Deep learning et authentification des textes

Étienne Brunet et 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 (*with a more detailed treatment planned for a subsequent publication*).

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

Résumé. *Deep Learning et authentification des textes*

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^{ème} 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 Garcie* 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. *Un développement plus étendu est prévu dans une publication ultérieure.*

Mots-clés : *Intelligence artificielle, Deep Learning, authentification des textes, paternité littéraire, distance intertextuelle, lexicométrie, roman au XX^{ème} 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).

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. La procédure que suit le *deep learning* respecte intégralement cette distinction, et dénonce immédiatement les forfaitures. 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

A – Reconnaître que deux textes sont d'un même auteur

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. 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 quelque temps un mystère, une rumeur, un soupçon, a fait l'objet d'une révélation

explicite qui ne permet plus aucun doute[1]. 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*). 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^{ème} 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 familier 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[2]. 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).

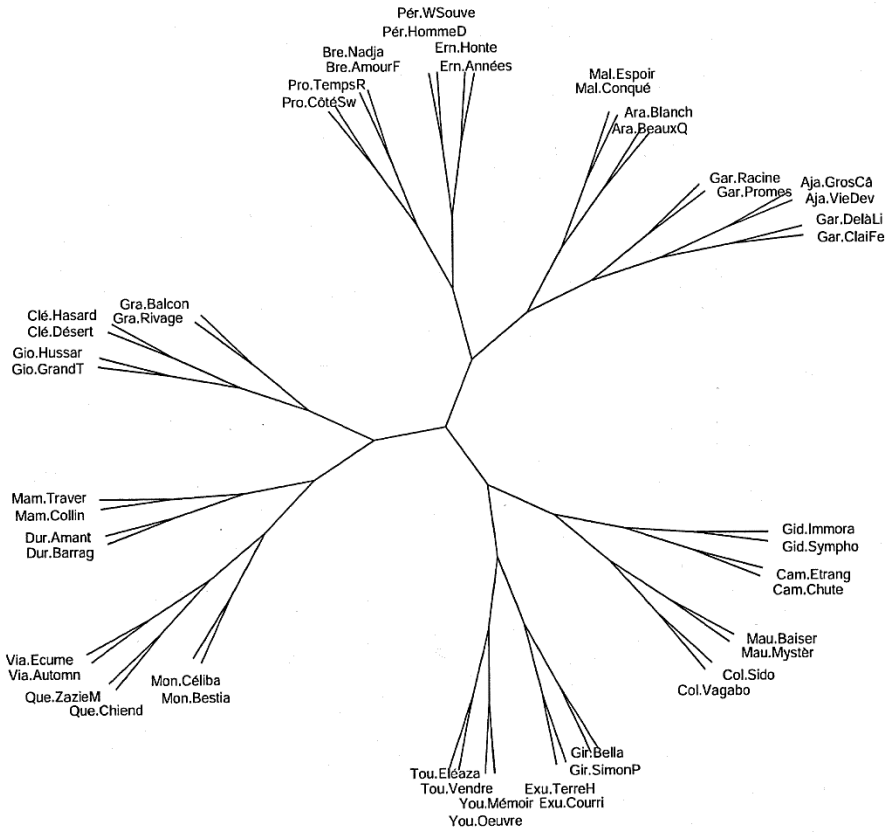
[1] On n'a pas retrouvé à la bibliothèque universitaire de Nice le mémoire d'une certaine Hélène dont parle Didier Van Cauwelaert dans le *Père adopté* (*Le Livre de Poche*, 2009, p. 241-245) et qui aurait soupçonné la collusion des deux noms à partir d'indices stylistiques et de clins d'œil à demi transparents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

[2] Voici la liste des 50 textes retenus :

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

2 – Les voies classiques

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 α le total marginal A+B,

par β le total marginal C+D,

par γ le total marginal A+C,

par δ le total marginal B+D.

La formule peut prend la forme d'un coefficient de corrélation ou celle d'un CHI2, l'une se dérivant de l'autre[3] :

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

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, 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.

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.

[3] 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 Muller, de Jaccard et de Labbé. On pourrait penser que l'algorithme d'Évrard, si efficace dans le graphique 1, l'emporte définitivement sur les autres. Mais il exige que les textes comparés soient de taille comparable. Quand l'étendue des textes est trop inégale, il vaut mieux faire appel aux autres méthodes. 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, installée dans le logiciel *SplitsTree*, est utilisée dans le graphique 1.

3 – La solution du *deep learning*[4]

L'intelligence artificielle suit la voie contraire. Elle 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[5]. 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[6]. 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.

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 3 représentent les titres retenus pour l'apprentissage, les lignes correspondant, dans le même ordre, aux romans soumis à l'épreuve de reconnaissance.

---

[4] 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.

[5] 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.

[6] 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.

	Gide.Symphonie	Proust.Swann	Giraudoux.Simon	Mauriac.Baiser	Montherlant.Célibataire	Malraux.Espoir	Exupéry.CourrierSud	Breton.Nadja	Colette.Sido	Giono.GrandTroupeau	Queneau.Zazie	Aragon.BlancheOubli	Camus.Chute	Vian.Ecume	Duras.Amant	Gracq.BalconForêt	Yourcenar.MémoiresHa	Mammeri.CollineOubliée	GARY1.RacinesCiel	Pérec.HommeDort	Tournier.Eléazar	Ajar.GrosCâlin	Gary2.ClairFemme	LeClézio.Hasard	Ernaux.Honte
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.RivageSyrtes	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
Gary1.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
Ernaux.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 3. 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[7]. 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).

[7] 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. »

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).

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

Série2	Gide	Proust	Giraudoux	Mauriac	Montherlant	Malraux	Exupéry	Breton	Colette	Giono	Queneau	Aragon	Camus	Vian	Duras	Gracq	Yourcenar	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
Dgiraudoux.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.Conquérants	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	11.1	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.BarragePacifique	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	5.42	11.18	17.63
Gracq.RivageSyrtes	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
Yourcenar.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.Choses	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 4. 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 25 paires reconnues.

	Gide	Proust	Giraudoux	Mauriac	Montherlant	Malraux	Exupéry	Breton	Colette	Giono	Queneau	Aragon	Camus	Vian	Duras	Gracq	Yourcenar	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.Ecume	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
Yourcenar.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 5. Attribution d'un titre à un auteur. Deuxième série.

C – 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 fournit la valeur maximale 79.62, qui ne dépasse guère celle du *Temps retrouvé* (68.65 dans le tableau 4) et celle du *Côté de Swann* (74.78 dans le tableau 6). 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 4 et 6. 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 8 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	0.56	2.23	3.49	4.2	2.68	**64.63**

Tableau 6. La reconnaissance des échantillons aléatoires.

II – Le théâtre classique

1 – 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, amateurs de complot, remuent encore les braises. L'affaire aurait pu et dû s'arrêter là, par un jugement autorisé de l'histoire littéraire[8]. 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[9]. 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.

[8] 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/

[9] 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.

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 avaient une ressemblance aussi étroite avec le modèle de Corneille[10]. Comme nous avons publié quelques articles sur cette question, on nous permettra d'y renvoyer le lecteur[11]. Nous nous contenterons de refaire le calcul lexicométrique avec les mêmes données, puisées au TLF. La figure 7 est proposée par le même programme d'analyse arborée qu'a emprunté Dominique Labbé[12]. 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 haute du graphique) et la prose (partie basse) et celle qui oppose parallèlement la tragédie (en haut) et la comédie (en bas). 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 du centre et de Corneille. La partie centrale est occupée par Corneille mais se divise en deux parties, suivant qu'il s'agit 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, même si deux facteurs, le genre et l'auteur s'y entrecroisent sans 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 parmi les tragédies de Corneille, 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 les invitent la chronologie et le genre, au centre des comédies en vers. 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 on devrait soutenir que les *Plaideurs* sont de Molière.

---

[10] 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.

[11] Etienne Brunet, « Où l'on mesure la distance entre les distances », in *Texto* ![en ligne], mars 2004, hptt://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.

[12] Ce programme, développé par Xuan Luong au sein du laboratoire CNRS *Bases, Corpus et Langage,* a été intégré au logiciel Hyperbase.

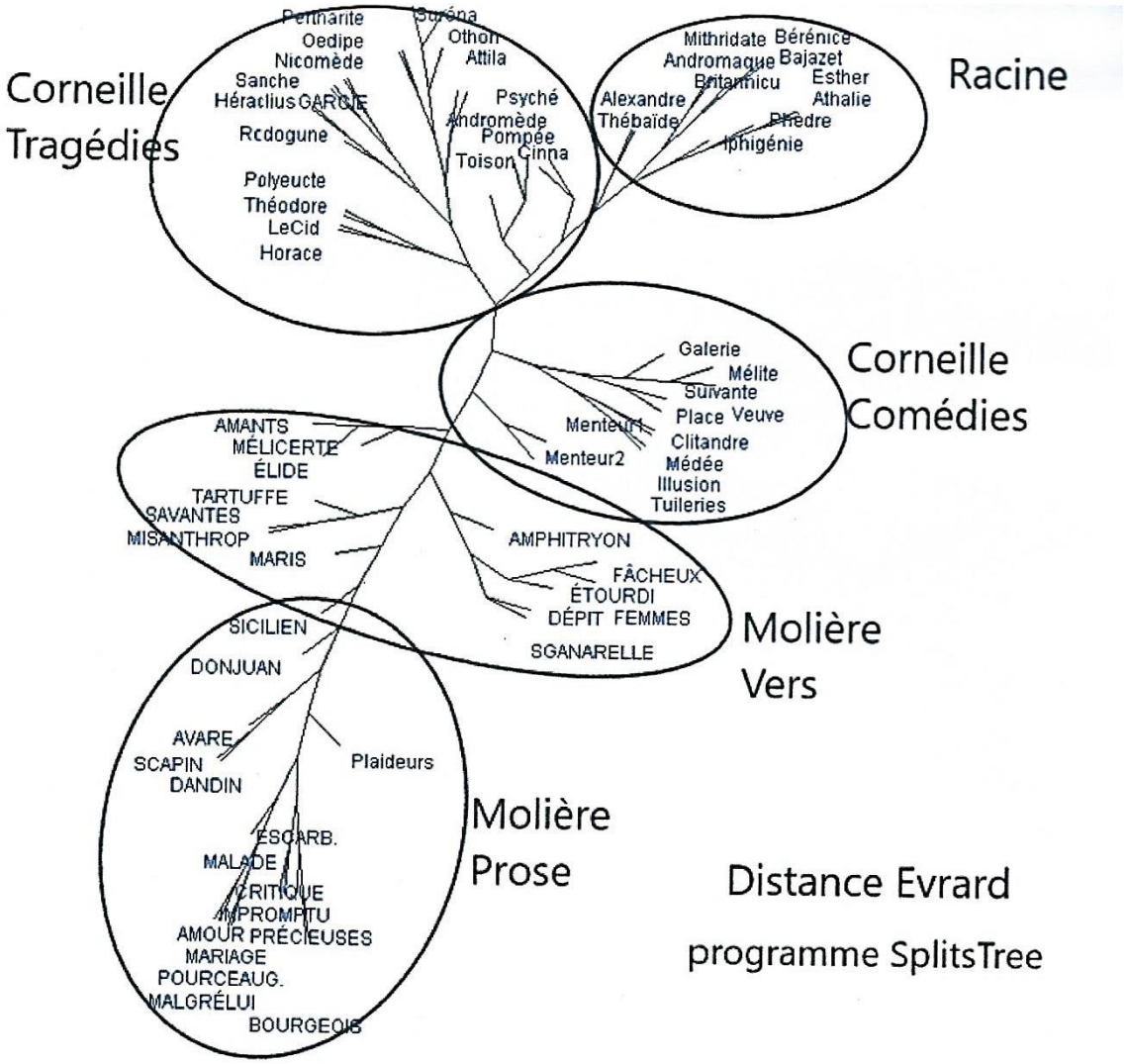


Figure 7. Analyse arborée du théâtre classique (méthode Évrard, appliquée aux lemmes)

2 – 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 8. 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[13]. Pour faire bonne mesure, ajoutons qu'en suivant une autre méthode, celle que prône André Salem pour le Tableau Lexical Entier, l'analyse factorielle aboutit à une décantation tout aussi claire. Si donc Corneille a écrit les pièces de Molière, il a bien caché son jeu, en se départissant de sa manière propre, pour imiter celle de Molière !

[13] 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.

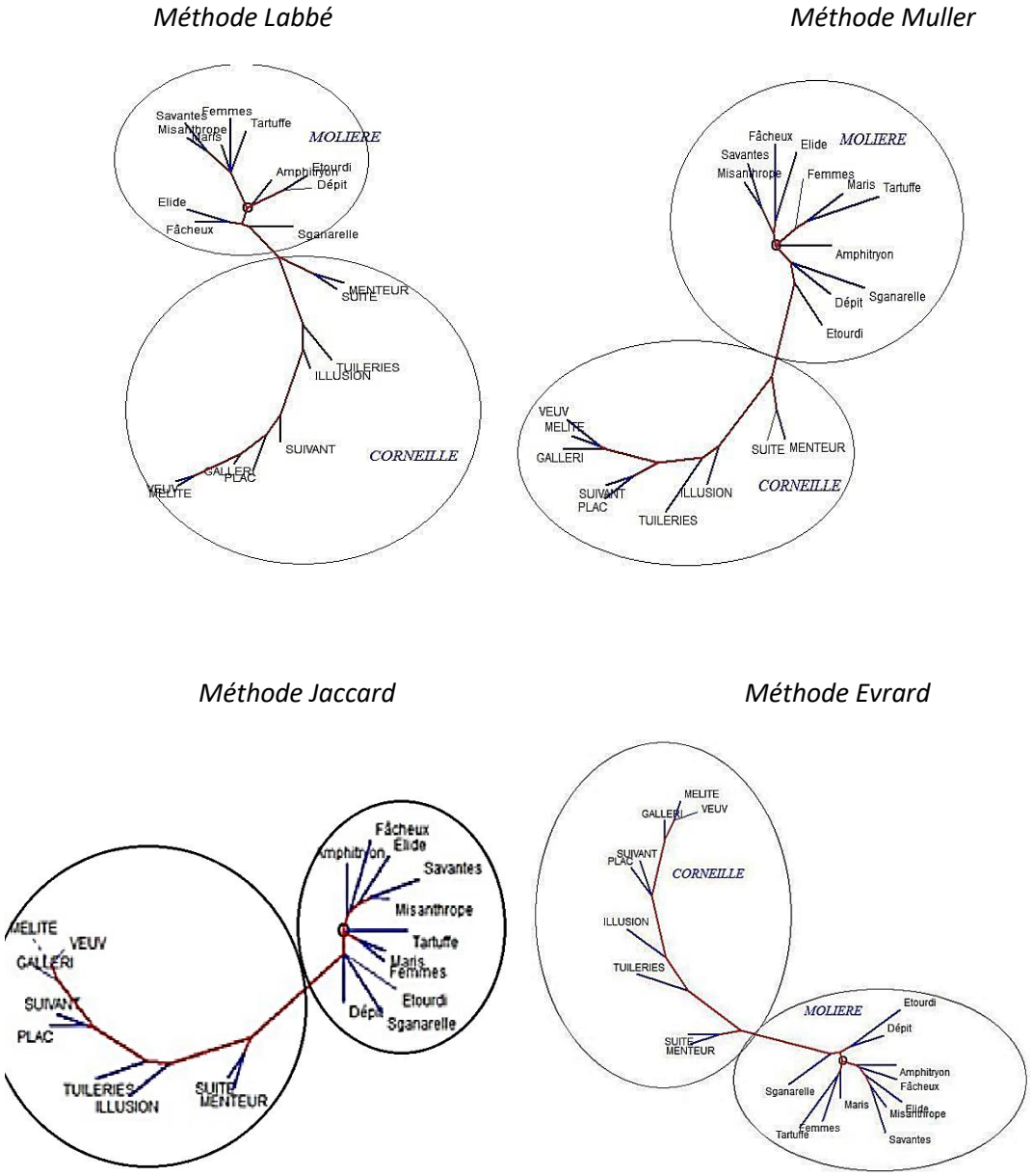


Figure 8. La distance intertextuelle des comédies en vers. Quatre méthodes convergentes

3 – 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.

Amphitryon, qui a d'abord éveillé les soupçons de Pierre Louÿs, est le premier d'entre eux dans le tableau 9. 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 Garcie 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 l'analyse classique de la figure 7 et loin d'en tirer argument contre le *deep learning* on lui fera crédit de cette sensibilité à la puissance du genre[14].

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

[14] *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.

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*[15] 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 %.

Tableau 9. Le jugement de l'intelligence artificielle dans l'affaire Molière-Corneille.

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

Tableau 10. Les comédies en vers de Molière.

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 10) 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%).

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.

[15] 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.

III – Essai d'explicitation

1 – Expérimentation humaine

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 marathonien 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 ?

2 – Imitation de la procédure du *deep learning*

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[16]. 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. 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 de son *Initiation à la statistique linguistique*. Il a le grand avantage de reposer sur la loi binomiale, ce qui permet le calcul du Chi2 et la mesure étalonnée de la distance intertextuelle[17].

---

[16] 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 que_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_m' Je_m'_endors m'_endors_2 endors_2_Et

[17] Muller a proposé sa méthode sous le nom de « connexion lexicale ». Les moyens matériels lui ont manqué à l'époque pour apporter une illustration concrète d'un calcul réputé long et complexe. L'implantation que nous en avons faite dans Hyperbase ne pose pas de problème, même pour un corpus de 4 millions de mots. Pour plus de détail voir :

Pourtant, appliquée aux mots individuels, la méthode de Muller ne surclassait pas les autres calculs, dans le classement de notre tableau 2. Son extrême précision en effet ne s'applique qu'aux basses fréquences. Or celles-ci occupent presque tout l'espace quand il s'agit de triplets. Sur plus de 2 millions de triplets différents, il n'y en a que 2441 dont l'effectif est supérieur à 50 (dont 833 dépassent la fréquence 100 et 12 seulement la fréquence 1000). Il suffit donc d'examiner le contenu des 1326 boîtes (c'est-à-dire la somme des entiers de 2 à 51) et d'appliquer à chacune la probabilité composée de la loi binomiale, compte tenu de la taille respective des deux textes en présence, pour lire la valeur globale du Chi2 et en déduire la distance qui sépare ces deux textes. Bien entendu l'opération est à renouveler pour chaque confrontation, c'est-à-dire $(n^2-n)/2 = 1225$ fois pour 50 textes. On trouvera ci-dessous les premières lignes du tableau des résultats.

GRAPHIQUE: clic sur un mot ou un texte	Gid.Immora	Pro.CôtéSw	Gir.SimonP	Gid.Sympho	Mau.Baiser	Mon.Bestia	Pro.TempsR	Gir.Bella	Mal.Conqué
95	Exu.Courri	Bre.Nadja	Col.Sido	Gio.GrandT	Mau.Mystèr	Que.Chiend	Mon.Céliba	Ara.BeauxQ	Bre.AmourF
	Mal.Espoir	Exu.TerreH	Cam.Etrang	Via.Automn	Via.Ecume	Col.Vagabo	Dur.Barrag	Gio.Hussar	Gra.Rivage
	You.Mémoir	Mam.Collin	Cam.Chute	Gar.Racine	Gra.Balcon	Que.ZazieM	Gar.Promes	Pér.HommeD	Ara.Blanch

Gid.Immora	6	18	18	16	24	22	21	21	22	26	23	23	31	22	24	24	18	24	20	23	31	30	33	22	29
Pro.CôtéSw	19	19	18	21	21	20	27	25	21	35	19	21	20	33	42	24	26	29	35	31	25	24	26	30	25
Gir.SimonP	1207 Gid.Immora																							
Gid.Sympho	18	1	16	19	21	18	11	17	21	23	17	19	29	18	21	18	15	19	19	21	29	29	28	20	24
Mau.Baiser	18	17	17	18	21	19	22	23	18	30	16	20	17	29	35	20	23	28	31	29	23	19	24	26	19
Mon.Bestia	1053 Pro.CôtéSw																							
Pro.TempsR	18	16	5	23	23	19	17	15	22	24	20	20	31	20	23	21	16	22	20	21	32	29	32	20	27
Gir.Bella	20	17	16	19	22	20	24	25	18	32	17	20	17	31	39	22	25	29	33	32	24	21	25	29	22
Mal.Conqué	1135 Gir.SimonP																							
Exu.Courri	16	19	23	6	31	24	21	26	23	30	24	29	33	23	24	24	18	25	20	26	38	31	36	23	30
Bre.Nadja	19	21	21	23	24	19	31	25	21	40	19	24	23	33	41	26	26	32	37	37	28	31	32	34	28
Col.Sido	1318 Gid.Sympho																							
Gio.GrandT	24	21	23	31	6	21	24	22	24	28	29	28	32	18	25	23	16	28	19	27	43	32	36	27	31
Mau.Mystèr	19	22	21	22	32	21	25	29	25	39	21	20	18	40	46	29	31	39	35	40	27	31	26	31	24
Que.Chiend	1351 Mau.Baiser																							
Mon.Céliba	22	18	19	24	21	3	19	18	20	21	21	21	25	18	19	14	13	22	16	21	32	24	25	21	24
Ara.BeauxQ	17	18	19	19	22	19	21	22	21	29	18	18	16	31	37	23	25	29	28	30	21	24	21	23	20
Bre.AmourF	1072 Mon.Bestia																							

Tableau 11 La distance de Muller appliquée aux triplets (extrait).

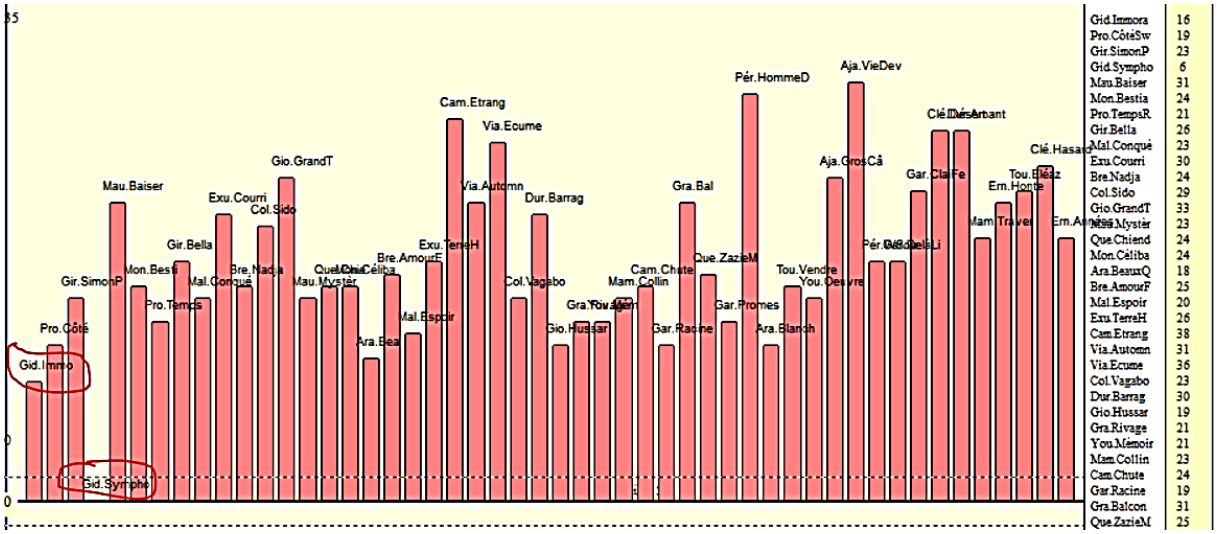


Figure 12. La distance de Muller appliquée à la *Symphonie pastorale*.

E. Brunet, « Une mesure de la distance intertextuelle : la connexion lexicale », in *Revue, Informatique et Statistique dans les sciences humaines*, n° 1 à 4, C.I.P.L., Liège, 1988, pp. 81-116.
E. Brunet, « Muller le lexicomaître », in *Mélanges offerts à Charles Muller pour son centième anniversaire*, CILF, Paris, 2009, pp. 99-119.

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 12 montre ainsi que la *Symphonie pastorale* reconnaît l'*Immoraliste* comme son voisin le plus proche. L'algorithme reconstitue correctement la plupart des paires mais 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 13 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 Pérec. 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 précedemment la palme revient à l'algorithme d'Évrard, qui obtient là aussi le plein succès.

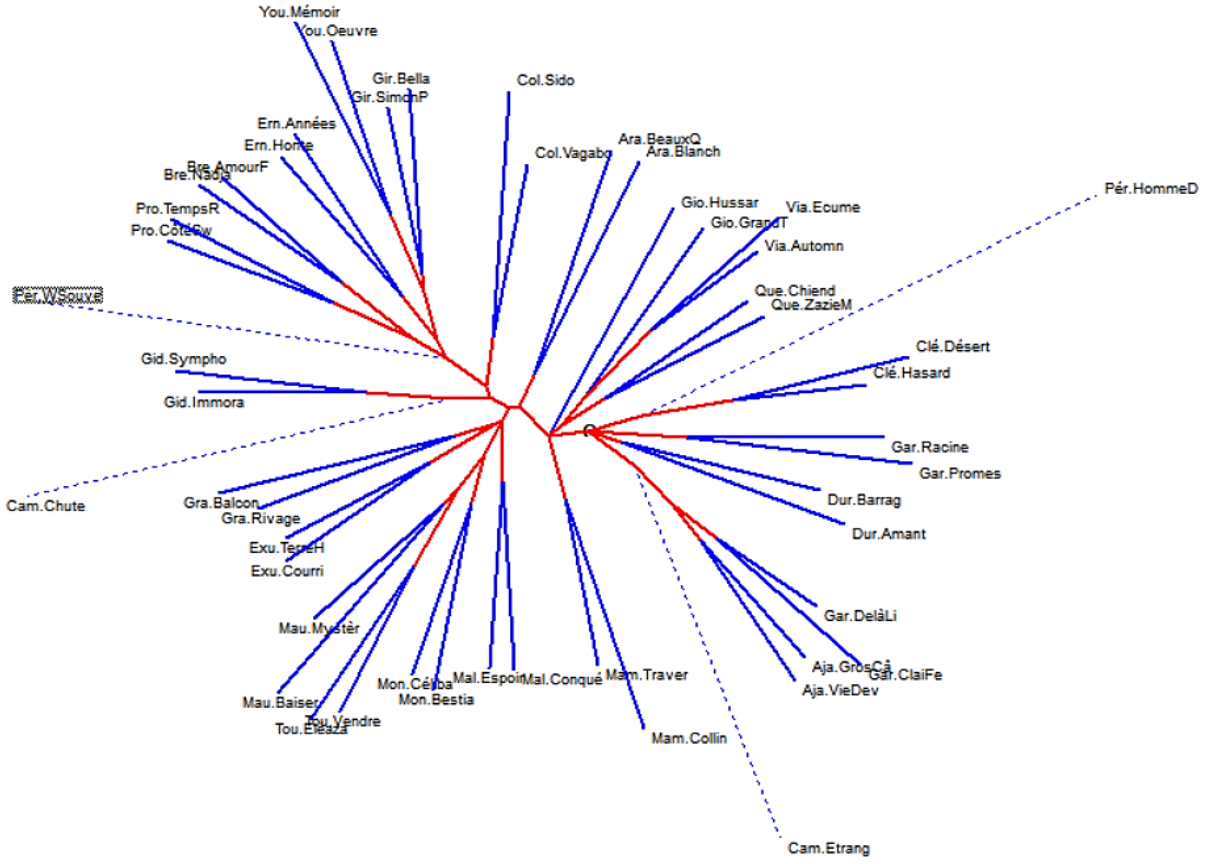
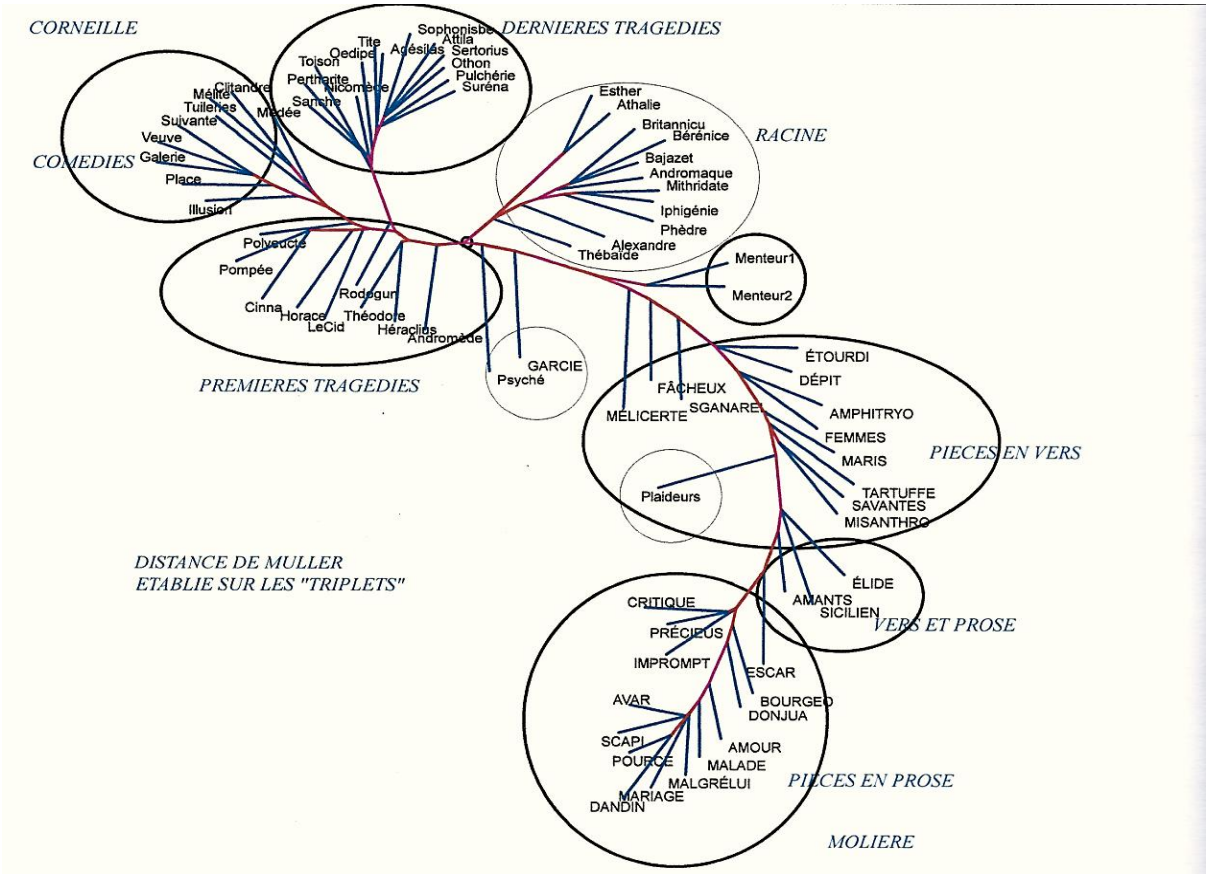


Figure 13. L'analyse arborée des triplets (à partir de la distance de Muller).
Les lignes en pointillé désignent les 4 textes mal identifiés de Pérec (*W ou le souvenir* et *L'homme qui dort*), et de Camus (*L'étranger,* et *La Chute*).

La même expérience des triplets peut être appliquée au théâtre classique. Le résultat figure dans le graphique 14 qu'on comparera au graphique 7 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*[18]), soit parce que les ballets ou intermèdes chantés et versifiés se mêlent à la prose du dialogue (*Les Amants magnifiques*).



Graphique 14. Analyse des triplets dans le théâtre classique (distance de Muller).

[18] Le *Deep learning* relevait aussi cette ambiguïté dans le tableau 12 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, nous n'attendions qu'à moitié une heureuse issue : les autres mesures de la distance n'avaient pas été encourageantes, car elles prêtaient peu d'attention aux événements rares, ce que la formule de Muller privilégie. Or 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.

3 – Intégration du *deep learning* dans un traitement classique

Ce qui gêne la comparaison des deux approches, 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, le *deep learning* a besoin de séparer apprentissage et reconnaissance et d'exclure du premier les textes destinés à la seconde. Pour atteindre la dimension maximale, il faut que le *deep learning* répète, 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 10 comédies en vers de Molière (voir tableau 10). Mais avec le corpus du roman, gros de 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. Réunissons dans un même tableau rectangulaire les résultats fournis ligne à ligne par le *deep learning* dans

les tableaux 9 et 10 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 15. 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).

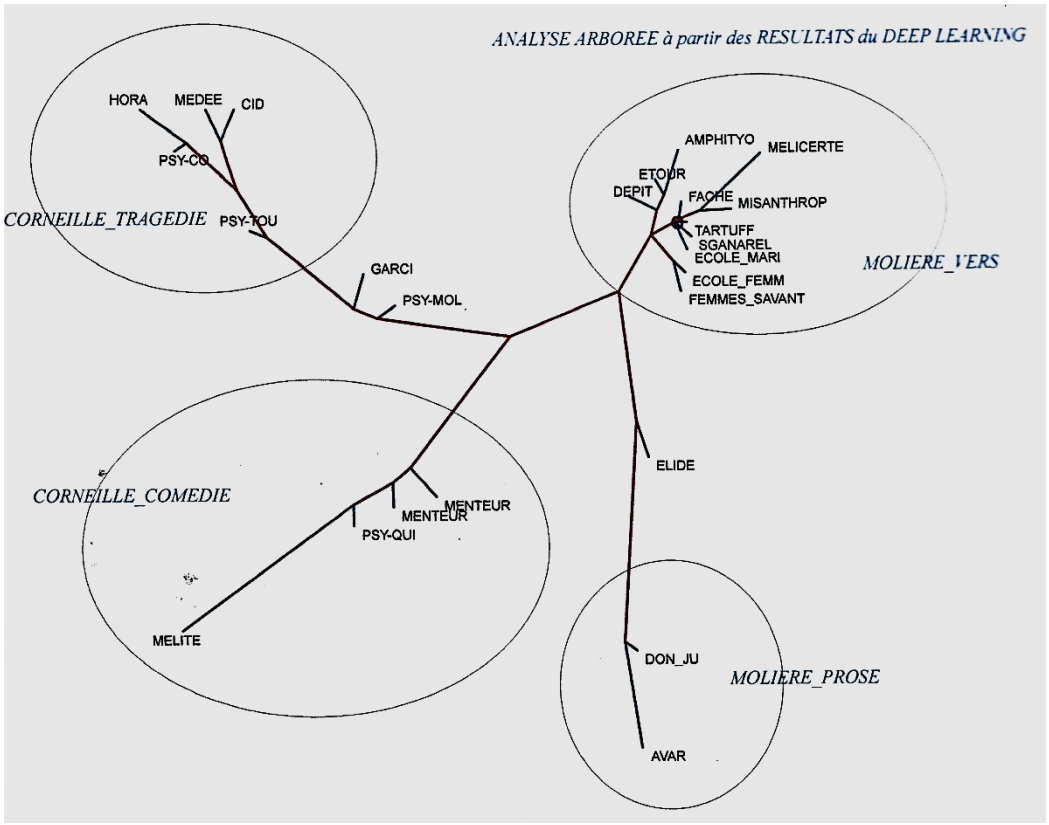


Figure 15. Analyse arborée appliquée aux résultats du deep learning (théâtre).

De la même façon dans les tableaux 4 et 5 relatifs au roman du XXe siècle, le *deep learning* associait 23 lignes à 23 colonnes. Pour s'aligner sur les résultats obtenus par la méthode Muller qui comptent 46 lignes et 46 colonnes dans une disposition symétrique obtenue d'un seul tenant, on peut ajouter l'un à l'autre les tableaux 4 et 6 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.

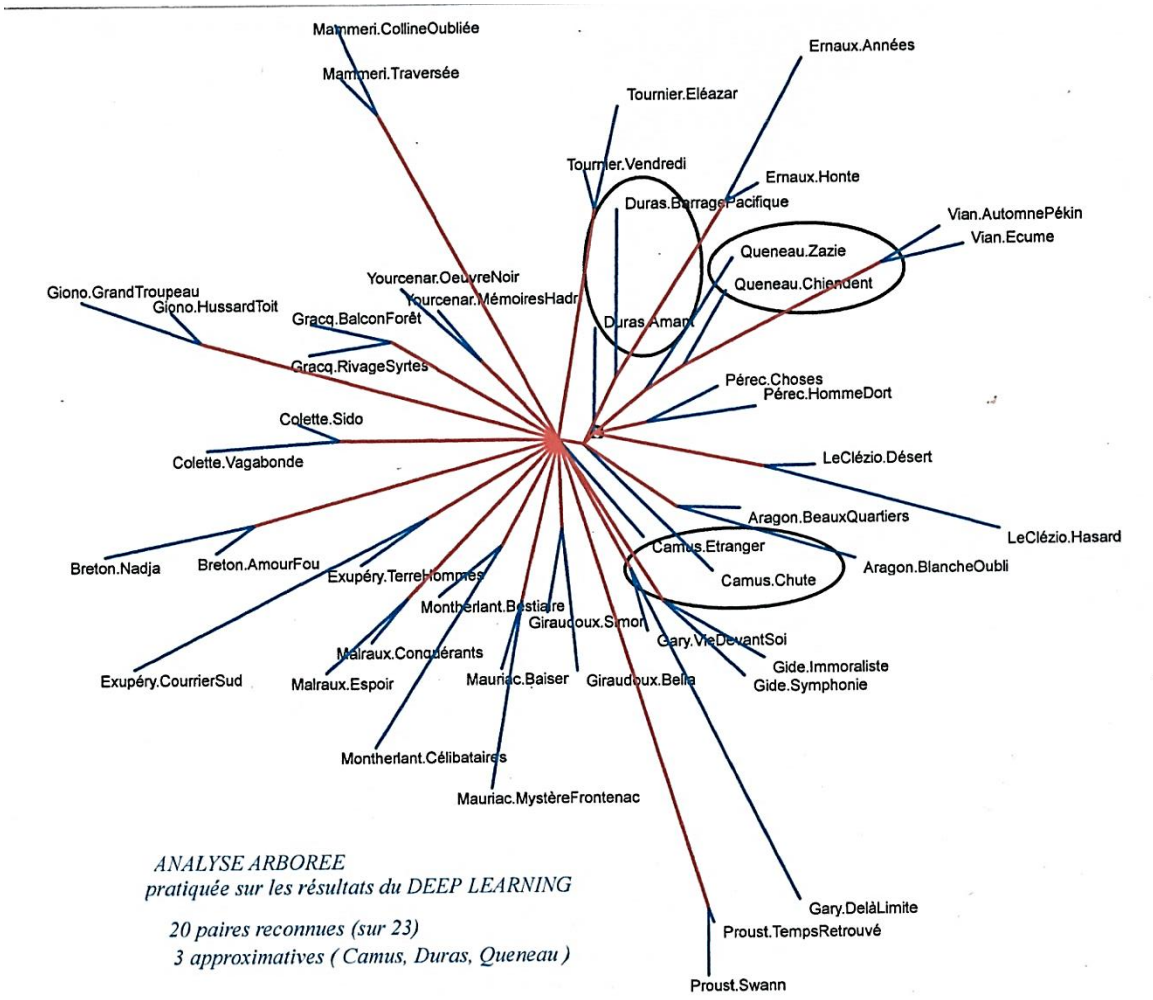


Figure 16. Analyse arborée du *deep learning* (données cumulées des tableaux 4 et 6).

À première vue la figure 16 est superposable à la figure 13. 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 4 et 5, 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.

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 de 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 13 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.

5 – 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[19]. 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. Pour *Le Menteur 1* on obtient plusieurs extraits dont

[19] 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.

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 (en rouge) ou ceux qui le sont moins (en bleu).

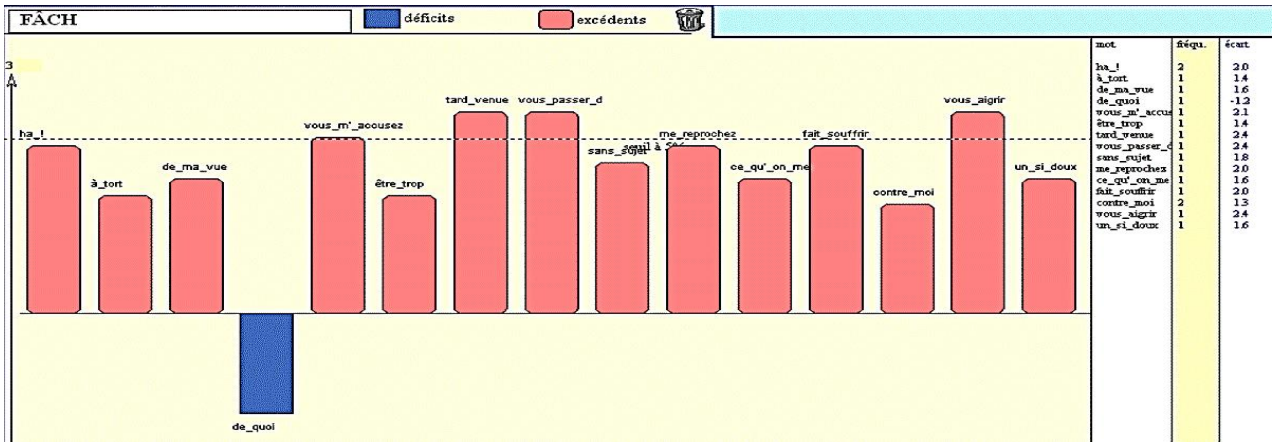
Plutôt du Corneille-Comédie :

" [...] 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'av ec 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)[20]. On ne s'étonne pas de trouver les mots *carrosse* et *dupe* dans une intrigue qui repose sur l'esbrouffe 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, nœud, 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 17). Ces mesures légères et sensibles semblent « peser des œufs de mouche dans des balances de toiles d'araignée », pour reprendre une expression de Voltaire.

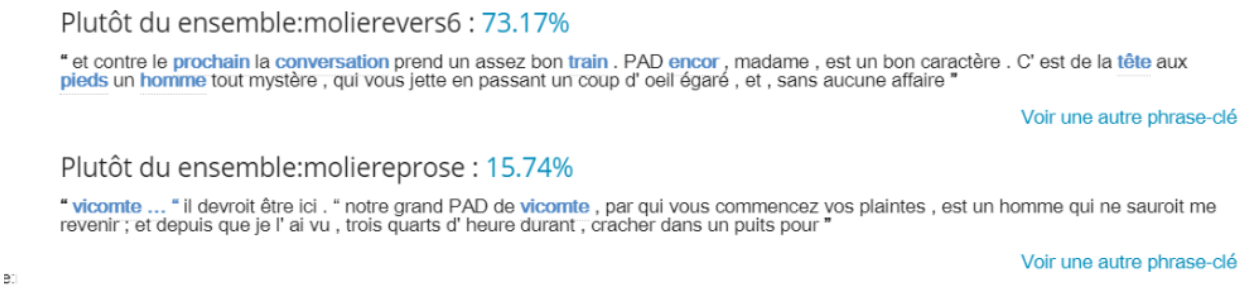


Graphique 17. Le passage tenu par la déconvolution comme spécifique des *Fâcheux*.

[20] 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.

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[21]. 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.[22]



En conclusion cette expérience, même quand elle s'accorde avec les résultats déjà acquis, est un peu frustrante pour qui a passé sa vie à écrire un code adéquat, à suivre les chemins balisés, à assurer chaque étape de la chaîne statistique, et parfois – prudemment – à tenter un modeste raccourci. Mais ici le raccourci est parfaitement direct et conduit sans attente[23] et sans détour à un résultat impossible à contester. On

[21] On a pourtant essayé, dans les deux corpus, de tirer parti des cooccurrences, en variant l'empan (page ou phrase), 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*.

[22] 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.).

[23] L'attente peut être longue : il a fallu toute une nuit de calcul à un ordinateur ordinaire, doté d'une carte graphique de bonne puissance, pour obtenir le résultat du tableau 8, qui mettait en œuvre un total de 12 millions de mots. Mais le résultat est donné d'emblée, sans qu'on vous fasse languir d'étape en étape, avec des contraintes qu'il faut respecter et des choix qu'il faut faire.

peut chicaner quand l'itinéraire est complexe et que la route n'est pas unique. Mais l'intelligence artificielle se dérobe à la chicane en menant d'un seul saut à l'objectif, comme fait un avion. Nul ne songe à parlementer avec le commandant de bord, sur le cap à tenir, d'autant que lui-même délègue le plus souvent aux commandes automatiques le soin de s'adapter aux conditions de vol. On parle souvent de boîte noire[24] 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, comme on désigne l'aéroport de destination au moment de l'envol. S'ensuivent des ajustements incessants et progressifs du processus, analogues aux corrections de trajectoire du pilote automatique. Mais s'il est possible de déchiffrer 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 *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 insuccès. 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.

[24] 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/*