

Dominique Labbé

Réponses à M. J.-M. VIPREY Corneille et Molière

**Avertissement
(mars 2008)**

Au printemps 2003, M. J.-M. Viprey – Maître de conférences à l'Université de Besançon – s'est autoproclamé « expert » et a placé sur sa page personnelle une série d'attaques contre nos travaux. Ces documents sont consultables sur :

<http://laseldi.univ-fcomte.fr/morneille.htm>

Nous y avons répondu immédiatement.
Ce dossier rassemble ces réponses classées par ordre chronologique.
Nous les reproduisons intégralement et sans aucune correction.

Sommaire

I. 16 mai 2003 Première réponse à M. Viprey	2
II. 19 mai 2003 Deuxième réponse à M. Viprey	5
III. Note du 22 mai 2003	14
IV. 10 juin 2003 M. Viprey confirme : Corneille est l'auteur des comédies en vers de Molière	17

Premières Réponses à M. J.-M. VIPREY

Le 14 mai dernier, j'ai pris connaissance du texte de M. Viprey concernant mes travaux d'attribution d'auteur (<http://laseldi.univ-fcomte.fr/morneille.htm>). J'ai immédiatement donné une réponse de principe dénonçant les contre-vérités, le ton et la méthode. On pourra lire cette réponse dans la première partie de ce document. Dès que j'ai eu un moment de libre, une étude plus attentive du texte de M. Viprey m'a convaincu qu'il était nécessaire de livrer une réponse au fond. En effet, la controverse sur les problèmes de corrélation et sur le rôle de l'AFC dépassent le simple problème de l'attribution d'auteur et concernent tous les usagers de l'informatique et de la statistique. Cette réponse au fond se trouve dans la seconde partie de ce document.

I . Réponse du 16 mai 2003

M. Viprey vient de mettre sur son site internet une note sur notre travail d'attribution d'auteur (Corneille et Molière)

1. Avant tout débat de fond, ce texte appelle quelques remarques :

- Mme Dumontet n'est pas "journaliste" mais universitaire. Constatant qu'elle ne fait pas partie de la rédaction du Monde j'ai demandé à la direction de ce journal de bien vouloir confirmer sa "mission" avant de la rencontrer.

Le 14 mai, deux heures après qu'une première version de ce texte ait été placée sur ma page personnelle, Mme Savigneau a bien voulu me confirmer, par mèl, que Mme Dumontet est bien chargée de ce travail. Entretemps, cette dernière m'avait fait savoir qu'elle ne souhaitait plus me rencontrer...

- La note de M. Viprey contient un certain nombre d'affirmations erronées. Par exemple :

- je n'ai jamais écrit que Corneille était le "nègre" de Molière. Il n'y a pas une ligne de moi qui va dans ce sens. On pourra le vérifier en lisant la conclusion de la "Réponse à mes contradicteurs" placée sur ce même site ;

- d'après ce contradicteur, je n'ai jamais présenté le calcul aux JADT. Denis Monière et moi-même l'avons fait à Lausanne en mars 2000. Cette communication a été étudiée par deux membres du comité scientifique, publiée dans les actes et elle est accessible

par internet (sur le site de la revue *lexicometrica*). Notre article du *Journal of Quantitative Linguistics* indique d'ailleurs d'autres congrès et d'autres travaux collectifs...

- mon contradicteur écrit au début de sa note : "je confesse avoir été jusqu'à maintenant totalement ignorant de cette affaire". Or, durant l'été 2002, M. Claude Blum lui a communiqué une copie de notre article du JQL. M. Blum nous a informé de cette communication par courrier...

Quelle confiance accorder à quelqu'un qui éprouve le besoin d'écrire de pareilles contre-vérités ?

2. Sur le fond :

M. Viprey n'aide guère son lecteur. Ses graphiques sont sans titre et avec des légendes obscures. Il parsème ses raisonnements formels d'explications de texte et d'invectives qui embrouillent le lecteur... Tout cela demande à être clarifié, contrôlé.

Dans l'immédiat, nous posons quelques questions qui peuvent éclairer la lecture de la note de M. Viprey :

Pourquoi ne pas avoir soumis le protocole qu'il comptait suivre et proposé de le réaliser ensemble, sous le contrôle de tiers neutres dans la querelle ? Les courriers envoyés par mon contradicteur sont aux antipodes de cette démarche logique. On pourra le constater en lisant la lettre qu'il a annexée à sa note.

Notre démonstration se fonde également sur la classification hiérarchique ascendante et sur l'analyse arborée. Mon contradicteur rejette-t-il aussi ces outils ? Sinon pourquoi ne pas en présenter les résultats sur Corneille et Molière ?

A la fin de son texte, notre contradicteur prend l'exemple de Maupassant et Flaubert. La "cuisine" à laquelle il se livre pour faire entrer ses résultats dans notre échelle ressemble au lit de Procuste sur lequel un de ses complices avait déjà couché Baudelaire et Rimbaud (voir sur notre site : "Baudelaire et Rimbaud victime d'une rumeur malveillante"). Pourquoi a-t-il choisi Maupassant et Flaubert ? Notre essai sur Corneille dans l'ombre de Molière permet de comprendre...

M. Viprey nous oppose la distance du Chi2 qu'il présente comme "universellement reconnue". Cet outil a-t-il subi les mêmes tests avec les mêmes fichiers ? Les résultats obtenus sont-ils stables et non-corrélés avec la taille des textes ? Pourquoi ne retenir que les 2.300 lemmes les plus fréquents ? L'influence de cette sélection sur les résultats a-t-elle été contrôlée ? Le lecteur doit-il faire confiance parce qu'on lui présente des dessins en couleur ?

D'ailleurs même avec cette procédure, notre contradicteur ne peut cacher totalement la curieuse position des Menteurs, de Dom Garcie et de quelques autres pièces. Et son interprétation est uniquement "visuelle". Est-ce un "progrès" ?

Enfin, M. Viprey a reçu mes fichiers lemmatisés, correspondant aux pièces de Corneille, Molière et Racine (1,1 millions de mots). Sa note annonce une "lemmatisation plus rationnelle". Nous attendons avec impatience la publication de cette avancée et la communication des données correspondantes. Qu'il ne m'oublie pas dans les destinataires : ce sera un juste retour...

Dans l'immédiat, cette polémique est stérile. Le texte de mon contradicteur soulève peut-être des points intéressants mais il est à l'opposé, dans le ton et dans le fond, d'un débat serein (dont la première condition est le respect mutuel). Il entre dans une campagne de harcèlement menée contre moi depuis six mois. La lecture du message annexé à la note de mon contradicteur éclairera le lecteur incrédule...

Dominique Labbé (16 mai 2003).

II. Deuxième réponse à M. Viprey (19 mai 2003)

Par respect pour les personnes qui s'intéressent à ce débat, j'ai pris un peu de temps pour une lecture attentive du texte de M. Viprey, en tentant d'oublier le ton et la manière... Ce texte soulève de très nombreuses questions auxquelles j'ai déjà répondu par ailleurs. Par exemple :

— les "rumeurs" sur la paternité des œuvres de Molière. Je redis que j'ai puisé cette information notamment dans... G. Forestier (voir "Réponse à mes contradicteurs" sur ce même site) ;

— la prétendue absence de faits matériels permettant de douter que Molière soit l'auteur de ces pièces. Dans la "Réponse à mes contradicteurs", j'en indique trois qui suffisent largement pour légitimer la question : "Qui a écrit les pièces de Molière ?"

— on conteste également la lemmatisation. En se reportant à ma bibliographie et aux documents placés en ligne sur ma page personnelle, le lecteur pourra lire les normes que j'ai adoptées, il y a vingt ans — j'ai suivi au plus près les conseils de C. Muller, le pionnier en la matière — et les intérêts de cette démarche illustrés avec de Gaulle, Mitterrand et divers autres corpus, notamment : Corneille, Molière et Racine. Aucun des arguments de mon contradicteur n'étant nouveau, il n'est pas utile de me répéter ;

— je ne nie pas mes insuffisances en anglais. L'essentiel de l'article du *Journal of Quantitative Linguistics* réside dans la notion de distance, les formules, les chiffres obtenus sur Corneille et Molière, la classification automatique et l'analyse arborée. Tout cela conduit à attribuer à Corneille 16 comédies de Molière (une synthèse en français est consultable sur ma page personnelle). De cela, je ne retire pas un mot. Un examen attentif des objections techniques de M. Viprey renforce plutôt ces conclusions. Je vais les exposer aussi simplement que possible de telle sorte que le lecteur cultivé, mais non statisticien, puisse se faire par lui-même une opinion.

1. Les différences de taille entre les textes comparés

C'est un des problèmes essentiels posés à la statistique appliquée aux textes littéraires. Par quels moyens vérifier que ces différences de taille n'influencent pas les calculs ? Mon censeur prétend que la distance intertextuelle ne résout pas ce problème et qu'il en apporte une preuve dans les pages 2 et 3.

En fait, ses tris et ses graphiques montrent simplement que, dans le corpus Corneille-Molière :

— toutes les farces sont courtes, toutes les tragédies sont longues et que, au milieu, il y a des comédies de longueurs variables...

— les farces sont très diverses par leur contenu, comme par leur style, et fortement décalées par rapport au reste du corpus ; les comédies sont déjà plus uniformes — celles en prose étant encore assez diverses et celles en vers, plus longues en moyenne que celles en prose, sont aussi plus homogènes — ; enfin, les tragédies sont très proches les unes des autres et se situent presque au cœur du corpus ;

— la partie gauche des graphiques 2 et 3 rend compte de la position spécifique des farces ; la partie droite, de celle des tragédies ; le milieu de celle de la plupart des comédies.

Les tris et les trois graphes présentés p. 3 ne disent rien d'autre !

En voici une vérification par l'absurde.

Les figures 2 et 3 indiquent que mon contradicteur pense avoir mis au jour une loi linéaire qui ferait diminuer la distance intertextuelle en fonction de la taille des textes. C'est pourquoi il trace des droites d'ajustement des deux séries. En prolongeant ces droites, elles devraient couper l'abscisse vers 50.000 mots (graphique 2) ou vers 80.000 mots (graphique 3) et ensuite - pourquoi pas ? - la distance devrait être négative... Avec *Mme Bovary* de Flaubert (125.000 mots), on devrait donc avoir certainement une distance nulle (voire négative en poussant le raisonnement linéaire jusqu'au bout). On verra, p. 11 de sa note, que mon censeur a été obligé de se livrer à un "bidouillage" pour amener la distance séparant *Mme Bovary* des romans de Maupassant au-dessous de... 0.25 !

Le lecteur doit savoir que ce censeur n'est pas allé chercher son exemple très loin : ayant eu connaissance des épreuves de mon essai *Corneille dans l'ombre de Molière*, il savait qu'il trouverait entre ces deux auteurs, l'un des exemples les plus frappants d'"intertextualité" de toute l'histoire littéraire française et que les distances séparant ces textes ne sont guère supérieures au seuil significatif. Il a pensé : "Un petit coup de pouce et le tour est joué". Il ne s'est pas rendu compte qu'il avouait ainsi que, au moins jusqu'à 125.000 mots, on ne constate pas de décroissance significative de la distance intertextuelle en fonction de la taille des textes considérés.

A trop vouloir prouver, il n'a démontré que sa mauvaise foi.

Le lecteur reste-t-il sceptique ?

Dans la note de ce censeur, la figure 7 p. 7 donne une représentation grossière — et probablement "manipulée", comme nous le verrons dans un instant — du "nuage" formé par le corpus (en quelque sorte, chaque texte est figuré dans ce nuage par un point dont les coordonnées relatives sont déterminées par les distances qui le séparent

de tous les autres textes). Dans cette figure 7, la représentation est grossière mais suffisante pour apercevoir près de l'origine (le centre de gravité du nuage), toutes les tragédies de Corneille et les plus longues des comédies. En revanche, les farces sont à l'extrême-gauche de ce même graphe, plus éloignées de ce centre de gravité (nous expliquons plus bas pourquoi elles ne le sont pas plus)... Le premier axe de son *analyse factorielle des correspondances* (AFC) classe donc essentiellement les pièces en fonction de leur taille : la majorité des longs se situent à droite ou juste à gauche de l'axe vertical, les moyens vers le centre ou le centre-gauche et les petites farces à l'extrême-gauche.

Si l'on rejette la distance intertextuelle à cause de cette histoire de taille, alors il faut écarter l'AFC pour la même raison !

Il me serait facile de prendre la pose comme M. Viprey et de tonner contre lui avec les mêmes épithètes. Je me contenterai de rappeler qu'une liaison statistique entre deux variables indique simplement une "covariation" et non pas un lien causal de l'une à l'autre. Un débutant peut oublier cette vérité de bon sens mais un "expert" ?

Dans le corpus Corneille-Molière, la liaison entre la taille et la distance (intertextuelle comme du Chi²) s'explique simplement par l'influence conjointe, sur ces deux variables, des genres et des auteurs.

2. L'AFC est-elle supérieure à la distance intertextuelle ?

En premier lieu, il faut rappeler que, contrairement à ce qu'affirme mon censeur, l'AFC est largement contestée, même parmi les statisticiens français ; quant à la communauté scientifique internationale... Une remarque au passage, l'AFC n'est pas "sophistiquée". Le calcul de la distance du Chi² proprement dit n'est guère plus complexe que celui de la distance intertextuelle et il n'y faut pas beaucoup plus de lignes de programme. Il me paraît exact de dire que les analyses multidimensionnelles peuvent avoir une valeur *exploratoire*. En contrepartie, elles sont grossières, difficiles d'emploi et demandent de l'expérience.

L'AFC est "robuste" et sans biais ?

Le lecteur peu au fait des subtilités de cette méthode aura certainement été surpris de voir M. Viprey commencer sa démonstration sur les "2.300 lemmes les plus fréquents". Pourquoi ne pas commencer par *tout le vocabulaire* ? Tout simplement par cette méthode en est incapable : soit on travaille avec les mots les plus fréquents, soit avec les rares, mais il est impossible de traiter les deux en même temps ! Nous laissons

à M. Viprey le soin d'expliquer à la galerie pourquoi sa méthode miracle est déstabilisée par le mélange entre mots rares et mots fréquents, alors que le langage naturel est justement fait de ce mélange.

Pire encore : il n'y a aucune règle évidente pour fixer la ligne de partage entre les mots à retenir et ceux à éliminer... L'opérateur consciencieux tentera de garder le maximum d'information, sans introduire de biais ; le maladroit produira facilement des monstres ; le malhonnête tâtonnera jusqu'à ce que la figure obtenue soit aussi proche que possible de ce qu'il veut "prouver" !

Signalons au passage que, en statistique, contrairement à ce que feint de croire notre "expert", "robustesse" ne signifie pas "solidité à toute épreuve" mais simplement qu'une modification marginale dans les données n'entraîne qu'une modification marginale dans les résultats. En revanche, quand on chamboule tout et que les résultats sont les mêmes, l'outil n'est pas "robuste", il est inutile...

L'AFC estime correctement les distances entre les textes ?

La première analyse porte sur les 2.300 vocables les plus fréquents. Le corpus en comporte au total : 9.994. Comment prouver que les 7.694 vocables qui sont écartés (les trois quarts du vocabulaire) n'ont aucune importance... Le postulat est loin d'être démontré. Ce serait plutôt l'inverse. Pour le corpus Corneille et Molière, nous tenons à la disposition du lecteur curieux la liste de ces 7.694 mots "sans importance" : la quasi-totalité des noms propres, plus des neuf dixièmes des substantifs et des adjectifs ! Nous tenons également à sa disposition une note, sur les propriétés de la distance intertextuelle, dans laquelle nous montrons que ces mots "rares" sont ceux qui génèrent le plus de "contraste" entre les textes.

Comme la sélection des vocables pertinents se fait sur le corpus entier, les différences de taille entre les pièces ont, pour le coup, des répercussions considérables. Les pièces les plus longues gardent quelques-uns de leurs personnages principaux et de leurs thèmes, les petites farces de Molière sont quasiment réduites à un squelette de mots outils et de verbes usuels. Sur les graphes d'AFC, elles apparaissent serrées les unes contre les autres alors qu'elles sont extrêmement diverses.

Une expérience comme celle-ci sera nécessairement grossière et les conclusions ne pourront aller très loin.

L'AFC doit être utilisée avec prudence

Les pages suivantes de M. Viprey montrent tout ce qu'il ne faut pas faire : l'opérateur élimine les cent premiers mots (pourquoi pas 99 ou 101 ?), puis les mots commençant par les lettres A à E, puis la quasi-totalité du vocabulaire et on ne sait

quel charcutage encore... Et il sort toujours la même chose de sa boîte noire ! Quand l'expérience est idiote, quelle signification peut avoir le résultat ?

Nos collègues littéraires peuvent à bon droit sursauter devant tant de désinvolture et nous partageons leur méfiance légitime envers ce genre d'opérations, menées à l'aveuglette, mais toujours "au bulldozer", surtout quand elles sont le fait de gens qui déclarent benoîtement se mettre à leur service...

Les graphes sont faciles à lire et à interpréter ?

En fait, la lecture est souvent "contre-intuitive". Il y a une erreur à ne jamais faire : considérer les axes comme des césure naturelles. C'est l'erreur que commettent tous les débutants et dans laquelle tombe également notre expert : Molière à gauche, Corneille à droite ! Le résultat, ose-t-il écrire, "va au-delà même de toute espérance" (il n'avait pas d'hypothèse mais des "espérances" !)

Pourtant, en AFC, on tombe très rarement sur un tel cas de figure. Il se produit quand deux conditions sont réunies en même temps :

- la zone centrale est vide, ou beaucoup moins densément peuplée que tout le reste. Or ce n'est pas le cas sur les figures 7 et 8 !

- le premier axe est beaucoup plus important que le second (dans ce cas, on peut négliger la dimension verticale et partager les données en fonction de leurs coordonnées sur l'axe horizontal). On peut le savoir grâce au "pourcentage d'inertie" qui doit figurer sur chaque axe. Surprise : cette information essentielle a été supprimée sur les figures 7 à 11 ! En fait, pour le corpus Corneille-Molière, le second axe du premier plan factoriel rend compte d'un pourcentage important de l'inertie totale.

Aucune des deux conditions nécessaires pour considérer l'axe vertical comme une ligne de partage n'est donc réunie. Dans un cas semblable (zone centrale peuplée et second axe important), les césures sont à rechercher dans les zones vides — ou très peu denses — et plutôt dans les diagonales...

Muni de l'information selon laquelle le second axe est presque aussi important que le premier, tout lecteur exercé à l'analyse factorielle n'aura aucun mal à lire la figure 7 (nous lui conseillons simplement d'imprimer le graphe sans les couleurs !) :

- la diagonale principale part du coin gauche, en haut de la figure, passe légèrement au-dessus de l'origine et rejoint le coin droit en bas de la figure. Au-dessus de cette ligne, on trouve les tragédies de Corneille et **Psyché** ; en dessous : les comédies ;

- une autre diagonale, perpendiculaire à la première, descend depuis le milieu du haut de la figure et isole à gauche la majorité des comédies en prose de Molière et toutes ses farces ;

— entre ces deux blocs homogènes, s'étend une zone ovale moins dense, étirée en dessous de la diagonale principale. Elle regroupe toutes les comédies en vers, de Corneille et de Molière ainsi que *Dom Juan* et *l'Avare*...

Chacun pourra tracer facilement ces droites de séparation et les frontières de ces 3 zones. On retrouve sans peine, les trois groupes évoqués plus haut... Tout cela est bien loin des "espérances" de M. Viprey !

Cette mésaventure montre le principal danger de l'AFC : la facilité apparente de la méthode favorise l'intuition molle, l'aveuglement par la doxa. La lecture des graphes demande du coup d'œil et de l'entraînement. Elle comporte une part inévitable d'interprétation et dépend beaucoup de la bonne foi de l'opérateur.

L'AFC repose sur la bonne foi de l'opérateur

Je n'ai aucun doute quant à la qualité du logiciel de notre collègue Lelu, mais beaucoup sur la manière dont il a été utilisé et sur le détail des graphes présentés. Cette question doit être posée car — outre la position curieuse des points 32 et 35 sur lesquels je reviens plus bas —, il y a bien eu intervention manuelle sur ces figures. D'abord, l'effacement du pourcentage de l'information totale dont rend compte chacun des axes. J'ai déjà donné l'explication principale (le coup de force visant à imposer la coupure verticale et à faire "oublier" la diagonale). J'ajoute que, quand deux outils sont en concurrence, ces pourcentages permettent de connaître le plus efficace : en l'occurrence, les lemmes (figure 7) ou les formes (figure 8) ? Pour avoir réalisé la même expérience, je peux affirmer que les lemmes se révèlent nettement plus performants que les formes graphiques. Tout à sa véhémence, ce censeur a voulu "parfaire" ma défaite... Il aura simplement montré au lecteur attentif combien il est peu crédible !

Surtout : c'est l'opérateur qui a colorié chaque point. Au passage, n'en aurait-il pas déplacé légèrement certains ? Par exemple, les points 32 et 35 ou 15 et 16... Pourquoi lui faire confiance alors que, par ailleurs, nous avons vu qu'il est capable d'écrire des contre-vérités, de "bidouiller les données", de cacher des informations essentielles pour l'analyse ?

La qualité de l'analyse dépend du soin apporté aux données

Le lecteur attentif aura certainement repéré que le corpus Viprey comporte deux fois la même pièce : **Psyché** (elle figure dans l'annexe 3 sous les numéros 32 et 35). Tout opérateur peut faire des bourdes mais, quand on s'érige en censeur, il vaut mieux éviter de se mettre dans une situation ridicule. C'est d'autant plus amusant que cette pièce, publiée sous le seul nom de Molière, a bien été écrite par Corneille : c'est un test avec les moliéristes, certains oublient **Psyché**, d'autres en bégayent !

Puisque les numéros 32 et 35 représentent probablement le même individu, il faudrait que, sur les graphiques issus de l'AFC (figures 7 à 11), ces deux points soient confondus. Or ils apparaissent nettement distincts sur toutes les figures (dans les tableaux 7 et 8, ils sont dans le quart nord-est près de l'origine). A-t-il fait une erreur en recopiant mes fichiers lemmatisés ? Le corpus a-t-il été abîmé ? Y a-t-il une erreur de numéro pour certaines pièces ? Le logiciel est-il bogué ? A-t-on déplacé manuellement, sur les graphes, ces deux points superposés (pour rendre la figure plus "lisible" ?)

Après la bordée d'injures que vient de m'envoyer M. Viprey, je me vois mal lui écrire : "Mon cher collègue, accordons nos violons et assurons-nous d'abord que nous travaillons bien sur les mêmes fichiers". Maintenant le mal est fait. Quelle que soit l'explication, je ne peux que lui dire : "Navré ! aucun crédit ne peut être accordé à votre raisonnement..."

L'incident me semble tout de même révélateur. Nous ne travaillons manifestement pas dans le même univers. Lui peut produire des tris et des graphes à la chaîne, les enrober d'un jargon pompeux, fulminer des excommunications... sans se rendre compte qu'il a une pièce de "trop" dans son corpus. A l'opposé, j'essaie de développer des outils et des traitements adaptés aux caractéristiques particulières de la langue et des textes français. Pour moi, le travail commence par une prise de connaissance approfondie des données. Le respect de la langue et des textes est l'article premier. Il m'arrive de faire des erreurs, mais on ne trouvera pas dans mes publications six graphes à la file comportant tous... un fantôme de texte ! Tout simplement parce que je contrôle chacun des résultats et que j'y réfléchis de manière approfondie. Si mon censeur avait réellement pensé au commentaire que méritaient ces figures, il aurait bien fallu qu'il les regarde d'un peu plus près.

Conclusions

Le lecteur me pardonnera cette discussion un peu longue. Elle lui aura permis de comprendre trois choses essentielles.

1. Les méthodes multidimensionnelles ne sont pas aussi sûres et puissantes qu'on veut bien le dire. Bien conduites, par un opérateur honnête et consciencieux, elles ont une valeur exploratoire. Sur le corpus Corneille-Molière, elles ne permettent pas d'identifier l'auteur mais simplement les trois genres principaux qui partagent le corpus. Comment aller plus loin ? Comment passer d'un jugement "à vue d'œil",

nécessairement grossier, à une réelle mesure scientifique ? Il est possible que la distance intertextuelle ne soit pas la seule réponse possible mais, en attendant, mon censeur ne propose rien d'autre...

2. De quelle manière aurait-on dû s'y prendre pour vérifier une éventuelle influence de la taille sur la distance intertextuelle ?

Il aurait fallu neutraliser les facteurs qui déterminent cette distance, c'est-à-dire, outre l'auteur : le genre, le thème et l'époque. Le corpus devrait comporter, pour les différentes longueurs, autant de textes de Molière que de Corneille, autant de tragédies que de comédies et de farces, etc...

3. Les conditions d'un véritable débat scientifique.

Dans cet échange, les conditions d'un vrai débat scientifique — notamment le respect mutuel et la transparence des procédures — ne sont pas réunies par la faute de mon censeur. En effet, j'ai fourni tous les éléments nécessaires ; il ne fait pas de même. Il avait tous mes textes de longue main ; je dois réagir sans délai à un pamphlet obscur, confus, plein de "coups de force" et de bourdes.

Dans une controverse scientifique, les partenaires doivent être à égalité. Il faut du temps pour que chacun puisse prendre connaissance des arguments des autres, demander des précisions, des vérifications, des expériences complémentaires...

De plus, il faut des arbitres qualifiés. Ceux-ci auraient empêché les mauvais coups, les manoeuvres diffamatoires et les erreurs stupides... Surtout ces arbitres auraient imposé un véritable protocole d'expérimentation. En l'occurrence, on aurait pu se mettre d'accord sur des procédures d'extraction d'échantillons aléatoires de différentes tailles, parmi plusieurs auteurs et genres. Naturellement, il faudrait que ce travail soit fait sous le contrôle d'arbitres impartiaux, que l'on se mette également d'accord au préalable sur les normes de dépouillement des textes et que l'on teste non seulement la distance intertextuelle mais aussi la distance du Chi2, puis les différentes techniques de représentation graphique, de classification, d'analyse arborée...

Beau programme scientifique ! Ce serait mieux que de m'envoyer des excommunications...

Répetons-le : une véritable discussion scientifique ne peut se dérouler ainsi sur les pages internet et encore moins dans la grande presse.

Mais au fait, pourquoi mes contradicteurs crient-ils si fort, s'ils sont sûrs d'avoir raison ? Leurs hurlements et leurs manoeuvres rendent impossible le débat de fond qu'ils craignent manifestement. Au passage, ils donnent une image déplorable de l'université française.

Le public cultivé et intéressé par ces questions pourra se reporter à notre essai sur *Corneille dans l'ombre de Molière* paru aux éditions "Impressions nouvelles". Pour ceux qui désireraient aller plus loin, je tiens également à disposition une note sur les principales propriétés de la distance intertextuelle (fournir une adresse postale).

Note du 22 mai 2003

Hier, M. Viprey a mis en ligne un texte (<http://laseldi.univ-fcomte.fr/morneille2.htm>) qui amende considérablement sa première analyse de notre travail sur Corneille et Molière (<http://laseldi.univ-fcomte.fr/morneille.htm>). Après lecture du second texte, je voudrais souligner deux points importants.

1. La distance intertextuelle

M. Viprey convient qu'il ne peut démontrer une supposée dépendance de la distance intertextuelle par rapport à la longueur des textes. Dans ce second texte, il rapporte une expérience intéressante sur les contes de Maupassant. A ce propos, il faut d'abord rappeler que, sur les textes très courts (quelques milliers de mots), l'influence des arrondis introduit une incertitude dans les calculs. Nous avons signalé ce problème dans notre article paru dans le *Journal of Quantitative Linguistics* et un texte en français aborde plus précisément cette question (je le tiens à la disposition des personnes intéressées). Malgré cela, l'expérience sur les contes de Maupassant, notamment le graphique 7, montre que, au moins jusqu'à des différences élevées entre les dimensions des textes comparés, il n'y a pas de décroissance significative de l'indice en fonction de la taille. Evidemment, pour les deux bouts de la distribution (lorsqu'on approche du rapport 1/10) : de son côté, il peut prétendre que l'indice est sensible aux différences de tailles ; de mon côté, je peux soutenir que l'indice est influencé par la nature un peu singulière des plus "longs" et des plus "petits" contes... En tout état de cause, ces deux "problèmes" (poids des arrondis du fait de petites tailles et grande différence de dimensions entre les textes comparés) ne concernent pas notre travail sur Corneille et Molière puisque, dans ce corpus, il n'y a pas de très petits textes et que l'échelle des pièces attribuées à Corneille est restreinte : au maximum de 1 à 4 entre *l'Avare* (21.000 mots) et *Mélicerte* (5.500 mots), la quasi-totalité des pièces comparées étant comprises entre 10.000 et 20.000 mots. Pour ces conditions-là, M. Viprey n'apporte aucun démenti à notre affirmation : l'indice mesure sans biais la distance intertextuelle entre les pièces considérées.

Naturellement, étant donné la suspicion dans laquelle me tiennent M. Viprey et ses amis, seuls des tests — effectués par des experts indépendants et selon des protocoles sérieux — pourront, à leurs yeux, valider le champ d'application de notre formule, l'incertitude relative dans laquelle sont inscrits les résultats, la pertinence de l'échelle que nous proposons. Mais dans ce cas : pourquoi rejeter a priori mon travail avant toute expertise sérieuse ?

2. A propos de l'AFC

M. Viprey et moi sommes d'accord pour considérer les analyses multidimensionnelles comme "exploratoires" et non comme un outil qui aurait réponse à tout (il resterait à s'accorder sur certaines règles de prudence et de lecture !). Nous avons donc déjà trouvé la première étape d'une méthodologie commune (et aussi les dernières étapes qui seront bien : la classification automatique et l'analyse arborée, comme nous l'avons fait sur Corneille et Molière). Il reste à comparer les différentes mesures imaginables pour la distance et à choisir la ou les meilleure(s) (voir la conclusion du point précédent). Pour l'instant, M. Viprey n'a rien à proposer.

En ce qui concerne la lecture des figures 7 et 8 de sa première note (et à nouveau dans la figure 1 de sa deuxième note), M. Viprey indique explicitement que son raisonnement porte sur le PREMIER AXE SEULEMENT. Je maintiens que cette lecture est GRAVEMENT ERRONEE. En AFC, il faut, au minimum, prendre en compte LES DEUX PREMIERS AXES SIMULTANEMENT : c'est-à-dire le PREMIER PLAN.

Dans les figures 7 et 8, il y a TROIS groupes nettement distincts :

- à droite et au-dessus de la première diagonale : les tragédies de Corneille ;
- à gauche à l'extrémité de l'axe : les petites comédies en prose de Molière (que j'ai appelées "farces" à la suite de M. Forestier, notamment : l'AFC lui donne raison et je lui concède volontiers que ces pièces ont une singularité par rapport au reste de l'oeuvre) ;
- au milieu, une zone allongée, en dessous de la première diagonale, rassemble les comédies de... de qui au fait ? Eh bien ! l'AFC n'est pas capable de répondre à cette question !

Les personnes sceptiques pourront interroger les spécialistes d'analyse factorielle. Elles le feront avec les deux graphiques de M. Viprey (ainsi on ne pourra pas m'accuser de partialité) : il leur faudra préciser à leur interlocuteur que le second axe pèse presque aussi lourd que le premier (elles pourront ajouter que l'examen des plans factoriels suivants est intéressant mais n'est pas indispensable pour cette question précise et que, dans le groupe central, les numéros renvoient à un classement chronologique).

Conclusions

Je tiens à disposition des chercheurs le corpus lemmatisé Corneille-Molière (avec trois fichiers pour **Psyché** : la pièce entière, les vers rédigés par Corneille, ceux de Molière) et les programmes correspondants. Naturellement, avant de se livrer à des expériences, il sera toujours préférable de s'assurer ensemble que tout est en ordre...

Pourquoi M. Viprey perd-il tant de temps à discuter un brouillon qui a été mis en ligne quelques heures seulement à la suite d'une mauvaise coordination avec notre webmestre ? Si ce texte a été immédiatement retiré, c'est qu'il n'était pas destiné à publication.

M. Viprey reconnaît d'ailleurs lui-même que les conditions d'un véritable débat scientifique ne sont pas réunies... N'aurait-il pas mieux valu s'en apercevoir avant de rédiger sa première note ? N'était-elle pas inconsidérée, voire calomnieuse ?

Ai-je eu tort de réagir vite ? Imagine-t-on ce qui se serait passé si j'avais gardé le silence comme j'ai eu le tort de le faire lors d'un précédent épisode de l'«affaire» ? (Voir à ce sujet : "Baudelaire et Rimbaud victime d'une rumeur malveillante" sur ma page personnelle).

Pour ce qui concerne l'«affaire» et le «dossier», je renvoie le lecteur à ma note : "La presse face à Corneille et Molière", sur ce même site. Evidemment, seul le dossier m'intéresse...

Enfin, il est exact que, derrière cet affrontement, il y a plusieurs débats intéressants, notamment sur le rôle des méthodes statistiques en sciences humaines ou, plus spécifiquement, sur la notion de probabilité. A ce propos, je reconnais avoir tendance, parfois, à utiliser le mot "preuve" là où, en toute rigueur, il faudrait écrire : "faisceau d'un grand nombre d'indices convergents". Une pareille lourdeur répétée à l'infini ne risque-t-elle pas de lasser le lecteur qui n'est tout de même pas idiot ? Je ne reviens pas ici sur ces indices qui — en plus des distances intertextuelles, de la classification automatique et de l'analyse arborée — convergent pour mettre en lumière le rôle de Corneille dans l'œuvre de Molière. Mon essai *Corneille dans l'ombre de Molière* en évoque un certain nombre. Maintenant qu'on s'intéresse au dossier (pas à l'affaire), on va certainement en trouver beaucoup d'autres...

M. Viprey confirme :
Corneille est l'auteur des comédies en vers de Molière
(10 juin 2003)

Dominique LABBE

M. Viprey vient de mettre sur son site internet le document suivant :
<http://laseldi.univ-fcomte.fr/morneille0.htm>

La seconde partie de ce document attribue à Pierre Corneille l'ensemble des pièces en vers de Molière.

En premier lieu, on le voit dans le graphique (que nous reproduisons en annexe de cette note). Une seule erreur dans cette figure : le point 14, correspondant au *Dépit amoureux* (1658), doit être en rouge puisque l'édition originale désigne Pierre Corneille comme étant l'auteur (voir *Réponse à mes contradicteurs* sur ma page personnelle). Il y a du Corneille des deux côtés de l'axe vertical !

Le tableau chronologique que nous plaçons en annexe aidera à la compréhension de cette figure.

Le premier axe (lecture horizontale, de droite vers la gauche, en considérant l'abscisse de chaque point sans prendre en compte son éloignement par rapport à cet axe) : le classement correspond à la chronologie. On voit nettement isolées à droite les 8 premières oeuvres de Corneille. Le reste forme un seul ensemble. Une seule exception : *Psyché* qui n'appartient pas tout à fait au même genre, comme on pouvait déjà s'en douter en lisant les graphes de la première note de M. Viprey.

Le second axe (lecture verticale, en considérant l'ordonnée de chaque point sans prendre en compte son éloignement par rapport à l'axe) : les deux oeuvres sont étroitement mêlées. Cependant, c'est Corneille qui occupe les deux extrémités de l'axe (n°9 et 12)... Certes pour affirmer que ce second axe doit être pris en compte, il faut connaître son "poids" (le pourcentage de l'inertie totale du nuage dont il rend compte). Normalement, cette information figure le long de chacun des axes. Pourquoi M. Viprey l'aurait-il supprimée si elle lui était favorable ?

Il s'agit donc de :

L'oeuvre d'un auteur unique classée par ordre chronologique.

Aucun spécialiste d'analyse factorielle ne me contredira. Au contraire, nous avons là un véritable "cas d'école" qui figurera certainement dans les manuels du futur...

En second lieu, le texte de M. Viprey confirme cette conclusion.

Avant de le lire, il faut rappeler que les pièces ne se succèdent pas régulièrement au cours du temps :

— de 8 (*l'Illusion comique*) à 9 (le *Menteur*), il s'écoule : 6 ans ;

— de 10 (la *Suite du Menteur*) à 13 (*l'Etourdi*) : 16 ans.

Enfin et surtout, il s'écoule plus de temps entre les *Menteurs* (n°10) et *l'Etourdi* (n°13) qu'entre ce dernier et les *Femmes Savantes* (n°23), c'est-à-dire entre les première et dernière œuvres en vers publiées sous le nom de Molière.

On attendrait donc logiquement un grand "trou" sur le premier axe entre les points 9-10 et 13. Si les auteurs étaient bien différents, ce "trou" devrait même être plus grand que l'espace séparant les points 13 et 23...

M. Viprey signale d'ailleurs la position "tout à fait remarquable des deux *Menteurs*" et il écrit que ces deux pièces sont :

"L'ultime tentative de Corneille dans le genre (comédie en vers), fortement *marquée* par l'influence désormais prédominante de Molière".

Vous avez bien lu : les *Menteurs* sont une **"tentative marquée par l'influence prédominante de Molière"**...

Statistiquement, cette bourde semble logique puisque, sur le premier axe du graphe, les points 9 et 10 (les *Menteurs*) sont nettement plus proches de 16 (*Dom Garcie*) et même de 13 (*l'Etourdi*) que de 6 et 8. Il n'y avait donc qu'un moyen de sauver Molière : affirmer que Corneille a été attiré dans son orbite ! Encore faut-il :

— limiter la lecture au premier axe (en "oubliant" le second) ;

— postuler que Molière est l'auteur des pièces 13 à 23 (en "oubliant" la n°14...) ;

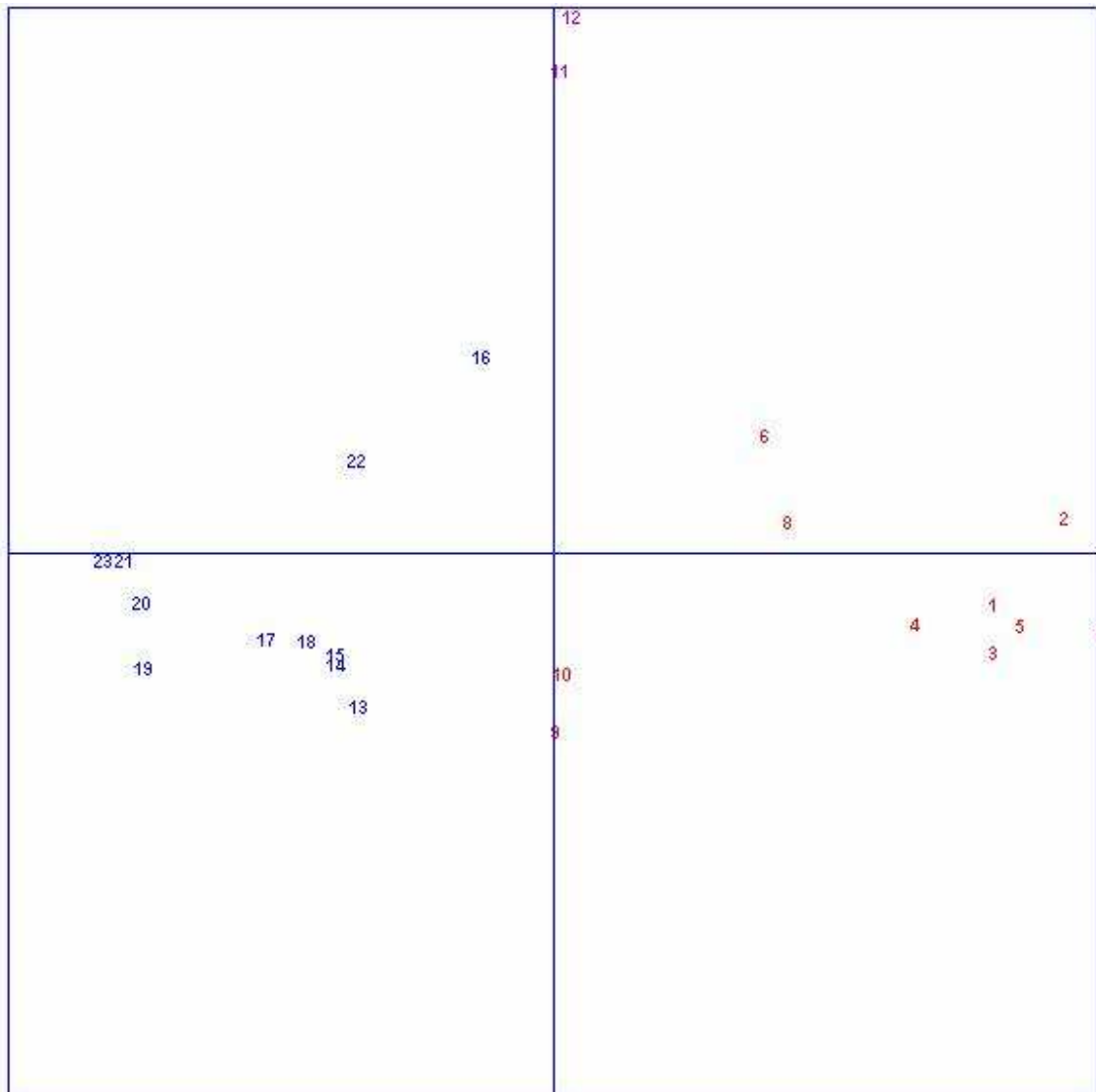
— considérer qu'avant même de monter sur les planches et d'écrire une seule ligne, Molière exerçait déjà une influence "prédominante" sur Corneille. En effet, Molière a fondé l'illustre théâtre en juin 1643 (un an après la création du *Menteur*) ; il a connu son premier grand succès (avec les *Précieuses Ridicules*) 17 ans plus tard...

Merci ! Voilà des arguments imparables... en faveur de Corneille.

Pour conclure, je suis flatté que la discussion se centre sur la distance intertextuelle puisque c'est un peu notre création. A ce propos, une note sur les propriétés de cette distance doit paraître à l'automne prochain (le manuscrit peut être communiqué sur demande). Lorsque seront instaurées les conditions d'un débat entre scientifiques, les nombreuses personnes qui ont participé à la mise au point de ce calcul témoigneront de son sérieux, de la solidité et de la fiabilité de notre indice.

En ce qui concerne Corneille et Molière — avant que M. Viprey nous apporte son concours — notre démonstration se fondait déjà sur beaucoup d'autres outils que la distance : la classification hiérarchique, l'analyse arborée, les groupes verbaux les plus fréquents, le sens des mots usuels, sans compter un certain nombre de faits historiques. C'est cet ensemble d'analyses — présentées au grand public dans notre essai *Corneille dans l'ombre de Molière* — qui permet de conclure que Corneille est l'auteur de 16 pièces de Molière dont les principaux chefs d'oeuvre (*Misanthrope*, *Tartuffe*, *Dom Juan*, *l'Avare*...) A tout cela, les "moliéristes" sont incapables d'apporter la moindre réfutation. Au contraire, leurs manœuvres, leurs gesticulations et leurs bourdes ne soulignent-elles pas leur désarroi et la faiblesse de leurs arguments ?

Analyse factorielle des correspondances sur les comédies en vers de Corneille et de Molière
 Premier plan factoriel (extrait de : <http://laseldi.univ-fcomte.fr/morneille0.htm>)
 Voir page suivante : table de correspondance des numéros des pièces et chronologie



NB : Le numéro 14 doit figurer en rouge puisque l'édition originale de cette pièce (le Dépit Amoureux) désigne Corneille comme l'auteur.

Tableau chronologique des pièces traitées par M. Viprey

1	Mélite	Corneille	1630
2	Clitandre	Corneille	1631
3	La Veuve	Corneille	1631
4	La Galerie du Palais	Corneille	1632
5	La Suivante	Corneille	1633
6	La Comédie des Tuileries	Corneille	1634
7	La Place Royale	Corneille	1634
8	L'Illusion Comique	Corneille	1636
9	Le menteur	Corneille	1642
10	La Suite du menteur	Corneille	1643
11	Psyché ¹ *	Corneille	1671
12	Psyché ² *	et/ou Molière	1671
13	L'Etourdi	Molière	1658
14	Le Dépit Amoureux**	Corneille	1658
15	Sganarelle ou le Cocu Imaginaire	Molière	1660
16	Dom Garcie de Navarre	Molière	1661
17	L'Ecole des maris	Molière	1661
18	Les Fâcheux	Molière	1661
19	L'Ecole des Femmes	Molière	1662
20	Le Tartuffe	Molière	1664
21	Le Misanthrope	Molière	1666
22	Mélicerte	Molière	1666
23	Les Femmes Savantes	Molière	1672

* La note de M. Viprey ne permet pas de savoir s'il s'agit de la pièce entière, de la partie attribuée à Corneille ou de celle attribuée à Molière...

** Cette pièce doit figurer en rouge puisque son édition originale désigne Corneille comme l'auteur.