les ressemblances et qu'inversement des auteurs différents puissent se rencontrer s'ils partagent un thème voisin ou la même école littéraire. Si l'ADN arrive à dissiper le brassage physiologique et à établir sans conteste les vraies paternités, les analyses lexicométriques n'offrent pas jusqu'ici la même garantie en matière de création littéraire, lorsqu'il faut distinguer la distance intra (entre les textes d'un même auteur) et la distance inter (entre les auteurs).

<sup>2</sup> On n'a pas retrouvé à la bibliothèque universitaire de Nice le mémoire d'une certaine Hélène dont parle Didier Van Cauwelaert dans le *Père adopté* (*Le Livre de Poche*, 2009, p. 241-245) et qui aurait soupçonné la collusion des deux noms à partir d'indices stylistiques et de clins d'œil à demi transparents : *Ajar* veut dire « braise » en russe et *Gary* « brûle », à rapprocher d'un autre pseudonyme de Gary : *Brûlard*.

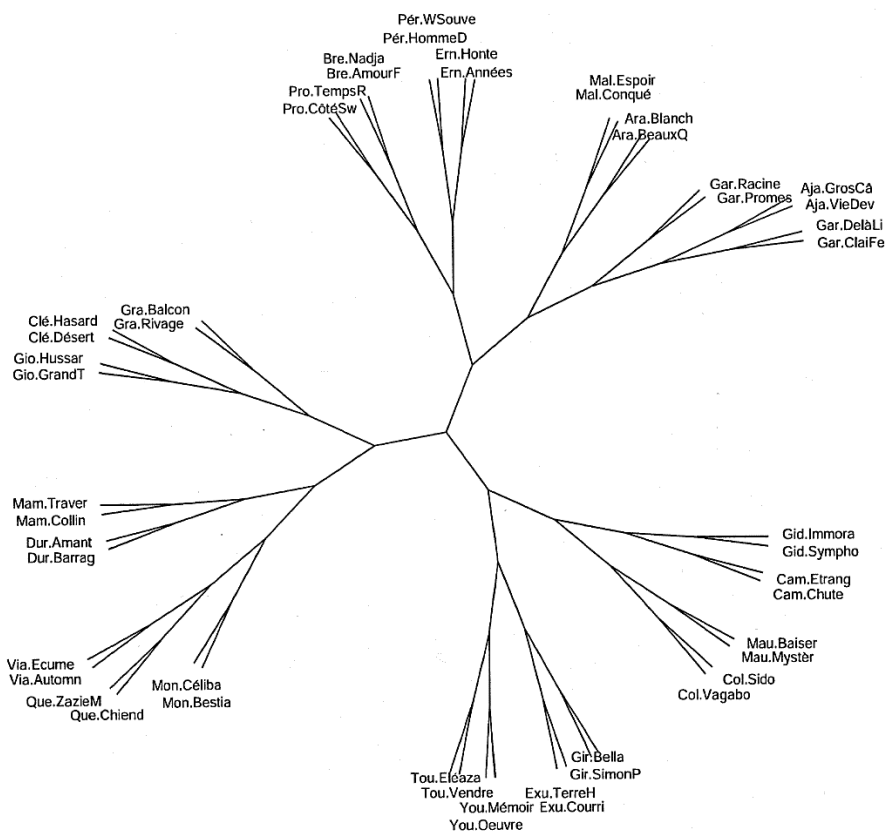
<sup>3</sup> Si la terminologie de ces deux pionniers a un peu vieilli (ils parlent plus volontiers de statistique linguistique plutôt que de lexicométrie, de stylométrie, de textométrie ou de logométrie), on verra dans la suite de cette étude que les algorithmes qu'ils ont mis au point il y a cinquante ans sont assez puissants pour rivaliser avec bonheur avec ceux de l'intelligence artificielle.

<sup>4</sup> Voici la liste des 50 textes retenus :

- |   |   |
|---|---|
| Gide : <i>La symphonie pastorale</i> et <i>L'immoraliste</i> ;            | Proust : <i>Du côté de chez Swann</i> et <i>Le Temps retrouvé</i> ;     |
| Giraudoux : <i>Simon le pathétique</i> et <i>Bella</i> ;                  | Mauriac : <i>Le Baiser au lépreux</i> et <i>Le Mystère Frontenac</i> ;  |
| Montherlant : <i>Les Célibataires</i> et <i>Les Bestiaires</i> ;          | Malraux : <i>L'Espoir</i> et <i>Les Conquérants</i> ;                   |
| Saint-Exupéry : <i>Courrier Sud</i> et <i>Terre des hommes</i> ;          | Breton : <i>Nadja</i> et <i>L'Amour fou</i> ;                           |
| Colette : <i>Sido</i> et <i>La Vagabonde</i> ;                            | Giono : <i>Le grand troupeau</i> et <i>Le bussard sur le toit</i> ;     |
| Queneau : <i>Le Chiendent</i> et <i>Zazie dans le métro</i> ;             | Aragon : <i>Les Beaux Quartiers</i> et <i>Blanche ou l'oubli</i> ;      |
| Camus : <i>L'étranger</i> et <i>La chute</i> ;                            | Vian : <i>L'Écume des jours</i> et <i>L'Automne à Pékin</i> ;           |
| Duras : <i>Un barrage contre le Pacifique</i> et <i>L'Amant</i> ;         | Gracq : <i>Le Rivage des Syrtes</i> et <i>Un balcon en forêt</i> ;      |
| Yourcenar : <i>Mémoires d'Hadrien</i> et <i>L'Œuvre au noir</i> ;         | Mammeri : <i>La colline oubliée</i> et <i>La traversée</i> ;            |
| Gary1 : <i>La promesse de l'aube</i> et <i>Les racines du ciel</i> ;      | Pérec : <i>W ou le souvenir d'enfance</i> et <i>Un homme qui dort</i> ; |
| Tournier : <i>Vendredi ou les limbes du Pacifique</i> et <i>Éléazar</i> ; | Ajar : <i>Gros-Câlin</i> et <i>La vie devant soi</i> ;                  |
| Gary2 : <i>Clair de femme</i> et <i>Au-delà de cette limite</i> ;         | Le Clézio : <i>Hasard</i> et <i>Désert</i> .                            |
| Ernaux : <i>La honte</i> et <i>Les années</i> ;                           |   |

## I.2 – Les mesures classiques de la distance intertextuelle

La matrice des distances peut être établie non seulement sur les formes graphiques, mais aussi sur les lemmes, les codes grammaticaux, les structures (c'est-à-dire des combinaisons de codes entre deux ponctuations), sur les bicodes ou tricodes (combinaisons limitées à deux ou trois codes) ou encore sur les classes de fréquences, la longueur des mots, les segments répétés, etc. Les effectifs obtenus peuvent aussi varier selon qu'on considère la présence/absence (méthodes Jaccard et Évrard) ou la fréquence de l'élément retenu (méthodes Muller et Labbé). Or, si les analyses convergent, aucune n'arrive à surpasser celle d'É. Évrard consignée dans la figure 1.



Graphique 1. Analyse arborée appliquée aux graphies de 50 romans.

Le calcul s'appuie ici sur les mêmes éléments que celui de Jaccard : pour deux textes en présence on désigne

- par A l'effectif des mots communs aux deux textes,
- par B l'effectif des mots présents dans le premier texte mais non dans le deuxième,
- par C l'effectif des mots absents dans le premier texte et présents dans le second,
- par D l'effectif des mots du corpus qui manquent dans les deux textes,
- par N le total général ( $A+B+C+D$ ),
- par  $\alpha$  le total marginal  $A+B$ ,
- par  $\beta$  le total marginal  $C+D$ ,
- par  $\gamma$  le total marginal  $A+C$ ,
- par  $\delta$  le total marginal  $B+D$ .

La formule peut prendre la forme d'un coefficient de corrélation ou celle d'un Chi2, l'une se dérivant de l'autre<sup>5</sup> :

$$r = \frac{AD-BC}{\sqrt{\alpha\beta\gamma\delta}} \quad X^2 = N \frac{(AD-BC)^2}{\alpha\beta\gamma\delta} \quad \frac{X^2}{N} = r^2$$

Le tableau des distances intertextuelles, ainsi réalisé en confrontant chaque roman aux 49 autres, est alors soumis à un programme d'analyse arborée qui trace une sorte d'arbre typologique de la population. Chaque élément terminal y est associé à l'élément dont il est le plus proche<sup>6</sup>, pour constituer une paire. Or, dans la totalité des cas la paire est une fratrie, où l'algorithme reconnaît la signature du même auteur (figure 1). S'agissant de Gary-Ajar, la cohésion de la famille est fortement soulignée sur la droite du graphique où les six membres se tiennent la main.

Le résultat est confirmé par les deux formules d'Évrard, tant au niveau des lemmes que des graphies. Mais d'autres approches ou d'autres observables, aussi légitimes, laissent en suspens quelques textes qui n'ont pas trouvé le partenaire naturel. Le tableau 2 restitue les résultats qu'on obtient lorsqu'on fait varier les paramètres de l'analyse arborée.

Méthode	Objet	Effectifs	Paires reconnues (sur 25 possibles)
Evrard	graphies	présence/absence	25
Evrard	lemmes	présence/absence	25
Jaccard (Brunet)	graphies	présence/absence	24
Jaccard (Brunet)	lemmes	présence/absence	22
Labbé	lemmes	fréquence	21
Labbé	graphies	fréquence	20
Muller	graphies	classes de fréquence	19
Jaccard (Brunet)	structures	présence/absence	11
Labbé	structures	fréquence	8
Jaccard (Brunet)	codes	présence/absence	6
Labbé	codes	fréquence	6

Tableau 2. Comparaison des analyses arborées établies sur des critères différents.

<sup>5</sup> L'exposé se trouve dans Étienne Évrard « Étude statistique sur les affinités de cinquante-huit dialectes bantous » in *Statistique et Analyse linguistique*, PUF, Paris 1966, pp.85-94. On a jugé utile d'exhumer cet article ancien et d'en démontrer les vertus dans É. Brunet, « Les affinités lexicales. Hommage à Étienne Évrard », 2014, <http://lesetudesclassiques.be/index.php/lec/article/view/320>.

Le calcul est implémenté dans le logiciel *HYPERBASE* comme ceux de Jaccard, de Muller et de Labbé. Quant à la représentation graphique, deux programmes concordants sont mis en œuvre : l'analyse arborée de Luong et celle de D.H. Huson et D. Bryant. Cette dernière, appelée *SplitsTree*, est utilisée dans le graphique 1.

<sup>6</sup> La notion de plus proche voisin peut prêter à confusion. Dans le tableau des distances, la valeur minimale observée sur la ligne A et désignant la colonne B correspond à la notion simple et locale de PPV (ou plus proche voisin). Mais l'algorithme des arbres additifs (ou analyse arborée) met en évidence des appariements beaucoup plus subtils. Il utilise non seulement les plus proches voisins, mais aussi la cohérence entre voisins de voisins. Ainsi plusieurs couples de textes, comme ceux de Duras ou de Colette, qui ne sont pas reconnus localement comme plus proches voisins, sont appariés en deuxième instance par l'analyse arborée, quand toutes les cartes sont sur la table et toutes les relations prises en compte.

Notons que l'efficacité diminue quand l'objet linguistique considéré perd en intension (ou compréhension) ce qu'il gagne en extension, ce qui est le cas des codes grammaticaux dont les effectifs sont considérables et la variété restreinte. La lexicologie s'est souvent engagée dans cette voie à ses débuts, parce que les données disponibles étaient de modeste dimension et que pour échapper à l'hypothèse nulle, elle recherchait un moyen de grossir les effectifs, en faisant des tas plus larges parmi les mots.

### I.3 – Méthodes fondées sur le TLE.

La plupart des méthodes précédentes reposent sur des effectifs, en nombre restreint, où les éléments du vocabulaire sont répartis comme les votes qu'on dépouille au sortir de l'urne. Que les mots soient riches ou pauvres, fréquents ou rares, les décomptes sont faits *per capita*, chaque élément de V (forme ou lemme) pesant le même poids, quel que soit le nombre de ses occurrences. Cela donne une importance démocratique aux mots de faible fréquence, tout en annulant le poids des mots opulents<sup>7</sup>.

Mais on peut contester ce système et explorer le dictionnaire des fréquences non plus en ordre alphabétique mais dans un classement hiérarchique (c'est le TLE, ou Tableau Lexical Entier), en commençant par les plus fréquents, et en tenant compte de l'importance économique de chacun, calculée en nombre d'occurrences. C'est ce que recommande un autre pionnier de la discipline : André Salem. Les décomptes ne reposent plus sur V, mais sur N, c'est-à-dire sur les occurrences, dont chacune importe. Chaque vocable (forme ou lemme) constitue une ligne dans un tableau cumulatif qui compte autant de colonnes que de textes dans le corpus. À l'intersection, la cellule élémentaire contient la fréquence du vocable A dans le texte B. En principe ce n'est pas un vote censitaire puisqu'en droit toutes les lignes (tous les vocables) sont prises en considération. En réalité l'analyse qui exploite ce tableau s'arrête avant d'atteindre les basses fréquences, dont la prise en compte ne modifierait plus les résultats acquis. Or cette analyse peut prendre deux formes.

a – La première est un **calcul direct**, par le canal de **l'analyse arborée**. Mais au lieu de traiter un tableau carré symétrique (en réalité, triangulaire), comme celui de la figure 1, le programme exploite une matrice rectangulaire, où les mots sont en x et les textes en y, et qui est convertie, par un calcul de Chi2, en un tableau (carré et

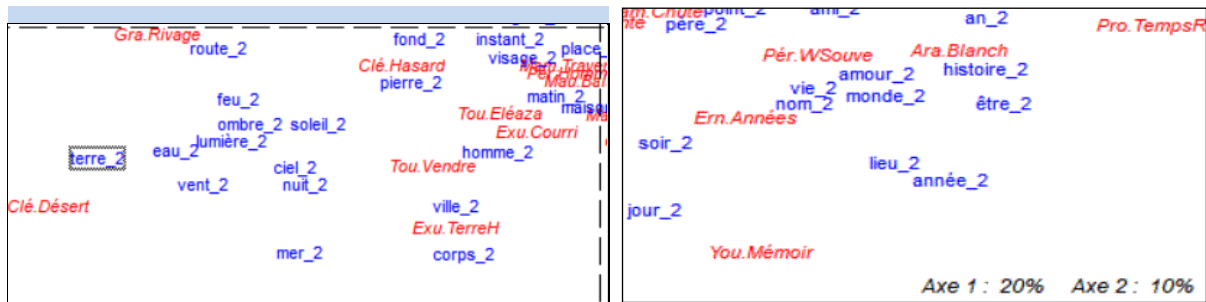
---

<sup>7</sup> Les mots fréquents ne peuvent guère échapper à la zone d'intersection où s'entassent les mots communs aux textes que l'on confronte. Leur effectif est constant d'un texte à l'autre du corpus, si bien qu'on peut les éliminer sans changer les résultats. Le calcul d'Evrard, qui aboutit à la réussite remarquable de la figure 1, a été refait en s'abstenant d'intégrer les 1000 formes graphiques les plus fréquentes, dont on pouvait penser qu'elles étaient présentes dans les 50 textes du corpus. Cela ne trouble aucunement la reconnaissance des 25 paires, alors que le poids des mots maintenus ne représentait plus que le quart de la surface imprimée. Même en ne gardant que les mots de fréquence inférieure à 100, soit le huitième du total des occurrences, la figure 1 ne bouge pas d'un pouce. C'est avouer que la distance d'Evrard, même si elle envisage la totalité du vocabulaire, n'est sensible qu'aux mouvements des basses fréquences. Il en est de même de la distance de Jaccard et aussi du calcul de Muller, lequel s'arrête à la fréquence 50.





un écrivain n'est pas tenu à exploiter toujours le même, et à se l'approprier en excluant les autres (les autres thèmes et les autres écrivains). Or le thème oriente principalement le choix des substantifs. Si l'on concentre l'étude sur les 94 substantifs qui ont plus de 1000 occurrences, l'analyse factorielle (extraits ci-dessous) distribue les thèmes dans deux directions opposées : à gauche les éléments de la nature occupent l'espace et c'est là que s'installent certains romans de Le Clézio, Tournier, Saint Exupéry, ou Gracq; à droite c'est le règne du temps (*soir, jour, an, année, histoire, vie*) et les romans qu'on voit dans ces parages invoquent le temps dans leur titre même (*Le temps retrouvé, Blanche ou l'oubli, Les années, Mémoires d'Hadrien, W ou le Souvenir*), comme c'est aussi le cas pour l'espace dans le camp opposé (*Désert, Le Rivage des Syrtes, Vendredi ou les limbes du Pacifique*). Et parmi les substantifs les plus contingents et les moins fortement liés à l'écrivain sont les noms propres, qu'ils désignent des personnes ou des lieux, qu'ils appartiennent à la fiction ou à la réalité.



L'espace et le temps dans l'analyse factorielle des substantifs (extraits)

Il convient donc de se méfier des mots-outils et des substantifs, ce qui est réalisé dans la figure 4, qui limite le traitement aux verbes du corpus, en neutralisant les marques du temps, du mode et de la personne. Privé de ces attributs, le verbe n'est plus qu'un matériau souple et transparent, qui sert de liant aux éléments de la phrase. S'il garde encore une valeur thématique, c'est sans raideur, sans exclusive, sans la radicalisation propre au substantif. Les verbes sont beaucoup moins souvent caractéristiques que les noms. Leurs valeurs-test ou écarts réduits (indices de spécificité qui ont 95 chances sur 100 d'être inférieurs à 2 en valeur absolue dans le cas d'une répartition au hasard) dépassent rarement 10, alors que celles des noms dépassent fréquemment 20, voire 30 pour certains romans. Il peut sembler paradoxal de s'appuyer sur des éléments plus neutres et plus constants pour obtenir une différenciation plus nette. Il en est ainsi en matière de biologie : la typologie des espèces délaisse souvent les caractéristiques extérieures pour retenir des critères moins visibles et plus solides. La figure 4 en apporte la preuve : toutes les paires sont clairement associées, à l'exception de deux écrivains, Camus et Pérec. On s'abstient provisoirement d'expliquer ces deux manquements, tant que d'autres méthodes ne les ont pas confirmés.

## ANALYSE ARBOREE (directe) de 1542 VERBES (fréquence &gt; 20)

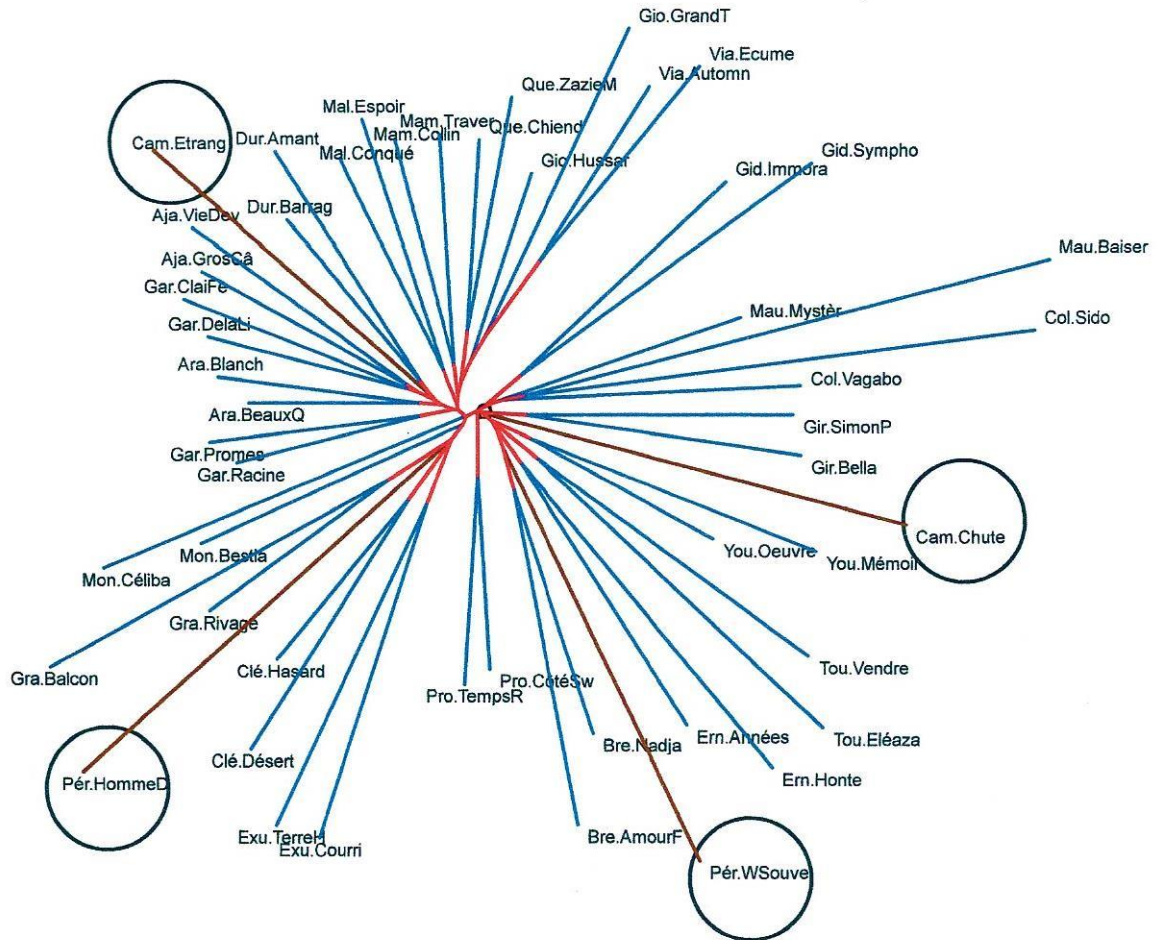


Figure 4. Analyse arborée (directe) de 1542 verbes lemmatisés (fréquence &gt; 20)

**b** - On peut aussi exploiter le TLE de façon **indirecte**, la recherche du plus proche voisin étant appliquée aux axes factoriels d'une analyse **factorielle préalable**<sup>8</sup>. Par le moyen de cette analyse (des correspondances) on accède à des informations sur l'organisation structurelle des mots et des textes – ce que l'analyse arborée ne délivre pas aussi précisément.

- L'approche la plus franche est l'analyse des formes graphiques sans aucun prétraitement, illustrée par l'arbre de la figure 5 (formes apparaissant plus de 50 fois). Avec cette approche rudimentaire, environ 32 romans (16 auteurs) sont assez

<sup>8</sup> Pour ce traitement plus complexe on a fait appel au logiciel DtmVic (*Data and Text Mining: Visualisation, Inférence, Classification*). Il s'agit d'un outil de Statistique Exploratoire Multidimensionnelle pour données complexes comprenant des données numériques et textuelles. DtmVic utilise les logiciels TreeTagger (Schmid, 1994) pour la lemmatisation et SplitsTree (Huson et Bryant, 2006) pour le tracé des **arbres additifs**. Ce logiciel gratuit, créé par Ludovic Lebart, est téléchargeable à l'adresse [http://www.dtmvic.com/05\\_SoftwareF.html](http://www.dtmvic.com/05_SoftwareF.html). Les données fournies à ce logiciel sont les mêmes que celles que traite le logiciel **Hyperbase**, à une exception près : le corpus comprend 51 textes, un roman ayant été ajouté à l'auteur Pérec pour compenser l'instabilité de *W ou le souvenir d'enfance*.

convenablement appariés (en vert). En rouge, les romans mal ou très mal classés. Ce sont les mêmes que ceux de la figure 3.

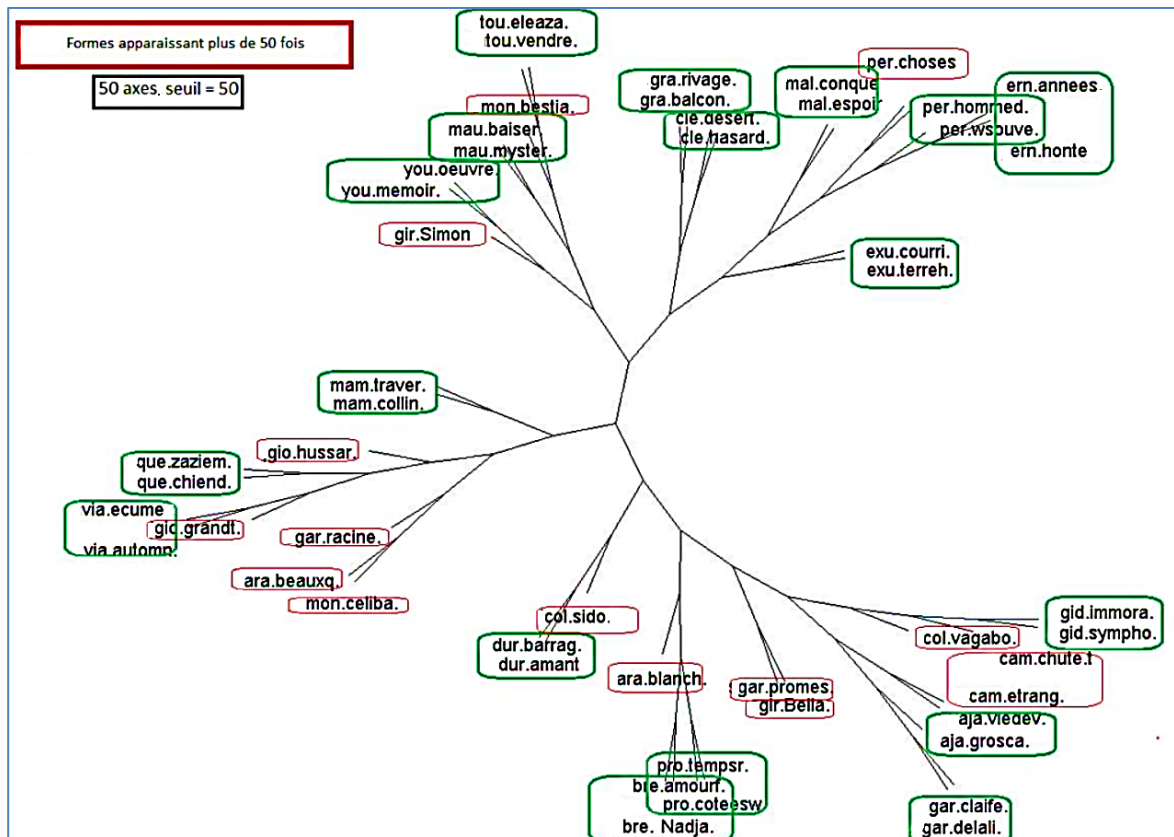


Figure 5. Formes graphiques apparaissant plus de 50 fois

- Il convient donc d'écarter certains mots-outils ou substantifs sur lesquels pèse le soupçon de collusion avec le genre ou le thème. L'exclusion parmi les lemmes est restée légère, et n'a concerné que les noms propres et les pronoms personnels. Cela suffit pourtant pour améliorer grandement le score dans la figure 6. Un seul auteur échappe aux regroupements: Montherlant. Avec un seuil de 100, la stabilité de l'analyse est confirmée, que l'on passe de 4018 lemmes à 2364 lemmes.

- Une troisième approche, plus radicale, concerne les seuls verbes. Ainsi en s'en tenant à cette seule catégorie (sauf *être* et *avoir*, un peu suspects), le résultat est très bon : 23 auteurs sont bien retrouvés (sauf Camus et encore Montherlant).

- On pourrait penser que la lemmatisation est une opération approximative et que les résultats s'en trouvent fragilisés. Or l'expérience présente utilise un lemmatiseur, *TreeTagger*, qui diffère de *Cordial*, sur lequel s'appuie la figure 4 consacrée pareillement aux verbes. Et pour faire bonne mesure, on a substitué aussi le logiciel *SplitsTree* à notre programme d'analyse arborée réalisé par Luong. Or le résultat, qu'illustre la figure 7, est parallèle au précédent, avec deux exceptions sur 25 couples.



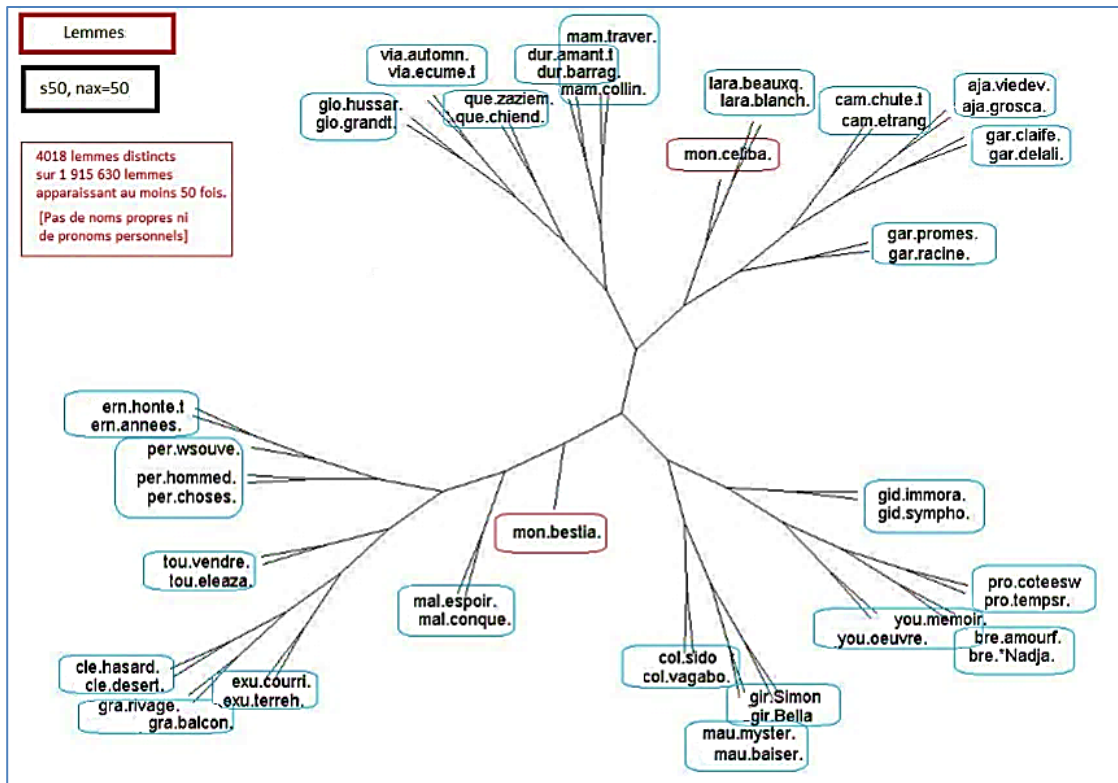


Figure 6. Lemmes sans noms propres ni pronoms personnels (seuil > 50).

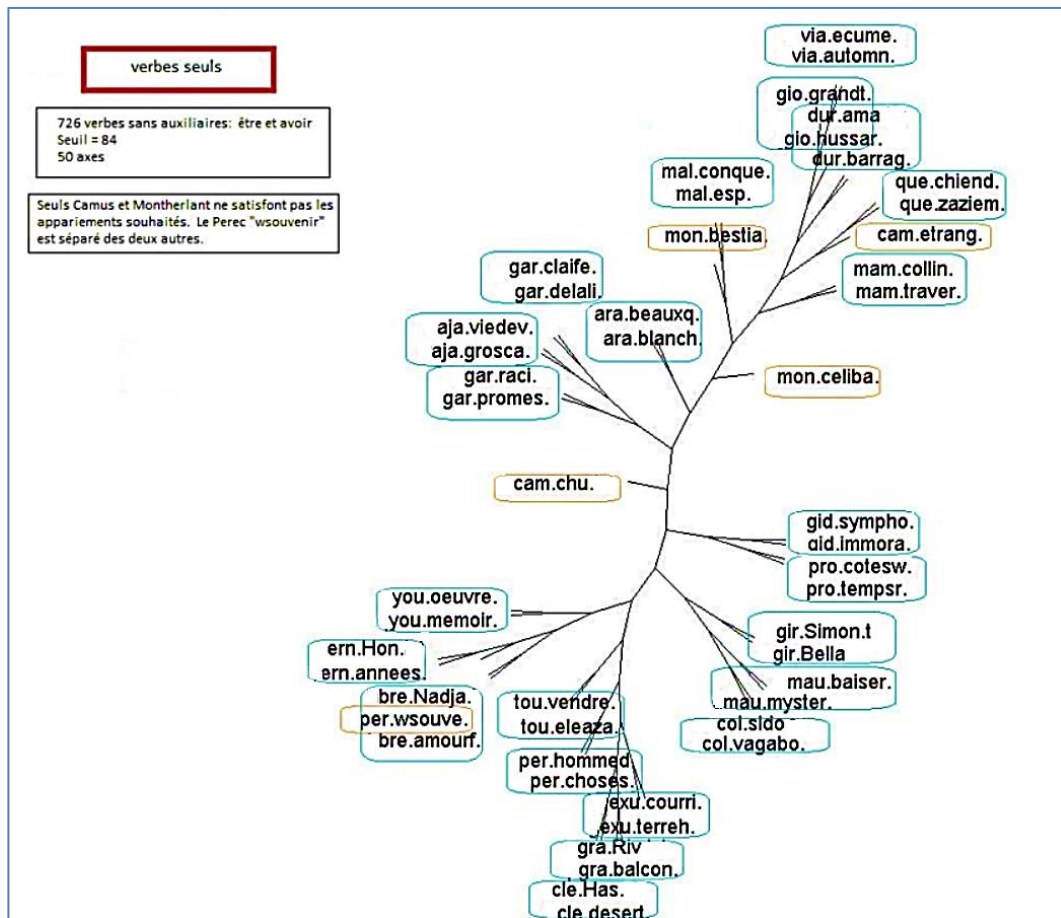


Figure 7. Analyse arborée de 726 verbes de fréquence supérieure à 84

c - Comme on l'a vu, la distance d'Evrard est liée au chi-deux et donc, dans le cas binaire (et seulement dans ce cas de codage « présence – absence ») liée au coefficient de corrélation classique.

Si l'on transforme le tableau lexical [graphies x romans] (qui a ici 40704 lignes et 51 colonnes, pour les graphies apparaissant au moins dans deux romans) en un **tableau binaire** (0 pour la fréquence "0", et 1 pour toute fréquence positive), le coefficient de corrélation entre deux romans coïncide alors avec celui qui intervient dans la distance d'Evrard.

On doit donc retrouver les mêmes distances à partir de tous les axes d'une analyse en composantes principales qui est fondée sur la matrice des corrélations de ce tableau binaire (40704 x 51). C'est bien ce qu'on observe dans la figure 8.

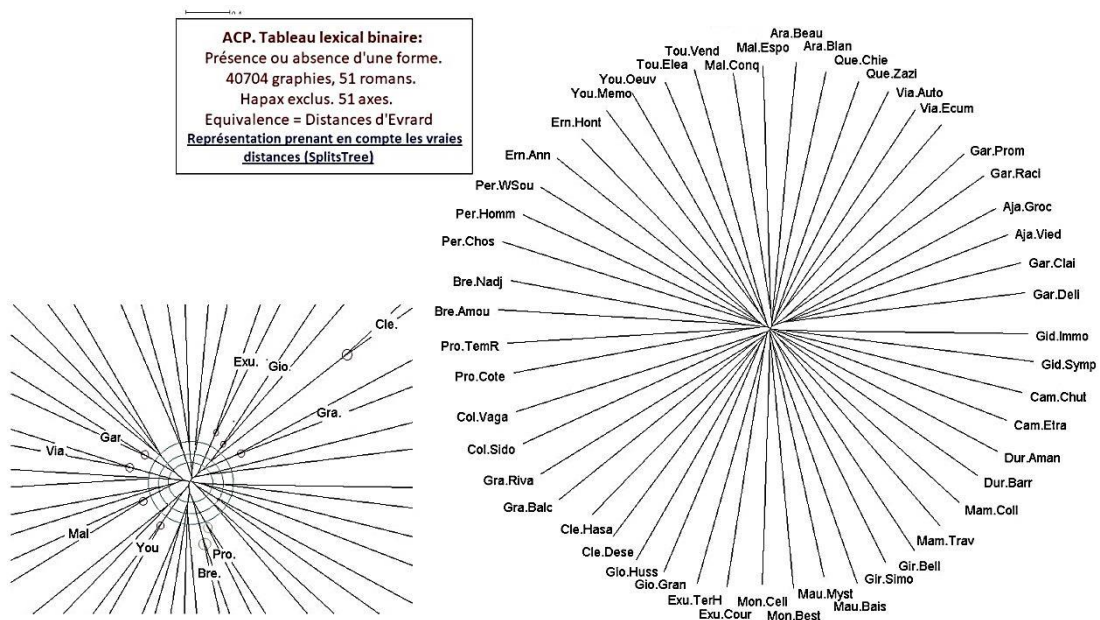


Figure 8. Analyse arborée du Tableau lexical binaire (après ACP)  
*(en médaillon, le zoom sur la partie centrale met en relief les paires les mieux constituées, qui s'éloignent plus tôt et plus loin de l'origine)*

On retrouve ainsi le même appariement complet pour les 25 auteurs, mais avec, en complément, toutes les visualisations intermédiaires permises par les paires d'axes factoriels successifs. Un des enseignements de cette reconstitution progressive des distances d'Evrard axe par axe est la mise en évidence de la typicité de certains auteurs, dont les romans apparaissent proches dès les premières dimensions et le restent jusqu'à la reconstitution totale des distances. On peut ainsi visualiser sur ces plans factoriels les paires de romans qui sont « voisins réciproques » pour la distance d'Evrard, c'est-à-dire des paires qui sont isolées dans la « forêt » des plus proches voisins, et constater ainsi

que l'appariement est plus précoce et robuste pour les auteurs concernés (cas de Proust, Yourcenar, Gary, Breton, Vian, St Exupéry, Mauriac, Queneau). La décomposition spectrale (qui concerne aussi bien les mots que les romans) met aussi en évidence pour les premiers axes des traits structurels non décelable dans le bilan global que constituent les analyses arborées. Citons simplement à titre d'exemple l'opposition sur l'axe trois entre l'emploi de vocabulaire familier, voire argotique (du côté de Ajar, Queneau, Vian) et l'emploi d'un langage dépourvu de ce type de vocabulaire (avec Yourcenar, Proust, Tournier, Gracq, Giraudoux). Chaque axe apporte ainsi sa contribution interprétative, jusqu'au bouquet final comprenant tous les axes et présentant un appariement parfait des romans par auteurs.

#### I.4 – La solution du *deep learning*<sup>9</sup>

L'intelligence artificielle ne nécessite aucune manipulation préalable, aucun codage grammatical, aucun regroupement ontologique. Elle n'a nul besoin que l'on distingue les homographes et que l'on procède à quelque lemmatisation<sup>10</sup>. Elle s'attache à l'environnement immédiat des mots en faisant glisser une fenêtre de trois mots à la fois dans un espace de cinquante mots<sup>11</sup>. Cela suffit dans bien des cas à dissoudre les ambiguïtés lexicales. Que l'on n'aille pas croire pourtant qu'elle accepte la négligence ou l'incohérence. Elle est très sensible aux artefacts qui résulteraient de choix peu constants ou peu rigoureux dans le traitement des ponctuations, des blancs, des majuscules, de l'apostrophe, des tirets, des noms propres. Nos cinquante textes ont donc été soumis à une préparation minimale, avec le seul souci d'un traitement homogène.

---

<sup>9</sup> Le lecteur qui souhaiterait un exposé plus détaillé du *deep learning* peut consulter plusieurs publications de Laurent Vanni au sein du laboratoire BCL, notamment :

Mélanie Ducoffe, Damon Mayaffre, Frédéric Precioso, Frédéric Lavigne, Laurent Vanni *et al.* *Machine Learning under the light of Phraseology expertise: use case of presidential speeches, De Gaulle -Hollande (1958-2016)*, JADT 2016 - *Statistical Analysis of Textual Data*, Nice, France, Volume 1, pp.157-168 <https://jadt2016.sciencesconf.org/>

Laurent Vanni, Damon Mayaffre, Dominique Longrée. *ADT et deep learning, regards croisés. Phrases-clefs, motifs et nouveaux observables*, JADT 2018, Juin 2018, Rome.

<sup>10</sup> Rien n'interdit cependant de proposer au *deep learning* des textes lemmatisés, où chaque graphie est accompagnée d'un code grammatical et du lemme correspondant, selon le format de TreeTagger, à raison d'une ligne de trois éléments pour chaque mot ou signe. En procédant ainsi on peut parfois acquérir plus de précision dans l'interprétation des passages significatifs et savoir si la « saillance » d'un mot vient de sa fonction ou de son sens. Voir l'expérience relatée dans l'article de L. Vanni et al., cité plus haut (JADT 2018). En revanche il y a peu d'espoir d'améliorer les résultats si l'on se contente de substituer le lemme à la graphie. Et la vieille querelle qui a longtemps opposé les partisans et les adversaires de la lemmatisation risque de se rallumer au bénéfice des seconds. Ajouter des étiquettes c'est certes préciser et enrichir les données. Mais réduire les mots à leur forme canonique est incontestablement un appauvrissement, comme toute simplification, et cela pénalise la puissance du *deep learning*, qui est capable de tenir compte de la variété fonctionnelle des graphies sans s'appuyer sur un codage explicite.

<sup>11</sup> Ces paramètres peuvent être ajustés à la situation. Précisons que d'autres paramètres interviennent dont le dosage peut varier. Cela n'influe pas sur la direction où s'engage la prédiction mais seulement sur sa précision et le degré de confiance que l'on peut accorder au résultat.

### I.4.1 - La reconnaissance des paires

Rappelons la distinction imposée par la procédure entre les textes que l'on propose à l'apprentissage et ceux qui sont présentés à la reconnaissance. Chacun des 25 auteurs a donc un représentant dans le premier lot de textes et un autre dans le second. Une fois l'apprentissage réalisé à partir du premier lot, le jeu consiste à présenter un élément du second lot et à demander à l'algorithme de choisir le modèle le plus proche parmi les 25 profils préalablement définis et supervisés. Les colonnes du tableau 9 représentent les titres retenus pour l'apprentissage, les lignes correspondant, dans le même ordre, aux romans soumis à l'épreuve de reconnaissance.

	Gide.Symphonie	Proust.Swann	Giraudoux.Simon	Mauriac.Baiser	Montherlant.Célibataire	Malraux.Espoir	Exupéry.CourtesSud	Breton.Nadja	Colette.Sido	Giono.GrandTroupneau	Queneau.Zazie	Aragon.BlancheOubli	Camus.Chute	Vian.Ecume	Duras.Amant	Gracq.BalconForêt	Yourcenar.MémoiresHa	Mammeri.CollineOubli	GARY1.RachisCiel	Pérec.HommeDort	Tournier.Eléazar	Ajar.GrosCâlin	Gary2.ClairFemme	LeClézio.Hasard	Ermux.Honir
Gide.Immoraliste	35.3	8.4	6.62	5.46	0.42	0.63	3.57	2	3.47	4.2	2.1	3.89	2.31	0.63	0.42	4.2	7.04	1.16	0.84	0.74	2.52	0	3.78	0.11	0.21
Proust.TempsRetrouvé	5.48	47.5	3.78	2.23	0.9	0.85	1.16	3.56	3.19	2.46	2.12	1.1	1.5	0.4	1.02	4.01	4.46	0.54	2.09	2.37	2.09	0.9	2.65	0.48	3.13
Giraudoux.Bella	11.1	5.49	15.1	1.93	1.42	0.81	5.49	1.5	5.49	1.29	1.72	2.53	1.46	14.7	1.03	3.69	7.08	0.99	2.57	4.63	1.97	0.3	5.4	0.6	1.67
Mauriac.MystèreFrontenac	4.06	2.03	3.74	38.1	2.99	0.64	1.71	0.21	0.53	3.63	4.16	2.03	0.11	1.28	0.43	0.96	19.3	2.03	0.64	4.48	2.88	0.11	0.53	0.21	3.2
Montherlant.Bestiaire	1.8	7.51	5.56	6.95	13.5	2.14	0.64	1.91	1.91	3.13	7.77	11.2	1.16	6.32	1.33	1.25	6.32	1.3	3.97	6.58	1.33	0.38	2.75	0.41	2.93
Malraux.Conquérants	2.83	0.9	1.88	2.81	6.97	11.8	5.17	1.62	0.88	7.63	6.56	4.42	0.95	8.01	1.43	3.24	3.14	6.95	6.17	8.29	2.18	0.25	1.75	2.44	1.75
Exupéry.TerreHommes	1.49	0.94	4.31	3.45	4.39	2.27	36.4	2.19	3.37	2.11	2.66	3.84	0.7	1.72	2.35	3.76	3.84	2.74	1.49	3.6	3.52	0.16	4.07	1.8	2.82
Breton.AmourFou	5.78	12.3	1.4	1.58	1.4	6.74	1.84	25	3.24	1.4	4.47	2.36	2.63	0.88	1.05	3.85	3.15	0.44	4.9	2.28	2.71	0.44	5.25	1.49	3.5
Colette.Vagabonde	4.81	2.4	7.42	3.55	2.19	1.88	6.17	2.09	14.6	1.67	3.76	5.22	1.25	0.52	0.52	4.39	5.02	2.19	3.55	7.63	5.85	0.31	7	1.36	4.6
Giono.HussardToit	1.95	0.03	0.09	8.74	3.74	2.74	3.52	0.09	1.26	28.9	7.04	6.79	2.04	5.09	6.32	1.51	3.49	3.33	1.79	0.82	2.33	0.66	0.91	6.32	0.47
Queneau.Chiendent	5.19	2.32	0.85	1.2	0.73	0.65	0.89	2.2	1.78	6.01	17.5	2.27	1.58	14.5	1.38	2.49	5.5	3.81	5.99	2.36	2.96	1.14	13.2	0.18	3.34
Aragon.BeauxQuartiers	0.6	1.82	2.7	2.7	2.18	1.17	1.09	0.92	2.96	4.27	6.41	29.2	1.08	4.66	2.29	2.07	2.96	2.19	9.83	4.71	1.22	1.18	7.58	1.1	3.11
Camus.Etranger	9.95	8.11	3.41	1.23	0.89	2.59	5.59	4.64	5.73	6.2	2.93	3.95	6.54	2.32	1.23	5.32	4.23	1.43	6	2.25	4.23	0.48	8.79	0.55	1.43
Vian.AutomnePékin	0.45	0.05	0.96	0.75	0.55	0.2	0.4	0.7	0.81	7.35	4.73	1.46	0.5	45.6	1.51	1.11	4.28	6.19	4.43	4.63	5.08	0.1	5.28	0.6	2.26
Duras.BarragePacifique	1.17	4.46	2.33	2.81	0.82	4.32	4.25	1.71	3.43	3.09	4.6	3.43	5.83	0.69	12.1	1.17	1.78	3.98	4.46	3.22	1.51	3.36	6.31	9.19	9.95
Gracq.RivageSyrtis	2.5	0.96	1.46	2.39	2.77	0.5	2	0.54	0.81	7.16	4.39	3.35	0.15	6.27	0.65	22.6	8.04	2.93	7.78	9.55	5.93	0.04	1.62	1.89	3.73
Yourcenar.OeuvreNoir	20.4	4.1	4.81	0.49	0.78	1.37	4.52	5.08	1.42	1.15	1.71	0.64	0.29	0.32	0.2	7.11	26.7	0.56	3.98	3.25	4.4	0.15	5.01	0.2	1.42
Mammeri.Traversée	5.25	3.1	2.74	3.13	0.67	0.63	2.78	0.43	1.41	6.7	3.37	1.84	0.67	4.55	1.57	2.98	4.11	29.9	3.96	3.68	5.6	0.35	5.37	2.82	2.39
Garyl.PromesseAube	5.15	5.78	3.56	0.88	0.23	1.42	1.95	2.38	2.22	2.2	2.01	1.47	1.86	1.42	2.68	5.15	2.85	0.73	16.5	2.36	4.71	1.76	27	0.52	3.28
Pérec.Choses	1.33	0.67	0.67	1.91	0.42	4.41	9.73	3.66	4.99	2.49	3.57	2.99	0.25	2.66	0.33	10.1	3.16	6.57	2.74	14.3	2.08	0.33	7.65	5.99	7.07
Tournier.Vendredi	1.02	0.37	7.44	8.09	1.58	0.47	1.3	0.56	0.09	4.19	3.91	7.63	0.09	1.86	0.37	2.88	11.4	3.63	5.12	6.14	24.8	0.28	1.12	1.4	4.28
Ajar.VieDevantSoi	1.05	5.29	0.77	0.24	0.2	1.86	0.65	2.34	1.66	4.2	5.25	3.47	2.58	4.4	1.86	1.74	1.49	1.01	14	0.81	1.66	10.2	30	0.4	2.87
Gary2.DelàLimite	4.71	1.92	1.74	1.55	2.05	3.04	2.6	1.12	2.42	6.82	4.84	3.35	3.41	4.34	2.17	2.29	2.36	2.85	8.37	1.05	1.18	3.66	29.1	0.99	2.05
LeClézio.Désert	0.15	1.25	2.14	1.07	4.61	0.96	2.91	0.18	0.77	2.84	4.98	5.72	1.03	3.65	6.09	5.13	2.36	7.45	4.94	6.79	6.64	1	3.43	12.5	11.4
Ermux.Années	1.75	6.18	2.27	1.03	0.1	0.93	1.34	4.74	2.57	2.16	5.56	2.27	3.4	1.34	1.96	1.65	3.71	1.24	2.16	4.33	3.3	1.13	4.02	1.44	39.4

Tableau 9. Les paires reconnues par le *deep learning*.

Pour lire la leçon du tableau, il faut s'attacher à la diagonale principale (en rouge), où se croisent les deux textes du même auteur. Par exemple le taux de proximité relevé pour les deux textes de Proust (deuxième colonne) s'élève à 47.5, très au-dessus des autres valeurs lues dans la même colonne, dont beaucoup ne dépassent pas 1 ou 2. Cet écart énorme met en relief l'originalité de Proust, que les méthodes classiques soulignent pareillement et qui est sensible à la lecture même. Certes les écarts sont plus faibles dans les autres colonnes mais partout la valeur la plus haute, et de loin, est celle qui se lit au croisement des textes du même auteur. Les taux les plus bas (6.54 et 10.2) sont observés dans les colonnes Camus et Ajar. Il y a loin en effet du premier roman de Camus au dernier, dans l'écriture comme dans le sujet, et l'innocence de l'*Étranger* ne



laisse pas prévoir la culpabilité de la *Chute*. En ce qui concerne Ajar, l'apprentissage a été réalisé sur le roman *Gros-Câlin*. Or, ce roman est un exercice de style où l'auteur, par un décalage systématique du vocabulaire, de la syntaxe et de la pensée, s'écarte non seulement de ses romans antérieurs mais de toute production littéraire<sup>12</sup>. La liaison avec *La vie devant soi* n'en est pas moins établie et n'est approchée, à distance, que par un texte de Gary *Au-delà de cette limite...* (3.66).

L'honnêteté nous oblige à reconnaître un seul échec. Il est dû à Pérec, le funambule de notre littérature. Nous étions conscients qu'il fallait éviter de proposer à l'expérience des modèles comme la *Disparition* ou les *Revenentes* qui violent l'usage ordinaire de la lettre *e*. Mais il aurait fallu se méfier aussi de *W ou le souvenir*, où se mêlent deux écritures alternées et diamétralement opposées, les chapitres en italiques relevant de l'imaginaire et les autres du réel. Devant cette alliance de la carpe et du lapin, l'algorithme s'est trouvé désarçonné et n'a pu trouver l'unité. On a dû pour certains traitements remplacer ce texte par un autre, les *Choses*, sans être sûr qu'un autre piège oulipien ne s'y trouve pas. Au bout du compte force est de reconnaître la solidité de l'intelligence artificielle qui dans cette première expérience n'est prise en défaut qu'une seule fois (sur 25).

#### I.4.2 – Attribuer un texte à un auteur

Fort de cette réussite, on peut tenter d'aller plus loin et de se placer dans la situation générale où un texte supposé inconnu doit être affecté à son auteur. La démarche est la même que précédemment, en revanche l'apprentissage y est plus sûr car il est établi non sur la base fragile d'un seul titre mais sur la production entière de l'écrivain, avec cette réserve que seul le genre romanesque est ici considéré. Nous garderons le même panel des auteurs, mais sans faire la distinction entre Ajar et Gary, qui n'a plus lieu d'être. Restent en compétition 23 écrivains, chacun ayant de 4 à 10 titres romanesques. Pour chaque apprentissage on a ajouté au texte unique de l'expérience précédente autant de romans que nous pouvions, sans atteindre toujours l'exhaustivité. Bien entendu un roman de chacun – le texte frère – était gardé en réserve pour être soumis à la reconnaissance.

Cette fois aucun échec n'est à déplorer. Le roman à identifier est toujours observé dans la case où on l'attend, à sa place sur la diagonale qui croise le texte et l'auteur. Et comme les corpus d'apprentissage sont quatre fois plus larges, les résultats sont plus sûrs et les taux de reconnaissance globalement plus élevés. Ainsi pour Proust le *Temps retrouvé* est mieux identifié (68.55) si on le compare à l'ensemble de la *Recherche* plutôt qu'au premier roman de la *Recherche* (47.5). Comme dans l'étude des paires, Camus est l'auteur le moins bien reconnu, car sa production romanesque est de faible étendue et l'on n'a pu ajouter qu'un seul texte, la *Peste*, au corpus. La base pour Pérec est suffisamment large mais la diversité, qui est ici maximale, affaiblit le taux de

---

<sup>12</sup> On est averti dès la première page : « Je dois donc m'excuser de certaines mutilations, mal-emplois, sauts de carpe, entorses, refus d'obéissance, crabismes, strabismes et immigrations sauvages du langage, syntaxe et vocabulaire. »

reconnaissance (9.06). Néanmoins l'authenticité des 23 textes est affirmée dans la totalité des cas (tableau 10).

Série2	Gide	Prus	Giraudoux	Mauriac	Montherlant	Malraux	Exupéry	Breton	Colette	Giono	Queneau	Aragon	Camus	Vian	Duras	Gracq	Yvesreuar	Mammeri	Gary	Pérec	Tournier	LeClézio	Ernaux
Gide.Immoraliste	33.19	4.2	3.57	6.09	5.36	2.73	1.26	4.41	3.36	3.47	1.79	0.53	1.05	0.63	0.21	1.58	3.05	0.11	1.68	14.29	5.78	1.16	0.53
Proust.TempsRetrouvé	2.46	68.55	1.89	0.76	2.4	0.25	0.06	2.2	2.23	2.48	1.58	1.24	0.51	0.68	0.14	2.88	2.23	0.4	1.04	1.02	2.32	1.38	1.3
Dgraudoux.Bella	7.76	7.55	17.75	9.35	7.38	2.96	1.54	1.03	6.39	2.7	2.49	4.72	0.51	0.99	0.73	2.49	5.23	0.77	3.6	1.5	4.63	4.12	3.82
Mauriac.MystèreFrontenac	5.34	0.64	0.53	29.99	4.59	4.16	0.32	0.43	3.09	4.06	7.58	5.23	1.49	5.23	5.02	1.81	3.63	0.53	0	2.35	12.59	0.85	0.53
Montherlant.Bestiaire	6.37	8.08	1.56	1.74	25.85	2.84	0.38	1.36	2.38	4.69	3.48	9.62	2	1.42	1.01	1.25	9.71	0.9	0.41	3.19	3.91	7.04	0.81
Malraux.Couquérants	1.29	0.59	1.37	0.75	7.9	33.93	1.38	0.78	0.57	3.6	10.61	12.69	2.73	1.89	3.42	1.33	1.12	1.6	1.54	1.61	3.62	4.82	0.88
Exupéry.TerreHommes	5.72	1.25	1.88	8.85	15.04	7.2	14.72	0.78	4.23	3.37	3.99	3.84	2.74	0.78	3.99	1.8	2.35	1.8	1.41	0.55	6.19	5.95	1.57
Breton.AmourFou	6.65	8.76	1.05	0.61	10.51	1.4	1.23	37.57	4.73	1.49	3.06	4.64	1.58	0.44	2.01	3.24	0.88	0.18	0.96	1.75	2.89	1.31	3.06
Colette.Vagabonde	4.18	3.55	1.36	2.82	6.17	1.25	0.84	1.25	39.81	5.02	3.76	3.13	0.52	1.25	1.25	1.78	2.09	0.63	1.67	1.57	7.73	4.08	4.28
Giono.HussardToit	8.21	0.16	0.06	0.6	1.82	3.58	0.57	0	1.29	56.76	10.35	3.68	1.38	2.48	1.07	0.22	0.16	0.72	0.13	0.22	2.52	3.49	0.53
Queneau.ChienDent	2.83	3.38	1.43	0.22	2.49	1.8	0.13	2.07	3.07	3.96	15.76	12.69	0.51	14.94	1.96	0.87	1.36	1.54	5.05	2.89	5.68	10.58	4.77
Aragon.BeauxQuartiers	1.96	4.65	1.66	0.42	2.71	2.79	0.61	5.04	3.38	2.62	5.33	31.45	0.43	0.7	1.34	1.06	1.4	0.98	2.26	3.83	5.85	12.6	6.94
Camus.Etranger	5.42	9.9	2.07	1.67	1.11	3.01	1.54	3.01	3.34	3.41	7.09	7.49	3.81	1.54	1.54	0.94	3.55	2.01	7.89	0.4	9.57	5.48	4.21
Vian.AutomnePékin	0.65	3.07	0.5	0	0.91	0.65	0.1	0.7	1.81	3.62	14.95	4.68	0.7	39.61	1.36	1.21	1.41	1.01	1.56	5.23	11.68	3.77	0.81
Duras.BaragePacifique	1.44	3.5	0.62	0.69	4.6	6.17	0.82	1.17	2.95	4.73	2.47	7.54	2.67	0.69	17.7	1.03	1.3	0.82	3.98	0.89	5.48	11.18	17.63
Gracq.RivageSyrtis	2	1.96	1.08	0.81	6.43	4.93	0.42	0.85	3.85	5.04	3.43	15.05	1.54	3.39	1.5	27.48	3.81	2.35	0.62	1.81	5.54	5.5	0.62
Yvesreuar.OeuvreNoir	7.01	8.5	5.91	1.86	4.42	4.84	1.56	6.74	3.57	1.49	3	2.59	1.39	1.44	0.34	5.06	20.52	1.25	2.76	2	10.06	2.61	1.07
Mammeri.Traversée	2.63	5.02	1.21	1.61	2.63	2.51	1.1	1.37	2.35	4.15	3.37	4.51	1.61	3.17	1.65	1.21	5.37	39.5	2.7	1.06	6.47	3.45	1.37
Gary.VieDevantSoi	10.97	2.87	2.07	0.69	5.52	3.7	0.65	2.58	2.72	2.76	8.17	5.01	1.34	2.36	1.56	0.76	2	1.09	23.46	2.21	7.37	5.3	4.83
Pérec.Choises	2.24	2.91	0.83	1.16	3.33	2.74	3.49	3.91	3.82	3.99	3.41	5.82	1.5	1	2.41	2.16	2	4.07	0.91	9.06	7.15	25.52	6.57
Tournier.Vendredi	1.3	2.7	0.84	0.19	1.12	3.16	0.84	2.05	0.19	3.07	2.79	2.7	3.91	3.07	0.56	2.23	2.42	1.95	1.4	3.81	46.88	12.28	0.56
LeClézio.Désert	1.03	2.1	1.11	0.41	6.64	2.4	0.33	0.41	0.52	2.4	1.14	9.59	2.1	1.88	3.84	0.96	1.55	1.36	2.47	3.32	4.35	46.48	3.61
Ernaux.Années	0.72	4.63	0.62	0.62	1.24	1.13	0.31	3.19	1.44	2.06	1.85	4.33	1.44	1.03	0.72	0.72	1.24	0.41	1.85	3.3	9.27	5.87	52.01

Tableau 10. Attribution d'un titre à un auteur. Première série.

Pour être assuré que ce résultat ne dépend pas du choix du texte soumis à l'examen, nous avons procédé à une seconde série de tests, en inversant les places : le roman que l'on avait séparé de son corpus pour établir son authenticité rentre dans le rang tandis qu'on détache du même corpus le texte-frère qu'on se propose maintenant d'authentifier. Là aussi le verdict du *Deep Learning* est impeccable avec 23 paires reconnues sur 23 (tableau 11).

	Gide	Prus	Giraudoux	Mauriac	Montherlant	Malraux	Exupéry	Breton	Colette	Giono	Queneau	Aragon	Camus	Vian	Duras	Gracq	Yvesreuar	Mammeri	Gary	Pérec	Tournier	LeClézio	Ernaux
Gide.Symphonie	36.86	4.32	2.83	4.52	3.14	2.57	0.72	2.06	4.99	4.88	4.37	0.67	1.18	1.03	0.51	5.6	4.99	2.11	2.42	2.06	6.12	1.49	0.57
Proust.Swann	0.96	74.78	0.8	0.19	2.89	0.32	0.29	2.99	0.83	0.59	1.82	2.56	0.56	0.43	0.27	1.44	1.39	0.24	1.1	0.85	2.16	1.2	1.34
Giraudoux.Simon	2.63	10.79	11.58	1.71	6.44	3.64	0.96	2.26	2.17	2.22	3.85	16.72	1.96	1.46	2.84	1.92	3.6	1.63	3.34	3.14	9.62	3.85	1.67
Mauriac.Baiser	5.88	2.81	1.92	19.34	9.45	6.47	0.6	0.98	2.6	4.9	1.83	19.12	2.47	0.94	1.45	1.36	2.94	0.64	0.81	2.04	8.43	2.13	0.89
Montherlant.Célibataires	1.56	4.33	1.12	0.47	52.23	2.8	0.49	0.77	1.65	4.03	3.18	7.27	1.26	1.43	0.71	1.48	1.78	1.59	0.69	1.56	5.79	3.24	0.58
Malraux.Espoir	4.96	1.17	0.87	1.14	6.87	39.99	0.94	4.36	1.71	2.75	2.08	6	3.05	1.37	1.48	1.71	1.31	0.87	1.14	1.64	4.99	7.84	1.74
Exupéry.CourrierSud	3.09	1.43	2.35	1.38	7.6	3.36	33.21	2.86	4.28	3.04	3.13	2.44	1.89	0.6	1.34	1.84	2.67	1.2	1.43	0.55	11.29	7.69	1.34
Breton.Nadja	2.72	8.33	0.64	0.32	3.36	1.56	0.46	48.46	1.84	0.78	4.1	7.5	0.28	0.64	0.41	2.62	1.2	0.41	1.43	1.7	6.72	2.16	2.35
Colette.Sido	7.18	3.43	1.13	0.91	10.97	2.94	0.94	2.07	29.45	2.91	4.72	6.99	0.45	1.33	0.71	2.62	1.97	0.74	2.72	1.72	5.99	4.01	4.08
Giono.GrandTroupeau	0.82	2.09	0.39	0.41	1.59	1.86	0.34	0.42	0.86	65.15	2.19	3.44	1.95	2.14	0.49	1.02	1.22	1.37	1.13	1.05	5.88	3.36	0.83
Queneau.Zazie	5.92	2.88	0.6	0.24	2.81	2.19	0.13	1.17	3.61	8.69	27.9	13.34	0.69	4.19	2.68	0.71	0.89	1.55	1.57	2.75	6.18	5.21	4.08
Aragon.BlancheOubli	3.17	1.77	0.76	1.03	6.04	2.34	0.55	0.94	3.19	3.83	4.15	54.09	1.39	1.08	1.4	0.83	1.23	0.82	1.04	1.26	4.24	3.15	1.68
Camus.Chute	4.95	6.82	1.16	2.25	6.05	3.28	0.77	1.03	0.64	10.35	3.15	5.08	14.28	2.64	2.32	1.09	2.12	0.64	5.34	0.32	6.75	13.57	5.4
Vian.Ecumé	0.85	1.41	0.47	0.47	1.5	2.32	0.09	1.27	0.85	7.59	5.62	4.33	1.29	40.94	1.21	0.85	4.65	2.83	2.5	2.62	7.77	5.89	2.68
Duras.Amant	0.74	2.8	0.76	0.44	4.79	3.54	0.37	0.84	0.54	8.73	12.56	12.46	3.07	3.37	6.91	0.84	0.69	1.6	3	1.87	3.54	23.6	2.95
Gracq.BalconForêt	4.29	6.23	2.41	1.13	3.26	2.16	1.31	3.28	6.86	3.07	3.2	4.94	0.87	1.88	1.54	24.76	9.12	1.54	2.75	1.15	8.45	4.51	1.29
Yvesreuar.MémoiresHadr	3.26	6.41	3.61	1.18	3.28	2.86	0.77	3.94	2.72	4.92	4.69	7.28	1.51	3.38	1.14	5.83	15.33	1.87	1	8.11	13.25	3.34	0.33
Mammeri.CollineOubliée	0.67	1.27	0.56	0.64	1.57	3.75	0.67	1.05	0.67	2.13	2.81	6.25	2.17	2.1	1.42	1.35	1.35	45.13	1.8	1.05	7.23	12.7	1.65
Gary.DelàLimite	0.9	3.19	0.31	0.03	1.35	0.38	0.14	0.24	0.45	7.48	4.12	4.43	0.94	2.39	1.35	0	1.56	1.8	48.77	1.21	3.5	7.97	7.48
Pérec.HommeDort	1.77	3.96	1.98	0.21	4.10	3.89	0.92	4.25	3.18	2.26	3.89	9.98	3.54	3.26	1.27	3.61	3.47	1.84	0.64	16.63	6.23	14.51	4.60
Tournier.Eléazar	1.06	2.59	1.40	0.34	2.49	1.84	1.06	2.56	0.97	1.56	2.31	1.78	2.03	2.34	0.75	3.77	1.12	0.53	0.72	2.03	59.38	6.11	1.28
LeClézio.Hasard	0.12	0.67	0.30	0.07	1.23	1.13	0.97	0.18	0.67	2.22	0.67	4.25	1.08	0.33	1.27	0.53	0.50	3.02	0.42	0.50	2.82	75.94	1.13
Ernaux.Honte	0.07	3.93	0.53	0.18	1.51	3.36	0.11	3.61	1.33	0.74	3.05	9.78	0.84	1.16	1.93	1.16	1.16	1.54	1.93	5.57	12.51	4.91	39.12

Tableau 11. Attribution d'un titre à un auteur. Deuxième série.

### I.4.3 – Échantillon et population

Soumis à l'épreuve du *deep learning*, aucun texte jusqu'ici ne s'est trouvé assez original et indépendant pour échapper complètement à l'attraction qu'exerce un écrivain sur tout ce qu'il produit. Mais il y a de grandes variations qui tiennent au sujet, au genre, aux circonstances de la rédaction et parfois à la volonté expresse de brouiller les pistes et d'échapper à soi comme aux autres, ce que l'on a constaté pour *Gros-Câlin* de Romain Gary. Il est difficile de considérer un roman particulier comme un échantillon aléatoire dont l'œuvre complète serait la population. Pour s'affranchir des particularités qui marquent chaque texte, on peut constituer un échantillon composite en détachant une page sur dix du corpus-auteur correspondant. La phase d'apprentissage s'exerce sur les corpus dont ces extraits sont exclus, lesquels sont présentés ensuite à l'épreuve de la reconnaissance. Chaque texte ainsi défini est un modèle réduit du corpus-auteur dont il est tiré et il suffit de peu d'intelligence, artificielle ou non, pour le prouver dans les faits. À quoi peut donc servir ce calcul apparemment gratuit et trivial ? Tout d'abord à affirmer la solidité du *deep learning* quand sont gommés les variables et accidents qui accompagnent la naissance des textes. Dans ces conditions idéales, la valeur lue dans la diagonale du tableau 8 ne s'abaisse guère au-dessous de 60%, en ne laissant que de très faibles taux, de 1 ou 2%, aux éléments rivaux, placés sur la même ligne ou la même colonne. De tels écarts excluent l'incertitude.

Cela sert aussi à établir une norme pour étalonner les résultats précédents, acquis sur les données naturelles. Ainsi pour reprendre l'exemple de Proust, le tableau de référence (tableau 12) fournit la valeur maximale 79.62, qui ne dépasse guère celle du *Temps retrouvé* (68.65 dans le tableau 10) et celle du *Côté de Swann* (74.78 dans le tableau 11). Cela prouve le caractère original et homogène de l'écriture proustienne tout au long de la *Recherche*. Inversement la plus faible valeur relevée dans la diagonale (57.3) est relative à Camus, comme c'était le cas dans les tableaux 10 et 11. Cela tient à la faible étendue du corpus Camus, qui ne comprend que 3 romans. En de tels cas les calculs du *deep learning* sont fragilisés. L'intelligence artificielle est liée aux *big data*. Ses résultats sont d'autant plus sûrs que les données sont plus larges. Le tableau 12 en apporte l'illustration : les quatre valeurs les plus élevées concernent Aragon, Le Clézio, Proust et Giono, dont les corpus sont les plus vastes, tandis que les valeurs les plus faibles sont relevées dans les plus petits ensembles romanesques : ceux de Camus et de Breton.

	Gide	Proust	Giraudoux	Mauriac	Montherlant	Malraux	Saint-Exupéry	Breton	Colette	Giono	Queneau	Aragon	Camus	Vian	Duras	Gracq	Yourcenar	Mammeri	Gary	Pérec	Tournier	Le Clézio	Ernaux
Gide	67.85	3.28	1.94	1.4	3.67	1.04	0.68	1.01	2.38	1.51	1.84	1.73	0.61	0.58	0.32	1.37	4	0.58	0.72	0.47	1.01	1.37	0.65
Proust	1.49	79.62	1.57	0.15	2.27	0.33	0.09	1.11	0.81	0.72	0.88	1.92	0.41	0.26	0.41	1.05	2.51	0.33	0.59	1.01	1.23	0.48	0.77
Giraudoux	3.07	5.58	58.42	0.74	2.14	1.49	0.84	0.65	1.67	1.3	1.3	5.67	0.47	0.56	0.28	0.37	4.37	0.37	1.86	1.67	3.91	2.05	1.21
Mauriac	4.42	1.51	2.67	60	2.56	3.26	0.93	0.58	2.44	5.47	0.81	2.56	0.7	0.35	0.81	1.51	3.02	0.7	0.23	0.58	3.49	0.35	1.05
Montherlant	3.51	4.4	2.84	1.11	58.84	2.31	0.75	0.89	1.95	2.71	1.2	4.88	0.44	0.31	0.62	1.33	2.31	1.02	0.75	0.89	1.95	3.51	1.47
Malraux	1.22	0.44	1.17	0.88	2.14	74.25	0.44	0.54	0.83	1.61	0.78	4.87	0.68	0.29	0.19	0.68	1.41	0.54	1.12	0.88	2.34	1.75	0.97
Saint-Exupéry	1.96	0.82	3.61	0.32	2.28	1.52	72.12	0.89	1.65	1.52	0.95	2.6	0.19	0.06	0.7	0.57	2.6	0.57	0.32	0.19	2.79	1.58	0.19
Breton	2.7	7.38	1.72	0.25	2.62	1.23	1.07	52.21	1.48	1.64	0.9	5.08	0.49	0.25	0.49	1.89	2.87	0.49	0.98	4.43	5.98	1.72	2.13
Colette	4.48	2.09	3.39	1.3	3.97	1.01	0.87	0.65	63.25	1.73	1.3	4.33	0.07	0.72	0.36	1.08	3.03	0.14	0.87	1.3	1.59	0.94	1.52
Giono	0.54	1.11	0.43	0.36	1.02	0.97	0.14	0.16	0.79	80.85	1.02	2.06	0.7	0.5	0.36	0.38	1.33	0.88	0.72	0.66	2.6	1.85	0.57
Queneau	1.8	1.28	0.88	0.11	1.58	0.55	0.51	0.51	0.99	2.13	73.58	3.7	0.29	0.77	0.99	0.62	0.88	0.73	1.58	1.65	2.93	0.95	0.99
Aragon	0.66	1.08	1.12	0.4	1.42	1.06	0.15	0.5	1.08	1.23	1.04	81.45	0.11	0.46	0.53	0.61	0.95	0.43	0.84	1.24	1.23	1.29	1.12
Camus	1.76	2.84	2.84	1.86	2.05	2.25	0.49	0.68	0.49	5.58	0.68	3.82	53.23	0.78	1.96	0.78	3.23	0.39	2.35	1.66	5.28	3.91	1.08
Vian	0.83	0.7	0.7	0.19	0.7	0.77	0.26	0.32	0.51	3.64	1.98	2.04	0.7	73.98	1.08	0.89	0.96	0.26	2.55	1.72	2.1	2.17	0.96
Duras	1.46	1.99	0.33	0.46	1.39	0.93	0.33	0.4	0.8	2.59	1.59	4.64	0.8	0.66	69.81	0.53	1	0.46	1.46	1.06	0.73	4.31	2.26
Gracq	1.7	3.13	2.02	0.69	2.34	1.01	0.32	3.61	2.44	1.54	0.9	3.72	0.85	0.32	0.53	57.3	5.1	0.96	1.06	3.08	3.24	3.03	1.12
Yourcenar	2.68	2.97	2.59	0.92	1.88	1.05	0.46	0.79	2.22	2.93	0.71	2.59	0.33	0.79	0.38	1.71	63.49	1	1.8	1.13	3.76	3.09	0.71
Mammeri	0.9	1.25	0.56	0.42	1.39	2.08	0.56	0.42	0.62	3.05	1.25	3.33	0.76	0.56	1.04	0.35	1.25	73.07	1.32	0.9	1.8	2.15	0.97
Gary	1.65	1.91	1.71	0.2	1.19	0.72	0.4	0.53	1.12	1.58	1.91	3.75	0.53	0.53	0.53	0.4	2.57	0.4	70.16	1.58	3.49	1.05	2.11
Pérec	1.03	2	1.23	0.19	1.55	1.16	0.32	1.55	0.26	1.1	2.06	3.48	0.32	0.77	0.64	0.58	1.48	1.1	1.81	67.76	3.29	3.48	2.84
Tournier	0.76	1.53	1.47	0.54	1.3	0.96	0.65	1.1	0.68	3.08	1.27	2.85	0.71	0.37	0.54	1.16	2.12	0.65	1.58	2.15	70.84	1.72	1.98
Le Clézio	0.34	0.81	0.57	0.09	1.29	0.67	0.22	0.48	0.41	1.88	0.77	2.6	0.38	0.67	0.95	0.65	1.15	0.65	0.62	1.81	1.67	79.4	1.91
Ernaux	0.91	2.28	0.71	0.76	1.82	0.61	0.1	1.16	1.92	1.92	1.27	4.15	0.76	0.66	1.67	0.56	0.96	0.65	2.23	3.49	4.2	2.68	64.63

Tableau 12. La reconnaissance des échantillons aléatoires.

## II – Le théâtre classique

Le soupçon que Corneille aurait pu écrire certaines pièces de Molière et en particulier *Amphitryon* est né en 1919, deux siècles après la mort des deux écrivains, dans l'imagination de Pierre Louÿs, lui-même auteur d'une supercherie littéraire, *Les Chansons de Bilitis*. Pierre Louÿs a beau multiplier ses efforts, il n'arrive pas à convaincre son ami Paul Valéry, non plus que l'opinion générale, même si de loin en loin certains esprits remuent encore les braises. L'affaire aurait pu et dû s'arrêter là, par un jugement autorisé de l'histoire littéraire<sup>13</sup>. Mais la lexicométrie s'est lancée dans le débat avec un article plutôt téméraire de Dominique Labbé qui avait construit un algorithme pour mesurer la distance intertextuelle et entendait l'appliquer au théâtre classique<sup>14</sup>. Nous avons eu l'occasion d'utiliser cette méthode et notre tableau 2 montre qu'elle s'accorde assez bien avec les autres, sans être pour autant la meilleure. Encore doit-on l'appliquer avec prudence et refuser tout barème absolu, pour tenir compte des forces qui rapprochent ou éloignent les textes et dont la principale est le genre littéraire. Il n'est guère étonnant que certaines pièces de Molière soient proches de celles de Corneille quand elles suivent les mêmes règles et partagent le même genre comique et la même forme versifiée. Charles Bernet a montré que d'autres dramaturges de la même époque

<sup>13</sup> Georges Forestier qui s'est placé sur le terrain de la vérité historique a fait le point sur cette question dans son site <http://www.moliere-corneille.paris-sorbonne.fr/>

<sup>14</sup> Dominique et Cyril Labbé, « Inter-Textual Distance and Authorship Attribution. Corneille and Molière », dans *Journal of Quantitative Linguistics*, vol. 8, n° 3, 2001, pp. 213-231.

avaient une ressemblance aussi étroite avec le modèle de Corneille<sup>15</sup>. Comme nous avons publié quelques articles sur cette question, on nous permettra d'y renvoyer le lecteur<sup>16</sup>. Nous nous contenterons de refaire le calcul lexicométrique avec les mêmes données, puisées au Trésor de la Langue Française.

## II.1 - Les basses fréquences

1.1 Les figures 13 (distance d'Évrard) et 14 (distance de Muller) ont été obtenues grâce à un programme d'analyse arborée semblable à celui qu'a emprunté Dominique Labbé. Seule diffère l'interprétation que l'on peut en tirer. Pas besoin d'être un expert pour lire sur le graphique la distinction radicale entre les vers (partie droite du graphique) et la prose (partie gauche) et celle qui oppose parallèlement la tragédie (à droite) et la comédie (à gauche). Après le genre s'impose la signature des écrivains. Les pièces de Racine font bloc à gauche, en s'écartant avec le temps vers la périphérie, les premières pièces restant proches de Corneille. La partie occupée par Corneille se divise en deux parties, qu'il s'agisse de tragédies ou de comédies (ou pièces assimilées). Reste Molière, nettement séparé du reste et lui aussi réparti en deux blocs, d'abord les comédies en vers, puis les comédies en prose. Il est rare que l'analyse arborée propose un résultat aussi lisible, et que la convergence soit obtenue à partir d'observables différents (présence/absence dans la méthode Évrard, fréquence dans la méthode Muller), alors même que deux facteurs, le genre et l'auteur, s'y entrecroisent sans trop se combattre.

Il y a cependant trois points de friction où le genre l'emporte sur la signature. D'une part *Dom Garcie de Navarre*, seule pièce sérieuse de Molière, se positionne entre les tragédies de Corneille et de Racine, au voisinage de *Psyché* dont le statut est particulier, puisque le canevas est de Molière et la versification en grande partie de Corneille. Inversement le clan de Molière reçoit une pièce transfuge, *Les Plaideurs*, que le programme n'a pas reconnue parmi les tragédies de Racine.

Quant aux deux *Menteurs* sur lesquels s'appuie l'argumentation de Labbé, ils se situent là où on les attend, au milieu des comédies en vers de Corneille. Rien ne permet d'en faire la pierre de touche et de proposer l'argumentation boiteuse : 1 - les *Menteurs* sont de Corneille, 2 - Les pièces en vers de Molière sont proches des *Menteurs*, 3 - donc Corneille est l'auteur de ces pièces. Avec pareil raisonnement le graphique inviterait à conclure que les *Plaideurs* sont de Molière.

<sup>15</sup> Charles Bernet, « La distance intertextuelle et le théâtre du Grand Siècle » [archive], in *Mélanges offerts à Charles Muller pour son centième anniversaire (22 septembre 2009)*, textes réunis par Christian Delcourt et Marc Hug, Paris, CILF, 2009, pp. 87-97.

<sup>16</sup> Etienne Brunet, « Où l'on mesure la distance entre les distances », in *Texte !* [en ligne], mars 2004, [http://www.revue-texto.net/Inedits/Brunet/Brunet\\_Distance.html](http://www.revue-texto.net/Inedits/Brunet/Brunet_Distance.html), texte repris dans E. Brunet, C. Poudat, *Ce qui compte*, Champion, Paris, 2011, pp. 331-351.

Etienne Brunet, « Muller le lexicomaître », in *Mélanges offerts à Charles Muller pour son centième anniversaire*, Conseil International de la langue française, Paris, 2009, pp. 99-119.

Etienne Brunet, « Le logiciel Franstat », in *Statistical Analysis of textual data*, JaDT 2016, Rome, juin 2016, vol.1, pp. 143-155.



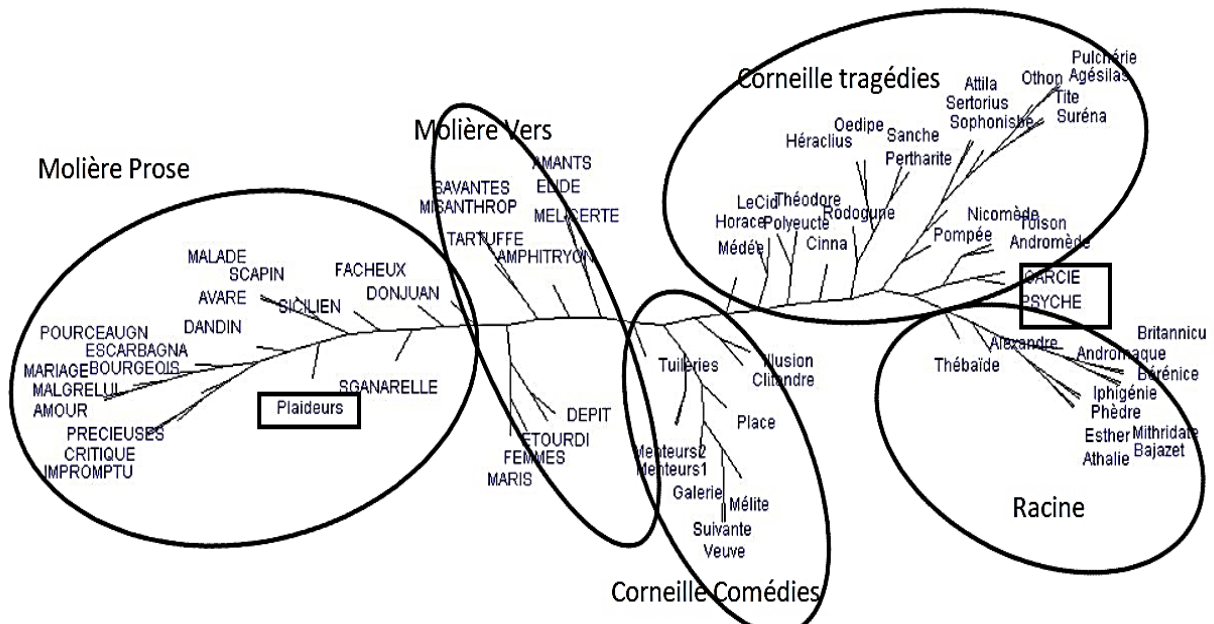


Figure 13. Analyse directe (méthode Évrard, présence/absence)

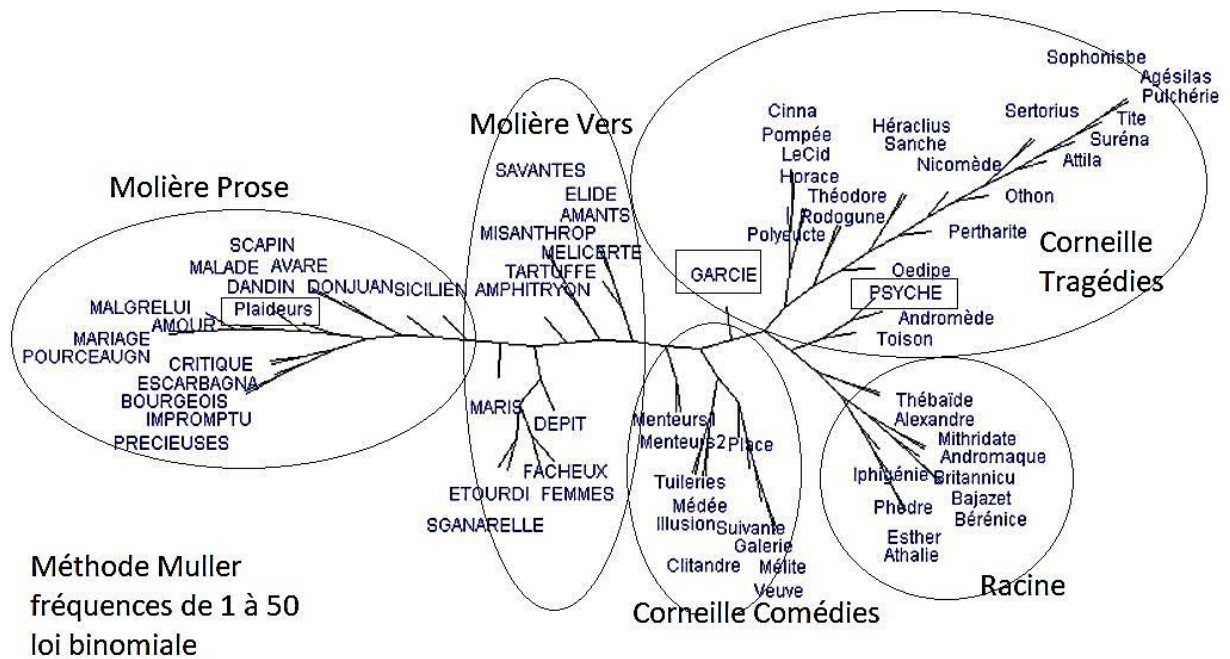


Figure 14. Méthode Muller, appliquée aux fréquences de 1 à 50

1.2 – Rappelons que la distance d'Evrard repose sur un calcul binaire où l'on ne retient que la présence/absence des mots dans les textes. Chaque mot y est intégré dans des effectifs globaux, mais son individualité n'est pas prise en compte. Si l'on veut détailler de profil de chacun, tout en s'abstenant de considérer les fréquences, on constituera un tableau binaire, de  $n$  mots sur  $t$  textes, dont chaque élément recevra un "1" quand le mot se trouve dans le texte considéré et "0" dans le cas contraire. Il n'y a pas lieu de retenir les mots présents partout, dont le profil ne contient que des "1", non

plus que les hapax qui n'en contiennent qu'un seul, la neutralité des premiers et la singularité des seconds n'ayant pas d'intérêt pour le calcul. Seules les lignes comportant au moins deux "1" sont gardées. Avec ces critères, il reste 9761 graphies, soumises à l'analyse en composantes principales. Le plan (2,3) de l'ACP est représenté dans la figure 15.

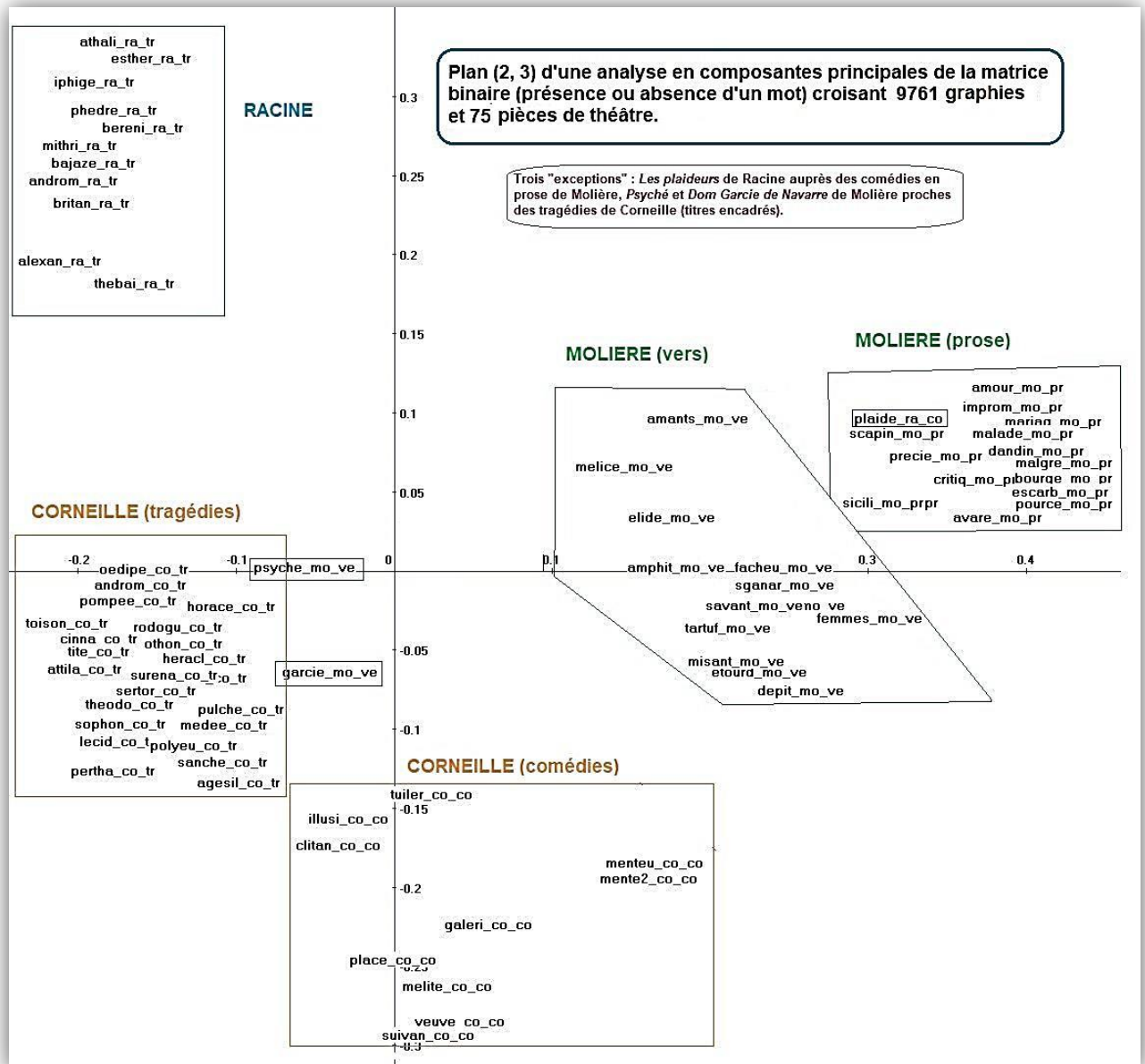


Figure 15. Analyse factorielle du tableau binaire (9761 graphies x 75 textes)

Cet espace, réduit à deux dimensions, contient, ce qui est exceptionnel, une information qui ne sera améliorée ni par la prise en compte d'autres dimensions ni par l'emploi d'autres techniques d'exploration de données. La preuve en est fournie par le panorama spectaculaire de la figure 16, qui concentre tous les axes de l'ACP précédente. Cette figure issue des données brutes, sans choix de seuil ni de nombre

d'axes, suffirait à elle seule pour invalider clairement et définitivement les thèses discutant l'auctorialité de Molière.

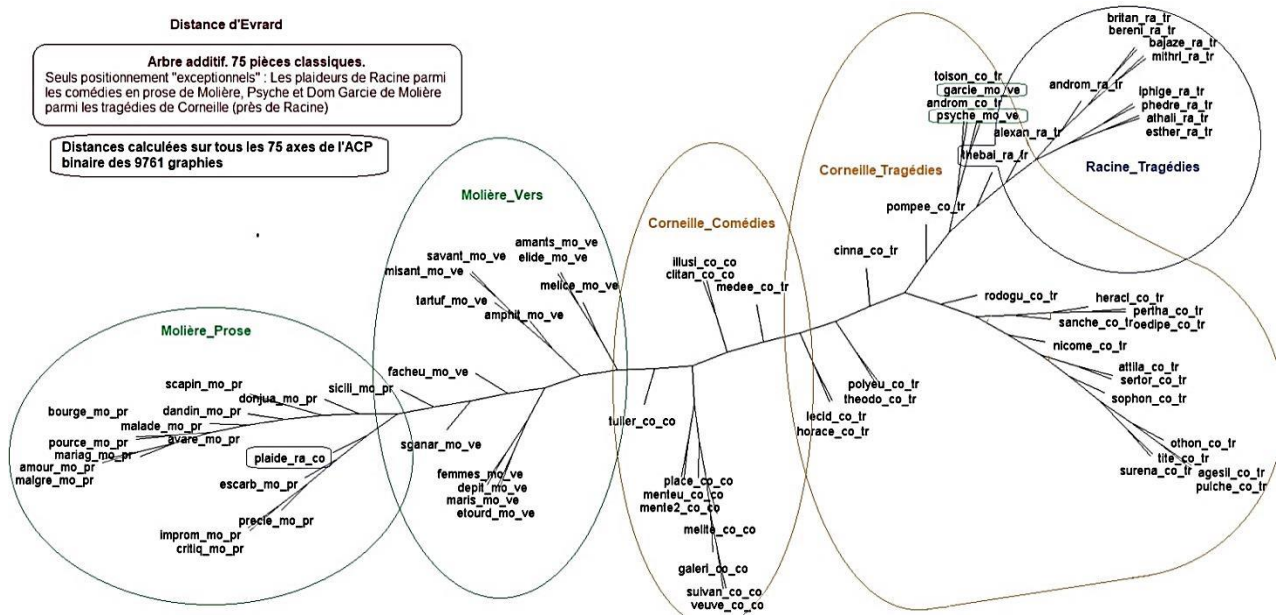


Figure 16. Arbre additif rendant compte des 75 axes de l'ACP

1.3 – Les résultats qui précèdent prennent en compte les simples graphies, sans aucun traitement préalable. Ils ne changent guère lorsqu'ils s'appliquent aux lemmes et l'on évitera de produire d'autres graphiques, qui reproduisent à l'identique ceux qu'on vient de voir. Mais on peut pousser plus avant l'analyse de la diversité stylistique en s'intéressant non plus au mot mais au code grammatical que le logiciel de lemmatisation accole à chaque unité. L'analyse de Cordial distingue plusieurs centaines de variétés, en distinguant le genre, le nombre, la personne, le temps et le mode des verbes, quand celle de TreeTagger se contente d'une trentaine de codes en ignorant genre, nombre et personne. On peut naturellement constituer un tableau avec les codes en ligne et les textes en colonne, et procéder à des analyses factorielles et arborées. Le résultat est à l'abri des influences thématiques qui influent grandement dans le tableau 13 mais il confirme pareillement, sur le plan grammatical, la séparation radicale des écrivains. Cependant en n'envisageant que les codes individuels, isolés les uns des autres, on donne une image très réductrice de la syntaxe, en ignorant que la syntaxe est avant tout une combinaison de codes, c'est-à-dire une structure. On a défini comme structure la suite des codes grammaticaux rencontrés entre deux ponctuations. Comme la combinatoire risque d'être trop étendue, on n'a considéré que les parties du discours, soit une dizaine de codes, chacun désigné par une lettre à l'intérieur d'une chaîne représentant le syntagme<sup>17</sup>. Tout le corpus devient alors une suite continue de structures dont le traitement est analogue à celui des mots. Or la diversité des structures

<sup>17</sup> Ainsi le dernier segment de l'hymne national « *Qu'un sang impur abreuve nos sillons* » serait représenté par la chaîne *cdnandn*, pour c=conjonction, d=déterminant, n=nom, a= adjectif, v=verbe.



l'emporte sur celle des mots : on en compte 68774 variétés différentes, soit trois fois plus que le nombre de mots différents. La plupart des structures relevées sont des hapax. Il convient donc de les traiter avec les outils adaptés aux basses fréquences.

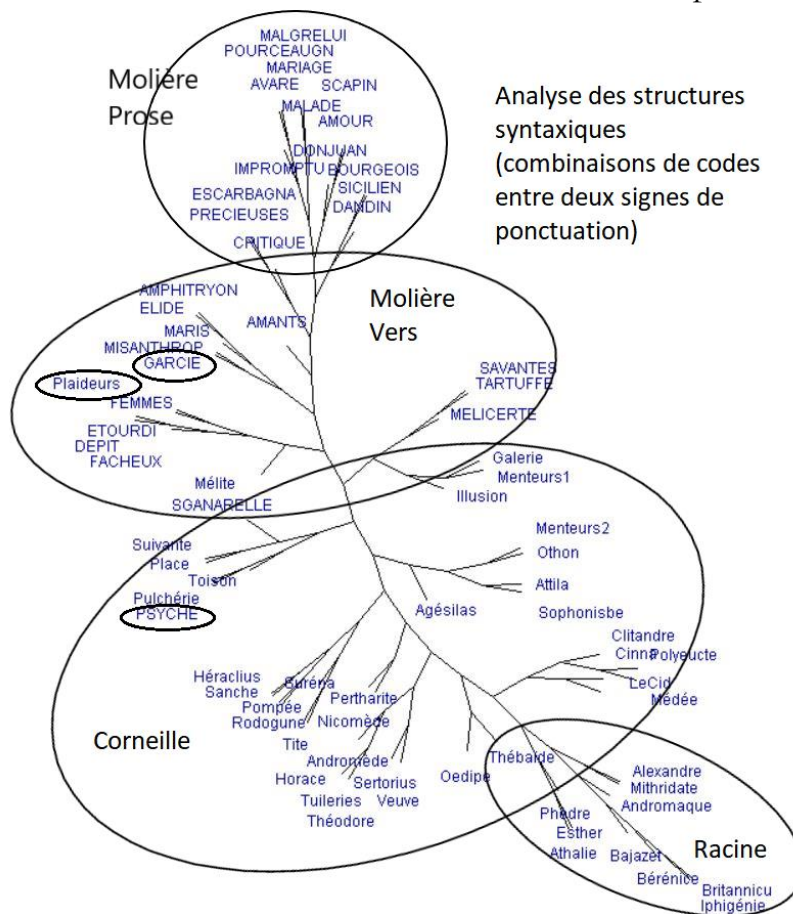


Figure 17. Les structures syntaxiques dans le théâtre classique.

Nous avons choisi la méthode Jaccard pour l'arbre additif de ces structures. On note peu de différence par rapport aux graphiques 13, 14 et 16 alors que le point de vue grammatical n'est pas nécessairement lié à la composante lexicale. La prose s'oppose aux vers et les écrivains ne mêlent pas leurs productions, sauf quand l'impérieuse voix du genre l'impose comme on le voit avec les *Plaideurs*. Le cas ambigu de *Psyché* n'appelle pas de commentaire. Mais celui de *Garcie* qui reste dans le camp de Molière provoque une heureuse surprise.

## II.2 - Les hautes fréquences

2.1 - Les démarches qui précèdent prennent appui sur les basses fréquences. Si l'on prend en compte les hautes fréquences<sup>18</sup>, les conclusions sont les mêmes et la figure 18, obtenue à partir de l'analyse arborée des 519 verbes (de fréquence égale ou

<sup>18</sup> C'est la méthode employée par J.M. Viprey, dès le début de l'affaire pour réfuter la thèse de Labbé ([http://elliadd.univcomte.fr/archives\\_laseldi/document/affaireMorneilleColiere/morneille.htm](http://elliadd.univcomte.fr/archives_laseldi/document/affaireMorneilleColiere/morneille.htm)),

supérieure à 50), est superposable aux figures 13, 14 et 16. On y retrouve, dans le même ordre et avec la même composition, les cinq groupes où la combinaison du genre, de la date et de l'auteur avait réparti les 75 textes du théâtre classique. Au centre de la figure les comédies en vers de Corneille et de Molière forment deux blocs voisins, mais parfaitement distincts. On y observe pareillement, hésitant entre Molière et Corneille et entre la comédie à gauche et la tragédie à droite, les deux pièces relevées précédemment parmi les exceptions facilement explicables : *Psyché* et *Don Garcie de Navarre*.

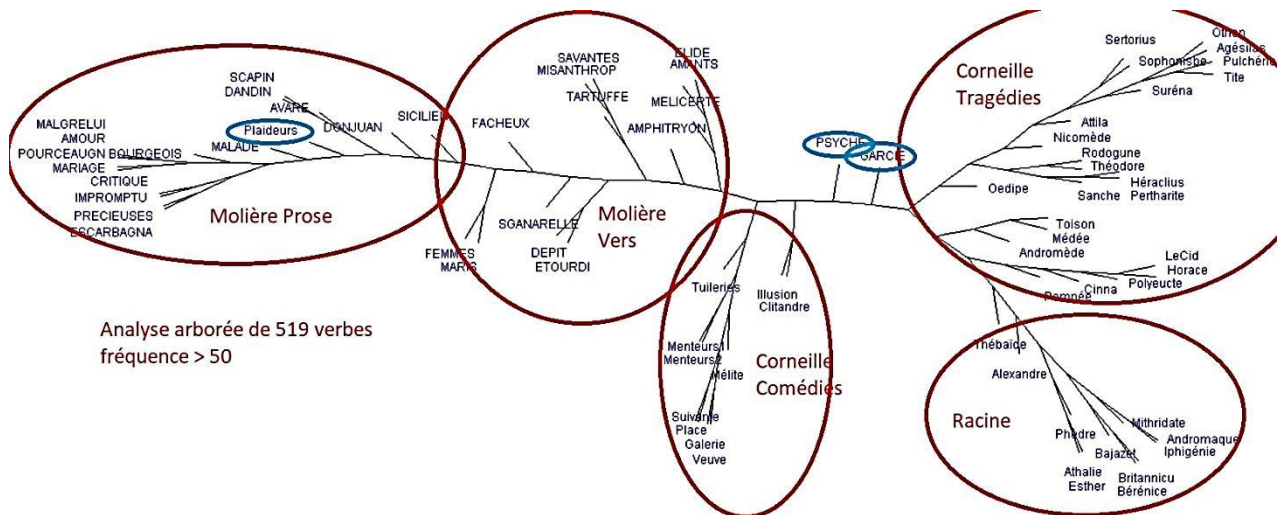


Figure 18. Le théâtre classique. Analyse des textes à travers les verbes fréquents ( $f \geq 50$ )

2.2 – Avant d'appliquer des techniques relevant de l'intelligence artificielle au théâtre classique, il peut être intéressant d'expérimenter une méthode également neuronale, que l'on désigne sous le nom de SOM (Self Organizing Maps) ou Cartes de Kohonen<sup>19</sup>. Il s'agit ici d'une classification (clustering) des 75 pièces en 25 classes (clusters), celles-ci étant astreintes à être positionnées sur une grille (ici un carré 5 x 5). Les positions sur la grille doivent refléter les proximités entre classes. Le résultat est aussi spectaculaire que celui des arbres additifs. Six classes sont vides: elles séparent des classes trop différentes pour être contiguës, en général entre auteurs différents, sauf pour *Mélicerte* de Molière, isolé dans une classe. Molière a droit aux six classes de la partie haute de la grille (et incorpore *Les Plaideurs*). *Psyché* et *Dom Garcie de Navarre* vont dans une classe de pièces de Corneille, mais c'est la seule classe qui est tangente à une classe de Molière. Racine, sans les *Plaideurs*, occupe la rangée du bas, mais *La Thébaïde* est dans une classe de pièces de Corneille, *Alexandre le Grand* étant proche de cette même classe. En bref, un algorithme radicalement différent nous confirme les résultats des diverses analyses précédentes.

<sup>19</sup> Pour le détail de cette technique, voir L. Lebart, B. Pincemin et C. Poudat, *Analyse des données textuelles*, Presses de l'Université du Québec, 2019, pp.279-282.

scapin_mo_pr bourgeois_mo	sicilien_mo_ precieuses_m impromptu_mo escarbagnas_ critique_mo_	<b>MOLIERE</b>	melicerte_mo	
pourceaugnac mariage_mo_p malgrelui_mo malade_mo_pr donjuan_mo_p dandin_mo_pr avare_mo_pr_ amour_mo_pr_	plaideurs_ra femmes_mo_ve elide_mo_ve_ amants_mo_ve	tartuffe_mo_ sganarelle_m savantes_mo_ misanthrope_ maris_mo_ve_ facheux_mo_v etourdi_mo_v depit_mo_ve_ amphitryon_m		suivante_co_ menteurs2_co menteurs1_co
<div style="border: 1px solid red; padding: 5px; color: red; font-weight: bold;">           Carte auto-organisée (carte de Kohonen)            Les 75 pièces sont caractérisées par les 4945            graphies apparaissant au moins 10 fois            [ 5 x 5 = 25 classes de pièces ]         </div>		théodore_co_ psyche_mo_ve polyeucte_co pertharite_c nicomède_co_ lecid_co_tr_ garcie_mo_ve andromède_co	tuileries_co illusion_co_ ciltandre_co	veuve_co_co_ place_co_co_ mélie_co_co galerie_co_c
file_co_tr_3 suréna_co_tr sertorius_co pulchérie_co attila_co_tr	sanche_co_tr héraclius_co	toison_co_tr thébaïde_ra_ rodogune_co pompée_co_tr oedipe_co_tr horace_co_tr cinna_co_tr_ alexandre_ra	<b>CORNEILLE</b>	médée_co_tr
sophonisbe_c othon_co_tr_ agésilas_co	bajazet_ra_t	<b>RACINE</b>	phèdre_ra_tr iphigénie_ra britannicus_ andromaque_r	esther_ra_tr athalie_raci

Figure 19. Carte de Kohonen, appliquée au théâtre classique

### II.3 - Neutralisation du genre

Si l'on veut comparer deux écrivains et éventuellement prouver qu'il s'agit d'un seul, il faut s'attacher à soumettre les textes aux mêmes conditions et à isoler un seul facteur en neutralisant les autres (et principalement le genre). Isolons donc les comédies en vers, qui sont soumises aux mêmes contraintes de genre, et voyons si celles de Molière et de Corneille se confondent. La réponse, évidente, est dans la figure 20. Qu'elle soit fondée sur la fréquence des mots (méthodes Labbé et Muller) ou sur la présence-absence (méthodes Jaccard et Evrard), l'analyse montre une séparation radicale des textes de l'un et l'autre écrivains<sup>20</sup>.

Comme ces quatre méthodes donnent la préférence aux basses fréquences, il convient d'y ajouter le témoignage des hautes fréquences. On a donc réduit le corpus classique à un ensemble de 23 comédies en vers, dont 10 de Corneille et 13 de Molière. L'analyse factorielle de correspondance, réalisée sur les lemmes de fréquence égale ou supérieure à 50, montre dans le graphique 21 une séparation franche entre Molière à gauche et Corneille à droite. Un calcul « bootstrap » de contrôle y est pratiqué sur certains éléments : l'étrécissement de l'ellipse qui enveloppe chacun d'eux montre la robustesse de la représentation. Si donc Corneille a écrit les pièces de Molière, il a bien caché son jeu, en abandonnant sa manière propre, pour imiter celle de Molière !

<sup>20</sup> Une analyse arborée, en tous points semblable, figure dans l'article de Ch. Bernet, p. 97. Établie sur un nombre moindre de textes et sur un objet plus précis (les mots à la rime), elle montre pareillement la différence entre les deux écrivains.





## II.4 – Le Deep Learning

Les batailles d'experts peuvent être interminables quand elles utilisent les mêmes armes. Mais elles doivent s'incliner quand le progrès offre une technique révolutionnaire qui met fin au débat. C'est le cas de l'ADN pour les recherches en paternité ou en criminalité. Or, l'intelligence artificielle est appelée à jouer de plus en plus ce rôle de dernier recours. Déjà dans la médecine son verdict, sans être infaillible, dépasse celui du praticien et dans tous les domaines, et plus particulièrement dans les sciences humaines, son pouvoir s'étend au-delà des capacités de la raison individuelle, pour peu que les données aient la taille suffisante. Voyons ce qu'il en est pour le problème qui nous occupe.

Il y a lieu d'établir trois corpus indépendants, voués à Corneille, Molière et Racine. Mais pour se prémunir des perturbations du genre on distinguera chez Molière la prose et les vers et chez Corneille la comédie et la tragédie. La nécessité de distinguer les phases d'apprentissage des corpus et de reconnaissance des textes nous oblige à mettre de côté les textes à soumettre au test. On en choisira deux ou trois dans chaque corpus, soit 14 pièces sur les 75 du théâtre classique.

	Corneille comédie	Corneille tragédie	Molière vers	Molière prose	Racine tragédie
AMPHITRYON	20.63	7.4	<b>52.35</b>	16.36	3.27
TARTUFFE	7.51	7.63	<b>65.91</b>	17.76	1.19
AVARE	3.81	1.65	26.6	<b>67.22</b>	0.72
DON JUAN	4.61	4.11	33.17	<b>57.23</b>	0.87
GARCIE	14.36	<b>43.99</b>	36.03	0.65	4.96
MELITE	<b>75.4</b>	11.44	11.3	1.33	0.53
MEDEE	31.64	<b>56.07</b>	4.15	0.31	7.83
MENTEUR1	<b>47.28</b>	11.1	30.29	9.27	2.09
MENTEUR2	<b>46.51</b>	19.96	24.85	7.22	1.47
CID	36.68	<b>51.17</b>	3.13	0.78	8.22
HORACE	19.11	<b>65.85</b>	5.56	0.68	8.81
PSYCHE-TOUT	19.56	<b>50.34</b>	18.74	2.74	8.62
PSYCHE-MOLIERE	14.29	<b>40.54</b>	35.52	5.02	4.63
PSYCHE-CORNEILLE	18.75	<b>60.73</b>	11.46	0	9.02
PSYCHE-QUINAULT	<b>43.86</b>	8.77	21.05	7.02	19.3
PHÈDRE	11.26	17.82	3.02	0.15	<b>63.75</b>
PLAIDEURS	6.03	0.48	30.02	<b>61.99</b>	1.21

Tableau 22. Le jugement de l'intelligence artificielle dans l'affaire Molière-Corneille.

*Amphitryon*, qui a d'abord éveillé les soupçons de Pierre Louÿs, est le premier d'entre eux dans le tableau 22. Le verdict est sans appel et range cette pièce à 52.35% du côté de Molière-vers, contre 20.63% du côté de Corneille-comédie. C'est pourtant ce texte qui est le moins éloigné de Corneille. L'écart est encore plus grand dans *Tartuffe*

(65.91 contre 7.51). Et il s'aggrave encore quand l'opposition de genre s'ajoute à celle des auteurs : 67.22 contre 3.81 pour l'*Avare* et 57.23 contre 4.6 pour *Dom Juan*. Comme aucune tragédie ne figure dans l'apprentissage de Molière, la pièce sérieuse *Dom Garvie de Navarre* apparaît étrange et l'algorithme hésite entre deux choix : celui du genre ou celui de l'auteur. Le genre l'emporte de peu et le texte est attribué à la tragédie de Corneille (43.99 contre 36.03). Le genre prédomine encore dans le cas de la comédie des *Plaideurs*, qui, le comique n'entrant pas dans l'apprentissage de Racine, est reconnue comme une comédie de Molière. Ces deux pièces étaient pareillement soumises à la domination du genre dans les analyses classiques (figure 13 et suivantes) et loin d'en tirer argument contre le *deep learning* on lui fera crédit de cette sensibilité à la puissance du genre<sup>21</sup>. Partout ailleurs des pourcentages écrasants désignent l'auteur réel. *Médée*, *Le Cid*, et *Horace* sont reconnus aisément comme des tragédies de Corneille, tandis que *Mélite* et les deux *Menteurs*<sup>22</sup> sont attribués sans hésitation à la comédie cornélienne. Quant à Racine, il reste étranger à toutes les pièces sauf à la sienne *Phèdre*, avec une préférence qui atteint 63.75 %.

Afin de compléter la démonstration, on a examiné les autres pièces en vers de Molière, en réordonnant à chaque fois le corpus afin de réincorporer le texte précédemment testé tout en excluant celui qu'on veut examiner. À chaque fois le corpus Molière-vers est soumis derechef à l'apprentissage en même temps que les autres corpus qui restent inchangés (mais on n'a plus jugé utile de faire intervenir Racine). Le constat (tableau 23) est cruel pour la théorie de Pierre Louÿs : aucune des pièces de Molière ne se compromet avec celles de Corneille, et les chefs-d'œuvre moins que les autres (*Misanthrope* 71% contre 4%, *Femmes savantes* 63% contre 5%).

	Corneille comédie	Corneille tragédie	Molière vers	Molière prose
ETOURDI	17.15	5.52	<b>59.49</b>	17.84
DEPIT AMOUREUX	15.8	8.62	<b>62.4</b>	13.19
SGANARELLE	7.8	7.09	<b>65.6</b>	19.5
ECOLE DES MARIS	11.3	4.74	<b>66.8</b>	17.32
FACHEUX	8.92	7.3	<b>63.38</b>	15.41
ECOLE DES FEMMES	9.04	4.52	<b>61.11</b>	25.32
ELIDE	3.42	17.87	<b>42.21</b>	36.5
MISANTHROPE	4.16	7.93	<b>71.91</b>	15.99
MELICERTE	5.84	14.4	<b>73.15</b>	6.61
FEMMES SAVANTES	5.24	6.39	<b>63.04</b>	25.32

Tableau 23. Les comédies en vers de Molière.

<sup>21</sup> *Psyché* représente un cas à part, puisque la majeure partie a été écrite par Corneille, Molière se réservant, outre le canevas, l'acte 1 et deux scènes. Le *deep learning* reconnaît la prédominance de Corneille, même là où Molière tient la plume (40 contre 35%). Mais ici les données sont de trop faible étendue pour que la puissance du programme puisse se déployer.

<sup>22</sup> Notons que les deux *Menteurs*, tout en se raccrochant solidement à Corneille (47 et 46%), sont relativement proches de Molière, avec des scores non négligeables de 30 et 24%. Cela explique, sans la justifier, la démarche de Labbé qui fait des *Menteurs* le centre du débat.

La cause est entendue. L'intelligence artificielle confirme avec autorité ce que propose la lexicométrie intelligente. Quant à l'opinion publique, qui se contente de lire le texte ou d'assister aux représentations, elle n'a besoin ni de l'une ni de l'autre pour rester fidèle à Molière.

### III – Essai d'explicitation et d'intégration

On aimerait qu'une troisième voie puisse départager les deux approches, et décider du sort de l'huître entre les plaideurs. Peut-on imaginer une expérimentation où l'homme soit confronté à la machine et invité à faire le tri sans autre truchement que la lecture. Le même lot de 50 romans serait livré à l'adresse d'un lecteur professionnel, comme Bernard Pivot. Ce marathonnier de la lecture – comme il se définit lui-même – aurait un livre à expertiser chaque jour, performance dont il est capable. Au bout de sept semaines plus un jour, il donnerait son verdict. Serait-il meilleur que celui que la machine livre en une heure ? L'homme a au départ des avantages qui sont refusés à l'ordinateur. Un bibliovore comme Pivot a probablement lu dans le passé les 50 titres du panel. Et, quoi qu'il dise de sa mauvaise mémoire, il a gardé des traces de ses lectures et, par exemple, il lui suffirait d'une minute pour reconnaître la manière de Proust. Et puis il dispose d'une culture historique, acquise non seulement dans la lecture antérieure de ces 50 ouvrages, mais aussi à l'extérieur, dans les milliers de pages parcourues au cours d'une vie, alors que la machine s'en tient aux seuls documents qu'on lui fournit. Que survienne un nom propre, comme Combray, Charlus ou Guermantes, cela suffit pour classer l'ouvrage. L'ordinateur n'y voit que des mots vides de sens et de référence et la majuscule ne l'autorise pas à consulter le Larousse. Et pourtant qui oserait parier que même un Pivot rendrait une dictée sans faute ? Certes il n'achopperait pas aux mêmes endroits que la machine et ne buterait pas sur Pérec ou Camus. Mais, dépouillés de la couverture, sans titre et sans auteur, certains romans, moins connus ou plus lointains dans la mémoire, manqueraient sûrement de conviction pour entrer dans un couple. Et que dire d'un lecteur, supposé innocent et inculte, qui, comme la machine, n'aurait lu aucun livre avant de se soumettre à l'expérience et ne connaîtrait que l'alphabet et les ponctuations, sans aucune notion du lexique, de la grammaire, de la sémantique et sans aucun accès à la vraisemblance ?

#### III.1 – Imitation de la procédure du *deep learning* : les triplets

Nous avons imaginé un automate, pareillement naïf et ignorant, qui imiterait la démarche du *deep learning*, tout en s'en tenant aux procédures de la statistique classique. Au lieu de s'élever à l'archilecteur, on s'est tenu à un rôle d'infra-lecteur, qui ânonne les lettres comme un écolier, le doigt appuyé sous le texte, en appliquant une méthode globale de trois mots à la fois. On a déjà dit en effet que le *deep learning* – au moins la

version à notre disposition –, ne s'intéresse pas aux mots individuels mais à des assemblages (en principe de trois mots dans un espace de 50). Grâce à une fenêtre glissante, on prépare le texte en collant à chaque mot celui qui le précède et celui qui le suit<sup>23</sup>. Les mots ainsi gonflés – qu'on appellera « triplets » – sont plus précis et plus rares que les mots individuels, et chacun livre plus d'information pour dissoudre l'entropie. Naturellement ces monstres lexicaux n'existent pas dans le dictionnaire non plus que dans les acquisitions mémorielles (en dehors d'expressions composées du type « pomme de terre » ou « chemin de fer »). On perd ainsi toute attache avec le lexique et conséquemment avec la syntaxe, la sémantique et l'expérience.

### III. 1.1 – La méthode Muller

Le traitement s'abstient de tout codage ou lemmatisation et fait appel à la version simple de notre logiciel Hyperbase (prévue pour les graphies). L'indexation catalogue tous ces triplets dont la plupart n'apparaissent qu'une fois. Le nombre considérable de ces hapax (1914040) contribue grandement à affirmer l'individualité exclusive de chaque texte et à l'éloigner de ses voisins. Ils constituent 83% du vocabulaire ( $V = 2298089$ ) et 48% du total des occurrences ( $N = 4004329$ ). Reste la part partageable du vocabulaire : dès qu'un mot (ou triplet) a plus d'une occurrence il peut se trouver dans plusieurs textes et contribuer à leur rapprochement. Le cas le plus simple est celui de la classe de fréquence 2. Si on limite l'étude aux textes A et B, les deux occurrences peuvent se rencontrer dans A, ou dans B, ou se partager entre A et B. La méthode Jaccard ne va pas plus loin et n'envisage que ces trois cas : les exclusivités de A (dont nécessairement les hapax), les exclusivités de B, et les mots communs à A et B, quelle que soit leur répartition. Or un relevé plus attentif au détail se complique dès la classe de fréquence 3, qui offre 4 configurations : 3 en A, 3 en B, 2 en A et 1 en B et enfin 1 en A et 2 en B. Il y a une case de plus pour la classe suivante, et la progression se poursuit de façon vertigineuse en s'arrêtant tout de même à la classe 50 qui distribue son contenu dans 51 boîtes. Pourquoi un dispositif aussi minutieux ? Parce qu'un calcul exact de la distance intertextuelle ne se satisfait pas de la fosse commune où la méthode Jaccard entasse les mots non exclusifs. Si le partage est léonin, par exemple 20 occurrences dans A et une seule dans B, on a tendance à croire que cela augmente la distance entre les deux textes, alors que le calcul de Jaccard conclut au rapprochement. Ce mode de calcul a été proposé par Charles Muller dès 1968 dans le dernier chapitre

<sup>23</sup> Les ponctuations entrent dans le chaînage, comme les mots simples. Mais comme elles ont souvent un rôle particulier dans un environnement informatique, on les a préalablement converties en chiffres. Voici comment apparaissent les premiers « triplets » de la *Recherche du temps perdu* :

```
Longtemps_1_je_1_je_me_je_me_suis_me_suis_couché_suis_couché_de_couché_de_bonne_de_bonne_heure
bonne_heure_2_heure_2_Parfois_2_Parfois_1_Parfois_1_à_1_à_peine_à_peine_ma
peine_ma_bougie_ma_bougie_éteinte_bougie_éteinte_1_éteinte_1_mes_1_mes_yeux_mes_yeux_se
yeux_se_fermaient_se_fermaient_si_fermaient_si_vite_si_vite_que_vite_que_je_je_n'_je_n'_avais_n'_avais_pas
avais_pas_le_pas_le_temps_le_temps_de_temps_de_me_de_me_dire_me_dire_3_dire_3_Je_3_Je_n'_Je_n'_endors
m'_endors_2_endors_2_Et
```





Un tableau de 50 par 50 n'étant guère lisible ni même présentable, on en aura une idée fragmentaire en mettant en relief n'importe quelle ligne ou colonne. L'exemple de Gide dans la figure 25 montre ainsi que la *Symphonie pastorale* reconnaît l'*Immoraliste* comme son voisin le plus proche. L'algorithme reconstitue correctement la plupart des paires, quoique avec moins de sûreté que le *deep learning*. Mais il faut reconnaître que les conditions ne sont pas égales : Le *deep learning*, pour chaque texte à apparier, n'avait que 25 candidats, alors que 49 se présentent dans le calcul de Muller.

Mais l'analyse arborée en enveloppant tous les profils d'un seul regard restitue au mieux les parentés dissimulées dans le maquis des lignes et des colonnes. Le résultat qu'elle obtient dans la figure 26 est superposable à celui du *deep learning* : sur les 25 paires, 23 sont reconnues. Les deux auteurs qui résistent à l'analyse sont précisément ceux qui avaient créé des difficultés au *deep learning* : Camus et Péric. On en a déjà évoqué les raisons. On comprendra le progrès accompli en passant des mots aux triplets, si l'on se réfère au tableau 2 où le même calcul de Muller n'avait reconnu que 19 paires à partir des mots individuels. Pourtant dans le cas des triplets comme précédemment la palme revient à l'algorithme d'Évrard, qui obtient là aussi le plein succès.

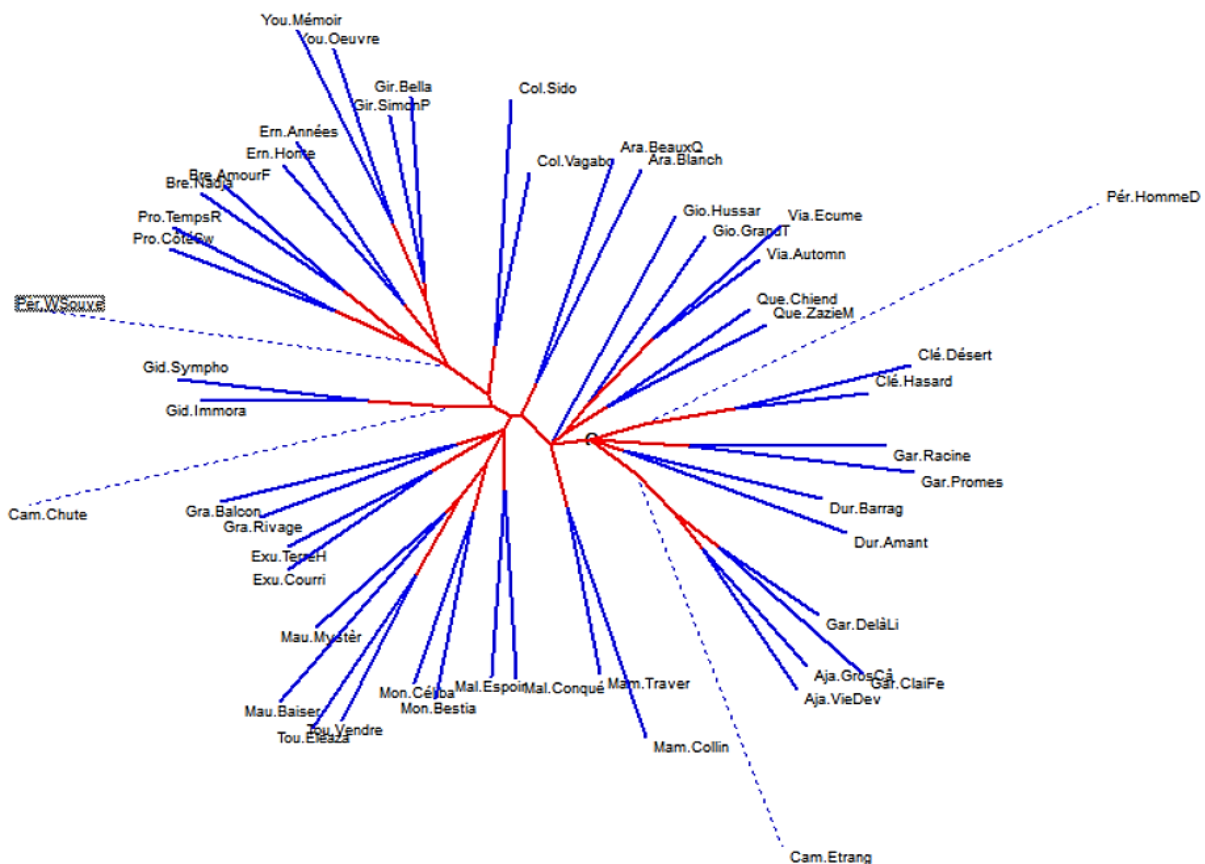
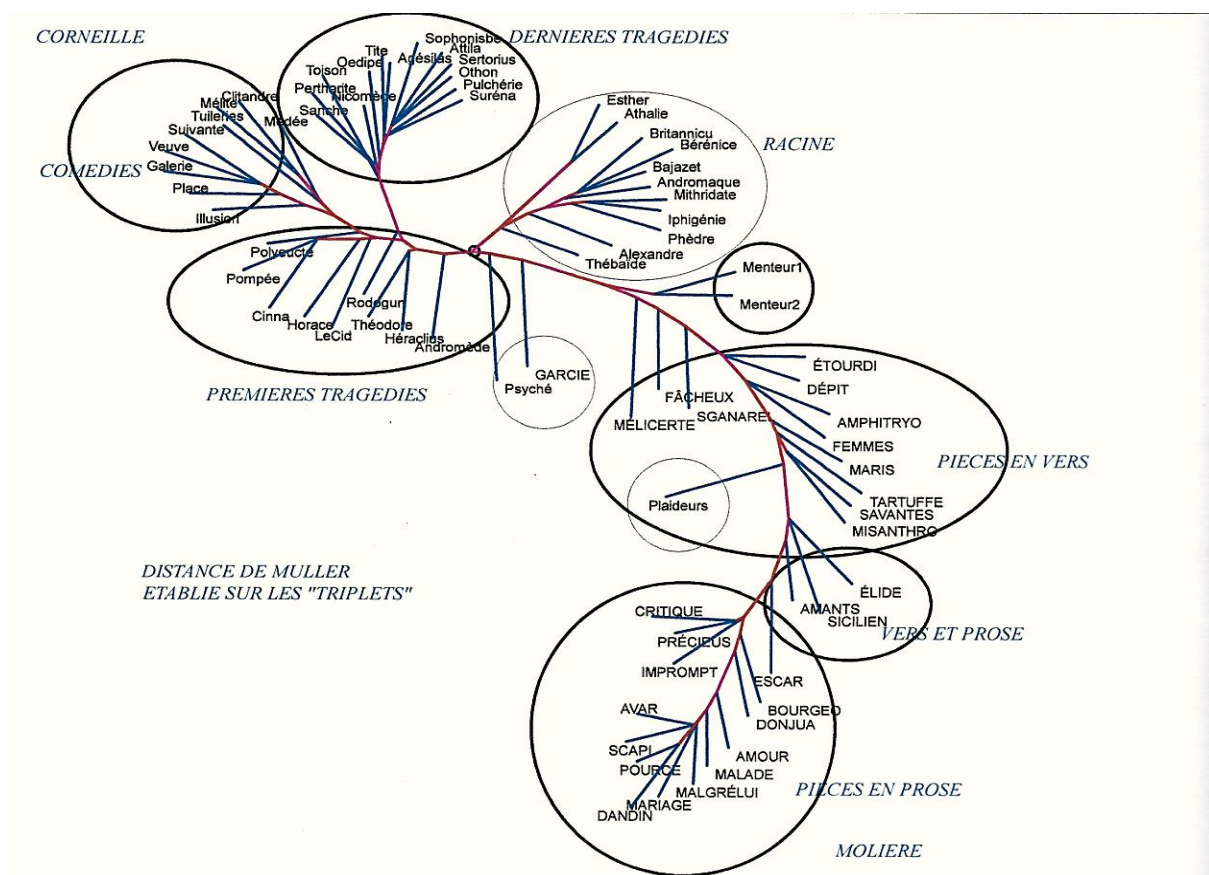


Figure 26. L'analyse arborée des triplets (à partir de la distance de Muller).

Les lignes en pointillé désignent les 4 textes mal identifiés de Péric (*W ou le souvenir* et *L'homme qui dort*), et de Camus (*L'étranger*, et *La Chute*).

### III. 1. 3 - Application au théâtre classique

La même expérience des triplets peut être appliquée au théâtre classique. Le résultat figure dans le graphique 27 qu'on comparera au graphique 13 établi sur les mots individuels. À première vue le parallélisme est frappant : même orientation d'ensemble, de Corneille à Molière (Racine est à l'extérieur), de la tragédie à la comédie, des vers à la prose. Mais pour chaque constellation polarisée par le genre ou l'auteur, le détail est plus précis. Cela est vrai pour Racine, dont l'évolution chronologique est marquée de la *Thébaïde* à *Athalie*. Cela est plus sensible encore dans le cas de Corneille dont la production est divisée en trois périodes distinctes : les premières pièces qu'on range habituellement dans le genre comique (même si certaines, comme *Médée*, relèvent plutôt d'une inspiration héroïque ou tragique), puis, du *Cid* à *Andromède*, la période des grands succès, et enfin les derniers titres que menace le déclin. Quant à Molière, l'algorithme est y plus sensible à la typologie générique qu'à la progression chronologique. Une ligne de démarcation sépare nettement les vers et la prose et met en relief, dans une zone intermédiaire, les pièces où les genres se mélangent, soit parce que Molière pressé par l'urgence n'a pas eu le temps de terminer la versification (c'est le cas de la *Princesse d'Élide*<sup>25</sup>), soit parce que les ballets ou intermèdes chantés et versifiés se mêlent à la prose du dialogue (*Les Amants magnifiques*).



Graphique 27. Analyse des triplets dans le théâtre classique (distance de Muller).

<sup>25</sup> Le *Deep learning* relevait aussi cette ambiguïté dans le tableau 16 qui range cette pièce pour 42% dans le genre versifié et pour 36% dans la prose.

Ainsi, en suivant la démarche du *deep learning* dans une expédition aventureuse au pays des triplets, et en passant des éléments simples aux combinaisons de ces éléments, on multiplie les variétés en divisant les effectifs. Les objets que traite la statistique sont alors plus volumineux, mais aussi plus précis, plus rares et, espère-t-on, plus spécifiques. De la même façon en s'intéressant aux molécules plutôt qu'aux atomes, la statistique a plus de chances de préciser la composition de la matière. Dans la population des triplets, les hapax, qui occupent la moitié de la surface imprimée, bénéficient de la rareté maximale puisqu'ils sont uniques, et si l'individualité de chacun n'est pas prise en compte (au moins dans le calcul de Muller), leur nombre importe grandement dans ce calcul. Quant aux triplets de fréquence supérieure à 1, ils sont le fruit d'une rencontre. Or, la rencontre de deux événements rares est elle-même potentiellement plus rare encore, au point de ressembler à un secret que l'on partage et qui resserre les liens. Enfin envisager les combinaisons plutôt que les éléments simples, c'est associer le séquentiel au fréquentiel, en appliquant le second au premier. Ce faisant, on grimpe d'un étage, avec une perspective plus large et plus précise sur le langage<sup>26</sup>.

### III. 2 – Réduction et intégration des deux approches, *procédurale et algorithmique*

Ce qui gêne la comparaison des deux approches, procédurale ou algorithmique, même quand l'une imite l'autre, c'est la dissymétrie initiale dans la saisie des données. Alors que la méthode classique aligne tous les textes sur la ligne de départ pour les soumettre ensemble et dans le même temps à la même épreuve et au même classement général, les méthodes d'apprentissage et en particulier le *deep learning* ont besoin de séparer apprentissage et reconnaissance et d'exclure du premier les textes destinés à la seconde. Il y a tout de même moyen d'accorder les deux approches, soit en réduisant la première, soit en étendant la seconde.

2.1 – L'analyse de correspondance peut, elle aussi, séparer ces deux phases en réduisant les colonnes actives aux seuls textes présentés à l'apprentissage, c'est-à-dire au premier roman de chaque paire. La base d'apprentissage est ainsi deux fois plus petite. Le second roman de chaque auteur figure en tant qu'élément supplémentaire, sans avoir participé à l'analyse. Malgré la sévérité de cette épreuve, 21 romans, projetés a posteriori, ont comme plus proche voisin un roman du même auteur. Les quatre exceptions concernent Montherlant, Camus, Colette et Giono (avec une liaison presque directe chez ce dernier écrivain). Quant au roman « W ou le Souvenir » il n'est pas loin

---

<sup>26</sup> C'est aussi ce qui se passe à l'étage inférieur, quand la statistique prend pour objet le mot plutôt que la lettre, le mot n'étant qu'une séquence de lettres. Et plus bas encore, la lettre est une séquence de points.

des deux autres romans de Perec qui eux, ont été actifs dans l'analyse, mais il s'agit d'un roman excentrique dont la présence est discutable dans cet exercice.

On rappelle que la projection des éléments supplémentaires est un cas particulier de régression multiple simultanée (et donc de perceptron à une couche cachée avec fonction de transfert identité). Mais c'est un cas particulier de méthode dite *neuronale* qui a une interprétation géométrique simple, et un formalisme analytiquement explicite (et non uniquement algorithmique).

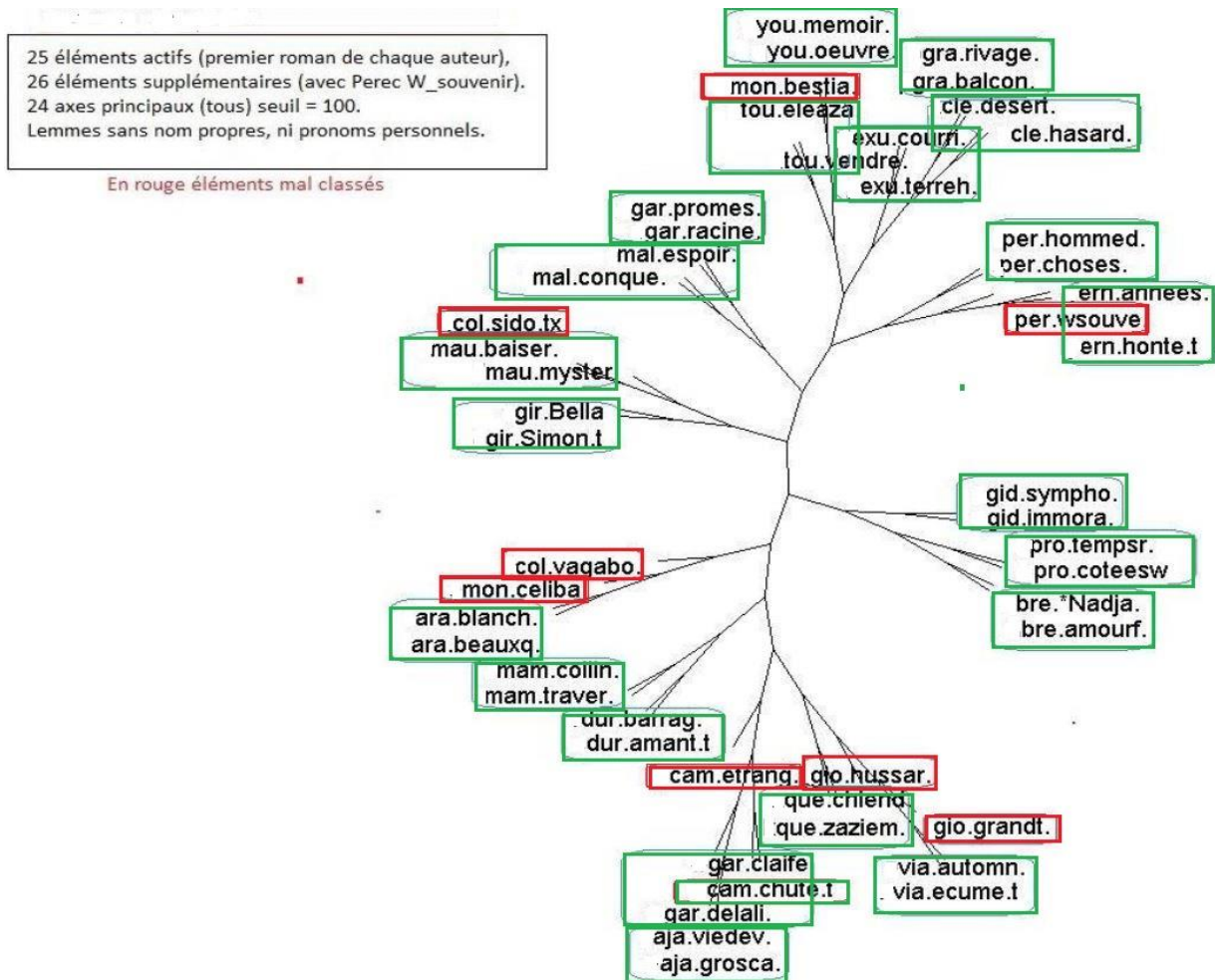


Figure 28. Lemmes sans noms propres ni pronoms. Seuil de fréquence : 100

*Seuls 25 romans (25 auteurs) sont analysés, les 26 autres sont projetés a posteriori en éléments supplémentaires*

2.2 – Inversement la procédure du deep learning pourrait atteindre la dimension maximale (le nombre de textes en présence). Il faudrait pour cela répéter, autant de fois qu'il y a de textes, la phase d'apprentissage, en enlevant à chaque fois un texte au corpus, pour le réserver à la phase de reconnaissance, avant de réintroduire ce dernier texte dans le corpus pour le coup suivant. Nous avons accepté cette procédure pour les



comédies en vers de Molière (voir tableau 10). Mais avec le corpus du roman, gros de 4 millions de mots et divisé en 50 partitions, l'opération s'est révélée trop lourde.

Il est toutefois possible d'intégrer les résultats du *deep learning* dans une procédure classique. Dans les tableaux que produit le *deep learning*, chaque ligne est acquise et fournie indépendamment des autres lignes. Toutes ont le même poids puisque ce sont des pourcentages dont le total est le même. Si chaque ligne a des éclats et des ombres accusés, le tableau lui-même reste assez opaque, si on ne lui applique pas un traitement de synthèse. En réalité le *deep learning* ne donne que des réponses isolées, qui se bornent à la question posée et proposent un classement sur cette question. C'est le chercheur qui construit le tableau en y ajoutant autant de lignes qu'il le souhaite.

2.2.1 - Réunissons dans un même tableau rectangulaire les résultats fournis ligne à ligne par le *deep learning* dans les tableaux 22 et 23 relatifs au **théâtre classique** (en excluant Racine dans les colonnes, et ses deux pièces, *Phèdre* et les *Plaideurs*, dans les lignes). Restent quatre colonnes dont deux concernent Corneille (comédies et tragédies) et deux autres Molière (vers et prose). On retrouve ces quatre sous-ensembles, nettement séparés, dans l'analyse arborée de la figure 29. Et la composition de chacun est ce que l'on attendait, même dans les cas d'hésitation (*Princesse d'Élide* entre vers et prose) ou de contestation quand le genre et l'écrivain sont en conflit (*Psyché* et *Garcie de Navarre* entre Molière et Corneille).

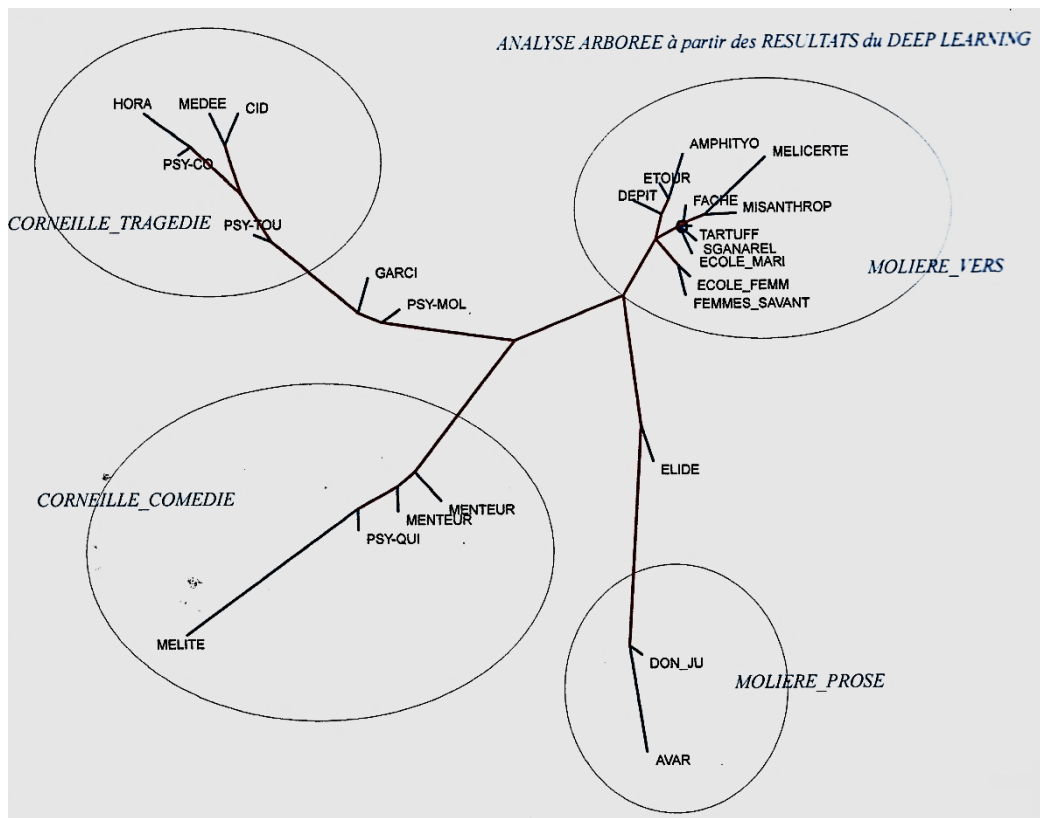


Figure 29. Analyse arborée appliquée aux résultats du deep learning (théâtre).

2.2.2 - De la même façon dans les tableaux 10 et 11 relatifs au **roman du XX<sup>e</sup>** siècle, le *deep learning* associait 23 lignes à 23 colonnes. Pour s'aligner sur les résultats obtenus par la méthode Muller qui compte 46 lignes et 46 colonnes dans une disposition symétrique obtenue d'un seul tenant, on peut ajouter l'un à l'autre les tableaux 10 et 11 et disposer ainsi de 46 lignes (la distinction entre Ajar et Gary étant abolie dans ces corpus). On obtient alors la totalité des auteurs en colonne et la totalité des romans en ligne. Et comme un tel tableau n'est pas symétrique, il compte finalement autant de cellules actives que celui de Muller.

Dès lors les résultats du *deep learning* cessent d'être un terminus ; ils peuvent s'inscrire dans une approche traditionnelle et par exemple servir de données aux programmes d'analyse arborée ou d'analyse factorielle. Cette fois c'est l'ensemble du tableau qu'on examine pour en tirer des enseignements qui dépassent la lecture linéaire des lignes ou des colonnes et la simple question de la paternité littéraire. Par ce biais le *deep learning* peut acquérir une cohérence, une lisibilité et une fonction descriptive qu'il n'a pas par lui-même.

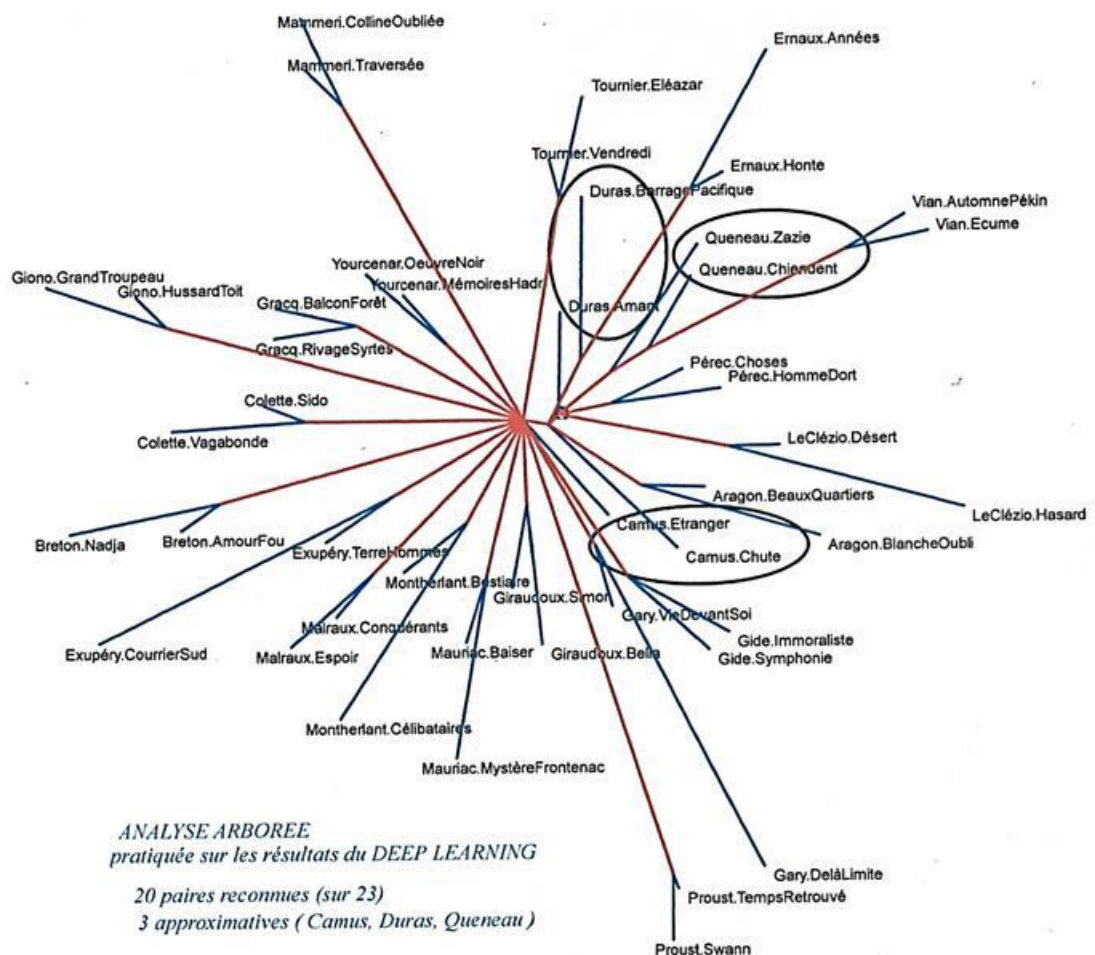


Figure 30 Analyse arborée du *deep learning* (données cumulées des tableaux 4 et 6).

À première vue la figure 30 est superposable à la figure 26. Les paires sont reconstituées sans hésitation, à l'exception attendue de Camus. Si Pérec rentre dans le rang, c'est parce qu'un texte plus assimilable, *Les Choses*, a été substitué au roman intraitable *W ou le souvenir*. Mais il y a un flottement léger, qu'on observait aussi dans les tableaux 10 et 11, à l'endroit de Duras, et un autre apparaît concernant Queneau. Mais dans ces deux cas, si la liaison directe n'est pas établie, le rattachement est à très courte distance. On a tout lieu de se satisfaire de cette convergence des deux graphiques et des deux approches.

2.2.3 - Mais à y regarder de près la **structure est différente** : la figure de Muller est une carte, avec des chemins et des routes, et celle du *deep learning* est une table d'orientation en étoile avec seulement des directions. Les points du paysage y sont désignés par des flèches mais rien ne permet de cheminer d'un point à l'autre. Du centre partent de longs rayons rectilignes, chacun conduisant à un auteur et, serrés l'un contre l'autre, aux deux textes qui sont de lui. Mais les écrivains n'ont pas de voisins, pas d'amis, pas d'adversaires. On en compte 15 sur 23 qui se rattachent directement au centre, les autres s'écartant un peu sur la droite pour former une autre figure rayonnante. En fin de compte, on ne trouve que trois écrivains, Vian, Queneau et partiellement Pérec, pour emprunter de conserve le même chemin, sans doute le chemin qui mène à l'*OuLiPo*, où Vian aurait eu sa place s'il avait vécu plus longtemps.

La figure 26 est beaucoup plus parlante. Certes les textes frères y sont moins solidement soudés (les lignes bleues sont plus longues) ; mais les chemins communs (en rouge) sont vraiment des voies de communications qui mettent les écrivains en relation. La structure s'allonge grossièrement dans le sens de l'histoire, le lointain à gauche, le récent à droite (mais les auteurs ayant des titres éloignés dans le temps, le mouvement chronologique en est perturbé). Et surtout la disposition en blocs rend compte des tendances ou des habitudes : un groupe se forme en haut à gauche autour de la littérature personnelle qui dit « je », un autre, plus bas, rassemble les représentants d'une narration classique, tandis qu'à droite un renouvellement des formes littéraires réunit des écrivains plus modernes.

Ainsi l'intégration du *deep learning* dans un dispositif de statistique traditionnelle ne lui donne pas d'emblée les vertus descriptives et explicatives qu'on pourrait attendre d'un outil si puissant. Il reste à les chercher dans sa propre démarche, en démontant, autant que faire se peut, les mécanismes de l'instrument, grâce à la déconvolution. C'est ce qu'on se propose de faire dans un dernier effort.



### III.3 – La déconvolution

Nous ne voulons pas ici approfondir les calculs qui permettent de rebrousser chemin et de retrouver la trace des marqueurs qui ont orienté les choix du *deep learning*. Ces précisions techniques ont été publiées ailleurs<sup>27</sup>. Si le *deep learning* reste discret sur ses critères, la déconvolution permet *a posteriori* de connaître certaines de ses décisions caractéristiques : une liste de passages exemplaires où les mots ou expressions qui ont motivé ses choix sont mis en relief.

3.1 - Pour *Le menteur 1* on obtient plusieurs extraits dont certains (majoritaires) sont attribués à Corneille et certains autres aux comédies en vers de Molière, ou à d'autres corpus invités à la comparaison. Un curseur permet de sélectionner les marqueurs les plus forts (on les a représentés en rouge) ou ceux qui le sont moins (en bleu)<sup>28</sup>.

Plutôt du Corneille-Comédie :

“ [...] à coiffe abattue, et sans les approcher il suit de **rue en rue** ; **aux couleurs**, au **carrosse**, il ne doute de rien ; to ut étoit à **Lucrèce**, et le **dupe si bien**, que prenant **ces beautés** pour **Lucrèce** et **Clarice**, il rend à [...] ”

Plutôt du Molière en vers :

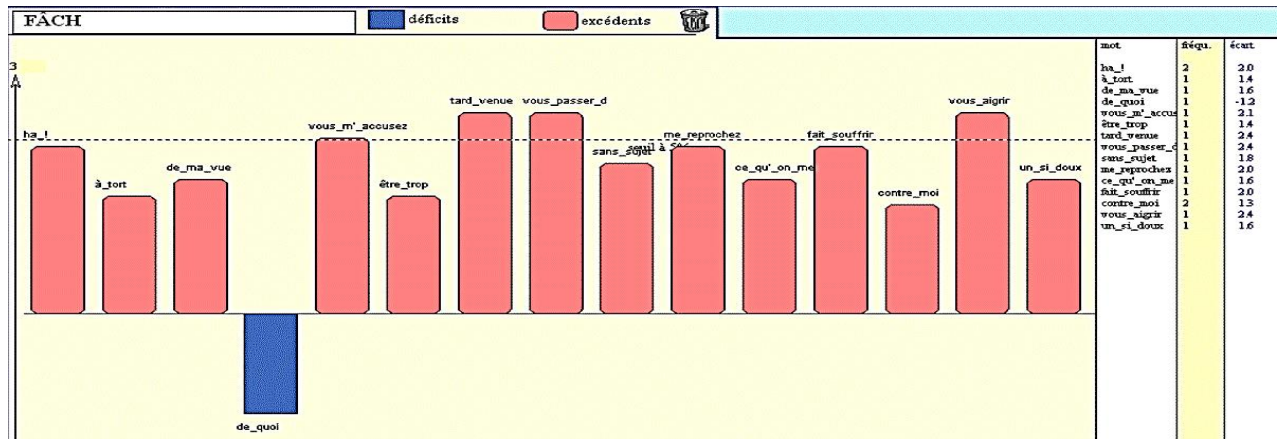
“[...] Mais, avant qu'avec moi **le noeud** d'hymen vous **lie**, vous **serez marié**, si l'on veut, en **Turquie**. Avant qu'avec toute autre on me puisse **engager**, je serai **marié**, si l'on veut, en **Alger**. Mais enfin vous n' [...]”

On constate sans surprise que les noms propres *Lucrèce* et *Clarice* contribuent grandement à la reconnaissance (sur 214 occurrences de *Clarice* rencontrées dans le théâtre classique, 177 se trouvent dans le *Menteur*, et 42 *Lucrèce* sur 50)<sup>29</sup>. On ne s'étonne pas de trouver les mots *carrosse* et *dupe* dans une intrigue qui repose sur l'esbroufe et le mensonge. Quant à l'expression galante de *ces beautés* on n'en trouve que trois occurrences dans le théâtre classique, toutes chez Corneille. On peut comprendre aussi que des mots comme *marié*, *hymen*, *noeud*, *lie*, *engager* orientent le second passage vers Molière, chez qui tant d'intrigues reposent sur les unions contrariées. Mais là s'arrête ce que la statistique classique peut aussi expliquer. La déconvolution propose des expressions hapax (par exemple *de rue en rue*, *il ne doute de rien*, *sans les approcher*), dont la statistique classique par définition ne s'embarrasse pas. Ainsi le passage jugé le plus caractéristique de la pièce des *Fâcheux* est fait d'expressions en exemplaire unique, qui peinent à atteindre le seuil fixé pour l'hypothèse nulle (figure 31).

<sup>27</sup> Laurent Vanni, M Ducoffe, D Mayaffre, F. Precioso, D Longrée *et al.* *Text Deconvolution Saliency (TDS): a deep tool box for linguistic analysis*, 56th Annual Meeting of the Association for Computational Linguistics, Jul 2018, Melbourne.

<sup>28</sup> Les couleurs n'ont pas le même sens dans le logiciel **Web Hyperbase** qui a fourni ces exemples : elles désignent le niveau où se situe ce qu'on appelle la « saillance » : graphie, lemme ou code.

<sup>29</sup> De même dans le second exemple la mention d'*Alger* dirige l'algorithme vers Molière parce que ce nom exotique apparaît trois fois dans les *Fourberies de Scapin*. Précisons que les noms propres qui désignent le personnage auquel on attribue la réplique ont été évacués à la saisie car ils ne font pas partie du texte qu'on entend à la représentation.



Graphique 31. Le passage tenu par la déconvolution comme spécifique des *Fâcheux*.

Ces mesures légères et sensibles semblent « peser des œufs de mouche dans des balances de toiles d'araignée », pour reprendre une expression de Voltaire. On soupçonne aussi des faits de rythme, des associations disjointes, des motifs liés au genre, ou à l'époque, toutes subtilités qui échappent à l'emprise directe de la statistique<sup>30</sup>. Le *deep learning* n'en est qu'à ses débuts dans le traitement des textes. Dans le traitement de l'image il peut imposer des filtres qui redressent un original déformé, par exemple une photographie bougée. On peut imaginer dans l'avenir qu'il pourra isoler ou neutraliser les faits linguistiques qui viennent du thème, ou du genre et qui se mêlent inextricablement au style propre de l'écrivain.

On en veut pour preuve une dernière illustration, relevée dans le *Misanthrope*. Cette pièce que Boileau considère comme le chef d'œuvre de Molière s'inscrit à 73% dans le moule de la comédie en vers, dont elle est l'archétype (ce seuil n'est jamais dépassé). Mais après cinquante pages versifiées une page en prose apparaît dans la scène finale pour révéler dans un billet compromettant les pensées secrètes de Célimène. Cela n'a pas échappé au *deep learning* qui cite un extrait de ce billet comme pouvant appartenir à la prose de Molière. Or, pas plus que Monsieur Jourdain l'ordinateur n'a appris à compter les syllabes et à distinguer la prose et les vers.<sup>31</sup>

<sup>30</sup> On a pourtant essayé, dans les deux corpus, de tirer parti des cooccurrences, en variant l'empan (page ou phrase), et en changeant aussi l'objet linguistique (tantôt substantifs seuls, tantôt mots pleins, y compris les verbes, les adjectifs et les adverbes). Les bases ainsi créées à partir de cooccurrences soudées l'une à l'autre sont certes en accord avec celles qui reposent sur les mots individuels, mais elles ne produisent aucune valeur ajoutée et sont loin de la précision des triplets et à plus forte raison de celle du *deep learning*.

<sup>31</sup> On pourrait penser que l'extrait contient des prosaïsmes qui suffisent à expliquer la décision. Mais le mot *flandrin* qui se cache derrière de symbole PAD est hors de cause. Et le mot *cracher* se rencontre plutôt dans les pièces en vers. Quant à l'expression pittoresque « *faire des ronds* » elle est sans autre exemple. En revanche si le *vicomte* (expressément mis en relief) n'a rien en soi qui le destine à la prose populaire, il est vrai que Molière le réserve aux pièces en prose notamment les *Précieuses ridicules* (10 occurrences), *Georges Dandin* (6 occ.) et la *Comtesse d'Escarbagnas* (39 occ.).

Plutôt du ensemble:molierevers6 : 73.17%

" et contre le prochain la conversation prend un assez bon train . PAD encor , madame , est un bon caractère . C' est de la tête aux pieds un homme tout mystère , qui vous jette en passant un coup d' oeil égaré , et , sans aucune affaire "

[Voir une autre phrase-clé](#)

Plutôt du ensemble:moliereprose : 15.74%

" vicomte ... " il devrait être ici , " notre grand PAD de vicomte , par qui vous commencez vos plaintes , est un homme qui ne sauroit me revenir ; et depuis que je l' ai vu , trois quarts d' heure durant , cracher dans un puits pour "

[Voir une autre phrase-clé](#)

B:

3.2 - En réalité les points d'ancrage les plus sûrs que la déconvolution permet de saisir recouvrent très souvent ce que la tradition désigne par **spécificité**. Quand un mot concentre ses occurrences dans un texte privilégié, cela n'échappe pas au *Deep Learning* et bien des passages sont restitués par l'algorithme, avec pour justification apparente la présence de noms propres ou de mots-clés qui ont moins de chance de se trouver ailleurs. Ainsi procède la lecture humaine : si l'œil se promène sur une page inconnue et y rencontre un *Rougon* ou un *Macquart*, on présume qu'il s'agit de Zola. On en vient à se demander si la prédiction du deep learning n'est pas fondée sur la somme des spécificités individuelles ou collectives. De même dans la reconnaissance d'une image de chat, ce sont sur des détails caractéristiques, comme la pupille verticale, que la présomption s'appuie. Nous avons donc mené une expérience pour évaluer le poids des mots spécifiques dans la reconnaissance des textes et des auteurs. La plupart des logiciels lexicométriques produisent des listes, alphabétiques ou hiérarchiques, de spécificités, points d'appui souvent précieux pour le commentaire et l'interprétation. Leur choix permet d'échapper au dilemme fréquences basses/hautes. Car le nombre absolu des occurrences importe moins que leur répartition, étale ou inégale. Et la sommation n'est pas celle des occurrences, ce qui donnerait l'avantage aux mots fréquents, mais celle des écarts, ou plutôt des écarts normalisés ou « réduits ». Quoique le calcul réel soit hypergéométrique, on a gardé l'échelle, bien commode, de l'écart réduit, emprunté à la loi normale. Nos listes s'arrêtent à la valeur 2, au moment où l'on a moins de 5 chances sur 100 d'écarter l'hypothèse nulle. La démarche consiste, pour deux textes dont on veut établir la distance, à repérer dans les 2 listes les éléments communs et à faire le total des écarts. Le calcul est à répéter pour les 50 textes et leurs 1225 combinaisons deux à deux. On obtient alors un tableau, carré et symétrique, non pas de distances mais de proximités, les plus proches voisins étant ceux dont le total est le plus élevé pour les éléments partagés. La lecture directe d'une ligne ou d'une colonne désigne généralement le texte-frère comme étant le plus proche voisin. Mais le recours à une vision de synthèse conforte cette observation. Que ce soit l'analyse factorielle (graphique 28) ou l'analyse en arbres additifs (graphique 29), le résultat est très semblable à ceux qu'on obtient du deep learning ou de la lexicométrie.

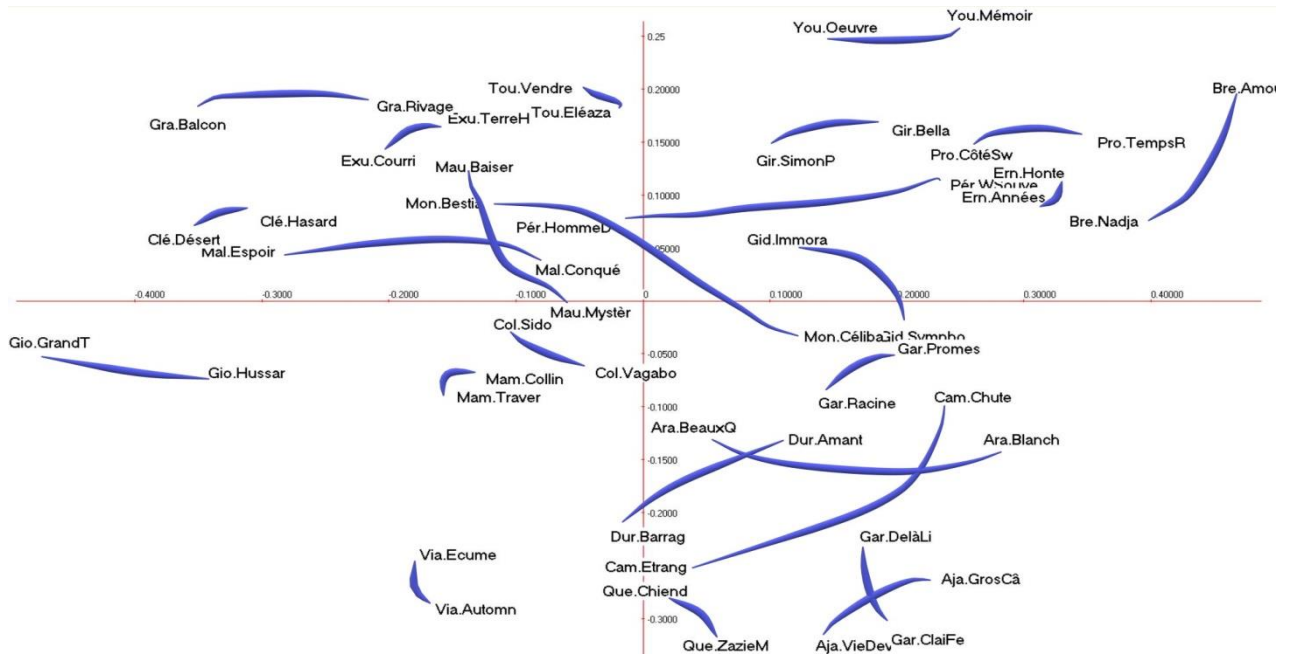


Figure 32. Analyse de correspondance des spécificités partagées

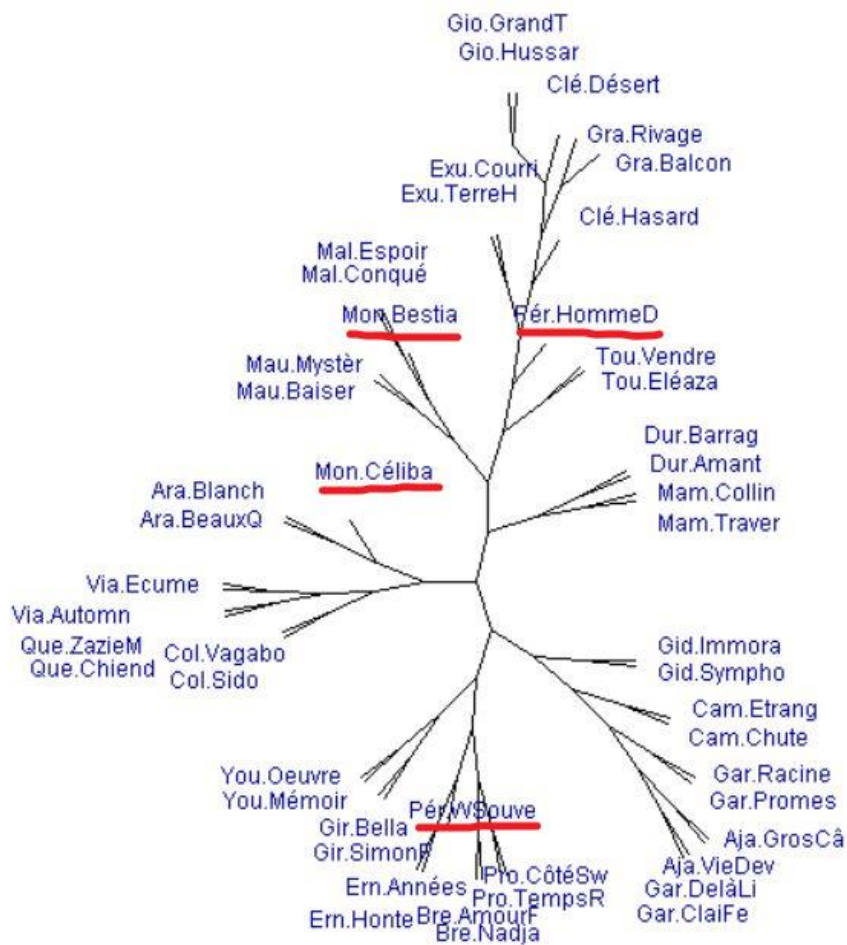


Figure 33. Analyse arborée des spécificités

Dans les deux cas, les textes de deux auteurs font difficulté: ceux de Pérec et de Montherlant. Ce sont les mêmes qui résistaient aux épreuves précédentes. Dans l'analyse factorielle ils chevauchent l'axe principal et dans l'analyse arborée ils n'admettent aucune proximité.

3.3 - Il faut avouer pourtant que les révélations de la déconvolution, même confirmées par le témoignage des spécificités, s'entourent d'incertitude et d'incomplétude. On parle souvent de boîte noire<sup>32</sup> quand on évoque le *deep learning*. On veut dire par là que la transparence fait défaut. Mais l'image évoque aussi l'inscription automatique de tous les paramètres qui orientent ou redressent le vol à chaque seconde. Dans l'apprentissage des textes, l'automate est supervisé en cela qu'on désigne l'auteur du texte proposé. S'ensuivent des ajustements incessants et progressifs du processus, analogues aux corrections de trajectoire du pilote automatique. Mais s'il est possible de déchiffrer intégralement les enregistrements de la boîte noire, quand du moins on la retrouve intacte après l'accident, les chemins empruntés par le *deep learning* ne permettent guère le contrôle total, *a posteriori*, qu'il y ait ou non accident. Dans la chaîne procédurale au contraire, le code suit pas à pas la théorie appliquée et la défaillance éventuelle est en principe facile à repérer : si une instruction a provoqué par exemple une division par zéro, elle avoue immédiatement et clairement son forfait.

En revanche le *deep learning* offre peu de justifications, ni pour expliquer ses triomphes, ni pour excuser ses succès. Albert Einstein disait avec humour: « La théorie, c'est quand on sait tout et que rien ne fonctionne. La pratique, c'est quand tout fonctionne et que personne ne sait pourquoi. ». Le *deep learning* est clairement une pratique obscure qui fonctionne sans trop savoir et dire pourquoi. Cette pratique est celle d'un réseau. Simple et progressive, elle s'oppose à la chaîne balisée des parcours procéduraux qui avancent par étapes. La progression du réseau est fondée sur la répétition ininterrompue d'un processus « neuronal » qui de proche en proche et de couche en couche, s'adapte continûment aux données, comme la courbure du roseau s'adapte continûment à la force du vent. Elle n'est pas sujette aux ruptures qui peuvent se produire quand un maillon cède dans la chaîne. La chaîne et le réseau, cela pourrait faire une fable.

---

<sup>32</sup> De la boîte noire au trou noir, il n'y a qu'un pas qui a été franchi dans un article paru en novembre 2017, sous le titre *Deep Learning, le grand trou noir de l'intelligence artificielle*, <https://www.maddyness.com/2017/11/13/ia-deep-learning-trou-noir-intelligence-artificielle/>



### III.4 - Analyse non-supervisée des « pages » (fragments de 50 lignes consécutives)

Dans nos exemples où s'applique une méthode classique de *deep learning*, on s'est tenu à une version supervisée du processus d'apprentissage, les étiquettes d'auteur étant clairement posées sur des textes bien définis, soumis à cette phase en toute transparence. Pour corser la difficulté et envisager le cas le plus général où les données sont morcelées et confuses, on pourrait ne pas prendre en compte cette information et proposer des textes en vrac, au kilomètre, sans partition explicite des textes, en forçant l'automate à agir en aveugle. On se propose de réaliser cet essai dans un proche avenir. Mais les conditions de l'expérience peuvent être préalablement explorées avec les procédures classiques. Voici comment.

4.1 - On reprend l'exemple du corpus de 51 romans. Au lieu de présenter les textes un par un en tant que partitions délimitées, successives et constitutives du corpus, on ne fournit à l'algorithme qu'un flot ininterrompu de **pages autonomes**, pourvues seulement d'un numéro d'identification, sans lien entre elles (l'ordre ne compte pas) et sans rattachement explicite à un texte particulier. Le traitement est donc entièrement non-supervisé. Il consiste en une analyse des correspondances des 3547 pages de 50 lignes du fichier lemmatisé des 51 romans français étudiés au début de cet article. Ces pages (qui correspondent plutôt à des doubles pages imprimées) peuvent être battues comme des cartes à jouer. Une fois que l'analyse factorielle a établi la typologie de ces pages, on obtient un nuage de 3547 points-page dans un espace de dimension élevée (ici 40 axes principaux). On positionne *a posteriori* dans cet espace les auteurs comme points moyens des pages qu'ils ont écrites. Au départ, l'analyse ne cherche pas à contraster les auteurs, mais à décrire ces pages sans idées préconçues. Il s'agit donc d'une épreuve très sévère<sup>33</sup>.

4.2 – À partir des **données brutes**, la figure 34 montre que 17 auteurs sur 25 donnent lieu à des appariements corrects. Par contre, Giraudoux est écartelé, comme Montherlant. Ajar de « la Vie devant soi » est très loin de « Gros Câlin » qui, lui, est proche des romans signés Gary. Dans les analyses précédentes (section 2) où les pages étaient agrégées par roman, la variabilité des pages « intra-roman » était neutralisée. On voit ici que chez six auteurs (surtout sur trois : Giraudoux, Montherlant, Ajar. Le phénomène est moins net pour les trois autres) il y a une forte variabilité interne des pages, au point de brouiller les cartes.

---

<sup>33</sup> On pourrait ici encore moduler la longueur des fragments en prenant seulement 25 lignes ou au contraire 100 ou 200 lignes.

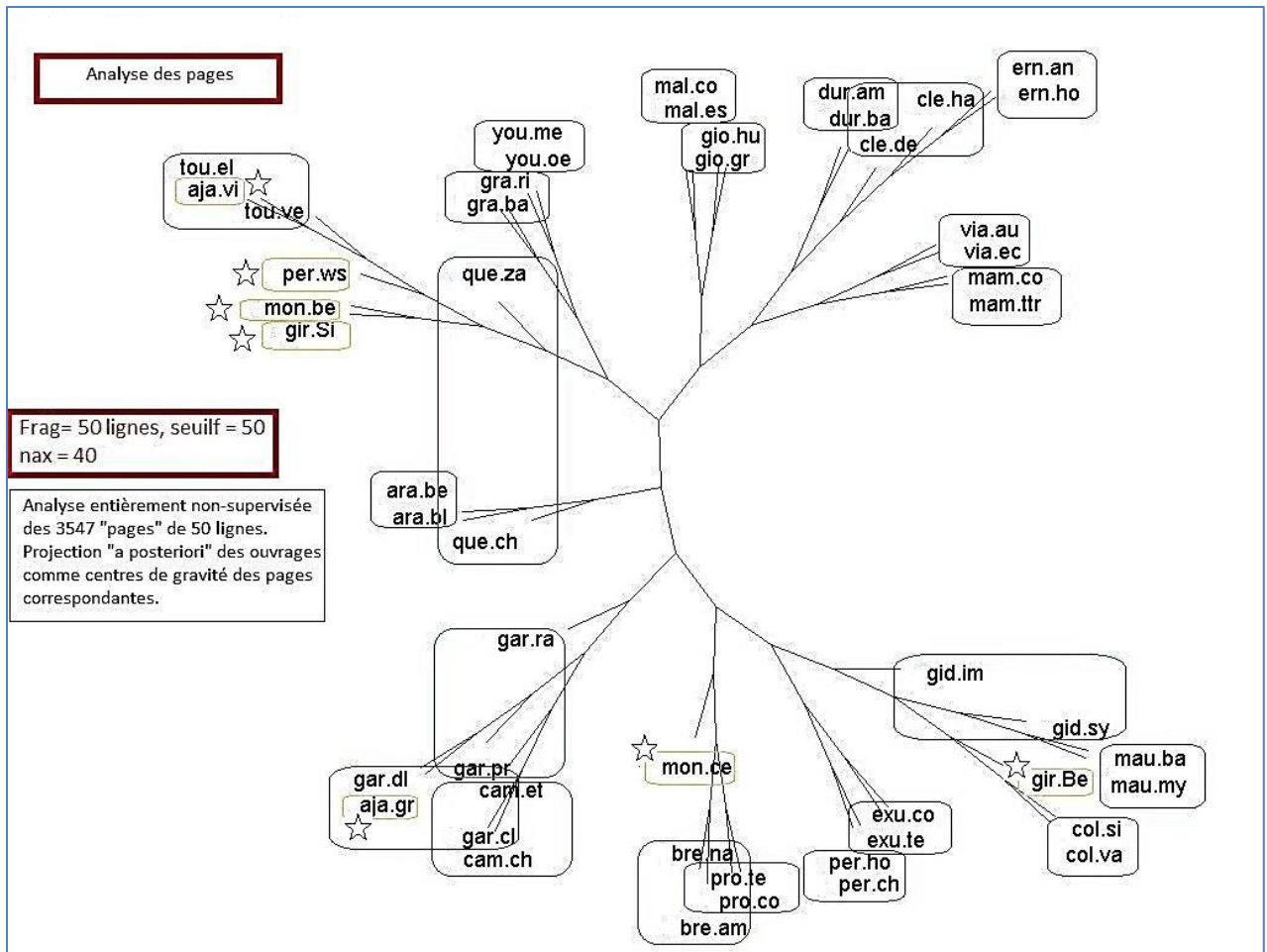


Figure 34. Analyse arborée à partir des 3547 « pages » de lemmes sans noms propres ni pronoms. Seuil de fréquence : 50. Les titres des romans sont projetés a posteriori comme points moyens de leurs pages. L'arbre est construit à partir de ces points-moyens.

Cette approche non supervisée n'est pas incompatible avec l'évolution en cours du *deep learning*<sup>34</sup>, qui recommande des représentations des données avec de multiples niveaux d'abstraction, après le constat, somme toute banal, que les processus d'acquisition des connaissances des animaux et des humains sont essentiellement non-supervisés et ne demandent pas de corpus étiquetés.

L'analyse des pages est sans doute plus proche d'un processus cognitif naturel que la seule prise en compte d'un bilan fréquentiel au niveau global d'un ouvrage. Elle est plus riche puisqu'elle étudie sa variabilité interne, exclue a priori des approches précédentes. Enfin, elle est plus convaincante puisqu'elle prend en compte la complexité de l'ouvrage, et prend le risque de confronter sa variabilité interne à la variabilité entre auteurs.

4.3 - Nous avons vérifié plus haut que les **verbes** étaient plus uniformément répartis dans les textes que les noms et adjectifs. Cela va avoir un effet stabilisant sur

<sup>34</sup> Cf. par exemple : LeCun Y., Bengio Y., Hinton G. (2015). Deep Learning. *Nature*, 521, 436-444.

l'analyse des pages précédentes. Dans la figure 31 on observe en effet que les pages de verbes permettent une bien meilleure prédiction des auteurs que les pages de mots (lemmes) en général. Ainsi, 23 auteurs sur 25 sont caractérisés par leurs pages de verbes. Il ne reste plus que Montherlant et Camus pour faire exception à ce classement (symboles « étoile » sur la figure 31).

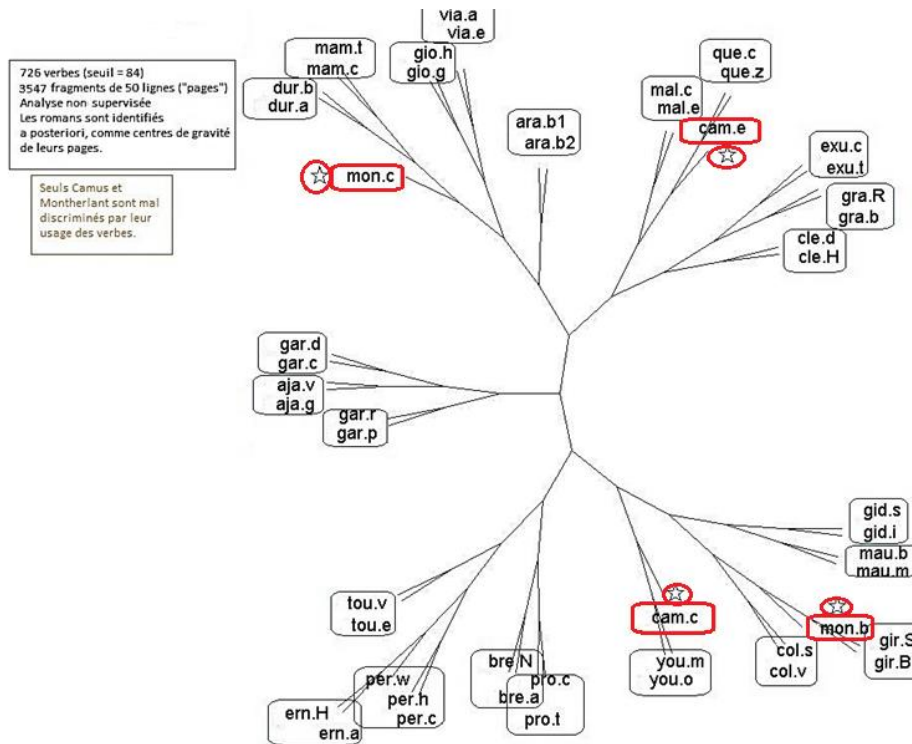


Figure 35. Analyse arborée de 3547 « pages » de verbes. Seuil de fréquence : 50.

*Les titres des romans sont projetés a posteriori comme points moyens de leurs pages de verbes. L'arbre est construit à partir des distances entre ces points moyens.*

#### IV. Conclusion (Éléments pour une ...)

On a vu lors de la section I dévolue aux romans du XXème siècle (50 textes et 25 auteurs) que les analyses des correspondances des tableaux lexicaux (graphies x auteurs) ou (lemmes x auteurs) permettaient de reconstruire progressivement, axe par axe, les distances entre auteurs, et il a fallu attendre les tous derniers axes pour obtenir une excellente discrimination entre auteurs (discrimination que l'on peut d'ailleurs obtenir directement à partir de la matrice des distances d'Evrard originale).

Cette situation est peu courante en analyse des données. Sans être trop technique, disons que les analyses structurales s'intéressent au haut du spectre (premiers axes, plus grandes valeurs propres) qui dégage des dimensions stables et qui permettent des visualisations. On dit souvent que le bas du spectre (plus petites valeurs propres) ne

récolte que du bruit (au sens statistique du terme)... quand il y a du bruit. Mais ici, le bruit, dénomination familière de l'aléatoire en théorie du signal, laisse la place à l'indéterminé, à l'inconnu qui échappent à toute formalisation. Les premiers axes mettent en évidence des traits structuraux généraux, des caractéristiques de genre, des niveaux de langage, alors que pour les derniers axes, on parlera plutôt d'idiosyncrasies, de jardins secrets, voire de tics.

Les techniques particulières de *deep learning* mis en œuvre fournissent les appariements souhaités avec aisance: leur spécialité, leur ADN comme on dit souvent maintenant, est de reconnaître, plus que de décrire. On ne peut plus les ignorer lorsqu'on manipule des données complexes, lorsqu'un panorama global ne suffit pas et lorsqu'on veut éprouver une information externe précise comme, par exemple : « ces deux romans ont été écrits pas un même auteur ».

A ce stade, nous sommes cependant loin d'avoir épuisé l'apport de cette expérience à notre compréhension du fonctionnement des méthodes quantitatives.

La section II dévolue au théâtre classique reposait, en partie, sur un corpus dont la structure est fort différente de la précédente : seulement trois auteurs pour 75 pièces. Trois points définissant un plan, on pouvait s'attendre à trouver des bons plans de visualisation. Mais la variable « auteur » est une information externe à la table lexicale analysée, qui ne connaît que les pièces et les graphies. Cependant, le résultat graphique obtenu sur un simple plan va au-delà des espérances. Tout est là : la distinction entre auteurs, entre genres (comédies, tragédies), entre vers et prose. Peu d'exceptions dans les attributions d'auteurs, largement explicables par les effets de genre. Quant aux procédures de *deep learning* utilisées pour le théâtre classique, elles ne font qu'une bouchée de ce corpus classique pour désigner les auteurs, sans toutefois pouvoir totalement neutraliser l'effet perturbateur du genre.

Les méthodes d'analyse statistiques exploratoires d'une part, et les méthodes d'apprentissage automatique (*machine learning*, comme leurs versions récentes : *deep learning*) d'autre part, sont certes complémentaires, mais sont aussi parentes d'un point de vue théorique, plus qu'on ne le croit généralement.

L'analyse des correspondances appliquée aux tables de contingence (ici : tableaux lexicaux) est ainsi le point de rencontre de nombreuses techniques d'apprentissage. Elle peut être décrite comme un perceptron multicouche supervisé particulier (dans ce cas, les couches d'entrée et de sortie sont respectivement les lignes et les colonnes du tableau analysé). L'analyse des correspondances est également un perceptron multicouche non supervisé particulier (dans ce cas, la couche d'entrée et la couche de sortie sont identiques et la couche cachée, plus petite, impose simplement une

compression des données)<sup>35</sup>. Enfin, lors de l'analyse des pages, nous avons positionné *a posteriori* les auteurs parmi les pages, comme des variables nominales supplémentaires. L'ajout d'éléments supplémentaires dans une technique d'axes principaux (analyses en composantes principales, des correspondances) constitue une variante descriptive de la régression multiple, elle-même une forme simple de perceptron. D'un point de vue géométrique, les deux situations sont en effet similaires.

En somme, parallèlement aux outils de *deep learning* éprouvés et validés, surtout dans les applications industrielles en temps réel, nous avons aussi puisé dans les boîtes à outils du *machine learning* des procédures plus exploratoires, transparentes, interprétables à chaque étape, dont les résultats pourraient être visualisés (plans, arbres) ou évalués via des procédures statistiques (zones de confiance *bootstrap* par exemple).

Il est clair que les méthodes exploratoires choisies ne représentent qu'une partie du potentiel de l'apprentissage automatique<sup>36</sup>. Mais c'était le prix à payer pour la transparence du calcul, les visualisations et la simplicité tant algébrique que géométrique du processus.

En utilisant l'arsenal des algorithmes d'apprentissage disponibles, la machine apprend. Les applications dans la vie réelle sont immenses et, de plus, en pleine évolution.

Lorsque nous utilisons de façon complémentaire le (grand) sous-ensemble des techniques exploratoires que nous avons sélectionnées, la machine apprend, certes, mais c'est surtout nous qui apprenons.

---

<sup>35</sup> Dans ces deux cas, les réseaux neuronaux utilisent la fonction d'identité en tant que fonction de transfert.

<sup>36</sup> Au plan technique, la fonction de transfert peut être une fonction plus générale que l'identité, permettant ainsi de délinéariser le modèle.