Cyril LABBE
Université Grenoble I
cyril.labbe@imag.fr

Dominique LABBE
Institut d'Etudes Politiques de Grenoble
dominique.labbe@iep.upmf-grenoble.fr

Corneille a écrit 16 pièces représentées sous le nom de Molière

Réponses à :

About Labbé's "Inter-textual Distance"
Journal of Quantitative Linguistics
13(2-3):265-283, 2006

Version préliminaire
Grenoble
Avril 2007

(en cours de traduction)

Dans ces affaires-là, il y a toujours des jalousies et des rivalités qui font perdre la mesure aux gens.

Guy de Maupassant, *Madame Baptiste* (1882).

N'étant pas moi-même statisticien mais utilisateur expert, je soumets cette question aux statisticiens.

Jean-Marie Viprey (Louvain-la-Neuve, 11 mars 2004)

Jean-Marie Viprey & Claude-Nicolas Ledoux
About Labbé's "Inter-textual Distance"
Journal of Quantitative Linguistics. 13(2-3):265-283, 2006

Résumé

Dans sa livraison de 2001 (vol 8-3, p. 213-231), MM. Dominique et Cyril Labbé ont publié un article sous le titre* "Inter-Textual Distance and Authorship Attribution. Corneille and Molière". *Dominique et Cyril Labbé (ci-après désignés par DCL) proposent une nouvelle formule pour le calcul de la dissimilarité entre textes, ainsi qu'une échelle des distances. Ils se proposent de les appliquer à la question de l'attribution d'auteur, et spécialement à un cas particulier : la controverse sur l'attribution de plusieurs pièces signées par Molière. L'objet de ce papier est de discuter leur raisonnement et leurs conclusions à la lumière de plusieurs expériences simples. Bien que nous citions largement l'article de DCL, une certaine connaissance préalable de son contenu est recommandé. Le lecteur sera aussi renvoyé à un livre écrit par Dominique Labbé (DL ci-après) en 2003 (Labbé 2003). Dans ce livre, DL recapitule pour un public plus large la plupart du contenu de cet article.

Abstract

In the 2001, Volume 8, Number 3, issue of the Journal of Quantitative Linguistics (pp. 213 – 231) M. M. Dominique and Cyril Labbé published a paper entitled "Inter-Textual Distance and Authorship Attribution. Corneille and Molière". Dominique and Cyril Labbé (hereafter referred to as DCL) propose a new formula for the computation of dissimilarity between texts, as well as a distances scale. They intend to apply it in the field of authorship attribution, and especially in one particular case: a controversy about the authorship of several plays signed by Molière. The object of this paper is to discuss their rationale and conclusions in the light of several simple experiments. Though we will quote DCL's paper throughout, some previous knowledge of its content is recommended. The reader will be also referred to a book written by Dominique Labbé (hereafter DL) in 2003 (Labbé, 2003). In that book, DL recapitulates for a larger audience most of the contents of the paper.

Cyril Labbé & Dominique Labbé

Corneille a écrit 16 pièces représentées sous le nom de Molière

Résumé

Un article paru en décembre 2006 dans le Journal of Quantitative Linguistics, remet en cause l'attribution à P. Corneille de toutes les pièces en vers de Molière et de deux de ses pièces en prose (Dom Juan et l'Avare).

Cet article présente des "expériences" qui ne concernent pas Corneille et Molière. Les calculs comportent de nombreuses erreurs. Par exemple, les ponctuations sont comptés comme des mots. Les exemples choisis se situent en dehors des limites de validité de l'échelle des distances. Les graphiques sont manifestement erronés.

La distance intertextuelle sort renforcée et la dernière de ces "expériences" confirme que P. Corneille a bien écrit les pièces en vers de Molière ainsi qu'au moins deux de ses pièces en prose.

Abstract

In its December 2006 issue, the Journal of Quantitative Linguistics published a paper who rejects our attribution to P. Corneille of the authorship of all the verse plays by Molière and two of his prose plays (Dom Juan and l'Avare).

This paper presents some so-called experiments which do not concern the Corneille-Molière case. Calculus are false. For example, ponctuations are counted as "words". The examples have been chosen far outside the validity limits given to the distance scale. Graphs are false or mistaken.

However, this confrontation leads to a reinforcement of the Intertextual Distance. It also gives a confirmation : P. Corneille did wrote the verse plays by Molière and two of his prose plays (Dom Juan and l'Avare).

Sommaire

Avant propos

Dans son numéro de décembre 2001, le Journal of Quantitative Linguistics (Vol. 8, n°3, p. 213-231), a publié notre article :

"Inter-Textual Distance and Authorship Attribution. Corneille and Molière"[1].

Dans lequel il est démontré que P. Corneille a écrit toutes les pièces en vers ainsi que le Dom Juan et l'Avare, pièces représentées sous le nom de Molière[2]. Cet article (ci-après *JQL01*) a été complété par plusieurs publications[3].

Dans son numéro de décembre 2006 (Vol. 13, n° 2-3, p. 265-283), le même Journal of Quantitative Linguistics a publié un papier (ci-après *JQL 06*), intitulé :

"About Labbé's “Intertextual Distance”"

Sous les signatures de MM. Jean-Marie Viprey et Claude-Nicolas Ledoux[4].

JQL06 n’a pas été communiqué avant sa publication aux deux chercheurs attaqués. D’une part, puisque leur travail est critiqué, leur sérieux et leur intégrité mis en doute, il aurait été normal de leur donner la possibilité de se défendre. D’autre part, l’auteur de *JQL06* a utilisé des fichiers, des données et un programme qui ont été mis gracieusement à sa disposition. Dans ce cas, il est d’usage de permette aux premiers auteurs de contrôler qu’il n’y a pas mésusage de ce don (d’ailleurs, les remerciements manquent à la fin de *JQL 06.*…).

Du seul fait de l’auteur de *JQL 06*, les voies traditionnelles du débat scientifique n’ont pas été respectées[5].

---

[1] Version française : htpp://halshs.archives-ouvertes.fr/halshs-00137675.
Version anglaise : htpp://halshs.archives-ouvertes.fr/halshs-

[2] La liste de ces pièces se trouve en annexe 1, p. 122.

[3] Notamment : Labbé & Labbé 2003 ; Labbé 2004a, 2004b, 2004c, 2007. Voir les références bibliographiques p. 119.

[4] Claude-Nicolas Ledoux : architecte français né en 1736 et mort en 1806.

Cela nous contraint à rendre public l'intégralité[6] de ce document avec notre réponse.

Dans la suite de ce dossier, l'article *JQL 06* est en italique et sur fond gris, nos réponses en caractères droits et sur fond blanc. La pagination originelle de *JQL 06* est remplacée par une numérotation des paragraphes : par exemple, [1.1] renvoie à la section 1, §1, etc. Naturellement, les personnes désireuses de citer les 2 articles du Journal of Quantitative Linguistics (2001 et 2006) sont priées de se reporter aux originaux.

Le débat principal porte sur la paternité de toutes les pièces en vers ainsi que du Dom Juan et de l'Avare qui sont représentées sous le nom de Molière (annexe 1, p 123 de ce dossier).

Ces pièces sont attribuées à P. Corneille grâce à un grand nombre d'indices concordants, notamment les distances extrêmement faibles séparant certaines pièces de Corneille et de Molière.

Ce dernier point est contesté par *JQL 06*.

Afin de clore cette discussion, l'annexe 4 (p. 131 de ce dossier) publie deux des nombreux tests effectués dans le cadre de la préparation notre article *JQL01*. Ces tests concernent la proximité remarquable des deux Menteurs (Corneille) avec certaines pièces représentées sous le nom de Molière. Ils démentent les affirmations de *JQL06*.

Le lecteur qui s'intéresse simplement au rôle de Corneille dans l'écriture des pièces de Molière peut donc se contenter de lire, outre l'annexe 4, notre synthèse (p. 19-20 de ce dossier), le § 8.6 de *JQL 06* (p. 78-79 de ce dossier), et les réponses à ce § (p. 82-83 de ce dossier).

Cependant, une lecture complète du dossier n'est pas inutile pour comprendre nos méthodes, vérifier leur solidité et l'inanité des multiples attaques lancées contre elles depuis plusieurs années[7]…

---

[5] Dès que nous avons eu connaissance de cet article, nous avons écrit au rédacteur en chef (annexe 12) qui a accepté immédiatement la publication et la mise en ligne d'un avertissement (annexe 13). Trois mois plus tard, la revue est parue sans cet avertissement et celui-ci n'a toujours pas été placé sur le site web. Le rédacteur en chef nous a envoyé une lettre d'excuses nous indiquant que le problème vient de la maison d'édition. Enfin, aucune des questions posées dans notre courrier de décembre 2006 (annexe 12) n'a reçu de réponse (au 23 avril 2007).

[6] La totalité de l'article conteste notre travail, il est donc impossible d'en citer seulement des extraits.

[7] Pour ne pas encombrer ce dossier avec des discussions secondaires, les réponses aux nombreuses attaques personnelles contenues dans *JQL 06* sont reportées en annexe 9, p. 153 sq de ce dossier.

Ce dossier se veut volontairement pédagogique. Au vu des nombreuses erreurs commises dans *JQL 06*, il a paru indispensable d'accompagner chaque notion de sa définition et des explications qui permettront à un lecteur, même non initié, de suivre l'essentiel des raisonnements[8].

Ce dossier suit le plan de *JQL 06*.

Après une introduction visant à préciser les termes du débat, il comporte deux parties principales.

La première partie discute les propriétés de la distance intertextuelle et son application à l'attribution d'auteur. Les différentes "expériences" présentées par *JQL 06* abordent les points suivants :

- la symétrie de l'indice de la distance et l'influence, sur cet indice, de la longueur des textes étudiés (chapitres 2 et 5) ;

- la robustesse de l'indice (chapitre 3) ;

- l'influence des normes de dépouillement (chapitre 4).

La seconde partie revient sur le cas Corneille-Molière. *JQL 06* présente des conclusions (chapitre 6) puis une expérience qui confirment la paternité de Corneille sur 16 pièces de Molière (chapitres 7 et 8).

[8] Le lecteur intéressé pourra compléter son information avec les ouvrages suivants : Harris & Stocker (1998) et Dodge (1993). Pour les tables, notamment celle des limites d'acceptation du coefficient de corrélation linéaire, et pour le calcul de ce coefficient, on peut se reporter à l'Aide mémoire du Cisia-Ceresta ou à n'importe quel ouvrage équivalent.

Partie I.
Voyages à la découverte de la distance intertextuelle

"Il faut réfléchir pour mesurer et non mesurer pour réfléchir"
(Gaston Bachelard. La formation de l'esprit scientifique)

Les trois premières sections de *JQL 06* - reproduites dans le premier chapitre de ce dossier - prétendent poser un autre cadre théorique, présenter différemment le calcul de la distance intertextuelle puis l'échelle de la distance.

Les trois chapitres suivants sont une discussion de trois propriétés de l'indice de la distance intertextuelle : symétrie, relation avec la longueur des textes, robustesse.

Dans tout cela, il est apparemment fort peu question de Corneille et Molière.

Comme Gulliver, le lecteur visitera successivement Lilliput, le pays des nains, Lagado, la ville gouvernée par de pseudo-savants, puis Brobdingnag, le pays des géants et, enfin, il fera une seconde visite à Lilliput…

Ce voyage sera utile puisqu'il prouvera, avec les "expériences" de *JQL 06*, que la distance intertextuelle n'a aucun des défauts qu'on lui prête mais, au contraire, beaucoup de vertus qu'on ne peut cacher.

Chapitre I. Introductions

1. Conceptual frameworks; text theory

[1.1] *DCL refer to three papers dealing with authorship attribution, as well as the works of four researchers in the field of lexical statistics. However, they do not refer to any particular theory, either of language, or of texts. This leads them to invoke several concepts, such as the concept of* actual distance *between texts, without providing the reader with an actual definition. Most researchers whose object of investigation is text will consider the term* inter-textual distance *itself to be inadequate to name their method, which consists of a single measurement, valued by a single scale. More modestly, Muller (1992a, b) talks of* lexical connexion. *This latter term makes it clear that massive vocabulary is just one component of textuality among others.*

[1.2] *Adam (1999), referring to Harris (1969) and overall to Bakhtine (1977) together with several works on discourse analysis, defines text as a combination of a* structure *and a* texture[9] *(for us:* microstructure). *The* vocabulary *of a text can no more be reduced to a lit of items, even if that list includes information on their frequency. It basically consists of a rhythm of occurrence* (macrodistribution) *which corresponds to thematic constitution and progression (variety, breaks, increasing), and of a network of* collocations.

[1.3] *Here is the first and most basic objection we put forward to DCL: by no means can comparison between lexical inventories of two texts be sufficient to draw conclusions about their kinship. Only as the end point of meticulous, interdisciplinary work, utilising a variety of approaches, including some statistical measurements, could we contemplate conclusions of general import.*

[1.4] *Furthermore, DCL use the terms* genre *and* theme *without referring to any literary and/or linguistic theory. The same can be said about the concept of* author *and*

[9] *We prefer the contrasting terms* macrostructure *and* microstructure (Viprey, 1997, 2002).

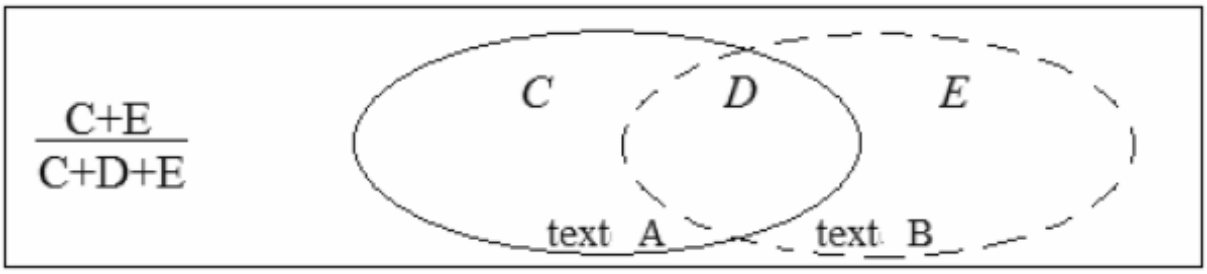
generally about their way of dealing with literary history, which is also deserving of criticism. Nevertheless, we will move on to a critical analysis of DCL's proposition.

2. DCL's formula

[2.1] *DCL propose a* distance *calculation* D(a,b) (a *referring to text* A, b *to text* B). *This is the quotient of the symmetrical difference between the vocabularies of the two texts, by a number describing the size of the global vocabulary. In set notation:*

$$\frac{VF_a \cdot \Delta : VF_b}{VF_a \cdot \cap VF_b.}, \quad where\ \mathrm{VF} = vocabularyfrequency.$$

Or, as presented in the figure and notations of DCL:



[2.2] *This* distance *is positive (it takes the value 0 when both texts share exactly the same lexical material, in the same proportions) and less than 1 (the value 1 means that they do not share any lexical item).*

[2.3] *If, exceptionally, both texts have the same length (N, overall number of occurrences), the* absolute distance *is the sum of the absolute values of differences between occurrences of items in* Va *and* Vb. *The* relative distance *is then the quotient of the* absolute distance *by the quantity* (Na + Nb) - Na∪b.

[2.4] *In the general case (where the texts have different lengths), DCL proceed in four steps:*

1. They modify the actual occurrence of the items of the longer text (B), by the coefficient U(a,b) = Na/Nb. *They thus obtain values that they denote (for each item* i) Eia(u). *At that stage, the sum of all these values is equal to* Na *(by construction).*

2. *For each item of A, they calculate* |Fia - Eia(u)|. *The sum of the absolute values of these differences falls into the calculation of the* absolute difference.

3. *The items of B which are missing in A (B\A) are taken into account only if* Eia(u) ≥ 1. *In that case, Eia(u) is added to the* absolute difference *(since Fia = 0,* |Fia-Eia(u)| = Eia(u)) *and also to* N'b [2]. If Eia(u) < 1, Eia(u) *is not counted into* N'b.

4. The intertextual distance (further: DI) is then the quotient of the absolute distance by Na + N'b.

[2.5] *Furthermore (2001, p. 218) DCL add a precision: they exclude from the summing of the numerator (in step 1) any individual absolute differences smaller than 0.5. They do not mention whether* Fia *must be then deducted from* Na.

[2.6] *The authors do, however, add several restrictions. They suggest that it is invalid to apply the formula to texts smaller than 1000 tokens, as well as to pairs of texts for which* Nb/Na > 10 *(i.e. pairs where B is more than 10 times longer than A). They also suggest that texts must be* normalized *and all tokens lemmatized, i.e., attached to their dictionary entries. This point will be discussed further below.*

---

2. represents the subset E of DCL's scheme, but by construction it is smaller than it. It will be used to establish the denominator of the final value.

3. DCL's "standardized scale"

[3.1] *The value thus obtained must then be interpreted. DCL therefore present what they coin the* Inter-textual Distance Standardized Scale, *which is reproduced here as Table 1. They claim (2001, p. 218) that this scale has been established via tests on several corpora, representing about 10 million tokens, from various genres and periods, including several novels from the last three centuries. On the other hand, in Labbé (2003, p. 14), DL reports that these tests have examined several thousand texts*

[3.2] *If "several" of these texts actually are novels, i.e., texts of at least 15,000 tokens, the average length of the* other *texts that have been tested must be about 2000 or 2500 tokens (whose number cannot be greater than 50, which already demands that we lower the average length of the other texts to 1000, the lower limit below which DCL refrain from applying their formula). So these must be very short texts, very specific. We must also suppose that the novels*[3] *have been only compared with other novels in the same genre.*

[3.3] *Neither theatre plays, nor poetry, essays, etc. are mentioned among the texts submitted to testing. This being the case, it seems difficult to consider DCL's "scale" as a scientific process, nor even an empirical one in the usual meaning of that term.*

[3.4] *Two major subtleties introduced by DCL's subsequent commentary must also be noted. Where the scheme indicates* same author, *the commentary asserts that "distances smaller than 0.20* usually[4] *do not exist between two different authors". And when the scheme states "sure authorship attribution", the commentary adds the adverb* quite *("quite sure"), which is equivocal in English and may be equivalent to either* entirely *or* somewhat). *But what DL keeps from this in all subsequent publications is what the scheme specifically says. In Labbé (2003, p. 14) he writes: "*Une distance inférieure ou égale à 0.20 désigne avec certitude un auteur unique. Même quand un écrivain en 'pastiche' un autre, la distance entre le pastiche et les originaux est toujours supérieure à ce seuil." *(A distance lower than or equal to 0.20 indicates with certainty a single author. Even when a writer makes a "pastiche" of another, the distance […] is always greater than that threshold.) We will show below (Section 6) that this is wrong.*

[3] *It is necessary to make so many suppositions* precisely because *DCL have not published any more precise details than those they give in their paper (and in Labbé, 2003).*

[4] *Our emphasis*

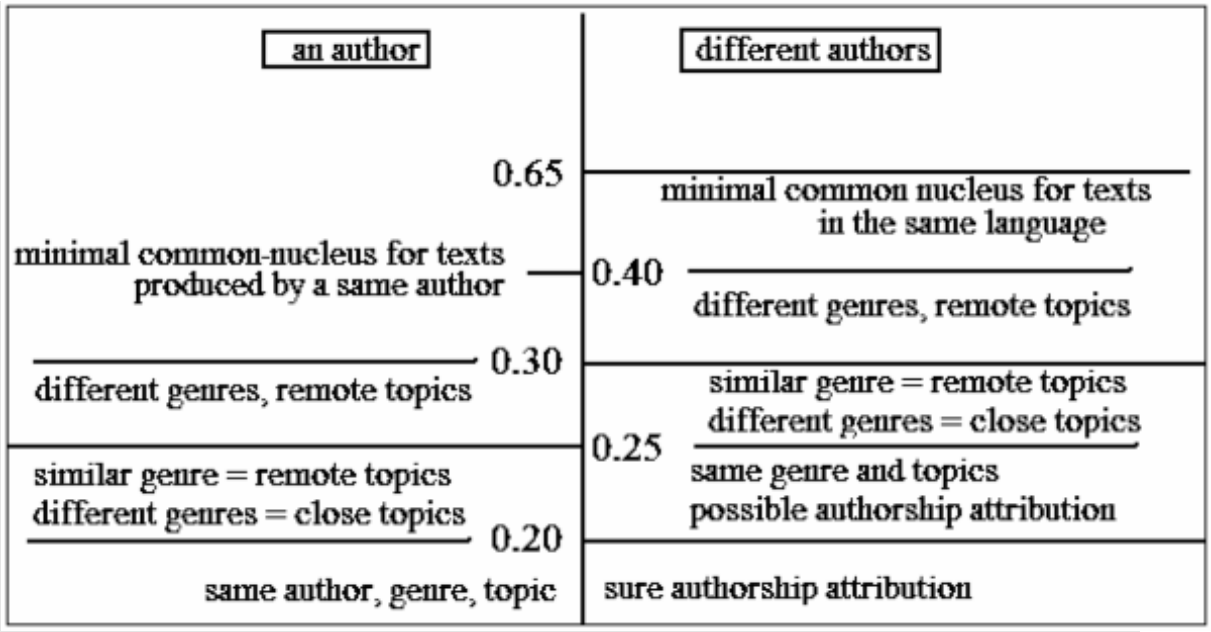


Fig. 1. Reproduction of the inter-textual standardised scale (Labbé, 2001).

[3.5] DCL claim (2001, p. 219) that "for the same author we always notice distances smaller than those existing between two different and contemporary authors (when they are dealing with the same topic)". This statement is again made unverifiable by the parenthesis. It is indeed thus far impossible to know when, and to what extent, two texts deal with the same topic, unless *we examine their lexical kinship. Similarly, in the following paragraph, DCL explain the problem of two texts, of known different authors, possibly having a ID inferior to 0.20. They write that "one of them was 'inspired' by the other". But what is the possible measurement of this* inspiration*? Which author is not* inspired *by another (and in general by many others)? Is this not a means of making the scale unfalsifiable, even in the face of very basic experiments such as we demonstrate below?*

[3.6] Furthermore, is it acceptable, from the scientific *point of view, to write on one side (at the right of the scheme) "sure authorship attribution" and, on the other side, at the same level, to nuance "same author" by the clause we have just quoted? Why should the inspiration of one author by another not perturb* attribution *certainty as well?*

[3.7] *To demonstrate the ultimate consequences of DCL's statements, we must again quote Labbé (2003, p. 15):* “Pour trouver l'auteur d'un texte douteux ou anonyme, il n'est pas nécessaire de rechercher tous les écrivains susceptibles de l'avoir écrit, il suffit d'en trouver un pour lequel la distance, entre une partie de son oeuvre et le texte analysé, sera inférieure aux seuils indiqués ci-dessus” *(To find the author of an uncertain or anonymous text, it is not necessary to search all the writers suspected of having written it; it is enough to find one writer for whom the distance between a part of his work and the analysed text is lower than the thresholds indicated above). Should not scientific cautiousness demand, on the contrary, that we foresee the case in which a third candidate might occur? By which empirical means did DL make sure that a single text could not be attributed, by his method, to two or more authors? What solution does he propose in that case, whose probability is by no means nil?*

[3.8] *More generally, a validation scale such as that proposed by DCL requires two properties which are obviously lacking here: it must be non-discrete (setting rigid and regularly spaced thresholds such that 0.20, 0.25, 0.30 is arbitrary), and formulated in probabilistic terms.*

Sur quoi porte le débat ?

JQL 06 soulève deux débats distincts :

- les propriétés de la distance intertextuelle et de l'échelle des distances ;

- qui a écrit les pièces en vers ainsi que le Dom Juan et l'Avare qui sont joués sous le nom de Molière?

L'introduction sera consacrée aux éléments indispensables pour la compréhension de ces deux questions : outils, principales définitions, calcul de la distance intertextuelle et propriétés de cet indice.

Attribution d'auteur[10]

L'attribution - à Corneille de toutes les pièces en vers de Molière, du Dom Juan et de l'Avare - repose sur la convergence d'une multitude d'indices : très grande parenté des textes, sens des mots les plus fréquents, styles et faits historiques[11]. Aucun de ces indices n'a été contredit jusqu'à maintenant.

Rappelons les principaux indices statistiques.

Notre article *JQL 01* présente les distances très faibles séparant certaines œuvres de Corneille et de Molière[12]. Une telle convergence ne se rencontre pas chez deux auteurs différents. Cela permet de conclure que l'auteur des deux Menteurs (Corneille) a aussi écrit : l'Etourdi, le Dépit amoureux, l'Ecole des maris, les Fâcheux, l'Ecole des femmes, la Princesse d'Elide, le Tartuffe, Dom Juan, le Misanthrope, Mélicerte, l'Avare et les Femmes savantes représentés jusqu'à nos jours sous le nom de Molière. La proximité de Dom Garcie de Navarre et de Psyché avec toutes les dernières pièces de Corneille indique également un auteur unique (annexe 3.2, p. 126 de ce dossier).

Deux algorithmes de classification (automatique et arborée) confirment ces conclusions en rattachant les deux Menteurs (Corneille) à l'ensemble des pièces en vers

[10] Sur la méthode d'attribution d'auteur, l'ouvrage de H. Love (2002) - paru après notre propre travail - présente un tableau d'ensemble qui correspond à la démarche que nous avons suivie dans le cas de Corneille et Molière. Voir : Labbé 2004b et Labbé 2004c.

[11] Notre réponse à la section 8 et l'annexe 10 rappellent quelques-uns des indices historiques qui plaident en faveur d'un tel examen.

[12] Le calcul de la distance est présenté en annexe 2 (p. 124 de ce dossier) et les principales distances caractéristiques en annexe 3 (p. 127 de ce dossier).

de Molière ; Psyché et Dom Garcie (Molière) étant rattachées aux pièces de Corneille dite de la "maturité"[13].

Il a été démontré, à Louvain-la-Neuve le 11 mars 2004, que les contextes des mots les plus fréquents dans les pièces de Molière et Corneille indiquent un auteur unique[14].

A de nombreuses occasions, il a été signalé que les combinaisons "verbe + verbe" désignent un même auteur[15]. Ainsi, "faire voir", "pouvoir être" et "pouvoir faire" sont les trois combinaisons les plus fréquentes (dans le même ordre), avec des densités très proches, chez Corneille et chez Molière. Ceci ne se rencontre pas chez deux auteurs différents. En revanche, c'est très fréquent chez un même auteur publiant sous son propre nom et sous le nom d'un autre[16].

Depuis décembre 2001 aucun de ces éléments n'a été remis en cause.

Rappelons enfin que, au début du XXe siècle, le poète P. Louÿs a souligné la forte parenté existant entre les deux œuvres (H. Wouters et de C. Ville de Goyer, dans leur ouvrage de 1990, donnent une synthèse des trouvailles de P. Louÿs). P. Louÿs a été beaucoup moqué mais n'a jamais été démenti.

C'est donc faussement que le § 1.1 de *JQL 06* (p. 13 ci-dessus) affirme que cette attribution repose sur une "seule mesure", mesure qu'il expose d'ailleurs de manière erronée.

Corpus, mots, vocables, lemmes…

En avril 2003, il a été remis à l'auteur de *JQL 06* le "**corpus Corneille-Molière**", c'est-à-dire 68 fichiers électroniques pour 66 pièces, l'une d'elles, Psyché, étant coupée en 3 selon les auteurs (voir annexe 1, p. 123 de ce dossier).

Ce corpus a les caractéristiques suivantes :

- **longueur** totale : 918 153 **mots** ("**tokens**" en anglais). Pour les longueurs en mots (notées N) des pièces, voir annexe 1. Chacun de ces **textes** est formé de la succession

[13] JQL 06 présente une troisième analyse (dite "des correspondances") qui aboutit exactement aux mêmes résultats (chapitre 7, p. 87sq de ce dossier).

[14] Labbé 2004a. Cette démonstration a été effectuée devant plus d'une centaine de chercheurs dont l'auteur de *JQL 06*.

[15] Voir en annexe 3.4 la liste des dix premières combinaisons (p. 129 de ce dossier).

[16] L'annexe 8 donne l'exemple de R. Gary et E. Ajar (p. 145 de ce dossier). Le même phénomène a été observé chez deux hommes politiques utilisant la même "plume de l'ombre" (Monière & Labbé 2006).

des N mots le constituant, c'est-à-dire ce qui est prononcé en scène par les acteurs. Toutes les notations marginales – comme les didascalies – sont neutralisées ;

- 28 982 **formes graphiques brutes** différentes ("current spelling"), telles qu'elles apparaissent dans le texte). Ces "formes" sont découpées selon les méthodes de segmentation automatique traditionnelles en analyse des données textuelles[17]. Par exemple, cette méthode aboutit à considérer que les "formes" suivantes sont toutes différentes : L', l', LE, Le, le, LA, La, la… De même, "aujourd'" et "hui" sont deux "formes", etc. En revanche, le pronom (la) n'est pas différencié de l'article (la) et de la note de musique (la)…

- 23 633 **formes graphiques** normalisées (on peut aussi employer le calque "**graphies standards**" pour "standard spelling"). Dans le cas précis, cette normalisation des graphies consiste essentiellement à ramener en minuscule la majuscule des mots communs placés en début de vers ou de phrase ("La" et "la" ne sont plus deux mots différents) et à rendre leur intégrité à des mots comme "aujourd'hui".

NB : la graphie originale du mot n'est pas modifiée, chaque mot reçoit une **étiquette** (en anglais "tag") comportant cette graphie normalisée.

- 9 995 **vocables** différents (en anglais "type") obtenus en rattachant chaque mot à son entrée de dictionnaire (par exemple "Le, art." ou "La, nom masculin")[18]. Cette entrée comporte le "**lemme**" ou "mot vedette" et la **catégorie grammaticale**. Ces indications sont ajoutées, après la graphie normalisée, dans l'étiquette attachée au mot. Par exemple, "L'" reçoit l'une des 2 étiquettes suivantes "l',le,pronom" ou "l',le,article".

La liste de ces vocables constitue le **vocabulaire** d'un texte ou d'un corpus.

Tous les calculs statistiques – notamment celui de la distance intertextuelle – portent sur les vocables qui sont déterminés en suivant toujours exactement les mêmes conventions. En se reportant aux chiffres cités ci-dessus, il est facile de comprendre que les résultats pourront être assez différents selon que l'on utilise les formes graphiques (brutes ou normalisées) ou les vocables.

[17] Voir par exemple, le début de l'ouvrage de Lebart, Salem & Berry (1998).
[18] Labbé 1990 et Labbé 2002b (à propos du corpus "Théâtre classique").

Connexion lexicale ou distance intertextuelle ?

La **connexion lexicale** mesure la différence entre deux vocabulaires sans tenir compte des fréquences[19]. Autrement dit, elle porte sur la présence ou l'absence, dans telle ou telle pièce, de chacun des 9 995 vocables constitutifs du vocabulaire du corpus Corneille-Molière. Peu importe que ce vocable soit présent 37 805 fois (comme le pronom "je") ou une seule fois (comme le substantif "zone"), il pèse du même poids dans ce calcul qui est fondé sur l'alternative "présence/absence" (0 ou 1).

La **distance intertextuelle** mesure le nombre de mots différents entre deux textes. Par exemple, un indice de 0.25 signifie que 25% des mots sont différents entre les deux textes comparés (ou encore qu'ils en partagent les trois quarts).

Autrement dit, la distance intertextuelle porte sur les 918 153 **mots** constitutifs du **texte** des pièces de Corneille et Molière et elle donne à chaque mot un poids équivalent à celui qu'il pèse dans chacune des pièces. Ainsi la distance intertextuelle prend en compte les différences considérables, d'une pièce à l'autre, de la densité d'utilisation de la première personne. A l'inverse, dans la connexion lexicale, cette première personne pèse du même poids dans tous les calculs puisqu'il y a au moins un pronom "je" dans chaque pièce.

Il est donc absurde de confondre – comme le fait *JQL 06* dès le § 1.1[20] - la connexion lexicale et la distance intertextuelle et, plus encore, de promouvoir la première contre la seconde. Est-ce parce la "connexion lexicale" est incapable de distinguer les auteurs (Hubert & Labbé 1998), contrairement à la distance intertextuelle qui le fait très bien ?

Indice et échelle de la distance intertextuelle

Les formules et schéma présentés par la section 2 de *JQL 06* sont erronés[21]. La présentation de la section 3 est tronquée.

[19] Brunet & Muller 1988 et Hubert & Labbé 1998. Labbé 2003 (p. 50-51) expose les intérêts de cette "connexion lexicale".

[20] Cette confusion se retrouve dans le § 1.3, puis du début de la section 4 à la fin de l'article, notamment dans la discussion des prétendus "biais" affectant la distance intertextuelle.

[21] Voir annexe 2 (p. 124 de ce dossier). Les principales erreurs de *JQL 06* sont listées à la fin de cette annexe, du moins quand elles peuvent avoir des conséquences sur ses calculs.

L'indice de la distance intertextuelle mesure l'influence de 4 facteurs : le genre, l'auteur, le thème et l'époque. Pour attribuer un texte d'origine inconnue ou douteuse à un auteur connu, il faut contrôler l'influence des 3 autres facteurs (genre, thème et époque). L'échelle de la distance intertextuelle (reproduite p. 17 de ce dossier) aide ces opérations.

Comme indiqué dans notre article *JQL01*, l'échelle de la distance intertextuelle s'applique à des textes présentant les caractéristiques suivantes.

- Les graphies ont été normalisées et chaque mot a été lemmatisé. Le calcul de la distance se fait sur les vocables (lemme et catégorie grammaticale de chaque mot). Il ne peut s'appliquer aux formes graphiques brutes comme le fait *JQL06* (§ 4.7-4.8, p. 46 ; § 6.1-6.3, p. 61-62). Dans ce cas, une autre échelle aurait dû être étalonnée, ce que ne fait pas *JQL06*.

- Les dates de rédaction sont aussi contemporaines que possible afin de neutraliser le facteur "temps". Le chapitre 6 (p. 83-85 de ce dossier) revient sur cette question à propos du corpus Corneille-Molière et l'annexe 3.4 à propos de Corneille et Racine.

- Les longueurs sont comprises entre 3.500 et 20.000 mots. Cette plage de validité est clairement indiquée dans notre article *JQL01*[22]. Ces dimensions sont très usuelles : entretiens sociologiques, discours politiques (une heure de discours public comporte entre 7 000 et 9 000 mots), articles de revue, recueils de poèmes, pièces de théâtre, nouvelles... Quant aux romans fleuves, on a utilisé des extraits.

Sous ces conditions, une distance inférieure ou égale à 0.20 (moins du cinquième des mots différents), signale avec certitude un auteur unique donc, en cas de deux signatures différentes, une collaboration ou un plagiat[23]. Une distance comprise entre 0.20 et 0.25 peut se produire exceptionnellement entre deux auteurs contemporains, travaillant dans un même genre, sur des thèmes proches. Si cet événement se répète, il y a une plume unique ou un plagiat. Aucun des résultats présentés par *JQL06* ne remet en cause ces conclusions, bien au contraire.

[22] Elle a été rappelée dans les communications d'Orsay (Labbé 2003c), de Dublin (2003b) et à Louvain-la-Neuve (2004a), en présence de l'auteur de *JQL 06* (voir annexe 9, p. 156 de ce dossier).
[23] Dans notre article *JQL01*, le mot "inspiration" est placée entre guillemets, car c'est toujours l'excuse des plagiaires et l'ultime argument en faveur de Molière…

Propriétés de l'indice de la distance intertextuelle

Comme expliqué dès la première publication,

> L'indice de la distance intertextuelle possède les propriétés classiques d'une distance (identité, symétrie, inégalité triangulaire, robustesse)[24], à condition de l'appliquer à des textes qui ne sont pas trop courts (la longueur doit en tous cas être supérieure à 1.000 mots) et dont les différences de longueurs ne sont pas trop importantes (en tous cas, le rapport entre le plus court et le plus long doit être inférieur à 1/10).

Ces limitations s'expliquent par deux propriétés qui sont connues, mesurées et publiées.

1. Le facteur longueur.

La relation entre la longueur des textes et leur distance a été signalée dès la première publication[25]. Elle tient aux caractéristiques de la langue - vocabulaire limité, diminution du poids relatif des mots de basse fréquence dans les textes longs - qui entraînent les conséquences suivantes : plus les textes s'allongent, plus leurs distances auront tendance à se réduire de manière extrêmement lente. La distance intertextuelle enregistre fidèlement cette tendance propre à tout texte ou à tout discours en langue naturelle.

Notre article *JQL 01* signale clairement que cette propriété est sans incidence sur notre attribution à Corneille de 16 pièces de Molière[26]. L'annexe 4 le démontre à nouveau (p. 131 de ce dossier). Le chapitre 2 y reviendra en détail.

2. La légère instabilité de l'indice lorsque le calcul est appliqué à certains petits textes (de longueur inférieure à 3 000 mots comme indiqué dans l'article de 2003). Le chapitre 3 définira cette notion et reviendra en détail sur cette seconde propriété. Elle tient

[24] Sur ces propriétés, voir l'introduction au numéro spécial de la revue Corpus (Luong 2003) et celle de notre article *JQL 01*.
[25] Elle est examinée en détail dans Labbé & Labbé 2003 (p. 104-106 et p. 114).
[26] Ceci a été répété à de nombreuses reprises dont une fois, publiquement, à l'auteur de *JQL 06* lors d'une table ronde, à L'Université de Louvain-la-Neuve, le 11 mars 2004, devant plus d'une centaine de chercheurs. Les deux interventions à cette table ronde sont consultables en ligne : Viprey (http://laseldi.univ-fcomte.fr/Archives/affaireMorneilleColiere/morneille6.htm) et Labbé 2004a (bibliographie p. 119).

également aux caractéristiques de la langue : les mots de basses fréquences - qui génèrent l'essentiel de cette instabilité - occupent une surface d'autant plus grande que les textes sont courts (Labbé & Labbé 2003, p. 113-114).

Cette légère instabilité ne concerne pas non plus le corpus Corneille-Molière (toutes les pièces ont plus de 3 500 mots).

C’est pourquoi, comme Gulliver, *JQL 06* part dans de longs voyages pour observer des nains (textes de moins de 3500 mots), des géants (textes de plus de 20 000 mots), puis des savants fous avant de revenir à Corneille et Molière. Ces voyages ne sont pas inutiles car ils confirment la solidité de nos méthodes et de l’attribution à Corneille de 16 pièces représentées sous le nom de Molière.

Chapitre 2.

Appareillage

La section 4 de *JQL 06* **porte sur deux sujets distincts dont chacun fera l'objet d'un des deux chapitres suivants.**

Le début, ci-dessous, porte apparemment sur Corneille et Molière. On y prétend que la distance intertextuelle comporte plusieurs "biais" qui "discréditent" toute la méthode.

Curieusement, cette démonstration, fort brève, s'appuie sur un seul graphique.

Notre réponse montre qu'il ne s'agit pas de "biais" mais de deux propriétés déjà connues que *JQL 06* présente de manière caricaturale et erronée.

La première propriété (symétrie de l'indice) est toujours vérifiée.

La seconde - relation entre l'indice de la distance et la longueur des textes - n'affecte pas le cas Corneille-Molière.

C'est pourquoi *JQL 06* **appareille rapidement vers des terres lointaines…**

---

4. The biases of the "intertextual distance"

[4.1] *For anyone who has dealt for some time with questions of lexical connexion, it is not difficult to suppose that DCL's formula will bring two orders of biases, even if some artifices allow these to be limited within certain zones of application.*

[4.2] *First, it is clear that the ID are inversely dependent from the length* (N) *of the studied texts. Indeed, the longer the texts are, the more the chancy part of the distributional differences muffles whatever is the ratio* Na/Nb. *We can show this phenomena with the help of two scatter diagrams (Fig. 2) showing ordinate* DI, *on abscissa,* Na *(left graph), and* Nb *(right graph). There are 2114 points, which represent all the possible pairs of texts in the lemmatised corpus Corneille-Molière provided by DCL.*

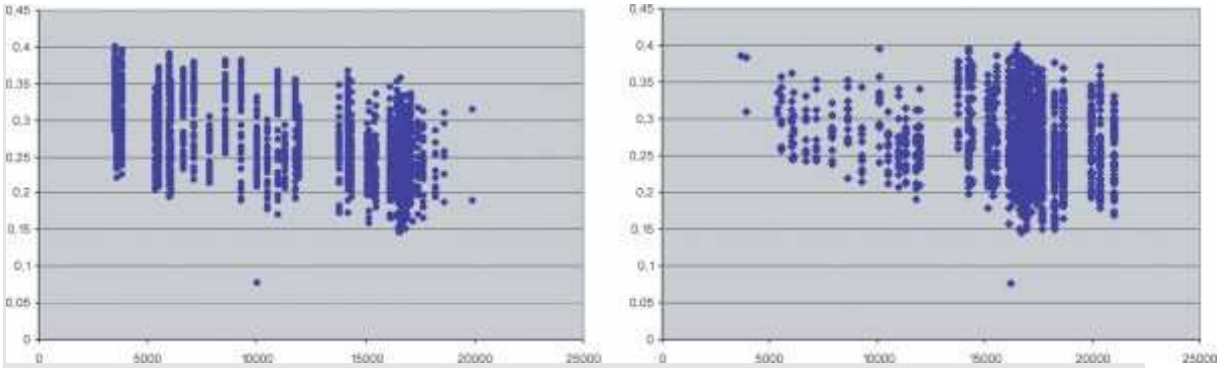


Fig. 2. ID and length of texts (left Na, right Nb).

[4.3] *This test can be applied to all sorts of corpora, whether by a single author or merged, and always gives this result. This bias is not a bias of the formula, but it affects the method as a whole. Overall it discredits the* distance standardized scale. *DCL do not explicitly mention this bias related to the quotient* Na/Nb. *They allude to it (as well as to the bias related to* Na *and* Nb) *with the following offhand comment (2001, p. 218): "It is convenient not to apply the calculation to too small texts*[5] *[…] and to avoid too large a scale of sizes*[6] *(around 1/10)".*

5. Bias related to N.

6. Bias related to Na/Nb.

Notes de Cyril et Dominique Labbé :

- Les deux notes 5 et 6 ci-dessus n'existent pas dans le texte cité et sont ajoutées par *JQL 06*. La dernière citation montre que notre article présente clairement les limites du calcul de la distance intertextuelle.
- La suite de la section 4 est reportée dans le chapitre suivant (p. 45 sq.)

Deux propriétés de la distance intertextuelle (le cas Corneille-Molière)

Les § 4.1 à 4.3 (p. 27-28 ci-dessus) affirment deux choses. D'une part, la distance intertextuelle présenterait des "biais" (asymétrie et dépendance rigide vis-à-vis de la longueur des textes). D'autre part, ces "biais" auraient faussé l'analyse du cas Corneille-Molière.

Ces deux affirmations sont infondées.

Qu'est-ce qu'un "biais" en statistique ?

JQL 06 parle beaucoup de biais sans s'expliquer sur le sens de ce mot.

Depuis C. F. Gauss, deux définitions sont acceptées. D'une part, il s'agit de l'erreur systématique engendrée par un procédé de mesure appliqué à un phénomène dont les paramètres sont connus avec certitude. D'autre part, le "biais" d'un estimateur est l'espérance de l'écart entre sa valeur et la valeur correspondante de la variable aléatoire qu'il estime.

Le terme "biais" n'a donc aucun sens ici : la distance intertextuelle n'est pas un procédé de calcul ou un estimateur mais une des multiples distances concevables[27].

Il fallait donc parler des propriétés de cette distance et montrer en quoi elles auraient pu influencer l'attribution à P. Corneille des pièces en vers de Molière et de deux de ses pièces en prose. Au lieu de cette démonstration, le lecteur se voit servir un argument d'autorité (*"For anyone who has dealt for some time…*[28]*"*) et un graphique obscur.

L'annexe 4 montre que la relation entre la distance intertextuelle et la longueur des textes comparés est sans influence sur l'attribution à Corneille de 16 pièces représentées sous le nom de Molière (p. 130 de ce dossier).

Cependant, l'étude de cette propriété est intéressante. La voici (on en découvrira tout l'intérêt dans le chapitre 8).

---

[27] Voir à ce propos, l'introduction de Labbé & Labbé 2003.

[28] *JQL 06*, § 4.1 (p. 27 de ce dossier) : "Pour tous ceux qui ont travaillé depuis un certain temps sur les questions de connexion lexicale, il n'est pas difficile de concevoir que la formule de Cyril et Dominique Labbé sera affectée de deux types de biais, même si ces biais peuvent être limités dans certaines zones d'application grâce à quelques artifices".

Sur quoi porte l'expérience présentée dans la figure 2 (§ 4.2) ?

Des 68 fichiers remis à l'auteur de *JQL 06* (annexe 1, p. 122 de ce dossier), deux semblent exclus de l'expérience présentée dans le § 4.2 ci-dessus : le prologue de Psyché (n° 36) et la Comédie pastorale (n° 56). En effet la figure 2 de *JQL 01* (p. 28 de ce dossier) ne comporte pas de points correspondant aux longueurs de ces deux textes. Cela fait donc 66 textes.

L'expérience se déroule de la manière suivante. Les textes sont rangés par longueurs croissantes : La Jalousie du Barbouillé (3 501 mots) devient le n°1 et L'Avare (21 033 mots) le n° 66. On calcule les distances entre chacun de ces 4290 (66*65) couples et les résultats sont rassemblés dans un tableau ("matrice") organisé comme le tableau II.1.

	1 Jalousie	2 Comédie des T.	3 Médecin volant	(…)	65 Toison d'or	66 L'Avare
Longueur (mots)	3 501	3 627	3 876	(…)	20 343	21 033
1 Jalousie	0	$d_{(1,2)}$	$d_{(1,3)}$	(…)	$d_{(1,65)}$	$d_{(1,66)}$
2 Comédie des T.	$d_{(2,1)}$	0	$d_{(2,3)}$	(…)	$d_{(2,65)}$	$d_{(2,66)}$
3 Médecin volant	$d_{(3,1)}$	$d_{(3,2)}$	0	(…)	$d_{(3,65)}$	$d_{(3,66)}$
(…)	(…)	(…)	(…)	0	(…)	(…)
65 Toison d'or	$d_{(65,1)}$	$d_{(65,2)}$	$d_{(65,3)}$		0	$d_{(65,66)}$
66 L'Avare	$d_{(66,1)}$	$d_{(66,2)}$	$d_{(66,3)}$	(…)	$d_{(66,65)}$	0

Tableau II.1 Schéma de principe de la matrice des distances entre les pièces du corpus Corneille-Molière, classées par longueur croissante.

En dessous de la seconde ligne (où les longueurs sont rappelées), chacune des cases de ce tableau contient la distance entre le texte en ligne et le texte en colonne. Par exemple, $d_{(1,2)}$ est la valeur de l'indice de la distance de la Jalousie du Barbouillé à la Comédie des Tuileries ; $d_{(2,1)}$ est la même distance de la Comédie des T. à la Jalousie, etc.

La distance d'un texte à lui-même est nulle (propriété notée : $d_{(a,a)} = 0$), c'est pourquoi la diagonale (cases en gras) est remplie de 0. Le tableau complet est trop grand pour être reproduit. Voici les valeurs correspondant au schéma de principe (tableau II.2).

	1 Jalousie	2 Comédie des T.	3 Médecin volant	(…)	65 Toison d'or	66 L'Avare
1 Jalousie	0	0,40	0,32	(…)	0,38	0,31
2 Comédie des T.	0,40	0	0,40	(…)	0,26	0,32
3 Médecin volant	0,32	0,40	0	(…)	0,36	0,25
(…)	(…)	(…)	(…)	0	(…)	(…)
65 Toison d'or	0,38	0,26	0,36	(…)	0	0,32
66 L'Avare	0,31	0,32	0,25	(…)	0,32	0

Tableau II.2 Extraits de la matrice des distances entre les pièces de Corneille et Molière.

Chaque texte est maintenant muni d'une valeur pour la variable "longueur" (en mots) et de 65 valeurs pour la variable "distance" (les distances aux 65 autres textes du corpus). La figure 2 de *JQL 06* est censée donner une représentation graphique de ce tableau complet. L'ensemble forme un "nuage de points". Les coordonnées de ces points sont : sur l'axe vertical les 65 distances de ce texte aux 65 autres pièces, sur l'axe horizontal, les longueurs de chacun de ces autres textes.

Ceci permet de comprendre que la figure 2 de *JQL 06* est erronée.

Une figure erronée

Le § 4.2 (p. 27 ci-dessus) affirme : "*2 114 points, which represent all the possible pairs of texts" (2 114 points, qui représentent tous les couples possibles de textes*).

Or le tableau complet comporte 66*65 = 4 290 cases non nulles. C'est donc le nombre de points qui devrait figurer sur la figure 2 (p. 28 ci-dessus).

Certes, on peut considérer qu'il y a 2 145 couples différents (4 290/2), à condition d'admettre que $d_{(a,b)} = d_{(b,a)}$, c'est-à-dire à condition de reconnaître que la distance intertextuelle est symétrique (*JQL 06* prétend démontrer le contraire).

En tous cas, 2 114 ne correspond à aucun effectif de tableau carré à diagonale nulle. Il y a donc une erreur certaine…

Enfin, il y a, sur la figure 2 (p. 28 ci-dessus), vers 10 000 mots (graphique de gauche) et vers 16 000 mots (figure de droite), un point "orphelin", en bas, très à l'extérieur du nuage - correspondant à une distance anormalement faible (inférieure à 0.10) par rapport au profil d'ensemble. Ce point étrange signale une erreur manifeste dans le traitement d'au moins deux textes.

Avec cette seule figure erronée, on prétend avoir mis en cause la symétrie de l'indice de la distance intertextuelle et avoir "démontré" la dépendance rigide de cet indice par rapport à la longueur des textes.

L'indice de la distance intertextuelle est toujours symétrique

Une mesure de distance entre deux individus (A et B) doit donner le même résultat quand elle est mesurée de A vers B (notée $d_{(a,b)}$) ou de B vers A (notée : $d_{(b,a)}$). Si cette "symétrie" n'est pas observée, la mesure n'est plus une distance…

JQL 06, dans la légende de la figure 2 (p. 28 ci-dessus), indique que, sur cette figure, le graphique de gauche représente les résultats du premier calcul (du petit vers le grand) et le graphique de droite les résultats du second calcul (du grand vers le petit). Et le § 4.5 (p. 45 de ce dossier) affirme que cette figure remet en cause la symétrie de l'indice.

Avec la formule de définition de la distance intertextuelle (voir annexe 2, p. 124 de ce dossier), le résultat est toujours le même que l'on considère d'abord le petit ou le grand texte. Outre le passage qui vient d'être cité, *JQL 06* en apporte deux autres confirmations dans la section 6 (p. 63 et p. 67 de ce dossier).

De ce fait, la figure de gauche devrait se superposer exactement sur celle de droite. Puisque ce n'est pas le cas, il y a des erreurs dans les traitements informatiques et/ou dans le graphique[29].

La symétrie de l'indice de la distance intertextuelle est toujours vérifiée sur tous les corpus (notamment sur Corneille-Molière).

Nous allons maintenant montrer, de nouveau, que les différences de longueur entre les pièces de Corneille et Molière n'ont pas d'influence sur la mesure de leurs distances et, par conséquent, sur l'attribution à Corneille de 16 pièces de Molière.

[29] Le tableau II.10 ci-dessous (p. 41 de ce dossier) présente la figure exacte.

La relation entre la longueur des textes et leurs distances

Cette propriété – qui a été signalée clairement et chiffrée précisément - peut être négligée quand les dimensions des textes étudiés ne sont pas trop différentes, ce qui est le cas pour Corneille et Molière (l'annexe 4 en apporte la confirmation).

Pourtant, le § 4.1 (p. 27 de ce dossier) affirme : "*il est clair que les distances intertextuelles sont inversement dépendantes de la longueur des textes étudiés*" (*"it is clear that the intertextual distances are inversely dependent from the length (N) of the studied texts*") et que cette dépendance serait toujours *(always)* observée "*dans toutes sortes de corpus, qu'ils soient composés d'un seul auteur ou de plusieurs*" *("in all sorts of corpora, whether by a single author or merged").*

L'adverbe "always" signifie qu'il s'agit d'une relation de type $y = f(x)$: les valeurs de la variable distance (y) sont expliquées par celles de la variable longueur (x). En l'occurrence, la distance diminuerait toujours quand augmentent la longueur des textes A et B (N_a et N_b) ou leurs différences de taille (ratio N_a/N_b)[30] selon une relation que *JQL06* n'explicite pas (linéaire, exponentielle… ?)

Dans le tableau II.2 ci-dessus, les textes sont rangés par longueur croissante. On devrait donc toujours observer des relations de ce genre : "Le texte 1 étant plus court que le texte 2 - lui-même plus court que le texte 3, etc. - la distance séparant le n° 1 du n°2 sera toujours plus grande que celle séparant le n°1 du n°3, etc., la distance séparant le n°1 du n°66 sera toujours la plus petite, etc". Ces relations peuvent s'écrire ainsi :

$$d_{(1,2)} > d_{(1,3)} > (\ldots) > d_{(1,66)}$$
$$d_{(2,1)} > d_{(2,3)} > (\ldots) > d_{(2,66)}$$
$$(\ldots)$$
$$d_{(66,1)} > d_{(66,2)} > (\ldots) > d_{(66,65)}$$

Le tableau II.3 ci-dessous résume les longueurs et distances observées au deux extrémités de la première ligne de la matrice des distances intertextuelles.

Textes	1 Jalousie	2 Comédie T.	3 Médecin V.	(…)	65 Toison d'Or	66 Avare
Longueurs (mots)	3 501	3 627	3 876	(…)	20 343	21 033
Distances de la Jalousie à :		0,40	0,32	(…)	0,38	0,31

Tableau II.3 Extraits de la première ligne de la matrice des distances intertextuelles

[30] En fait, il s'agit de la même relation et non de deux différentes comme le prétend JQL 06 (Voir annexe 4, p. 133-135).

Le texte n° 1 (la Jalousie du Barbouillé, Molière) comporte 3 501 mots, le n°3 (le Médecin volant, Molière) en a 3 876 alors que le n° 65 (la Toison d'or, Corneille) en compte 20 343. Cette dernière pièce est donc 5,25 fois plus longue que la n°3. D'après *JQL 06*, il est impossible que sa distance à la n°1 puisse être plus élevée que celle enregistrée entre les n°1 et 3 qui sont de longueurs quasiment égales. C'est pourtant ce qui se produit. Il ne s'agit que des deux extrémités de la *première* ligne. Toutes les lignes suivantes comportent des "anomalies" comparables.

Généralisation

Pour une vérification d'ensemble, deux opérations sont combinées : le jugement sur graphique et le calcul de corrélation (pour une présentation de ce calcul, cf. annexe 4, p. 130 de ce dossier).

Premièrement, chaque ligne (ou colonne) de la matrice des distances est convertie en un graphique présentant la relation entre distances et longueurs. Les 2 graphiques ci-dessous représentent respectivement les première et dernière lignes de la matrice.

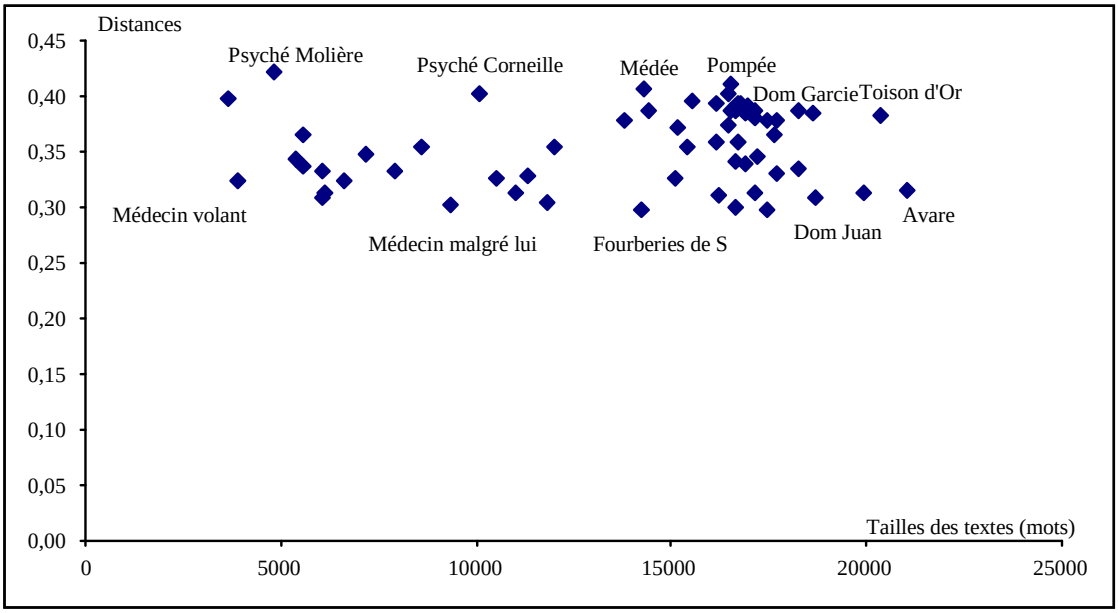


Tableau II.4 Les 65 distances séparant la Jalousie du barbouillé (3 501 mots) des autres pièces du corpus Corneille-Molière.

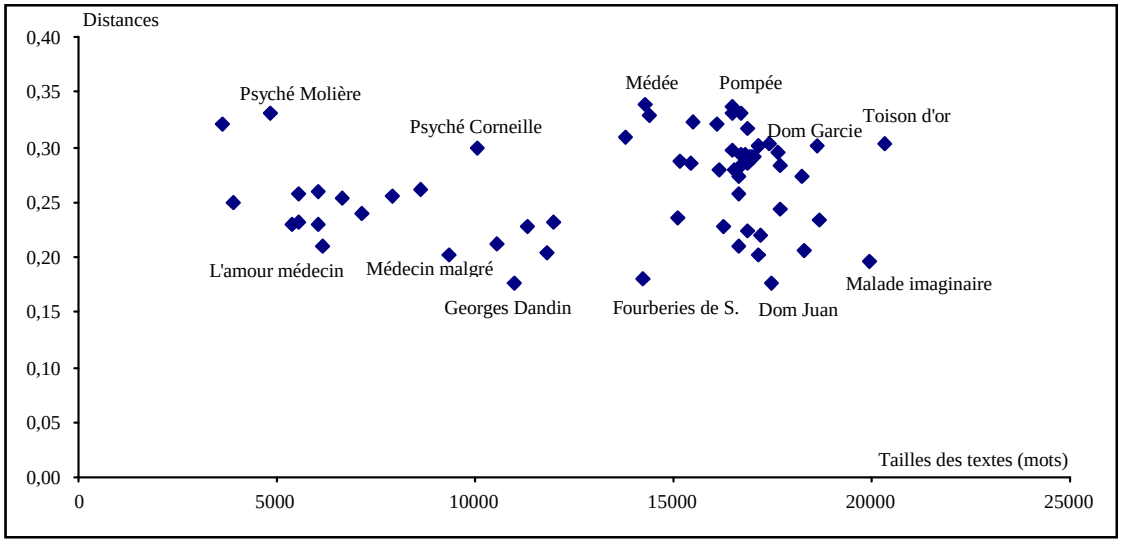


Tableau II.5 Les 65 distances séparant L'Avare (21 033 mots) des autres pièces du corpus Corneille-Molière.

Chaque point correspond à la distance de la pièce considérée (la Jalousie puis l'Avare) à l'une des 65 autres pièces. Sur l'axe vertical (ordonnée) sont portées les valeurs de l'indice de la distance dans les couples ainsi formés. Sur l'axe horizontal (abscisse), est portée la longueur des pièces correspondantes à l'autre élément du couple. Par exemple, sur le graphique II.4, les coordonnées du point le plus à droite sont : la longueur de l'Avare (21 033 mots) et la distance de la Jalousie… à l'Avare (0,31).

Il est impossible de nommer chaque point - les textes situés le plus à la périphérie du nuage sont identifiés - mais l'orientation du nuage est ici la chose importante. En effet, ces deux pièces sont celles qui mettent le plus en valeur une éventuelle relation entre distances et longueurs (longueurs extrêmes, donc plus forts ratios N_a/N_b).

Si *JQL 06* a raison, on doit observer :

- un nuage orienté vers le bas : du haut à gauche (petites longueurs donc grandes distances) vers le bas à droite (grandes longueurs donc petites distances)…

- un nuage plus dispersé à gauche et se resserrant vers la droite (convergence des vocabulaires).

A titre d'exemple, les tableaux VIII.1 et VIII.2 (p. 103-104 de ce dossier) présentent exactement ces deux caractéristiques. En revanche, rien de tel ne se produit sur les deux graphiques II.4 et II.5 ci-dessus.

Deuxièmement, un calcul de corrélation complète nécessairement le jugement sur graphique[31]. Le coefficient de corrélation mesure la force de la liaison existant entre deux variables pour lesquelles on dispose d'un grand nombre de mesures conjointes. Ce coefficient varie entre 0 (absence de liaison) et 1 (liaison rigide et variation dans le même sens) ou –1 (liaison rigide mais variation inverse). Plus on s'approche de ±1, plus cette liaison est forte.

Dans le cas du tableau II.4 ci-dessus, le coefficient de corrélation linéaire entre les deux variables est égal à +0,325 (avec 64 degrés de liberté, il y a moins de 1% de chances de se tromper en acceptant cette corrélation). Il existe donc une covariation positive. Autrement dit, il y a une tendance à l'augmentation des distances avec l'allongement des textes. C'est exactement l'inverse de ce que prévoit *JQL 06*…

Puisque les 65 valeurs de la variable longueur (notées x_i) semblent "expliquer" les 65 valeurs correspondantes de la variable distance (notées y_i), il est possible de tracer une droite, dite "droite d'ajustement de y en fonction de x". Cette droite, d'équation $y=ax+b$, passe par le "barycentre" du nuage dont les coordonnées sont la longueur moyenne ($\bar{x} = 13\,881$) et la distance moyenne ($\bar{y} = 0{,}352$). Le coefficient directeur de cette droite (ou coefficient de régression de y en x) est donné par la formule :

$$a = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = 3{,}7e^{-6}$$

Cette droite coupe l'origine à : $b = \bar{x} - a\bar{y} = 0{,}301$.

On trace la droite d'ajustement (tableau II.6).

La droite d'ajustement ne passe pas "au milieu" du nuage. Cela est dû à ce que, dans la partie supérieure, vers 16 000 – 18 000 mots, un amas de points - correspondant aux distances de la Jalousie du Barbouillé aux tragédies de Corneille et à Dom Garcie (Molière) - accentue la pente de la droite et l'écarte excessivement de l'horizontale

[31] Sur ces calculs, voir l'annexe 4.

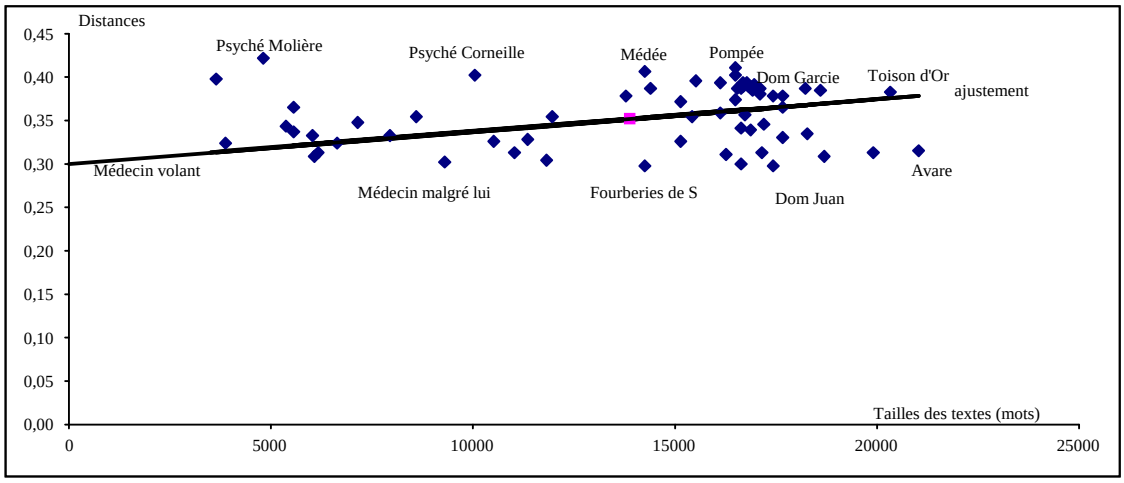


Tableau II.6 Ajustement linéaire des 65 distances séparant L'Avare des autres pièces du corpus Corneille-Molière.

Dans ce cas, on a le choix entre deux solutions : recommencer l'expérience en retirant les éléments qui semblent la perturber, ou renoncer. Cette seconde décision revient à conclure que la liaison entre les deux variables n'est pas prouvée et que l'expérience est perturbée par un "troisième facteur" : il s'agit ici clairement des différences de genre entre la Jalousie du Barbouillé et les tragédies de Corneille (plus Dom Garcie).

Si l'examen du nuage avait été favorable (droite passant "au milieu" du nuage), il aurait alors fallu évaluer, par un nouveau calcul, la qualité de cet ajustement[32] avant de pouvoir conclure à l'existence d'une **liaison** entre les deux variables. Naturellement, une liaison entre deux variables ne signifie pas une dépendance causale entre elles. La conclusion de ce chapitre revient sur cette discussion.

Dans le deuxième cas (distances de l'Avare aux 65 autres pièces), le coefficient est nul (+0,004) : les deux variables évoluent de manière indépendante, ce que traduit bien l'allongement horizontal du nuage des points sur le tableau II.5.

NB : il s'agit des deux extrêmes du corpus Corneille-Molière et des deux cas pour lesquels la dépendance entre distances et longueurs devrait être la plus visible, du moins si l'on accepte le postulat posé par *JQL06* dans son § 4.1 (p. 27 de ce dossier).

[32] Pour la procédure et les formules, voir : Cisia-Ceresta, 1995, p. 204.

La relation entre la distance intertextuelle et les longueurs des textes comparés se manifeste tendanciellement (et non mécaniquement).

La décroissance de l'indice, en fonction de l'allongement des textes, est très lente[33]. C'est pourquoi cette propriété peut être négligée quand les textes comparés ne sont pas de longueurs trop différentes. L'échelle de 1/6 - qui est celle du corpus Corneille-Molière - est relativement favorable de ce point de vue.

Dans le corpus Corneille - Molière, d'autres facteurs plus puissants sont à l'œuvre et l'"auteur" ne fait pas partie de ces facteurs puisque, dans les trois graphiques ci-dessus, on trouve des pièces de Molière dans les parties inférieures et supérieures de ces deux nuages (spécialement Psyché ou Dom Garcie). Un tel mélange est bien rare, ce qui soulève déjà quelques doutes concernant la dualité Corneille-Molière…

En fait, les pièces en prose (toutes signées Molière) figurent dans la partie inférieure du nuage et, dans la partie supérieure, on trouve toutes les tragédies (de Corneille) ainsi que la plupart des comédies en vers (de Corneille comme de Molière)…

Les principaux sous-groupes dans le corpus Corneille-Molière

Les comédies en prose sont en moyenne plus courtes que les pièces en vers et les comédies en vers sont en moyenne plus courtes que les tragédies. Ceci explique la tendance à l' "ouverture" des deux nuages de points ci-dessus vers la droite.

Prenons la situation la plus défavorable : les textes les plus éloignés du point de vue de leurs longueurs comme de leurs genres, c'est-à-dire les comédies en prose de Molière et les tragédies, toutes en alexandrins par Corneille. Le tableau II.7 présente le résultat de cette expérience. Il s'agit de la superposition des nuages correspondant à chacune des pièces dont la dimension est indiquée par un point sur l'axe des abscisses. De nombreux points sont confondus, mais le phénomène d'ensemble reste lisible.

Entre 3 500 et 21 000 mots, les distances ne présentent aucune tendance à l'augmentation ou à la baisse en fonction de la longueur des textes. Le calcul de corrélation vérifie ce constat : il n'y a aucune liaison linéaire entre les deux variables. Cette configuration était pourtant la plus défavorable.

[33] Les prochaines sections de *JQL 06* en apportent plusieurs confirmations (notamment p. 61-62 de ce dossier). Nous revenons sur la mesure de cette propriété en conclusion de ce chapitre.

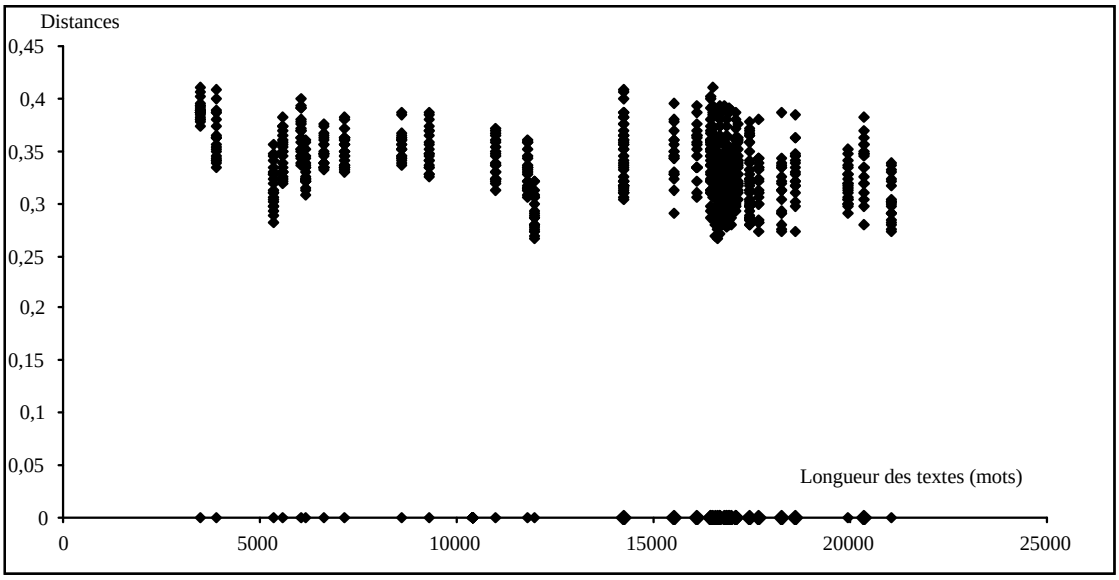


Tableau II.7 Nuage de points correspondant aux indices de la distance intertextuelle séparant les comédies en prose de Molière des tragédies en alexandrins (Corneille) classées en fonction de leurs longueur en mots.

Le tableau II.8 présente le cas le plus favorable : les distances des tragédies entre elles. Là encore, il n'y a aucune corrélation entre les valeurs de l'indice et les tailles qui sont d'ailleurs fort peu différentes. Mais l'intervalle est situé plus bas (0,16 ; 0.28).

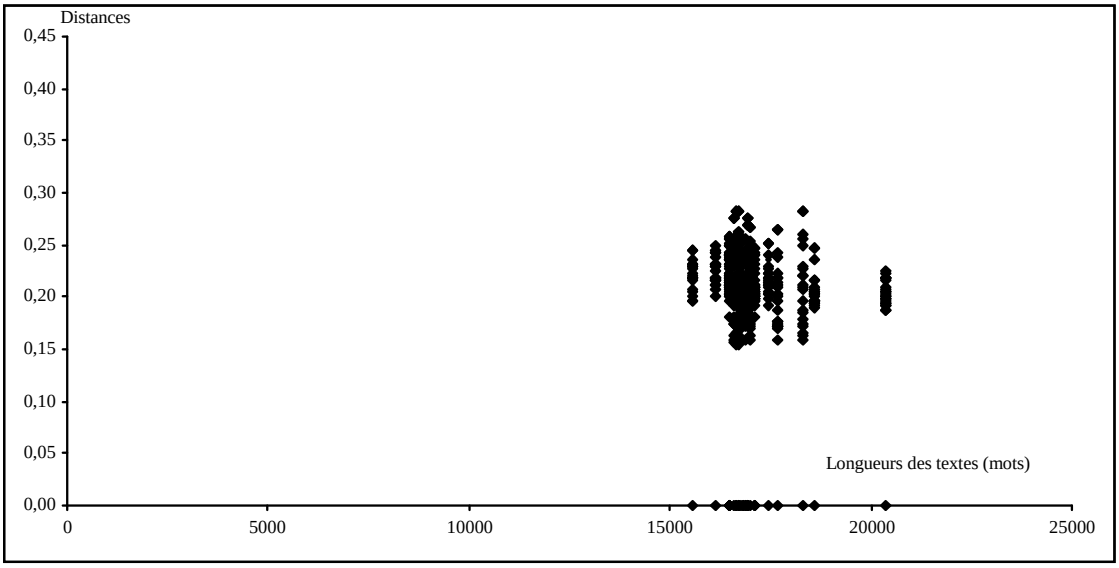


Tableau II.8 Nuage de points correspondant aux indices de la distance intertextuelle séparant entre elles les tragédies en alexandrins (Corneille) classées en fonction de leurs longueurs en mots.

Le tableau II.9 superpose les deux graphiques. Cette opération soulève évidemment une objection : il manque les distances séparant les comédies en prose entre elles. Mais le lecteur doit se souvenir que l'on cherche à comprendre comment a été obtenue la figure 2 de *JQL 06* (p. 28 de ce dossier).

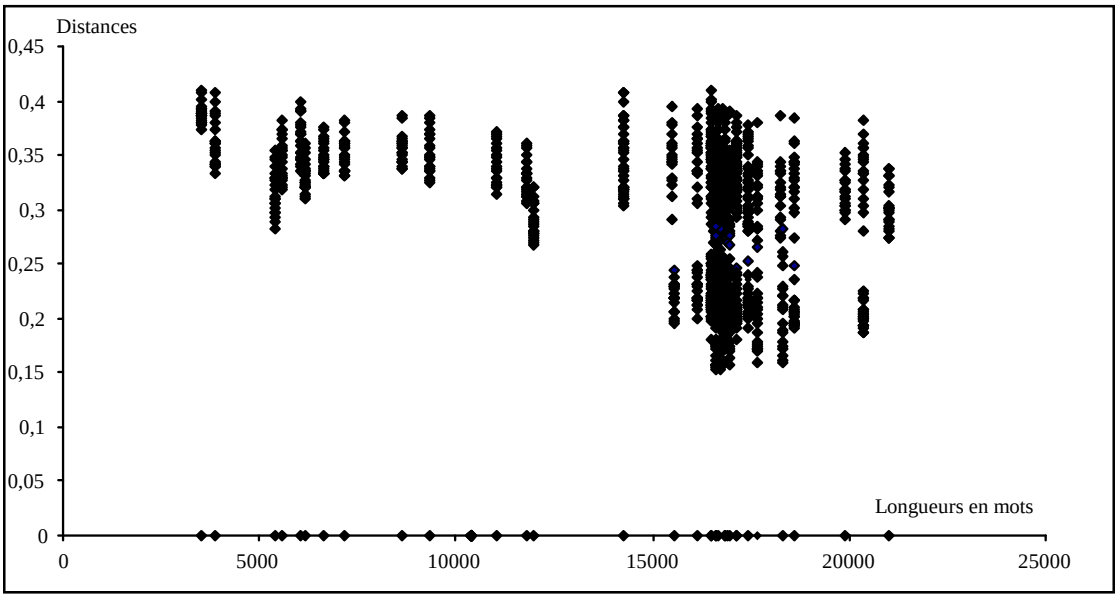


Tableau II.9 Superposition des tableaux II.6 et II.7

Sur la figure II.9, l'orientation vers le bas de ce nuage résulte uniquement du mélange de deux séries (et de l'absence de certains points) et non pas d'un prétendu "biais".

Si la longueur était la variable "explicative", on devrait d'ailleurs observer une convergence des distances – donc un rétrécissement du nuage - au fur et à mesure de l'allongement des textes. C'est l'inverse qui se produit : plus les tailles augmentent plus le nuage s'ouvre… Cela montre que l'explication proposée ne "colle" pas avec les observations.

Le tableau II.10 ci dessous superpose les 66 nuages (notamment ceux des tableaux II.4 et II.5), contenant chacun 65 points correspondant aux distances séparant chaque pièce aux 65 autres. Il est impossible d'avoir autant de formats différents pour les points et malgré leur finesse, beaucoup sont confondus dans de petits amas grisâtres…

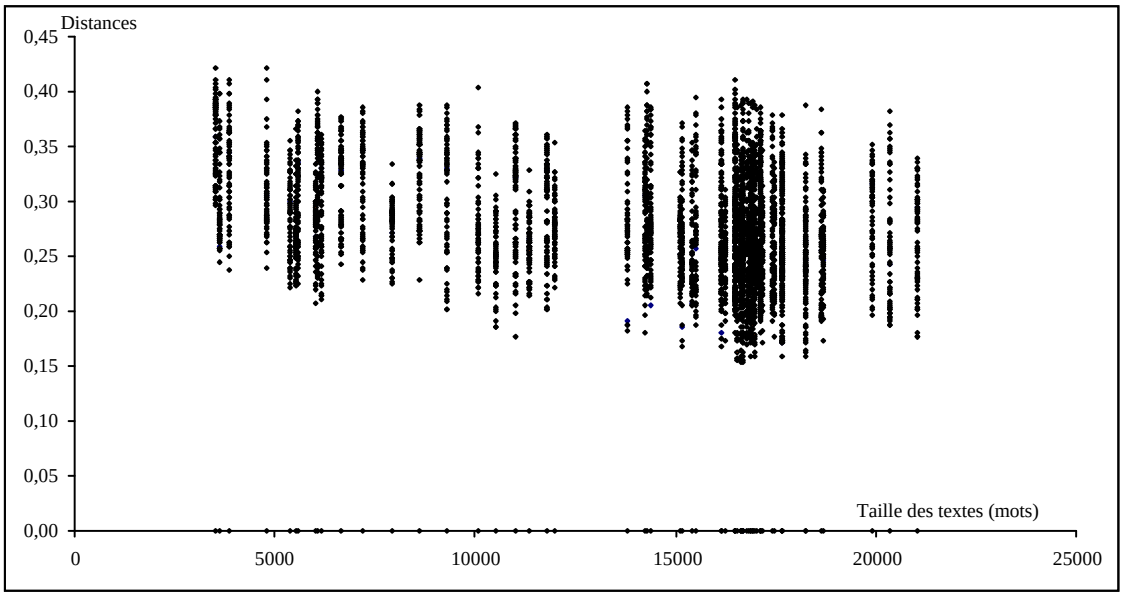


Tableau II.10 Distances entre les 66 pièces du corpus Corneille-Molière classées par longueurs croissantes.

On comparera ce tableau avec la figure 2 de *JQL 06* (p. 28 de ce dossier).

Ce nuage pourrait suggérer une très légère tendance à la baisse, du moins dans la partie gauche (et uniquement à cause d'une décroissance des minima). On sait maintenant que cela résulte essentiellement du mélange de plusieurs groupes de textes dotés de longueurs et de distances différentes…

Les principaux groupes de pièces dans le corpus Corneille-Molière

Pour rendre le phénomène lisible, deux opérations sont possibles.

1. Représenter chaque pièce un point "moyen" qui a pour ordonnée la moyenne arithmétique des distances de ce texte à tous les autres et comme abscisse la taille de ce texte (Tableau II.11 ci-dessous).

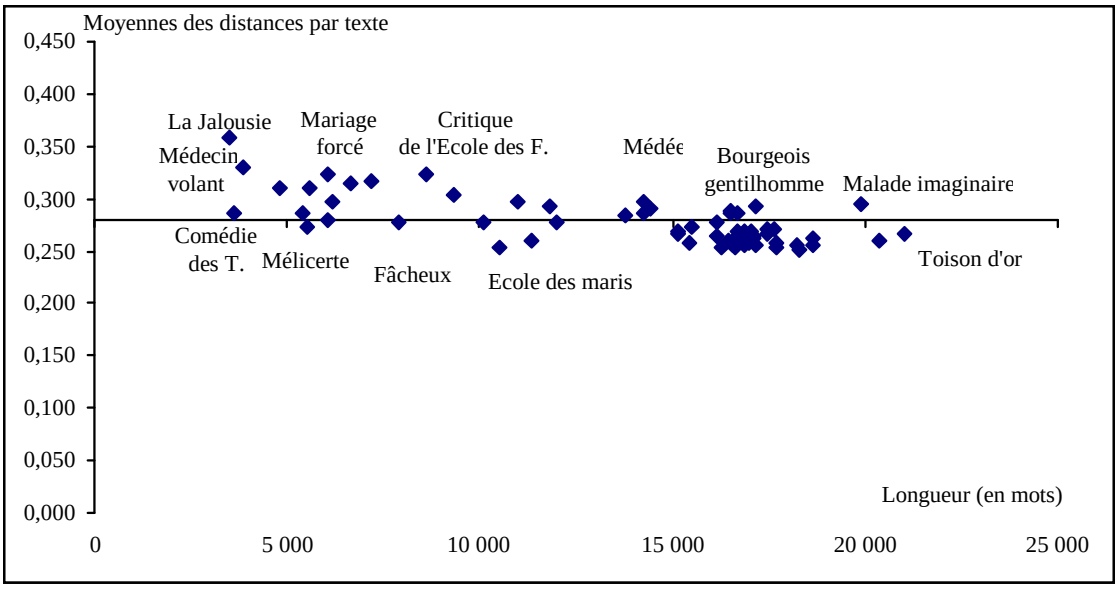


Tableau II.11 Ajustement du nuage de points du tableau II.10 par les moyennes de chaque pièce à toutes les autres

L'axe des abscisses est placé à la moyenne de toutes ces moyennes (0,279). Toutes les pièces en prose se trouvent au dessus de cette moyenne (quelles que soient leurs longueurs) et toutes les pièces en vers sont à la moyenne ou en dessous (sauf Médée, première tragédie de Corneille qui apparaît assez décalée dans tous les classements).

Là encore, la très légère tendance à la baisse (dans la partie gauche du nuage), s'explique par l'hétérogénéité du corpus comme on le vérifie grâce la seconde opération.

2. Représenter chaque groupe de distances entre les pièces classées selon leurs genres grâce au point moyen correspondant (tableau II.12). A côté de ces points, figurent entre parenthèses les trois paramètres : longueur et distance moyennes, coefficient de variation relative autour de la distance moyenne[34].

[34] Coefficient de variation relative : rapport de l'écart type à la moyenne arithmétique exprimé en pourcentage. Par exemple, le troisième chiffre entre parenthèse pour les comédies en prose signifie que les deux tiers d'entre elles sont séparées par des distances comprises entre 0.219 et 0.281 (0,255 ± 9,9%).

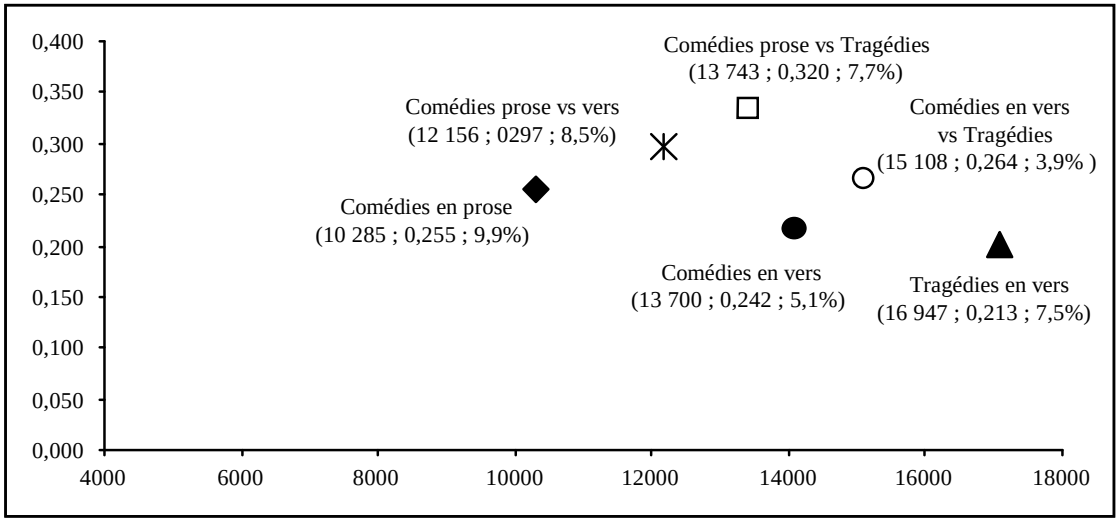


Tableau II.12 Points moyens des 6 principales sous-populations de distances dans le corpus Corneille-Molière.

La longueur moyenne des comédies en prose est de 10 285 mots, contre 13 700 et 16 947 pour les comédies en vers et les tragédies. Les premières sont nettement plus distantes entre elles que les secondes et les troisièmes (distance moyenne : 0.255 contre 0,242 et 0.213). La confrontation des comédies en prose avec les tragédies conduit à une taille moyenne de 13 743 mots et à une distance moyenne de 0,320, etc. Ces caractéristiques suffisent à expliquer la forme du nuage dans le tableau II.10.

Conclusions du chapitre 2

La relation de la distance intertextuelle avec la longueur des textes (le "facteur longueur") a été mesurée grâce à de nombreux tests[35]. Pour une échelle de 1 à 5 dans la longueur des textes comparés, la distorsion introduite par ce facteur, au deux extrémités de l'intervalle, conduit à une déviation maximale de l'ordre de ±5% par rapport à l'hypothèse d'indépendance entre distances et longueurs (dans 90% des tests)[36]. Cette limite est dépassée principalement dans deux cas. D'une part, quand au moins l'un des deux textes est fortement hétérogène ou contient une proportion notable de "jargon" ou de mots étrangers (le prochain chapitre permettra de le vérifier). D'autre part, quand les

[35] La méthode est présentée dans Labbé & Labbé 2003. Voir également pour l'anglais : Labbé 2007.
[36] Ces "performances" ne doivent pas être jugées dans l'absolu mais par rapport à celles des autres distances (voir par exemple, le chapitre 8 de ce dossier).

deux textes ont des vocabulaires extrêmement proches (leur convergence peut alors être assez rapide).

Pour le corpus Corneille et Molière, où la plupart des textes ont des longueurs proches, cette propriété peut être négligée sans inconvénient.

Cette discussion amène également trois remarques.

Un graphique qui mélange de nombreuses séries statistiques se prête mal à l'analyse, mais facilement aux manipulations.

Le jugement sur graphique doit toujours être accompagné de calculs vérifiant les hypothèses. Si l'on renonce à ces calculs, on accepte implicitement l'hypothèse d'indépendance entre les variables.

Le calcul de corrélation donne une certitude raisonnable quand il conduit au rejet de l'hypothèse. A l'inverse le fait de ne pas rejeter cette hypothèse ne la valide pas pour autant. La co-variation des deux variables peut très bien s'expliquer par un troisième facteur.

Ici le troisième facteur est constitué par l'existence de plusieurs groupes différents de pièces dans le corpus Corneille-Molière.

Enfin, tous les tests statistiques confirment que les différences de longueurs entre les pièces n'influent en rien sur l'attribution à Corneille des pièces en vers de Molière et de deux de ses pièces en prose.

Chapitre 3

Voyage à Lilliput

Dans la suite de la section 4, reproduite ci-dessous, *JQL 06* quitte le cas Corneille-Molière et s'intéresse à autre chose : de très petits textes d'un tout autre genre littéraire et d'une autre époque (100 nouvelles de G. de Maupassant) qui ont été choisis pour se situer à la limite inférieure de validité du calcul de la distance intertextuelle.

On prétend ainsi démontrer l'asymétrie de l'indice de la distance intertextuelle (en fait c'est la robustesse de l'indice qui est discutée), à l'aide de deux textes présentés comme de même longueur.

Notre réponse va montrer qu'une cascade d'erreurs entachent cette "démonstration".

4. The biases of the "intertextual distance"

[4.4] *Indeed, except in the case where* Na = Nb, *the calculation conjures away all the hapaxes of B (for which* Eia(u) < 1). *If* Na/Nb > 2, *then it conjures away all items whose frequency is 1 or 2. And so on, until items whose frequency is 10 are discarded when* Na/Nb = 10 *(if we respect DCL's limit for* Na/Nb).

[4.5] *The second bias is very underhand, since it causes no visible dependence between* ID *and* Na/Nb. *In such conditions,* ID *may have no reliable worth. The achievement of its symmetrical property (the requirement that* D(a,b) = D(b,a)) *is questionable.*

[4.6] *In addition, conjuring away lots of items introduces a destructive threshold effect. Indeed, in the special case where the pair of texts presents the identity* Na = Nb, *their* ID *will then be calculated from the whole of both their frequency vocabularies. If, for any reason (or for the purpose of an experiment), one token (one single token) is deleted*

from A, then all the hapaxes of B suddenly get left out of the calculation, which inevitably provokes a collapse of ID. *Let us demonstrate this experimentally on a selection of 101* Contes *of Maupassant, as published by Conard in 1929 (table of contents in appendix).*

[4.7] In the non-lemmatised corpus, Na *is strictly equal to* Nb *for 3 pairs of texts, for instance* Le Remplaçant *and* La Tombe *(1647 tokens). ID between these two texts is 0.507. If, in the vocabulary of* Le Remplaçant, *we reduce the frequency of a single type, by a single token (*me, *from 37 to 36 tokens),* ID *falls to 0.451. We are dealing here with an outstanding example of a threshold effect.*

[4.8] *The discovery of this perturbation also reveals the actual asymmetry of DCL's calculation. This asymmetry is only hidden by the implicit assumption that* Na *should always be different from* Nb. *Indeed, if we conversely now modify* Vb *by the same tiny quantity that we did previously for* Va (me, *from 17 to 16 tokens),* ID *again falls drastically, but this time, to 0.465. As stated above, the symmetry claimed here could be illusory.*

[4.9] In authorship attribution criticism, to which DCL hope to see their formula applied, one is often in possession of fragments of texts; cutting such fragments may be highly hazardous. It cannot be allowable that a distance calculation, otherwise so global, might be so sensitive to whether the calculation is made from A towards B, or from B towards A (once we have noticed that the calculation can only be done in one of these two ways).

Discussion de la robustesse de l'indice (les nouvelles de G. de Maupassant)

Pourquoi ne pas utiliser les pièces de Corneille et Molière ou des extraits de celles-ci (si l'on voulait travailler sur des textes plus courts) ? En effet, l'annexe 6 (p. 140 de ce dossier) montre que les nouvelles de G. de Maupassant sont extrêmement courtes et, pour la plupart, beaucoup plus courtes que les pièces de Corneille et Molière.

Quel rapport ont ces très petits textes de la fin du XIXe avec les pièces de Corneille et Molière ? On va comprendre dans un moment les raisons de ce choix, mais auparavant, il faut relever de graves anomalies.

De graves anomalies

Outre l'utilisation d'un calcul erroné[37], trois anomalies méritent d'être signalées.

Premièrement, les éditions de référence n'ont pas été utilisées. Le recueil cité est absent du catalogue de la Bibliothèque Nationale comme de celui du réseau des bibliothèques universitaires[38].

Certaines de ces nouvelles ont eu plus d'une version et plusieurs portent le même titre : 2 Clair de lune, 2 La Peur, 3 Confession, 2 Le Souvenir et 2 Le Père… De ce fait, il est impossible de reproduire exactement le corpus utilisé[39].

Enfin, G. de Maupassant a écrit plus de 300 nouvelles (répertoriées). Sur quels critères ont été sélectionnés les 100 textes[40] constitutifs de ce corpus ?

Deuxièmement, il existerait dans ce corpus, trois paires de textes comptant exactement le même nombre de mots (§ 4.7). L'existence de 2 longueurs égales est déjà

[37] Cf. encadré dans l'annexe 2, p. 126 de ce dossier.

[38] Ces catalogues sont consultables en ligne : on n'y trouve pas de recueil de textes de Maupassant, publié par les Editions Conard sous le titre Contes, en 1929…

[39] Nous avons donc traité 106 nouvelles. Les indications concernant les doublons et le triplet sont données en bas de l'annexe 5. Nos fichiers sont à la disposition des chercheurs qui souhaiteraient prolonger notre travail.

[40] *JQL 06* annonce 101 textes, mais La Bécasse est éliminée car elle compte moins de 1 000 mots.

assez improbable[41] et, de fait, les 106 nouvelles ont des longueurs différentes (annexe 6, p. 141 de ce dossier)…

Troisièmement, *JQL 06* affirme avoir traité ces textes selon nos propres méthodes (cf. aussi le § 6.7 ci-dessous, p. 63) sans citer aucun des ouvrages et articles où ces méthodes sont exposées ni préciser le logiciel utilisé. En fait, les rares indications présentées dans les §4.7 et 6.7 prouvent que notre méthode n'a pas été suivie (tableau III.1).

	JQL 06 Tokens	La réalité : Tokens	Ponctuations	Tokens + ponctuations
La Folle	1 322	1 137	185	1 322
Le Remplaçant	1 647	1 316	311	1 627
La Tombe	1 647	1 393	257	1 650
La Maison Tellier	12 212	10 455	1 781	12 236
Corpus entier (106 textes)	300 000	258 991	49 175	308 166

Tableau III.1 Les indications de longueur données par *JQL 06* et la réalité

Pour atteindre les tailles indiquées dans *JQL 06*, il faut considérer que les signes de ponctuation sont des "mots". Jusqu'ici tout le monde était d'accord pour ne pas considérer les points ou les virgules comme des "mots". Comme il y a, dans ces textes, près d'une ponctuation tous les six mots en moyenne, cela change complètement les résultats.

Enfin, les différences de longueurs entre La Tombe (1 393 mots ou 1 650 en comptant les ponctuations) et Le Remplaçant (1 316 ou 1627) ne peuvent s'expliquer par un ou deux mots composés comptés différemment (voir à ce sujet le § 6.7, p. 63 de ce dossier). Comme il n'existe qu'une seule version de ces deux textes, les dimensions sont erronées.

[41] La probabilité d'un tel événement est faible mais non négligeable. Il survient d'ailleurs une fois dans le corpus Corneille (Théodore-Pertharite). Dès lors, pourquoi ne pas avoir utilisé ces deux pièces ? On remarque aussi que *JQL 06* présente le calcul sur une seule paire de textes et ne nomme pas les deux autres censées avoir mêmes longueurs. Si elles avaient existé pourquoi ne pas avoir présenté les résultats obtenus sur ces deux autres paires ?

Que signifie l'"expérience" présentée dans les § 4.5 à 4.7 ?

Deux notions sont confondues : la symétrie et la robustesse. La symétrie de l'indice de la distance intertextuelle est toujours vérifiée (cf. chapitre précédent, p. 32-33). Qu'en est-il de sa robustesse ?

En statistique, un indice est robuste quand il n'est pas affecté par une très légère modification apportée à l'une des variables. Dans le cas inverse, l'indice est instable. Ceci se marque souvent par un "effet de seuil" ("*threshold effect*") : autour de certaines valeurs, l'indice est affecté de variations trop importantes par rapport à ce qui est attendu.

Dès l'origine, il a été signalé deux circonstances qui peuvent générer une légère instabilité de l'indice de la distance intertextuelle :

- lorsque les textes comparés sont de petites tailles (inférieures à 3 000 mots : Labbé & Labbé 2003, p. 103 et Labbé 2003, p. 54). Comme indiqué dans notre article *JQL 01* (passage cité dans l'annexe 2, p. 125 de ce dossier), le poids important des basses fréquences dans les textes courts explique l'essentiel de cette instabilité ;

- lorsque les textes contiennent une proportion notable de mots étrangers ou de "jargon" (voir le passage de notre article *JQL 01* reproduit à la fin de l'annexe 2 - p. 125 de ce dossier - et Labbé & Labbé 2003, p. 106).

JQL 06 a sélectionné certaines nouvelles de G. de Maupassant justement pour cela : textes très brefs avec, pour certains, une proportion importante de jargon. Des distances aussi fortes que celles citées[42] signalent simplement que ces textes ne sont pas écrits dans le même langage. L'annexe 5 (p. 136 de ce dossier) reproduit les deux textes (La Tombe, Le Remplaçant) sur lesquels porterait l'expérience. La Tombe (écrit dans le style soutenu des plaidoiries d'assises) et Le Remplaçant (écrit dans un jargon de corps de garde) confirment la règle. De telles "étrangetés" se prêtent mal à l'analyse statistique. Voilà la principale raison pour laquelle ces textes ont été retenus.

JQL 06 prétend avoir réalisé l'expérience suivante. Il aurait mesuré la distance entre ces deux textes (prétendument de "même longueur"). Puis il aurait retiré un seul mot au Remplaçant et recommencé la mesure. Il aurait observé à cette occasion, une chute

[42] La Tombe et le Remplaçant seraient séparées par une distance de 0,507. Il s'agit d'une étrangeté radicale, même si l'échelle de la distance n'est pas valable sur des textes aussi courts (les repères devraient être relevés).

"drastique" de l'indice et une absence de symétrie dans les résultats quand il aurait réédité l'opération en enlevant un mot à la Tombe.

En admettant que ces deux textes soient de même longueur (ce qui est faux), les résultats sont-ils crédibles ? La distance relative entre ces deux textes serait de 0.507. On en tire la distance absolue (D) :

$$0.507 = \frac{D}{1\ 647 * 2} \Rightarrow D = 1\ 670\ \text{"}tokens\text{"}$$

Dans le § 4.7, *JQL06* affirme que, en enlevant un mot au Remplaçant, la distance diminuerait à 0.451. Il aurait donc disparu du numérateur et du dénominateur :

$$0.451 = \frac{1\ 670 - X}{(1\ 646 + 1\ 647) - X} \Rightarrow X = 337\ \textit{vocables retirés du calcul}$$

Si l'on a suivi le procédé exposé dans notre article *JQL01*, il y aurait, dans le Remplaçant, 337 vocables qui n'apparaissent qu'une seule fois, tout en étant absents de la Tombe. Cela représente 20,5% de la surface totale du Remplaçant (en admettant qu'il compte réellement 1 647 mots). Le tableau III.2 donne les chiffres exacts.

	La Tombe	Le Remplaçant
Longueur des textes ("tokens")	1 393	1 316
Total vocables différents	488	394
Total vocables de fréquence 1	330	254
Vocables de fréquence 1 absents de l'autre texte	282	203

Tableau III.2 Caractéristiques du vocabulaire comparé de la Tombe et du Remplaçant.

Même si le baragouin du dragon Siballe n'a rien à voir avec le plaidoyer de l'avocat Courbataille - raison pour laquelle *JQL06* a sélectionné ces deux textes (ils n'ont rien en commun, même pas leurs longueurs) -, les chiffres annoncés ne pourraient être approchés qu'en enlevant du calcul tous les vocables de fréquence 1 de la Tombe qu'ils soient ou non utilisés dans le Remplaçant. Cela ne correspond pas du tout à ce que propose notre article *JQL01*, ni au programme informatique remis en avril 2003.

En réalité, sans la ponctuation, la distance entre ces deux textes est de 0.490, ce qui est déjà considérable, mais s'explique bien par la petite taille de ces textes et par

l'étrangeté des deux vocabulaires. Le fait d'enlever un vocable - quel qu'il soit - à l'un des deux textes ne change rien à la valeur de l'indice.

Dans ce corpus, quelques textes ont des tailles très proches (voir annexe 6, p. 140 de ce dossier). Il est donc possible de les utiliser pour tester la stabilité de l'indice. Il suffit d'enlever 3 ou 4 mots au plus long pour le faire passer sous la dimension de son voisin de longueur immédiatement inférieure. Mais, les vocabulaires étant moins étrangers, on n'obtient pas de résultats "spectaculaires"…

En fait, la légère instabilité de l'indice, observée sur certains textes très courts - caractéristique que nous avons nous-mêmes signalée clairement - devient négligeable bien avant 3 000 mots (sauf présence de jargon ou de mots étrangers en proportion importante). Sur des textes comme ceux du corpus Corneille-Molière, l'indice est robuste. Sans s'en rendre compte, *JQL 06* en apporte deux preuves éclairantes.

- Il y a dans le corpus Corneille deux pièces de même longueur (Théodore et Pertharite). Il est donc facile de faire l'expérience sur ces deux textes. L'indice est parfaitement robuste.

- La figure 3 (section 5, p. 54 de ce dossier) présente une courbe parfaitement lisse qui correspond aux distances calculées avec la formule de définition sur les textes lemmatisés du corpus Corneille-Molière : aucun saut, même minime, aucune rupture de pente qui seraient l'indice d'une instabilité…

JQL 06 confirme que, sur les textes de Corneille et Molière, l'indice de la distance intertextuelle est parfaitement robuste.

Chapitre IV
Voyage à Lagado

La section 5 reproduite ci-dessous, conteste les normes de saisie et de dépouillement des textes et rejette les conventions jusqu'ici admises en lexicographie et dans les recherches sur le langage.

Notre réponse montre que ces choix sont intenables, tout comme le graphique qui est produit dans cette section. En fait, *JQL 06* présente une pièce supplémentaire en faveur de nos méthodes mais aussi en faveur de la paternité de Corneille sur un certain nombre de comédies de Molière.

---

5. Lemmatization

[5.1] *DCL then stipulate that texts must be* normalized *(without any reference to give any precise meaning to that term) and, "from [their] point of view […] tagged". We see here that DCL consider* tagging to *be the same as* lemmatization, *when they are distinct operations (Habert et al., 2000) and the English actually use the verb lemmatize. And indeed, the results of* ID *are noticeably different depending on whether we utilize, for the corpus Corneille-Molière, raw or lemmatised text (DL provided us with his lemmatized corpus). This is shown by Figure 3, where the smooth curve represents the* ID *between lemmatized texts (sorted by increasing order), and the bumpy curve, the* ID *between rough texts. The maximum difference between the two results is 0.07, the minimum 0.025, the mean difference 0.042. This difference is not proportional to the values.*

[5.2] *It is not our intention to tease apart, in these gap variations, what comes from properties of texts themselves, and what comes from the lemmatizing process. Crucially, we would note and completely agree with this statement: "One can see that the distance calculation implies a prior agreement on standards" (2001, p. 218). Such an agreement*

being difficult to achieve, the proposed method is unlikely to be generalized or submitted to experiment.

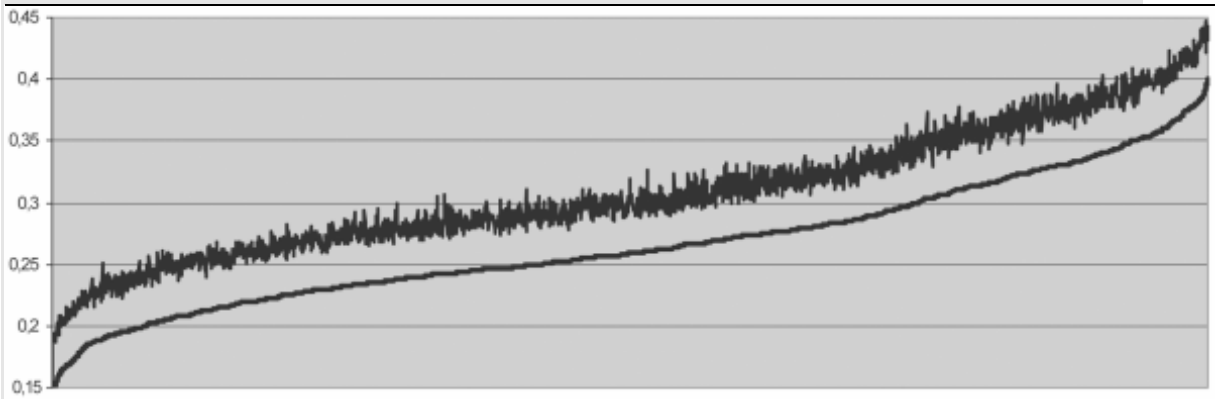


Fig. 3. Rough parallelism of ID depending on whether texts are lemmatised (smooth curve) or not.

[5.3] *In any case, such an agreement could not be reached with regard to the lemmatization standards used by DL. For instance, for the whole corpus (928,000 tokens, 9947 different lemmas found by DL), only 16 lemmas are compounds (plus 12 named entities). Compounds as frequent as* afin que, afin de, bien que, sans cesse, *are systematically counted as two lemmas;* cesse *is considered to be a noun, etc. Admittedly most French corpora, submitted to diverse lexical statistic operations, have suffered from insufficient reference to modern lexicological and semantic theories and to computational linguistics. Admittedly, these mistakes and naiveties are widespread. Nevertheless, that should not conceal the main point: some inaccuracy can be compensated for using probabilistic statistics, but nothing is probabilistic in the DL approach.*

[5.4] At *any rate, if necessary for an experiment, the calculation of* ID *may be performed on non-lemmatized texts. This is worth noting, because if scientists intend to test the method at the desired scale (thousands of texts, hundreds of millions of tokens), demanding that the text should be lemmatized (and, what is more, following a specific norm), is exorbitant.*

De l’art de passer à côté des évidences

L’objectif de la section 5 de *JQL06* semble être de montrer que l’indice de la distance intertextuelle peut être calculé sur des "textes bruts" (non lemmatisés) et que ces résultats peuvent être comparés avec les nôtres. En revanche, cette section n’explique pas ce qu'est un "texte brut" ("raw text", voire : "rough text"[43]) ni comment les mots ("tokens") sont découpés ni avec quel logiciel cette opération est réalisée. Cette expérience ne peut être reproduite.

D'après le §5.1, deux expériences ont été réalisées : l'une avec nos fichiers lemmatisés, l'autre sur les textes "bruts". La figure 3 (ci-contre) ne contiendrait pas deux courbes mais deux nuages dont chaque point correspondrait à la distance intertextuelle séparant un couple de textes ("bruts" pour le nuage supérieur et lemmatisés pour le nuage inférieur), ces distances ayant été rangées par ordre croissant.

Ce graphique confirme la solidité de notre méthode et de nos conclusions.

L’ombre de Corneille

La "courbe" inférieure de la figure 3 reproduit les données qui ont été remises à l’auteur de *JQL 06* en avril 2003 (distances sur les pièces lemmatisées). Le profil est remarquablement régulier - sans aucune rupture de pente ni "effet de seuil", ce qui invalide la prétendue démonstration de la section précédente (avec les nouvelles de Maupassant) et les conclusions de la section 7 (p. 64-65 de ce dossier).

Les points épousent un profil en S caractéristique d'une fonction de répartition continue d'une variable aléatoire, fonction qui est habituellement représentée par un "histogramme des fréquences" (tableau IV.1).

Comme dans la figure 3 (ci-contre), les distances entre couples sont rangées par ordre croissant, mais elles sont ensuite classées par intervalles égaux (ici 0,01) dont les effectifs sont comptés. Les valeurs centrales des classes sont portées sur l’axe horizontal et les effectifs de chacune de ces classes sur l’axe vertical. Par exemple, à l'extrême gauche, la première classe contient toutes les distances inférieures à 0,17 (soit 11

[43] Synonyme de "draft" (brouillon)…

couples de pièces), la seconde les distances comprises dans l'intervalle {0,16 < $d_{(a,b)}$ < 0,18}, soit 26 couples de pièces, etc.

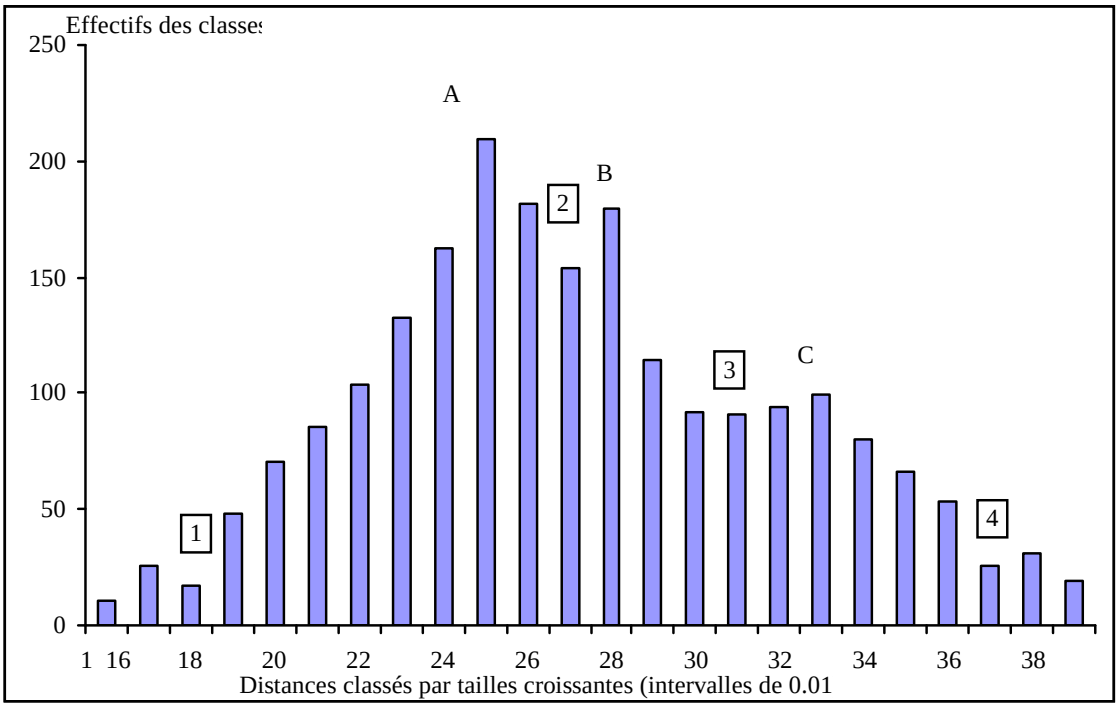


Tableau IV.1 Histogramme des fréquences des distances entre tous les couples de pièces du corpus Corneille-Molière rangées par valeurs croissantes (effectifs absolus).

Le graphique suggère l'existence de 3 sous-ensembles principaux indiqués par des lettres au-dessus des trois "modes" (les fréquences que l'on a le plus de chances de rencontrer)[44] et grâce aux points d'inflexion (1, 2, 3, 4).

Ces trois sous-ensembles concernent, de gauche à droite :

- Mode A (0,25). Autour de ce mode, en regroupant toutes les distances comprises entre 0.19 et 0,26, on délimite un premier sous-ensemble (le plus nombreux). On y trouve les distances observées entre pièces de même genre : les tragédies entre elles ; les premières comédies de Corneille entre elles, les comédies en vers de Molière entre elles (plus les deux Menteurs de Corneille, l'Avare et Dom Juan) ;

- Mode B (0,28). Autour de ce mode et grâce aux points d'inflexion 2 et 3, on délimite un deuxième sous-ensemble {0,26 < $d_{(a,b)}$ < 0,31}. On y trouve principalement

[44] Le tableau II.11 (p. 42 de ce dossier) a déjà présenté les distances moyennes observées sur chacun des sous-ensembles de distances entre pièces (tragédies, comédies, vers, prose).

les distances dans les couples formés par une tragédie avec une comédie en vers (plus l'Avare et Dom Juan), quel que soit l'auteur officiel ;

- Mode C (0,33). Ce dernier sous-ensemble contient les distances comprises entre 0,31 et 0.36. Elles séparent les pièces en prose (autres que l'Avare, le Dom Juan) avec la plupart des pièces en vers.

Enfin, deux petits groupes se signalent aux deux extrémités du graphique (points d'inflexion 1 et 4).

A l'extrême gauche, les plus faibles distances (inférieure à 0,18) concernent les dernières tragédies de Corneille entre elles (de Nicomède à Suréna, à l'exception de la Toison d'or) plus : Tartuffe-Misanthrope (0,171), Etourdi-Dépit amoureux (0,174), Veuve-Galerie du Palais (0,175), Dom Juan-Avare (0,177) et, enfin, Menteur - Suite du Menteur (0,1796).

A l'extrême droite, les plus fortes distances séparent trois pièces de Molière (la Jalousie du barbouillé, le Médecin volant, le Mariage forcé) avec les tragédies de Corneille et avec… Dom Garcie ainsi qu'avec les passages de Psyché censés avoir été écrits par Molière ("Psyché Molière"). La plus forte distance observée sur ce corpus (0,422) sépare d'ailleurs Psyché-Molière de la Jalousie du Barbouillé… censée avoir été également écrite par Molière.

Ce profil de distribution est classique en statistique. Des tests peuvent aider à décider laquelle des deux hypothèses doit être rejetée : ensemble unique perturbé par quelques individus aberrants et/ou, comme ici, deux ensembles distincts se recouvrant partiellement (Tableau IV.2, ci-dessous).

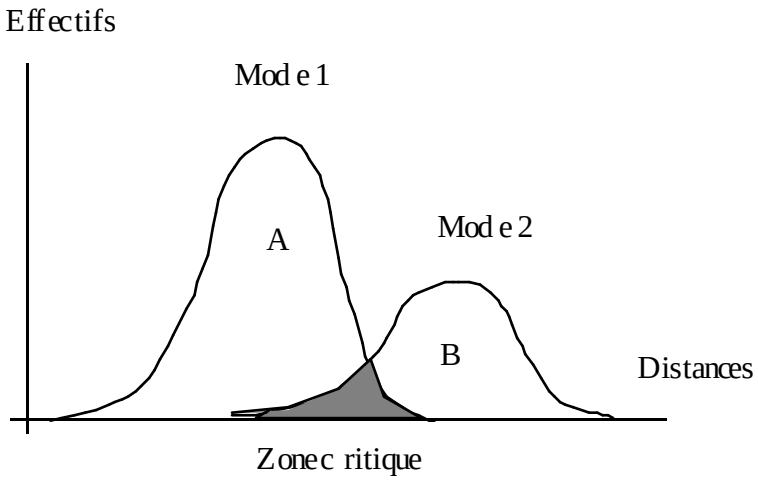


Tableau IV.2 Deux ensembles de pièces différentes se recouvrant partiellement

L’ensemble A correspond aux distances séparant toutes les pièces en vers entre elles, plus quelques pièces en prose (notamment les comédies en vers avec le Dom Juan et l’Avare). L’ensemble B contient les distances des pièces en vers à quelques comédies en prose (Jalousie du B., Médecin volant, Escarbagnas, Sicilien, etc.)

Il demeure une zone "critique" (le recouvrement entre les deux courbes, en gris sur le schéma ci-dessous). Certaines des pièces, impliquées dans les distances comprises dans cette zone, ne peuvent être rattachées avec certitude à l'une ou l'autre des deux sous-populations… Pour le corpus Corneille-Molière, c'est dans cette zone que l'on trouve la majorité des distances séparant spécialement le Bourgeois gentilhomme et le Malade imaginaire des deux Menteurs de Corneille. Telle est la raison pour laquelle - malgré leurs évidentes affinités avec les comédies en vers de Molière et avec les deux Menteurs (et avec l'Avare et Dom Juan, dont elles sont également très proches) – elles ont été laissées dans la zone critique…

Cette méthode statistique standard a été appliquée sur un grand nombre de corpus (auteur unique, duos, triplets, etc. – dont certains en "double aveugle" (nous ignorions les auteurs et les titres)[45]. C'est ce qui permet d'affirmer qu'on ne rencontre pas de distribution semblable en mélangeant 33 et 32 textes écrits par deux auteurs différents (hors Psyché, il y a 33 pièces de Corneille et 32 de Molière). En revanche, ce profil correspond au mélange d’une cinquantaine de textes par un auteur avec une douzaine d’un autre…

Cette méthode a été utilisée pour tester la distance intertextuelle et étalonner l’échelle des distances. Le travail a été fait avec soin, en suivant les procédures habituelles. Cela répond aux imputations du § 3.8 : nous ne fixons aucun seuil ("*rigid and regularly spaced thresholds*"), mais des repères pour baliser un continuum. Ces repères permettent d'émettre un premier jugement sur une série statistique et de n'examiner que certaines zones critiques parmi le grand nombre des distances.

L’académie de Lagado

Sur la figure 3 (p. 54 de ce dossier), le nuage supérieur correspondrait aux distances intertextuelles calculées selon la méthode présentée dans la section 2 de *JQL 06* et sur

[45] Pour deux exemples, voir : Labbé 2002 et Labbé 2007, p. 50-52.

les "textes bruts". Ce nuage présente un profil heurté, voire chaotique. L'amplitude des "sauts", d'un point à l'autre, approche parfois le cinquième des valeurs portées en ordonnées, ce qui est considérable. Cela signale au moins trois choses.

Premièrement, cela démontre la supériorité de la lemmatisation sur le dépouillement "grossier" (comment traduire autrement "rough" ?).

L'absence d'une norme claire pour le dépouillement des textes conduit au chaos.

Deuxièmement, il est impossible que la distance intertextuelle, mesurée sur des pièces voisines du corpus Corneille-Molière, puisse enregistrer des fluctuations supérieures ±10%, en passant d'une pièce à l'autre, car ce corpus est très homogène. Une instabilité aussi considérable signale que le programme de calcul est erroné.

Troisièmement, puisque le nuage monte et descend, en allant de gauche à droite, c'est que les distances, calculées sur les textes "grossiers", n'ont pas été classées par valeurs croissantes (ou du moins, il y a des erreurs dans le classement).

Il est également confondant de lire au § 5.3 que "Cesse" n'est pas un nom féminin. Tous les dictionnaires français, depuis Furetière, classent ainsi ce mot. Aucun de ces dictionnaires ne comporte d'entrées à la lettre D (pour "De cesse") ou à la lettre S (pour "Sans cesse"). Les lexicographes considèrent donc que ces locutions ne sont pas des "vocables" (plus petites unités du lexique d'une langue). Il en est de même pour : "Afin de", "Afin que", "Bien que". Tout le monde est d'accord sur ces conventions…

Au passage : *JQL 06* n'a trouvé que cela à critiquer dans un travail qui porte sur près d'un million de mots. Ces mots ont donc été attachés sans erreur à plus de 9 900 lemmes. D'ailleurs, l'expérience, présentée dans la dernière section (graphique 6, p 84 de ce dossier), utilise nos lemmes et non pas le dépouillement "grossier"…

Ce sont des hommages involontaires qui ont beaucoup de poids.

Les normes lexicographiques du français ont été scrupuleusement respectées. Ces règles, qui régissent l'élaboration des dictionnaires, ont été esquissées par Furetière puis reprises notamment par Littré ou Hatzfeld et Al. (1898). A l'époque contemporaine, elles ont été adaptées à l'informatique notamment par : P. Guiraud (1955-1964), C. Muller (1967), C. Bernet (1983), B.-M. Kylander (1995), etc.

Les programmes de normalisation (standardisation) des graphies et de lemmatisation des textes sont calqués, le plus fidèlement possible, sur ces conventions pragmatiques et respectueuses de la langue. Ils ont été testés avec la collaboration de nombreux chercheurs qui, depuis plus de 30 ans, ont signé avec nous des communications, des articles, des livres…

Naturellement, tout cela peut être discuté, mais il faut proposer autre chose qu'un dépouillement "grossier" et sans méthode.

Enfin, à la fin du § 5.3 et dans le suivant (p. 54 de ce dossier), *JQL 06* confond le calcul des "probabilités" et les "incertitudes" pesant sur une mesure du fait d'imperfections possibles dans les observations. Pour les expériences cruciales, ces imperfections doivent être éliminées grâce à des contrôles rigoureux. Dans les sciences expérimentales, le fait d'admettre qu'une mesure est inscrite dans une certaine "incertitude" - qu'il est toujours préférable de réduire au maximum - ne permet pas de faire n'importe quoi. En particulier, il est absurde de prétendre traiter des "centaines de millions de mots" sans aucune méthode, comme revendiqué dans le § 5.4 de *JQL 06*.

Au fond, la seule chose certaine c'est qu'il faut suivre toujours les mêmes conventions et les mêmes procédures. Il est indispensable que les mots soient toujours comptés de la même manière. C'est pourquoi les positions exprimées dans les § 5.2 - 5.3 (p. 54 de ce dossier), sont la négation de toute science expérimentale. Il n'y a pas d'expérience sans conventions partagées, sans unités de mesure communes, sans procédures standardisées et sans le respect des protocoles usuels.

Si l'on décide qu'un nuage de points est une courbe ; qu'un graphique suffit à prouver une "loi" ; que l'empilement d'observations, classées n'importe comment, forme une "courbe cumulative" ; que les virgules et les points sont des "mots" ; qu'un même mot doit être compté pour 10 parce qu'il est écrit de 10 manières différentes dans les éditions de référence (et non les "rough texts") - une fois avec une majuscule initiale, une autre en lettres capitales, une autre encore en minuscules, puis sous diverses abréviations et jargon… - on s'installe alors à Lagado, bien connue grâce aux voyages de Gulliver…

Chapitre 5

Nouveaux voyages de Gulliver

La 6e section de *JQL 06* revient sur la relation entre la distance intertextuelle et la longueur des textes analysés grâce à quelques-uns des plus longs romans de la littérature du XIXe siècle : G. Flaubert, G. de Maupassant, A. Dumas, H. de Balzac… puis elle retourne à Lilliput (les nouvelles de G. de Maupassant).

La 7e section affirme que ces expériences "discréditent" l'ensemble de la méthode.

Notre réponse indique à nouveau que l'échelle de la distance intertextuelle n'est pas étalonnée pour de pareilles longueurs et que ces "expériences" vérifient au contraire la solidité de cet outil. Au passage, on rencontre encore bien des erreurs…

---

6. Experiments[7]

[6.1] *When we tried to verify DCL's assertion that* ID *below 0.20 indicates a single author, we thus first began work on non-lemmatized texts. We thus established that, for instance, Flaubert's* Madame Bovary *shows an* inter-textual distance *of 0.223 with Maupassant's* Une Vie, *and also 0.223 with the same author's* Fort comme la mort. *Mean* ID between Madame Bovary *and Maupassant's eight novels is 0.241. It is far higher in the case of* Salammbô*: 0.348. See Table 1.*

[7] Flaubert's and Maupassant's works have been established from the Conard editions, Dumas' works from the Calmann-Levy editions.

Table 1. ID between three novels of Flaubert and Maupassant's eight novels (1. Une Vie; 2. Bel-Ami; 3. Mont-Oriol; 4. Pierre et Jean; 5. Fort comme la mort; 6. Notre cœur; 7. L'Âme étrangère; 8. L'Angelus).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	***Moyenne***
Madame Bovary	0.223	0.239	0.231	0.242	0.223	0.238	0.280	0.250	0.240750
Salammbô	0.321	0.351	0.340	0.356	0.346	0.358	0.358	0.350	0.347500
L'Education sentimentale	0.245	0.234	0.245	0.244	0.237	0.246	0.288	0.268	0.250875

[6.2*] In a second step, we then set about lemmatizing* Madame Bovary *and* Une Vie, *adhering most closely to DCL's "word for word" technique. The result confirmed our expectation: between the lemmatized texts,* ID *decreases to 0.197 – in other words, below the threshold of 0.20 under which DCL rule out the existence of two different authors.*

[6.3] Among numerous possible tests, we also noticed an exceptionally low ID *between the non-lemmatized texts of Balzac's* Père Goriot *and Dumas'* Comte de Monte-Cristo, *as well as between diverse other novels of those two writers in particular. See Table 2.*

Table 2. ID between three novels of Balzac and five of Dumas (1. Fernande*; 2.* Le Comte de Monte-Cristo*; 3.* Joseph Balsamo*; 4.* Le Collier de la Reine*; 5.* Les mille et un fantômes).

	1	**2**	**3**	**4**	**5**	***Moyenne***
Histoire des Treize	0.221	0.240	0.241	0.241	0.232	0.2350
Le Père Goriot	0.239	0.218	0.221	0.224	0.230	0.2264
Le Médecin de campagne	0.250	0.238	0.251	0.256	0.220	0.2430

[6.4] *At least some of those pairs would necessarily fall under 0.2 if lemmatized. This simply demonstrates that DCL's tests were too incomplete to claim scientific status.*

[6.5] *If we apply* DI *inside the whole of the* Comédie humaine *(CH), we see that many pairings give a* DI *far greater than that obtaining between several of Balzac's novels and several of Dumas'. This empirical evidence contradicts the claims of DCL about their own scale (see Section 3 above).*

[6.6] *Moreover, this observation throws light on the bias related to* N, *by a commonly admitted and very robust statistical test: Spearman's rank correlation (Kendall, 1962). As CH consists of 86 texts, 3655 pairs are possible. Only 3148 are permitted since 507 have a quotient* Nb/Na > 10. *Of these 3148, 888 have a* DI *greater than 0.3. Those 888 pairs involve 77 of the 86 texts. All of remaining nine texts are among the 20 longest texts, and have more than 100,000 tokens. We therefore have counted the number of times when each of the 86 texts of CH is involved in a* DI *superior to 0.3: from 0 times (nine texts) to 58 times (*Jésus-Christ en Flandre, *7838 tokens). Then we established the Spearman correlation index between the ranking of the 86 texts, by decreasing length on the one hand, and by increasing frequency of involvement in a* DI > 0.3 *on the other. The result is clear: 0.842 for 86 items, which suggests a very probable inverse correlation between* N *and* DI.

[6.7] *Those first tests were conducted on large texts, i.e., "favoured" by one of the two biases inherent in DCL's* ID. *We then tested* ID *on very short texts, Maupassant's 101* Contes *listed in the appendix. We entirely lemmatized that corpus (more than 300,000 tokens), following again DL's "word for word" technique (except for some compounds DL identifies in his own Corneille-Molière corpus). Except for* La Bécasse, *which we excluded on DCL's recommendation (it has only 859 tokens), lengths (*N*) of the 100 others go from 1322 (*La Folle*) to 12,212 (*La Maison Tellier*), with a mean of 2938. So, all of them have more than the 1000 token-threshold beneath which DCL find it convenient not to calculate* ID.

[6.8] *Results are in accordance with our predictions, i.e., irremediably affected by the bias described. The mean for the 4950 pairs is established at 0.371, i.e. very close to the fateful limit of 0.4, which DCL consider the* minimal common nucleus for texts produced by a same author. *The mean for the couples with* Na + Nb < 4000 *is 0.401. See the scatter diagram in Figure 4, showing dependence of* ID *towards* Na + Nb.

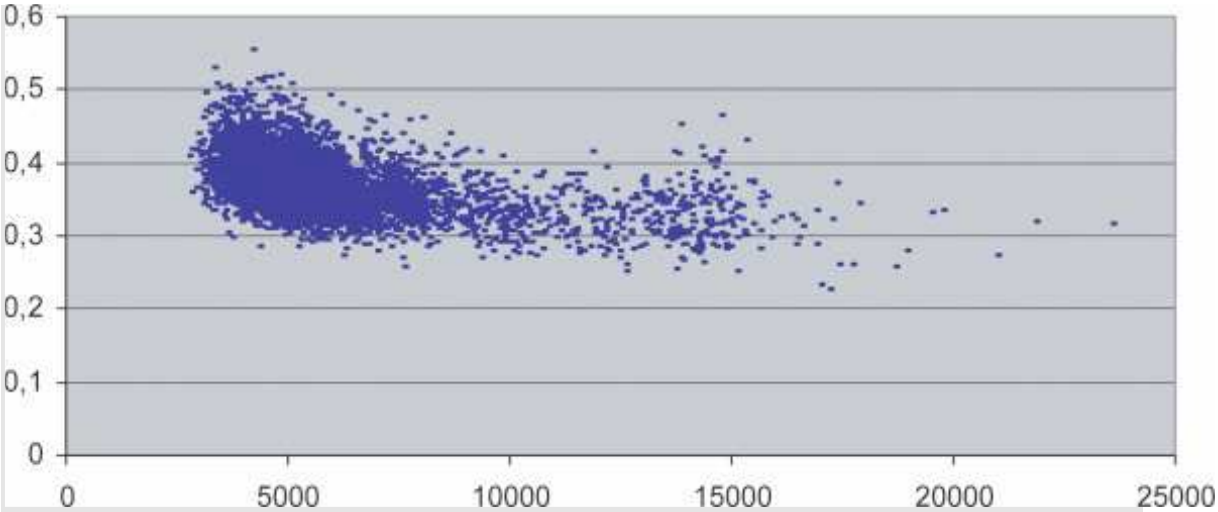


Fig. 4. ID and cumulative length of texts – Maupassant's Contes.

[6.9] *If we were to trust* ID *and its* standardized scale, *we should conclude that those 100 texts have been written by several different writers. Furthermore, we would remain unable to determine any more precise attribution unless some substantial amendment were made to the interpretation process. The huge number of incompatibilities (couples whose ID > 0.4), the frequency of which increases as fast as* Na + Nb *falls, discourages any reasonable clustering.*

7. First conclusions

[7.1] *Before examining how first DCL, and then DL, applied their proposition to the case of Corneille and Molière, let us summarize our prior observations.*

[7.2] *We have shown*

* *in Section 1, that when they advance their thesis, DCL do not lean on any theoretical reference concerning the key notions they deal with;*
* *in Section 3, that the proposed Inter-textual Distance Standardized Scale corresponds to no scientific standard, either in its making, or in its presentation;*
* *in Section 4, that the formula for ID includes two major biases, one patent – dependence towards Na, Nb, and Na + Nb – the other one underhand: uncontrolled conjuring away of distance factors (the least frequent items of the longer text). That second bias moreover involves a threshold effect, strong enough to discredit the whole method;*

** in Section 5, that the suggestion by DCL that texts should be lemmatised (which became a strong demand in Labbé, 2003) is formulated in such a way that it throws doubt upon any generalisation of this method, and upon any verification done by third parties;*

** in Section 6, that if we apply ID to actual cases, we are led to unacceptable and/or absurd conclusions. This is due to its biases on the one hand, to the naively discrete character of its interpretation scale, on the other hand.*

Voyage à Brobdingnag et retour à Lilliput

Les § 6.1 - 6.5 et la figure 4 sont de nouveaux voyages d'abord chez les géants (Brobdingnag), puis chez les nains.

Sachant que l'échelle de la distance s'applique à des textes dont les longueurs sont comprises entre 3 500 et 20 000 mots, *JQL 06* se place largement au-delà de la limite supérieure en utilisant des textes très longs (tout en n'expliquant pas les raisons de ces choix et en cachant la plupart des tailles au lecteur).

Voyage à Brobdingnag

Voici les longueurs des textes utilisés au début de cette section 6.

- Le comte de Monte-Cristo (A. Dumas) : 505 935 mots ;
- Joseph Basalmo (A. Dumas) : 469 457 mots ;
- l'Education sentimentale (G. Flaubert) : 152 890 mots ;
- Madame Bovary (G. Flaubert) : 122 660 mots ;
- Histoire des treize (H. de Balzac): 115 508 mots ;
- Bel Ami (G. de Maupassant) : 112 729 mots ;
- le Père Goriot (H. de Balzac) : 91 686 mots,
- le Médecin de campagne (H. de Balzac) : 87 696, etc.

La plus longue pièce du corpus Corneille-Molière compte 21 000 mots… *JQL 06* est clairement hors sujet, mais ce qu'il dévoile involontairement est bien intéressant.

Parmi ces œuvres les plus longues de la littérature française, on affirme avoir "trouvé" une distance inférieure à 0.200 entre deux textes d'auteurs différents : Madame Bovary (G. Flaubert) et Une vie (G. de Maupassant), textes qui auraient été lemmatisés suivant nos conventions (sans citer aucune source ni aucun logiciel).

Pour avoir traité tous les romans de G. Flaubert et de G. de Maupassant - notamment Madame Bovary et Une vie - nous pouvons affirmer que la ponctuation a été intégrée dans ce calcul. Sans la ponctuation, l'indice de la distance intertextuelle - calculé avec la formule de définition donnée dans l'annexe 2 sur les textes lemmatisés de Madame

Bovary et de Une vie - est égal à 0.208 : c'est la plus basse obtenue entre tous les couples possibles d'auteurs différents pour les textes cités dans *JQL 06*[46]. D'ailleurs, toutes les distances affichées dans les tableaux 1 et 2 (p. 62 de ce dossier) sont inférieures à la réalité et ne peuvent être approchées qu'en comptant les ponctuations comme des "mots".

Etant donné la longueur considérable des textes utilisés, ces "expériences" :

- ne concernent pas le cas Corneille - Molière ;
- sortent radicalement de la plage de validité de la distance intertextuelle ;
- confirment que la décroissance de l'indice de la distance intertextuelle - en fonction de l'allongement des textes - est extrêmement lente et ne justifie aucunement les conclusions de la section 7 ci-dessus (p. 64-65 de ce dossier) ;
- ne tiennent pas compte de la conclusion de notre présentation (Labbé & Labbé 2003, p. 114-115) selon laquelle, pour des textes de grandes tailles, la méthode doit être aménagée, du moins si l'on souhaite se servir de l'échelle des distances.

L'annexe 7 (p. 144 de ce dossier) donne les distances entre les œuvres romanesques de G. Flaubert et de G. de Maupassant ainsi que la classification arborée sur ce corpus. Contrairement à ce que suggèrent les § 6.1 et 6.2 de *JQL 06* (p. 61-62 de ce dossier), G. Flaubert se distingue très bien de G. de Maupassant. Ce corpus donne un bel exemple d'"inter-textualité" et apporte une réponse pragmatique à la question – posée dans le § 3.5 de *JQL06* (p. 17 de ce dossier) - de savoir comment mesurer l'influence d'un auteur sur un autre. Ce cas a été signalé dans : Labbé 2003 (p. 79). *JQL 06* n'a pas été capable de trouver d'autres exemples…

Une étrange comédie

L'"expérience" présentée dans le § 6.5 et 6.6 (p. 62-63 de ce dossier) - étude de la Comédie humaine de H. de Balzac - révèle trois choses.

Premièrement, *JQL 06* reconnaît que la distance intertextuelle est symétrique. En effet, avec 86 textes on forme 7 310 couples (86*85). Pour en afficher seulement 3 655, comme le fait JQL 06 dans son § 6.6, il faut diviser le nombre de couples par deux, c'est-à-dire admettre que $d_{(a,b)} = d_{(b,a)}$. (le résultat est le même de A vers B et de B vers

[46] Nos fichiers sont à la disposition de qui voudra vérifier ces informations.

A). Toute la section 4 - spécialement le § 4.9 (p 45 de dossier) - est invalidée.

Deuxièmement, faute d'indications bibliographiques (dans les § 6.5 et 6.6), il est impossible de refaire cette expérience[47].

Troisièmement, le coefficient de corrélation des rangs[48] requiert des conditions précises qui ne sont pas réunies ici.

- C'est un test "non paramétrique" utilisé quand il est impossible de calculer la moyenne et l'écart-type de l'une ou des deux variables étudiées (cas des variables nominales ou cardinales). Ici les deux variables (longueur des textes et indices de la distance intertextuelle) sont numériques : il est facile de calculer les moyennes et les écarts-types.

- Il est réservé aux cas pour lesquels on ne dispose pas d'hypothèse concernant le facteur explicatif d'une éventuelle co-variation des deux variables. Ici, on est exactement dans la situation inverse : il s'agit de vérifier l'hypothèse selon laquelle la longueur des textes explique les valeurs de la variable "distance".

- Quand l'une au moins des deux variables est numérique, il faut que ses valeurs s'échelonnent à peu près régulièrement, car le fait de remplacer ces valeurs par des rangs revient à postuler qu'elles sont séparées par des intervalles égaux. Les longueurs des romans de Balzac et leurs distances mutuelles ne remplissent pas cette condition.

- Il faut deux classements seulement pour chaque texte : un rang pour la longueur et un rang pour la distance. Or ici, pour chaque roman, il y a 85 distances différentes (avec tous les autres romans de ce corpus). Comment classer dans un rang unique chaque texte en fonction de 85 dimensions différentes ? En postulant que la moyenne des distances est représentative des 85 mesures et en classant les textes en fonction de cette moyenne. Pourquoi ne pas avoir suivi cette procédure recommandée par tous les manuels ? Est-ce parce que le résultat était décevant ?

Le coefficient de corrélation linéaire - dit de "Bravais-Pearson", présenté dans l'annexe 4 (p. 130 de ce dossier) - s'imposait, d'autant plus que, en cas de liaison avérée entre la longueur (variable explicative) et les indices de la distance (variable expliquée),

[47] Il y a eu du vivant de H. de Balzac trois éditions de la Comédie Humaine avec un nombre différent d'œuvres, puis plusieurs éditions posthumes. Aucune ne comporte 86 titres. Le catalogue de l'édition de 1845 (dernière de H. de Balzac) comporte 138 titres (voir : http://www.v1.paris.fr/musees/balzac/furne/).

[48] *JQL 06* prétend avoir utilisé le coefficient de Spearman mais, curieusement, il renvoie à un ouvrage qui présente un calcul assez différent : le coefficient des rangs de Kendall.

ce calcul permet de connaître la pente de la droite d'ajustement du nuage en fonction de la variable explicative (voir chapitre 1, p. 36-37 de ce dossier). Ce calcul permet aussi de juger de la qualité de l'ajustement.

Retour à Lilliput

Dans les § 6.7 et 6.8 (p. 63 de ce dossier), *JQL06* revient aux nouvelles de G. de Maupassant[49]. Dans ce corpus la taille médiane est de 1 991 mots, c'est-à-dire que la majorité des distances sont calculées sur des très petits textes dont un bon nombre contiennent du "baragouin" (comme Le Remplaçant).

La figure 4 (p. 64 de ce dossier) est erronée : il faut considérer les signes de ponctuation comme des mots pour atteindre les longueurs indiquées sur l'abscisse et générer l'essentiel de la "pente" apparente du nuage de points.

Le commentaire de *JQL 06* reproduit l'erreur commise à propos de Corneille et Molière dans la section 4 (p. 27-28 de ce dossier). Voici à titre d'illustration, le nuage de points des 105 distances du Remplaçant aux autres nouvelles (tableau V.1)[50].

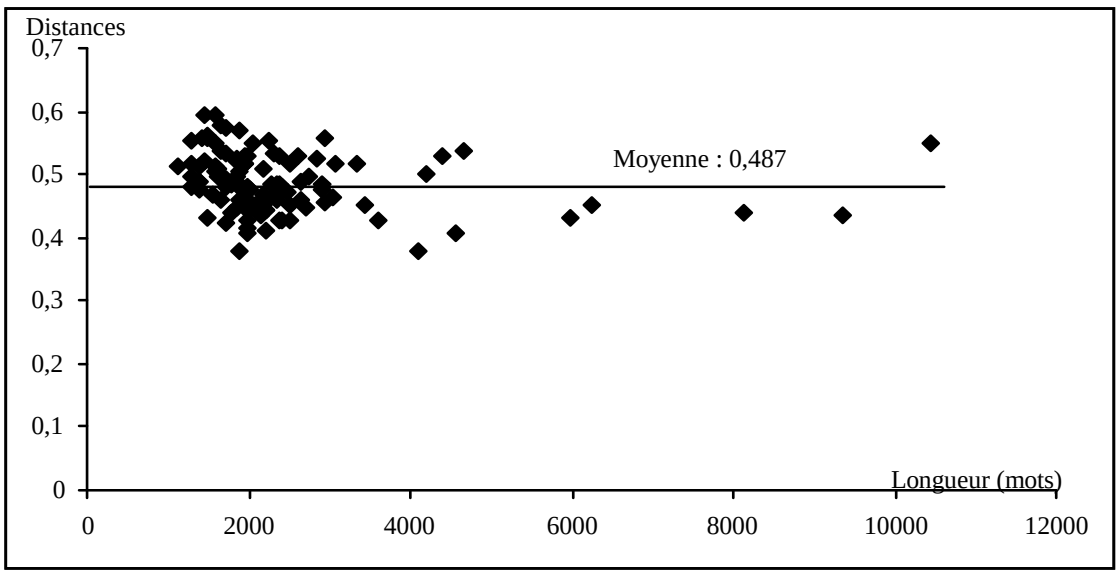


Tableau V.1 Distances entre Le Remplaçant (1 316 mots) et les 105 autres nouvelles de G. de Maupassant classées par longueurs croissantes

[49] Au passage, le § 6.8 affiche 4 950 distances, pour 100 textes, et non pas 9 900 (100*99), confirmant à nouveau la symétrie de l'indice.

[50] La construction des quatre tableaux ci-dessous obéit aux principes présentés dans le chapitre 2 (notamment, p. 34-35 de ce dossier).

Le coefficient de corrélation linéaire entre les longueurs et les distances est égal à – 0,069. Pour 104 degrés de liberté, la limite inférieure d'acceptation pour un tel coefficient est de ±0.195 (avec un risque de première espèce égal à 5% de chances de se tromper). Il n'y a donc aucune corrélation entre les longueurs des textes et leurs distances mutuelles[51]. Ce cas est pourtant au-delà des limites du raisonnable : texte très court, échelle des longueurs proche de 1/10 et étrangeté radicale par rapport aux autres…

Suivant la procédure déjà adoptée dans le chapitre 2, les deux cas extrêmes seront examinés. Le tableau V.2 présente le nuage de points correspondant au texte le plus court : La Folle (1 137 mots) et le tableau V.3, celui du texte le plus long (La Maison Tellier (10 455 mots). Ce sont les deux cas qui devraient mettre le mieux en valeur une hypothétique dépendance de la distance par rapport aux longueurs des textes comparés.

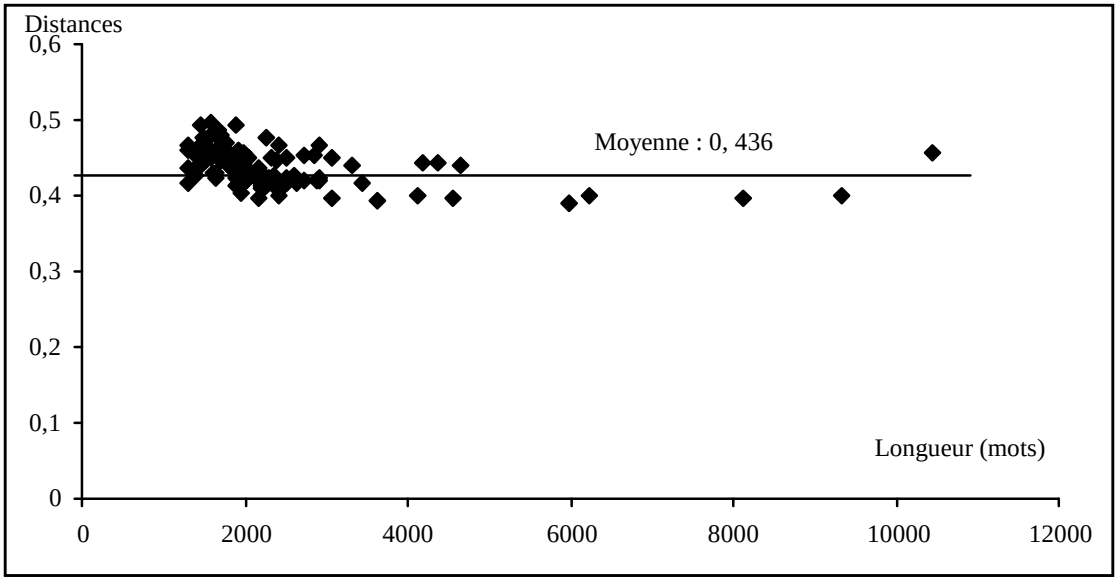


Tableau V.2 Distances entre La Folle (1 137 mots) et les 105 autres nouvelles de G. de Maupassant classées par longueurs croissantes

Ce texte extrêmement court est à la limite du rapport des longueurs avec le plus grand texte (1/10). Il n'y a pourtant pas de corrélation entre la longueur des textes comparés avec la Folle et les indices de leurs distances (tableau V.2) : le coefficient est

[51] Vu leur dimension, les tableaux de calcul. ne sont pas reproduits mais sont à la disposition des chercheurs.

égal à -0,112 alors qu'il devrait être au moins de -0,195 pour pouvoir examiner une éventuelle co-variation entre les deux variables.

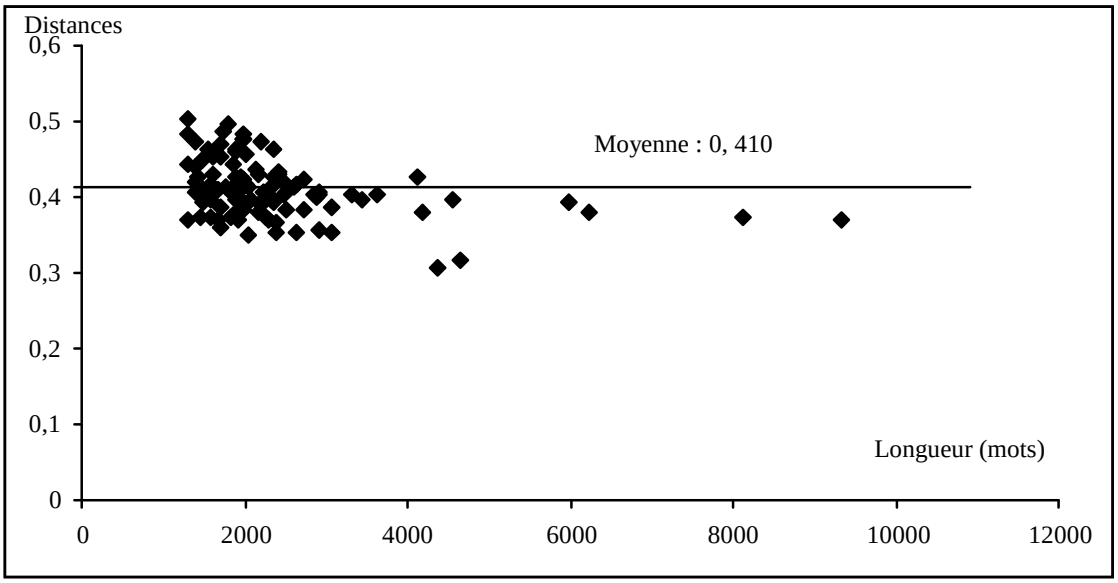


Tableau V.3 Distances entre la Maison Tellier (10 455 mots) et les 105 autres nouvelles de G. de Maupassant classées par longueurs croissantes

L'indice de corrélation linéaire correspondant aux données du tableau V.3 est égal à –0,03, ce qui conduit à la même conclusion que ci-dessus : pour la plus longue des nouvelles sélectionnées par JQL 06, il n'y a aucune liaison avérée entre les distances la séparant des autres et leurs longueurs respectives.

En superposant les 106 nuages de points, identiques à ceux qui viennent d'être présentés, on obtient le tableau V.4.

Les commentaires sous les tableaux II.10 et 11 (p. 40-41 de ce dossier) s'appliquent à ce graphique : forte dispersion, mélange de plusieurs populations et influence perturbatrice, dans les petites tailles, de quelques textes "étrangers" aux autres. Le Remplaçant (et ses semblables jouent ici le rôle que jouaient le Médecin volant, la Jalousie du barbouillé ou le Mariage forcé dans le corpus Corneille-Molière). A contrario, à l'extrême gauche, le plus petit texte – 1 137 mots dont du "franco-prussien" - se situe à un niveau moyen inférieur à la majorité des autres. Si la dépendance de l'indice par rapport à la taille était telle que *JQL 06* le prétend, une telle situation serait impossible…

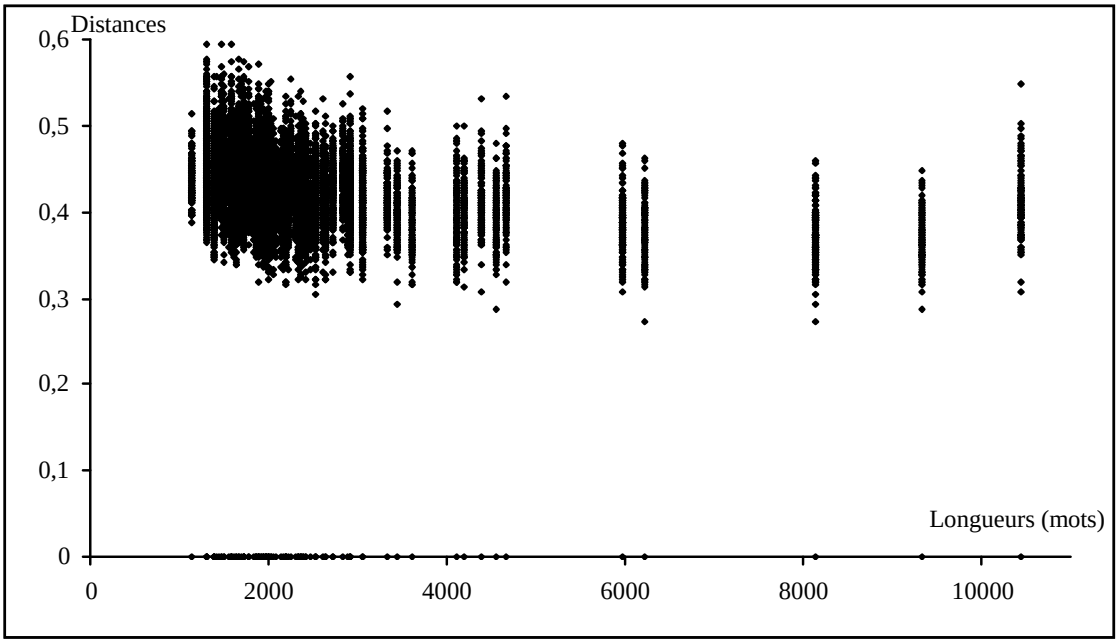


Tableau V.4 Distances entre les 106 nouvelles de G. Maupassant classées par longueurs croissantes.

Le tableau V.4 ci-dessus doit être comparé à la figure 4 de *JQL 06* (p. 64 de ce dossier). Sur l'axe horizontal de cette figure 4, les coordonnées sont les sommes des longueurs des deux textes ($N_a + N_b$) – augmentées de la ponctuation et baptisées à tort "longueurs cumulées". Ceci a pour effet d'étirer le nuage, de combler les "vides" à droite et, combiné avec l'introduction de la ponctuation dans le calcul, d'accentuer la "pente" apparente de ce nuage…

Contrairement à ce qu'affirme *JQL06* à la fin du § 6.9 (p. 64 de ce dossier), loin d'interdire une classification, ces résultats, solides et fiables, l'appelaient évidemment, en l'accompagnant d'une analyse du vocabulaire caractéristique des différents groupes. Malgré la petite taille de la plupart de ces textes, la présence de quelques individus aberrants, et l'hétérogénéité de l'ensemble, cette analyse aurait fait apparaître les thèmes favoris de G. de Maupassant et n'aurait pas été inutile à une meilleure connaissance de cet auteur.

Conclusions de la première partie

Dans deux des quatre "expériences" – sur Corneille et Molière (§ 4.2) puis sur les nouvelles de G. de Maupassant (§ 6.8) – *JQL 06* présente des graphiques (erronés) mais se garde du moindre calcul : pas de corrélation et aucun ajustement du nuage de points. Il n'a donc aucun résultat significatif à présenter concernant Corneille et Molière ou concernant les nouvelles de G. de Maupassant. En effet, le fait de renoncer à ces calculs standards signifie que l'on accepte l'indépendance des deux variables et – en cas de nuage étiré parallèlement à l'un des deux axes - que l'on admet que l'orientation de ce nuage ne s'écarte pas significativement de l'horizontale (ou de la verticale) ou - en cas de nuage curviligne – qu'aucune courbe ne l'ajuste correctement.

JQL 06 confirme qu'il est incapable, sur le corpus Corneille-Molière, comme sur les tous les autres, d'établir un lien significatif entre l'indice de la distance intertextuelle et les longueurs des textes (dans les limites que notre article *JQL 01* leur assigne).

En revanche, dans les § 6.5 et 6.6, une troisième "expérience" (sur la Comédie Humaine de H. de Balzac) établirait une corrélation (à l'aide d'un calcul erroné et incontrôlable). Le graphique n'est pas présenté, il n'y a aucun ajustement du nuage de points ni évaluation de la qualité de cet ajustement. Là encore, le fait de renoncer à ces procédures standards signifie que l'auteur n'est pas parvenu à établir la co-variation des deux grandeurs.

Au surplus, ces "expériences" ne concernent pas le cas Molière-Corneille. Logiquement, il aurait fallu utiliser d'autres pièces de théâtre, par exemple, J. Racine puisque ses pièces lemmatisées ont été mises dans le domaine public[52]… Tout au contraire, on est allé chercher des exemples le plus loin possible – tant par le genre, l'époque, que par la dimension des textes. Et aucun de ces exemples ne débouche sur quoi que ce soit de probant.

[52] Cf. annexe 3, p 127 de ce dossier.

JQL 06 confirme donc que :

La relation entre la distance intertextuelle et la longueur des textes n'a aucune influence sur l'attribution à Corneille de 16 pièces représentées sous le nom de Molière.

Partie II.

Enfin au port : Corneille et Molière

Dans la section 8 de *JQL 06*, il est enfin question de Corneille et Molière.

Cette section porte sur deux sujets distincts. C'est pourquoi elle est ici découpée en deux chapitres.

La première partie de cette section discute nos conclusions concernant Corneille et Molière. On y voit développée la thèse selon laquelle "on savait déjà" et confirmer ainsi l'étrange proximité entre les deux "auteurs" (chapitre 6).

Dans la seconde partie de la section 8, *JQL 06* présente une sorte de contre-épreuve "sans aucun biais" portant sur le corpus Corneille-Molière amputé de Psyché. Cette "expérience" confirme à nouveau nos principales conclusions concernant l'étrange proximité entre ces deux auteurs (chapitre 7).

Cette "expérience" reste inachevée. Nous l'achevons. Elle confirme la paternité de Corneille sur 16 pièces de Molière. Ce sera aussi l'occasion de présenter les raisons qui ont conduit à l'indice de la distance intertextuelle (chapitre 8).

Chapitre 6.

On le savait déjà ?

La section 7 de *JQL 06* affirmait : "La distance intertextuelle ne vaut rien".

La section 8 change brusquement de thèse et soutient que :

"Tout le monde savait ce que les Labbé pensent avoir découvert"…

Notre réponse montrera que ce n'est pas du tout exact et, au passage, d'autres anomalies seront dévoilées.

---

8. Application to Corneille and Molière

[8.1] *Given all this, we may suspect that the application of DCL's method to the case of Corneille et Molière, which is mentioned in the very title of their paper, should be questionable. That application is laid out in two successive chapters:* Molière's plays, *then* Corneille and Molière.

[8.2] *In* Molière's plays, *DCL begin with a selection of eight plays, which they present as Molière's best-known plays (further "main masterpieces"). This cannot fail to astonish a scientific mind, since no notoriety criterion is made explicit. Moreover, in the table presented, DCL fail to mention that, of those eight plays, four are written in verse and four are not; they do not use that distinction for their results. Mentioning this fact would have made apparent the strong influence of versification upon ID. This influence is linked to the lexical restriction which occurs in the rhyme position, even more particularly when constrained by a genre and a century. If we want a clear idea of that effect, we just have to sort the plays* according to that criterion. *Then we establish the mean values of every block. This works even within the limits of DCL's restrictive selection.*

Table 3. Rearrangement of DCL's Table 2, putting the versified plays together (MI: Le Malade imaginaire).

	T	**M**	**FS**	**DJ**	**A**	**BG**	**MI**
Ecole des Femmes (EF)	0.183	0.194	0.198	0.205	0.200	0.231	0.223
Tartuffe (T)		0.167	0.170	0.199	0.199	0.230	0.219
Le Misanthrope (M)			0.173	0.204	0.210	0.239	0.239
Les Femmes savantes (FS)				0.219	0.214	0.234	0.226
Dom Juan (DJ)					0.170	0.207	0.205
L'Avare (A)						0.194	0.187
Le Bourgeois gentilhomme (BG)							0.196

[8.3] *Mean* ID *is, between the versified plays, 0.181; between those in prose, 0.193; for the other pairs, 0.218. That observation seems important enough to be noted.*

[8.4] *Then DCL present another table, containing the* overall distances *(which would be more correctly named* mean distances *as it is done in the paper itself). They quickly draw the following summary conclusion: "except for these few plays [those presenting the highest mean ID], it is quite*[8] *certain that all the work is from a single author".*

[8.5] *Then, in* Corneille and Molière, *DCL do essentially two sorts of things. On the one hand, they apply to the* ID *matrix for the 67 plays of the joint corpus (33 of Corneille's, 32 of Molière's, and both versions of* Psyché*) two synthetic analysis methods*: cluster analysis *and* tree classification. *We may remind the reader that the data submitted to those analyses are biased.*

[8.6] *Let us carry on regardless of those biases, and consider the results. They tell specialists on 17th century theatre precisely what they already know*

1. that Molière's play Dom Garcie de Navarre, *because of its genre (it is his only* heroic comedy), *is related to Corneille's texts of the same genre*;

2. *that Corneille's last two comedies (the* Menteurs) *are more related to Moliere's comedies (especially to his versified ones) than to Corneille's early comedies;*

[8] *Once again, we may note the highly equivocal adverb (see Section 3).*

3. *that both versions of* Psyché, *which strongly intersect, are rather eccentric (being written by several hands, a well recognised fact).*

[8.7] *But when DCL write (2001, p. 228) that on the tree-analysis graph, the "Menteurs (15-16) stand quite*[9] *in the centre of Molière's works", they overvalue a pure graphical artifice. Actually, even with their biased data, DCL should limit themselves to the statement that 15 and 16 are attached to the* Molière's verses *cluster. Indeed, both arrangements (Fig. 5) are strictly equivalent as graphs from the same tree-analysis: the left one is the one published by DCL in JQL, the right one is a graphic variant from exactly the same results. One can see that the right one* does not *suggest that the Menteurs are "in the centre" of Molière's works.*

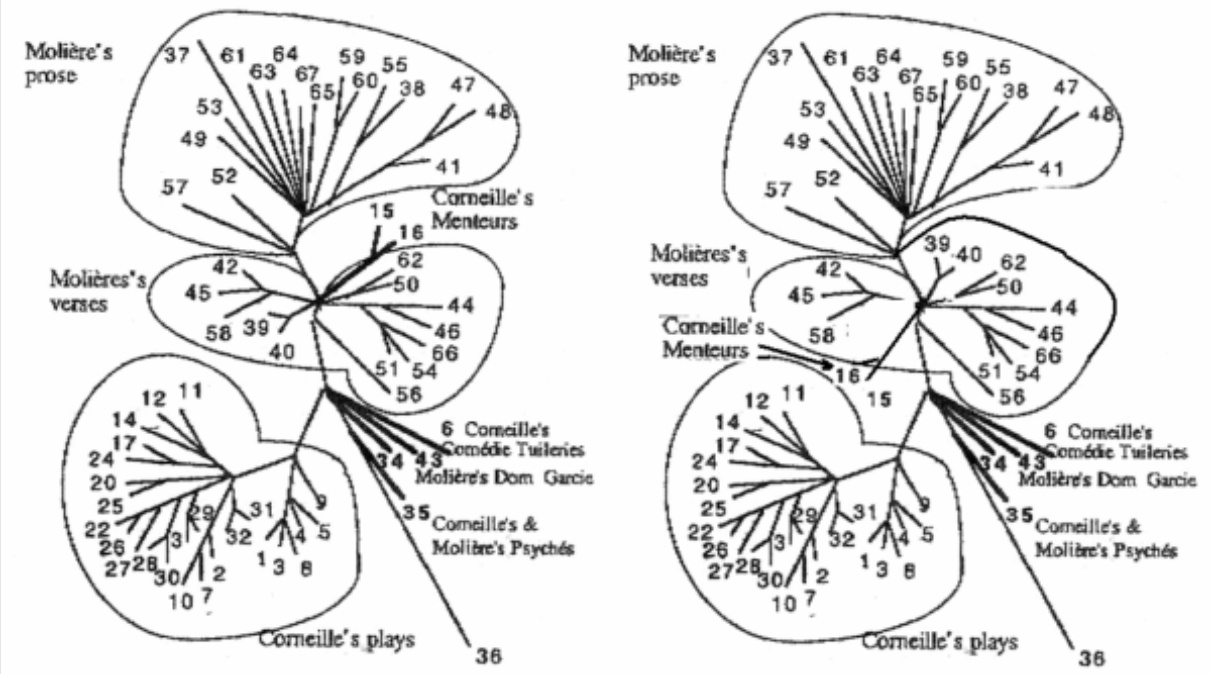


Fig. 5. Two equivalent graphs (but one more suggestive than the other) of DCL's tree-analysis.

[8.8] *Only one serious conclusion could be drawn from this analysis, if the matrix data were not biased. The* Menteurs *are indeed significantly nearer, following only that criterion, to Molière's versified comedies than to his prose comedies and than to*

[9] *Ter relaps, see note 8.*

Corneille's early comedies. As for interpreting that graph as DCL do: "in other words, the Menteurs *authorship is clearly the same as most of Molière's masterpieces" – this is akin to a conjuring trick.*

[8.9] *Where are the control contrasting analyses? Why did DCL not contrast, for instance Flaubert's and Maupassant's or Dumas' and Balzac's works? Did they clearly prove that such a phenomenon cannot be met elsewhere, among distinct but notoriously related authors, to various extents?*

Suite de la section 8 : p. 87 de ce dossier.

Etrange revirement

La section 8.14 de *JQL 06* (reproduite dans le prochain chapitre), prétend qu'il existe des "Moliéristes" et des "Corneillistes". La lecture du § 8.6 ne fera peut-être pas très plaisir à ces gens, du moins aux "Moliéristes", car ce paragraphe reprend l'essentiel de nos constatations concernant les proximités étonnantes entre Corneille et Molière[53].

Palinodies

Les paragraphes 8.2 à 8.5 et le tableau 3 de *JQL 06* (p. 77-78 de ce dossier) utilisent la distance intertextuelle pour mesurer (jusqu'à la troisième décimale) l'influence du "genre" sur la parenté relative des pièces les unes par rapport aux autres. Les valeurs moyennes données dans le § 8.3 coïncident avec celles prévues par l'échelle normalisée des distances (reproduite p. 17 de ce dossier). Ainsi,

JQL 06 confirme : l'indice et l'échelle la distance intertextuelle sont fiables et leur application aux œuvres de Molière et Corneille est correcte.

Les alinéas 1 et 2 du § 8.6 (p. 78 de ce dossier) soutiennent que les proximités entre Corneille et Molière seraient dues aux "genres" des pièces concernées (comédies en vers et comédies héroïques) :

JQL 06 utilise les termes genre *et* thème *sans renvoyer à aucune théorie littéraire et/ou linguistique. La même chose peut être dite à propos du concept d'*auteur *et plus généralement à propos de la manière criticable dont est traitée l'histoire littéraire*[54].

Enfin, les spécialistes du théâtre du XVIIe connaissaient déjà toutes nos conclusions… Pourquoi ne pas citer ces "spécialistes" ?

[53] La section 8 comporte de nombreuses attaques infondées (notamment § 8.9). L'annexe 9 y répond.

[54] *"Furthermore, JQL 06 uses the terms* genre *and* theme *without referring to any literary and/or linguistic theory. The same can be said about the concept of* author *and generally about his way of dealing with literary history, which is also deserving of criticism" (JQL 06,* § 1.4, p. 13-14 de ce dossier).

Trois confirmations éclatantes

Si aucun spécialiste du 17e n'est cité, c'est parce que personne n'a écrit quelque chose de comparable aux trois conclusions énoncées dans le § 8.6 (p. 78-79 de ce dossier), du moins avant la parution de notre article en décembre 2001.

1. PERSONNE n'a écrit que Dom Garcie de Navarre (représenté sous le nom de Molière) est "apparenté" aux comédies héroïques de Corneille. Trois auteurs seulement (G. Couton, C. Despois et F. M. Warren) ont souligné quelques ressemblances - les noms des personnages et quelques expressions - entre Dom Garcie et le seul Don Sanche (de Corneille). Tous les autres spécialistes ont jugé superficielles ces ressemblances et les ont expliquées par une sorte d'imprégnation inconsciente, Molière acteur ayant beaucoup joué les pièces de Corneille. Seule B.-M. Kylander (1995, p 183-194) a signalé certaines proximités lexicales et stylistiques avec les "comédies héroïques" de Corneille, mais elle rejoint tous les autres spécialistes pour estimer que Corneille n'a même pas été une source d'inspiration pour Molière.

2. NULLE PART, il est écrit que les deux *Menteurs* (Corneille) sont "apparentés" aux comédies en vers représentées sous le nom de Molière. Si l'on en croit M. Bourqui (1999), qui présente une synthèse des travaux sur les sources de ces pièces de Molière, il existe toutefois un article (paru en 1985 dans le Kentucky Romance Quaterly), suggérant que le valet sentencieux et couard de Dom Juan (Molière) serait inspiré du même personnage des Menteurs (Corneille). Le parallèle a été rejeté, comme trop banal, par l'ENSEMBLE des spécialistes.

3. Il y a UNE SEULE "version" de Psyché et personne n'a écrit que cette pièce est "plutôt excentrique". En revanche, l'annexe 10 (p. 160 de ce dossier) nomme les "différentes mains" que *JQL 06* maintient maladroitement dans l'anonymat : Quinault, Molière et **Corneille**, ce dernier ayant versifié les deux tiers de la pièce. Cette pièce a été le plus éclatant succès de Molière de son vivant. On lira partout qu'elle possède une grande unité qui est généralement attribuée à Molière (officiellement, Corneille a simplement achevé la versification).

Le prochain chapitre expliquera pourquoi *JQL 06* cache ces informations tout en confirmant nos principales conclusions.

Influences comparées du genre et du temps dans le corpus Corneille-Molière

Le tableau 3 (p. 78 de ce dossier) utilise les données qui avaient été classées par ordre chronologique (tous les tableaux de notre article de *JQL01* sont classés dans l'ordre des éditions de référence qui est à peu près l'ordre chronologique).

La réorganisation fait ressortir que les différences de genre - y compris prose *vs* vers – augmentent les distances sans effacer l'identité de l'auteur (contrairement au présupposé du § 8.6). En revanche, cette réorganisation masque une dimension importante : l'influence du temps.

Le temps est un facteur aussi important que le genre, spécialement dans le cas de Corneille et Molière, car la création de ces pièces s'étend sur près de 45 ans. Pour illustrer l'influence du temps dans l'œuvre de Corneille, on peut prendre quelques exemples de pièces contemporaines et régulièrement échelonnées (tableau VI.1). Les dates de création figurent entre parenthèses.

	Menteur	Suite du Menteur	Pompée	Dom Garcie	Misanthrope
Menteur(1642)	0	0,180	0,281	0,289	0,252
Suite du Menteur (1643)	0,180	0	0,280	0,275	0,233
Pompée (1642)	0,281	0,280	0	0,263	0,301
Œdipe (1659)	0,254	0,250	0,218	0,223	0,269
Toison d'Or (1661)	0,258	0,254	0,216	0,221	0,268
Sertorius (1662)	0,245	0,238	0,222	0,230	0,247
Tite et Bérénice (1670)	0,248	0,232	0,249	0,227	0,234
Pulchérie (1672)	0,235	0,223	0,250	0,230	0,230
Surena (1674)	0,244	0,230	0,250	0,216	0,237
Etourdi (1660)	0,205	0,206	0,289	0,252	0,234
Dépit amoureux (1660)	0,215	0,212	0,299	0,243	0,222
Dom Garcie (1661)	0,289	0,275	0,263	0	0,230
Ecole des Femmes (1662)	0,226	0,217	0,303	0,261	0,205
Misanthrope (1666)	0,252	0,233	0,301	0,230	0
Femmes savantes (1672)	0,260	0,248	0,347	0,247	0,183

Tableau VI.1 Extrait de la matrice des distances pour les deux Menteurs, Pompée (Corneille), Dom Garcie et Misanthrope (Molière).

Corneille a écrit le Menteur (comédie en alexandrins) et Pompée (tragédie en alexandrins) durant l'hiver 1641-1642 et la Suite du Menteur dans les mois qui suivent.

Ces deux comédies sont très éloignées de Pompée (tragédie) et leurs distances sont exceptionnellement élevées (0,280 et 0,281) pour des pièces en alexandrins écrites par un même auteur, en même temps, dans des genres différents[55].

Il s'écoule ensuite :

- 17 ans entre Pompée et Œdipe : la distance entre ces deux tragédies en alexandrins, toutes deux d'un même auteur (Corneille), est de 0,218 ;

- 18 ans entre le Menteur (Corneille) et l'Etourdi (Molière) : la distance entre ces deux comédies en alexandrins, supposées être d'auteurs différents, est de : 0, 205.

Il est très rare de trouver des distances aussi faibles, chez un même auteur, entre des œuvres écrites dans un tel intervalle de temps. Par exemple, il s'écoule :

- 12 ans entre l'Etourdi et les Femmes savantes : la distance entre ces deux comédies en vers du même auteur officiel (Molière) est de 0,238 ;

- 30 ans entre Pompée et Pulchérie : distance entre ces deux tragédies (toutes deux en alexandrins par Corneille) : 0,250 ;

- 30 ans entre la Suite du Menteur (Corneille) et les Femmes savantes (Molière), comédies supposées être d'auteurs différents : 0.248.

On constate également :

- des proximités significatives entre Dom Garcie ou le Misanthrope (comédies, Molière) et les tragédies contemporaines de Corneille (Sertorius, Tite, Pulchérie, Suréna) ;

- un rapprochement progressif entre les Menteurs (Corneille, 1641-1642), les tragédies tardives de Corneille - de Œdipe (1659) à Surena (1674) – et Dom Garcie (pour Molière, 1661), le Misanthrope (pour Molière, 1666) et Psyché (pour Molière, 1671), au fur et à mesure que l'auteur unique vieillit…

Tout se passe donc comme si le travail pour Molière "déteignait" sur l'œuvre officielle de Corneille. Le même phénomène est observable dans les derniers romans "officiels" de R. Gary, progressivement "contaminés" par ceux qu'il publiait, à la même

[55] Le tableau 3 de *JQL 06* en témoigne a contrario. Voir également, l'annexe 3.4 (p. 129 de ce dossier).

époque, sous le nom de E. Ajar[56].

A titre de comparaison, voici les distances séparant les 2 premières et les 3 dernières tragédies de Racine (tableau VI.2). Il s'écoule 13 ans entre la Thébaïde (première pièce de Racine) et Phèdre ; 25 ans entre la Thébaïde et Esther et 27 ans entre la première et la dernière pièce (Athalie).

	Thébaïde (1664)	Alexandre (1665)
Phèdre (1677)	0.275	0,266
Esther (1689)	0.317	0,315
Athalie (1691)	0.295	0,295

Tableau VI.2 Extrait de la matrice des distances entre les tragédies de J. Racine[57].

Etant donné le temps écoulé entre la création des Menteurs et celles des pièces représentées sous le nom de Molière, les distances enregistrées entre les 2 Menteurs (Corneille) et les comédies en vers représentées sous le nom de Molière (ainsi que le Dom Juan et l'Avare) sont les plus petites que l'on puisse rencontrer chez un auteur unique pour des textes écrits dans un même genre.

[56] Voir annexe 8 (p. 144 de ce dossier). Quelques résultats ont été publiés dans : Lafon & Peeters 2006, p. 309-313.
[57] Le tableau complet concernant l'œuvre de J. Racine se trouve dans : Labbé & Labbé 2006, p. 323-324.

Chapitre 7

Couvrez Psyché que je ne saurais voir…

TARTUFFE.
Couvrez Psyché que je ne saurois voir :
Par de pareils objets les âmes sont blessées,
Et cela fait venir de coupables pensées.
(D'après Pierre Corneille, *L'Imposteur*)

La seconde partie de la section 8 de *JQL 06* présente une analyse graphique "sans aucun biais" portant sur le corpus Corneille-Molière. Ce corpus a été amputé des deux seuls textes qu'il ne fallait pas retirer : les deux parties de Psyché officiellement écrites par Molière et… Corneille.

Les deux points correspondants étaient des témoins gênants : leur proximité signale une fois de plus que Corneille et Molière ne sont en réalité qu'un même et unique auteur.

Notre réponse rétablit les deux points et achève l'analyse. Ce sera aussi l'occasion de rétablir une autre vérité : dès les premiers succès de Molière, des voix informées ont affirmé clairement qu'il n'est pas l'auteur des œuvres jouées et imprimées sous son nom.

---

8. Application to Corneille and Molière

[8.10] *We will compare the result obtained so indirectly and hazardously by DCL with one from a classical Correspondence Analysis. For the CA principles in the context of text analysis, see for example Lebart, Salem and Berry (1998). Here, the submitted matrix is a very large table (6200 lines, 65 columns). It contains the distribution of all lemmas (hapaxes excepted) in the 65 plays in question (*Psyché *in its two versions has*

been withdrawn). The analysed data are therefore strictly observed frequencies and we could thus integrate 99.7% of all occurrences.

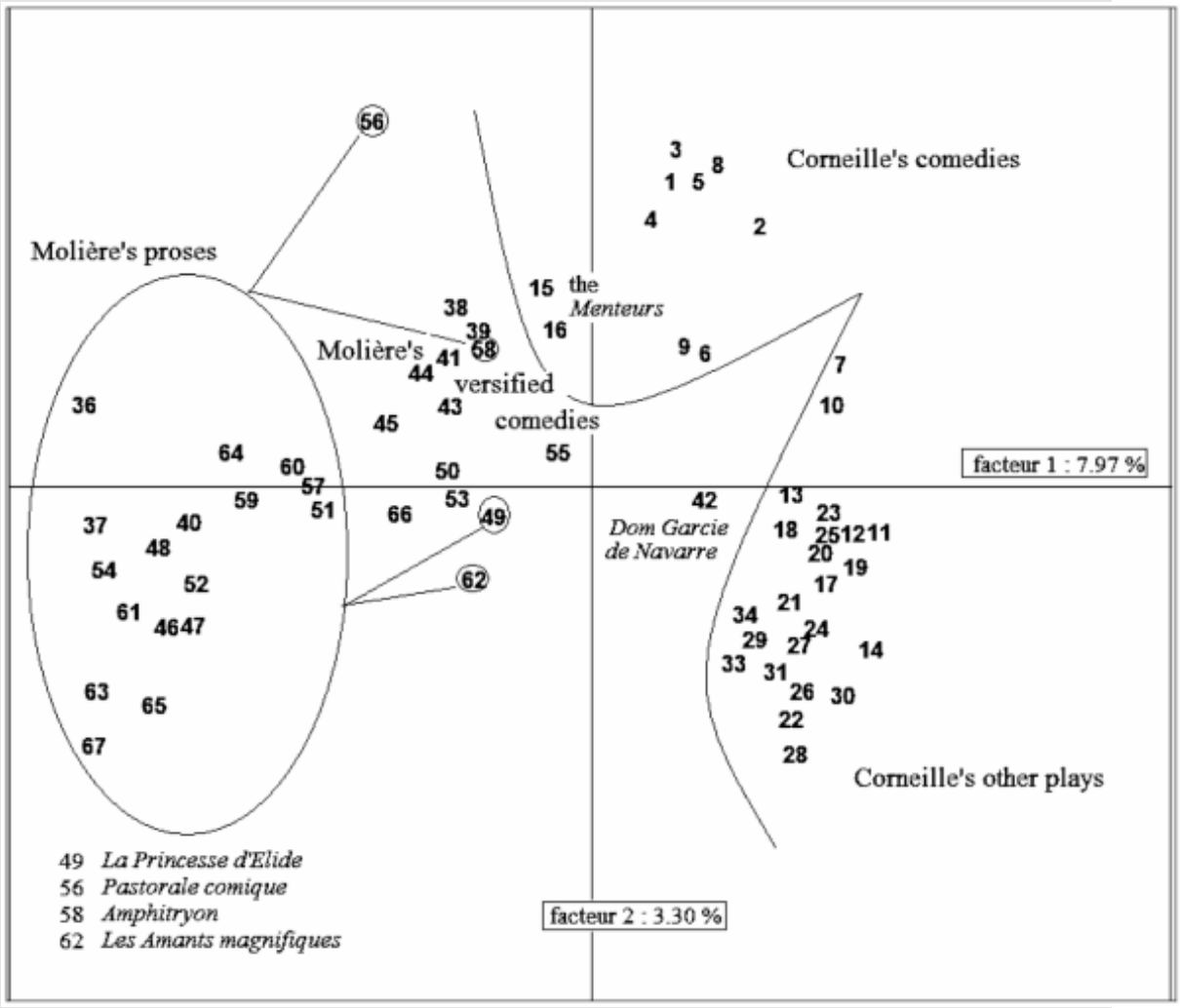


Fig. 6. CA graph of distribution of all lemmas in all the plays of the corpus (columns only shown).

[8.11] *This graph indicates very well the medium position of the* Menteurs, *kinship of* Dom Garcie *with Corneille's tragedies and tragi-comedies, and even an interesting position of* Mélicerte (55).

[8.12] *It is worth noting the locations of* Dom Juan *(51) and* L'Avare *(60), which DCL attribute with certainty to Corneille. This graph presents, just more clearly and without any bias, the data which are* grosso modo *on DCL's tree-analysis graph. Who would interpret this as a proof of Corneille's authorship of 16 plays of Molière?*

[8.13] *On the other hand, DCL produce a table (5) of* ID *between the* Menteurs *and each of Molière's plays. What essential would those data show if they were not biased? That the* ID *are regularly spaced from 0.205 to 0.341 with* Le Menteur, *from 0.206 to 0.331 with* La Suite du Menteur. *We particularly notice that no* ID *is lower than 0.2. That does not prevent DCL from concluding in favour of a sure attribution to Corneille of all Molière's versified plays, as well as of* Dom Juan, *and* L'Avare.

[8.14] *In order to justify their spectacular intervention into the field of literary studies, DCL claim (2001, p. 220) that "From the very beginning, it was rumoured that Molière was not the writer of his plays." Overall, they claim that "Since then, the problem has been discussed many times." Indeed, it was raised three times in total: at the beginning of the 20th century by Pierre Louÿs, a French poet; in 1957 by Henry Poulaille, a French writer; in 1990, by two lawyers, Hippolyte Wouters and Christine de Ville de Goyet. Their theses are, moreover, fairly different one from another. Overall, DCL omit this key point: so far not a single specialist scholar of French classical theatre, or even of the 17th century or of theatre in general, in France or in the whole world, ever validated those hypotheses. Labbé (2003) evokes a silent plot, organized by Molierists and/or Corneillists. That suspicion would perhaps be more justified if the "problem" was more recent and if relevant specialists were not counted by hundreds, all around the world.*

On a retrouvé Psyché !

La section 8 de *JQL 06* présente une analyse du corpus Corneille-Molière - "analyse des correspondances classique" et "sans biais"[58] - qui débouche curieusement sur l'affirmation qu'il ne peut pas y avoir de réponse.

Cette analyse a été refaite et rétablit Psyché qui avaient disparu dans la figure 6 de *JQL06*. Quelques questions préalables permettront de comprendre les raisons de cette disparition.

Cachez ces détails…

La légende de sa figure 6 (p. 88 ci-dessous) indique que l'analyse porte sur : "all lemmas in all the plays of the corpus". C'est doublement faux : il manque près de 4 vocables sur 10 et deux textes ont été retirés.

Première contre-vérité : "all lemmas".

Dans le § 8.10 (p. 87 ci-dessus), il est écrit : "hapaxes excepted (…) a very large table (6 200 lines, 65 columns)". Dans ce tableau, les vocables – que *JQL 06* appelle "lemmes" - sont en ligne (il y en a donc 6 200) et les textes en colonnes. Or le corpus Corneille-Molière comporte 9 947 vocables[59]. On a donc enlevé :

9 947 - 6 200 = 3 747 vocables soit 38% du vocabulaire…

Deuxième contre-vérité : "all the corpus". Dans la citation ci-dessus, il est affirmé que le tableau de calcul comporte 65 colonnes. Or il y a 67 (ou 68) textes dans le corpus… Dans ce même paragraphe 8.10, on lit : "Psyché in its two versions has been withdrawn". Il manque donc deux textes dans le corpus. De plus, il n'y a qu'une version de Psyché, publiée du vivant de Molière et indiscutée depuis lors, et l'on sait précisément ce que chacun des trois auteurs (Molière, Quinault, Corneille) est censé

[58] Le programme utilisé n'est pas mentionné, ce qui rend difficile la reproduction de cette "analyse des correspondances".
[59] Ce chiffre (9 947 vocables) est indiqué par *JQL 06* au § 5.3 (p. 54 de ce dossier).

avoir écrit (annexe 10, p. 160 de ce dossier).

Les deux (ou trois) parties de Psyché étaient donc les seuls textes qu'il fallait absolument conserver car elles pouvaient départager les deux thèses en présence :

- si Corneille et Molière sont deux auteurs différents, un algorithme de classification "sans biais" doit leur rendre, sans équivoque, les passages qu'ils ont écrits ;

- si Corneille et Molière sont un même auteur, les deux parties seront rattachées à celui qui a effectivement tenu la plume…

On peut "soupçonner" la réponse en voyant la manière dont *JQL 06* sort Psyché de l'analyse - subrepticement, sans aucune explication, et après avoir maladroitement caché en § 8.6 les noms des co-auteurs.

Pour comprendre la figure 6 (p. 88 ci-dessus) et suivre son commentaire (§ 8.11-8.12), le lecteur ne dispose d'aucune table pour établir une correspondance entre les points de cette figure et les pièces. Or il y a des anomalies évidentes dans cette figure 6.

Par exemple, la manière dont sont tracées les lignes séparant les groupes de pièces : ces lignes ne respectent pas les principaux espaces entre les nuages de points. Il est évident que le n° 42 appartient au groupe des tragédies de Corneille et que les Menteurs de Corneille (n° 15 et 16) se rattachent aux comédies en vers de Molière[60]. De plus, les espèces de tentacules, à la gauche de la figure, signalent un défaut rédhibitoire dans l'analyse (le prochain chapitre explique ces anomalies).

Autre exemple, on prétend que l'analyse porte sur 65 pièces - 32 pièces de Molière et 33 de Corneille (voir annexe 1, p. 122 de ce dossier) - or il y a :

- un point n° 34 dans le "cluster" Corneille (dans le quart sud-est au dessus du point 29) ;

- un point n° 66 (dans le quart sud-ouest, au milieu et au plus près de l'axe) et un point n° 67 (le point le plus en bas et le plus à l'ouest).

Deux pièces inédites ?

En fait, les points n° 32 et 35 manquent sur la figure 6 (p. 88 de ce dossier).

L'annexe 1 (p. 122, colonne "n° *JQL 06*"), identifie chaque point du graphique. Les points n° 32 et 35 correspondent aux parties de Psyché censées avoir été rédigées respectivement par Corneille et Molière.

[60] Confirmation plus bas, p. 94-96 de ce dossier.

Pourquoi *JQL 06* les a-t-il retirés de la figure 6 ?

Cachez ces deux points…

La même analyse a été refaite à l'aide du logiciel R (version 2.4)[61]. Les hapax (les vocables n'apparaissant qu'une fois dans le corpus total) ont été retirés, comme dans l'expérience de *JQL 06*. En revanche, les deux fichiers correspondant aux deux parties de Psyché sont maintenus (n° 32 : Psyché par Corneille ; n° 35 : Psyché par Molière). Le tableau VII.1 (page suivante) présente le graphique issu de cette analyse.

Par rapport à la figure 6 (p. 88 de ce dossier), les graduations ont été rétablies et le barycentre se trouve à sa vraie place (la figure est nettement étirée vers la gauche et vers le haut pour des raisons que le prochain chapitre permettra de comprendre). En négligeant ces détails importants, la figure 6 de JQL 06 et le tableau VII.1 sont semblables : les points sont situés aux mêmes emplacements, les pourcentages d'inertie, inscrits sur les deux axes, sont comparables.

MAIS LE GRAPHIQUE VII.1 COMPORTE 67 POINTS…

Car sur la figure 6 (p. 88 ci-dessus), les points n° 32 et 35 se trouvent exactement SOUS l'étiquette "Dom Garcie" (n° 42) !

Les deux parties de Psyché (par Corneille et Molière) sont donc très proches l'une de l'autre, comme de Dom Garcie de Navarre (officiellement par Molière) et des tragédies de Corneille dites "de la maturité".

Dans le troisième alinéa du § 8.7 de *JQL 06*, on lit :

Les deux versions de Psyché, qui se recouvrent fortement, sont plutôt excentriques (ayant été écrites par plusieurs mains, ce qui est un fait bien connu).

Comment l'auteur sait-il que les deux parties de Psyché sont fort proches s'il n'a pas fait l'expérience ? Il a donc décidé de retirer ces points gênants… mais il a laissé, dans la numérotation, deux vides très parlants.

JQL 06 valide nos conclusions : P. Corneille est l'unique auteur de Psyché et de Dom Garcie.

[61] Téléchargé à partir du site : http://www.R-project.org. Le script de l'expérience est à la disposition des chercheurs qui le désireraient. Le prochain chapitre présente l'analyse des correspondances.

Tableau VII.1 Analyse des correspondances sur les 67 textes du corpus Corneille et Molière (pour les numéros des pièces, se reporter à l'annexe 1, p. 123, colonne "n° *JQL 06*").

Cachez cette classification…

Lebart, Salem & Berry expliquent que la "classification automatique" est un "complément indispensable" à toute analyse des correspondances (1998, p 81-82). Une méthode est spécialement recommandée : la classification automatique ascendante ("hierachical cluster analysis", ibid p. 82-100), la constitution des classes étant retracée par un "dendrogramme" (c'est-à-dire un arbre). C'est exactement la méthode appliquée dans notre article de *JQL01* (p. 224).

Pourquoi *JQL 06* n'a-t-il pas effectué cette classification alors que le seul manuel cité en souligne longuement l'importance ?

Le tableau VII.2 (page suivante) présente le résultat de cette classification, réalisée sur la matrice des distances dans le premier plan factoriel, à l'aide du logiciel R. Les résultats de cette classification sont reportés sur le tableau VII.3 (les points sont les mêmes que sur le tableau VII.1). Les principaux groupes de pièces y sont délimités - non pas selon les choix de l'opérateur, comme sur la figure 6 (p. 88 ci-dessus) - mais suivant le meilleur classement possible.

Sur le dendrogramme, chaque pièce (ici représentée par son numéro[62]) est symbolisée par un trait vertical. L'algorithme classe les textes, puis les groupes de textes, en fonction de leurs distances du Chi2. Plus le trait horizontal joignant ces pièces est bas, plus elles sont proches, plus le trait est élevé, plus les textes (ou groupes de textes) qu'il joint sont éloignés donc hétérogènes.

Deux "clusters" principaux s'opposent.

A gauche, le groupe A comporte les 16 pièces en prose de Molière, du n° 36 (Jalousie du Barbouillé) au n° 54 (Médecin malgré lui). Il est extrêmement loin de tout le reste (la barre verticale qui l'unit à l'autre ensemble est très élevée).

A droite, un vaste ensemble formé d'un groupe très homogène (B) et d'un sous-ensemble de deux groupes moins homogènes (E).

[62] Nous respectons la numérotation probable de *JQL 06* (voir annexe 1, p. 123 de ce dossier et légende du dendrogramme p. 96).

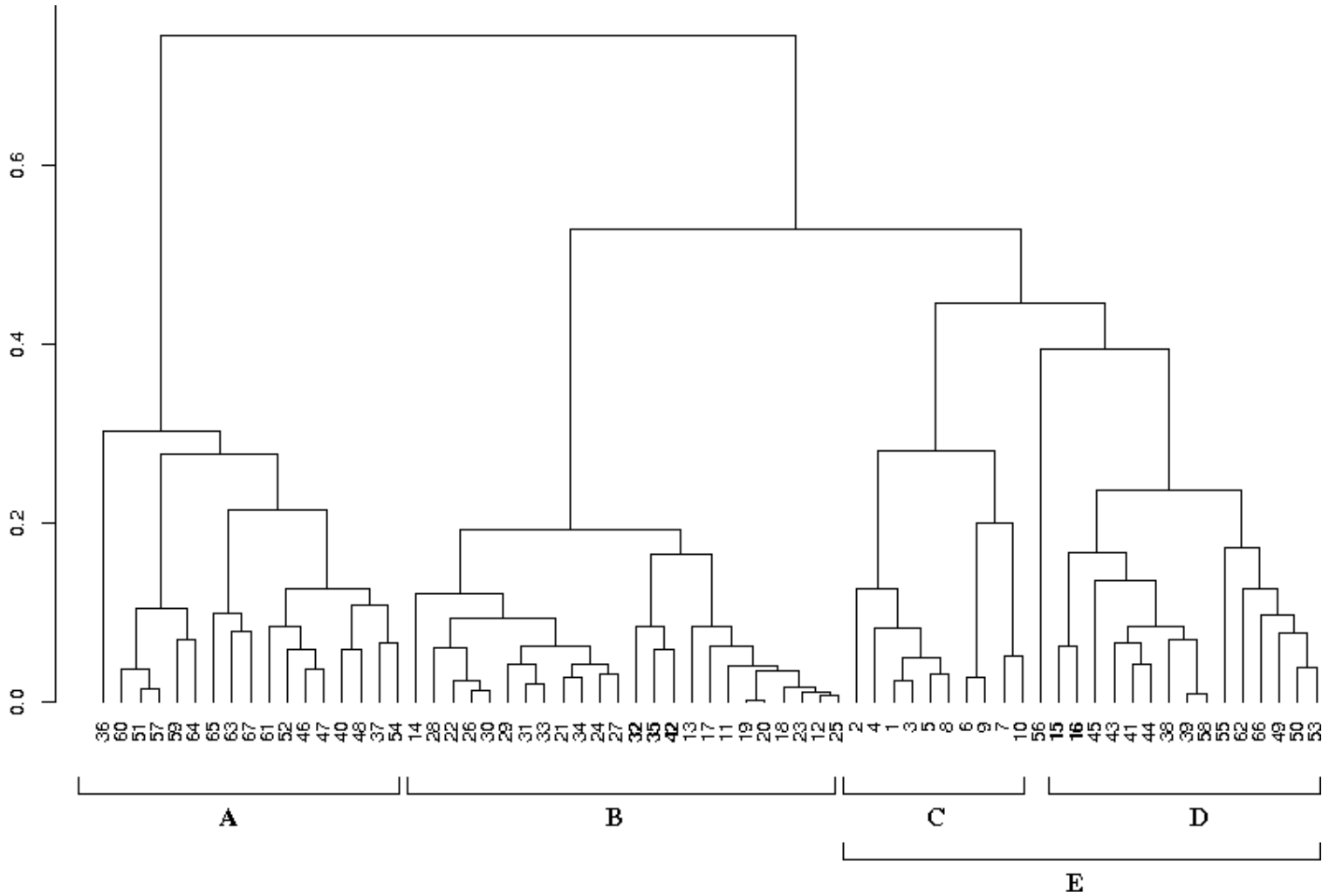


Tableau VII.2 Classification hiérarchique ascendante sur les points du premier plan factoriel (Corneille-Molière, 67 pièces, hapax exclus, n° *JQL 06*)

Légende de la classification automatique (tableaux VII.2 et VII.3). Les "anomalies" sont en gras.
De gauche à droite sur le dendrogramme :

Groupe A	Groupe B	Groupe C	Groupe D
Molière	Corneille et **Molière**	Corneille premières pièces	Molière et **Corneille**
36. Jalousie du B.	14. Pompée	(1630-1636)	**15. Menteur (Corneille)**
60. Avare	28. Othon	2. Clitandre	**16. Suite du Menteur (Corneille)**
51. Dom Juan	22. Nicomède	4. Galerie du Palais	45. Ecole des femmes
57. Sicilien	26. Sertorius	1. Mélite	43. Ecole des maris
59. Georges Dandin	30. Atilla	3. Veuve	41. Sganarelle
64. Fourberies de Scapin	29. Agésilas	5. Suivante	44. Fâcheux
65. Escarbagnas	31. Tite et Bérénice	8. Place royale	38. Etourdi
63. Bourgeois gentilhomme	33. Pulchérie	6. Comédie des T.	39. Dépit amoureux
67. Malade imaginaire	21. Don Sanche	9. Illusion comique	58. Amphytrion
61. Pourceaugnac	34. Suréna	7. Médée	55. Mélicerte
52. Amour médecin	24. Oedipe	10. Cid	62. Amants magnifiques
46. Critique de l'Ecole	27. Sophonisbe		66. Femmes savantes
47. Impromptu de V.	**32. Psyché (Corneille)**	*56. Comédie pastorale (Molière)*	49. Princesse d'Elide
40. Précieuses ridicules	**35. Psyché (Molière)**		50. Tartuffe
48. Mariage forcé	**42. Dom Garcie (Molière)**		53. Misanthrope
37. Médecin volant	13. Polyeucte		
54. Médecin malgré lui	17. Rodogune		
	11. Cinna		
	19. Héraclius		
	20. Andromède		
	18. Théodore		
	23. Pertharite		
	12. Horace		
	25. Toison d'Or		

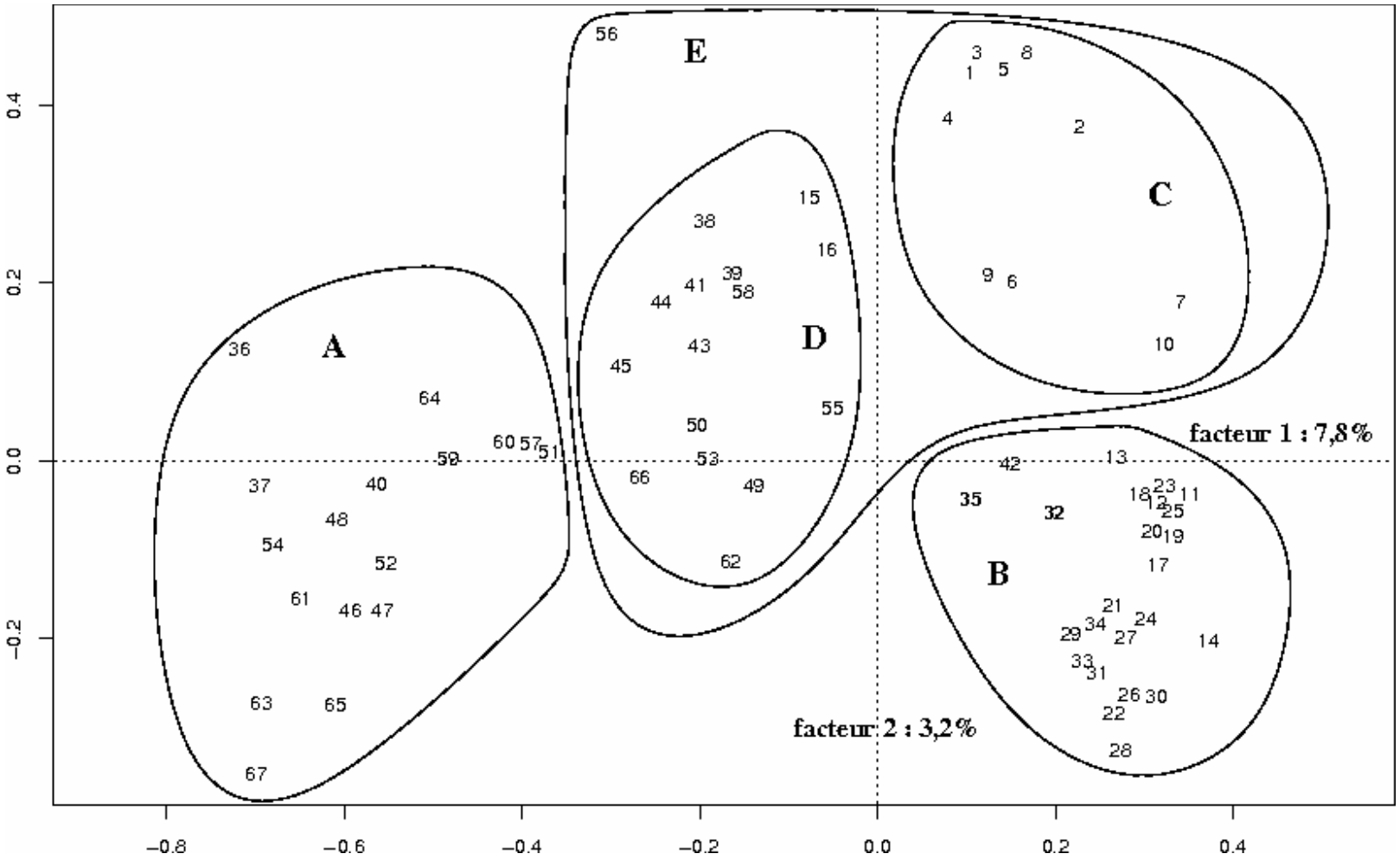


Tableau VII.3 Principales classes (“clusters”) dans le corpus Corneille-Molière. Analyse des correspondences - premier plan factoriel.

Le groupe B est le plus homogène (les traits de jonction sont situés très bas). Il rassemble toutes les pièces dites de la maturité et de la vieillesse de Corneille (de Cinna à Suréna). Au milieu, de ce cluster les numéros 32, 35, 42 (en gras sur le tableau) : Psyché (par Corneille), Psyché (censée écrite par Molière) et Dom Garcie (Molière).

Le sous-ensemble E rassemble :

- le groupe C : premières pièces de Corneille (de Mélite au Cid) ;

- le groupe D, c'est-à-dire toutes les pièces en vers de Molière avec les deux Menteurs de Corneille (n° 15 et 16 en gras au début du groupe D).

Ces résultats sont reportés sur le graphique VII.3. Les traits délimitent chacun de ces ensembles et groupes repérés par la classification. On pourra comparer ces traits avec ceux de la figure 6 de *JQL 06* (p. 88).

Notre article *JQL01* (p. 223-226) commente ainsi notre propre classification automatique – effectuée sur les distances intertextuelles.

"Cet outil détecte quelques 'anomalies' :

- Une des pièces de Molière se trouve au milieu de celles de Corneille, Dom Garcie. Cette pièce est très probablement de Corneille. D'ailleurs, elle est très proche de celle qu'il écrivait à l'époque où Dom Garcie a été créé.

- Les deux Psyché sont placées ensemble dans l'œuvre de Corneille.

- Deux des comédies de Corneille (le Menteur et la Suite du Menteur) se trouvent au milieu des pièces en vers de Molière. Cette classification est très surprenante car ces comédies (les dernières officiellement écrites par Corneille) datent de 1642-43, alors que les premières pièces de Molière sont supposées avoir été écrites au plus tôt en 1656 et ont été jouées à Paris seulement à partir de 1660. C'est pourquoi, Corneille étant l'auteur indiscuté des deux Menteurs, il a probablement aussi écrit les pièces qui se trouvent [proches d'elles] sur le dendrogramme : Tartuffe, le Misanthrope, les Femmes savantes, l'Etourdi, le Dépit amoureux, l'Ecole des maris, Sganarelle, Amphytrion, la Princesse d'Elide, Mélicerte et les Fâcheux, c'est-à-dire toutes les pièces en vers de Molière".

Ces conclusions sont donc entièrement validées par l'expérience de *JQL 06*

Corneille est l'unique auteur de Dom Garcie et de la totalité de Psyché. Il est aussi celui des 2 Menteurs et de toutes les comédies en vers publiées sous le nom de Molière.

Cachez cette histoire…

Enfin, le § 8.14 (p. 89 de ce dossier) conteste une phrase de notre article *JQL01* : *"From the very beginning, it was rumoured that Molière was not the writer of his plays"* (Depuis le tout début, le bruit a couru que Molière n'était pas l'auteur de ses pièces).

Tout d'abord, il a été rappelé à plusieurs reprises que les premiers éditeurs de trois des pièces, imprimées sous le nom de Molière, ont indiqué que celui-ci n'en est pas l'auteur : le Dépit amoureux, Psyché et Dom Juan (voir annexe 10, p. 159). Pour le Dépit et Psyché, P. Corneille a été désigné nommément ou par une allusion transparente. Pour Dom Juan, versifié par Thomas Corneille (le jeune frère de P. Corneille), Molière est désigné comme étant celui sous le nom duquel la pièce est représentée (il s'agit d'un avertissement de l'éditeur et non de T. Corneille).

La célèbre remarque de Boileau - qui dépossède Molière-Scapin du Misanthrope – a aussi été rappelée…

L'annexe 11 (p. 161 de ce dossier) donne quelques exemples, choisis au tout début des succès de Molière. On verra que cette petite phrase n'a pas été écrite à la légère : plusieurs contemporains de Molière, bien informés de la vie littéraire, ont indiqué que celui-ci n'est pas l'auteur (ou pas l'auteur principal) de "ses" pièces.

> En l'absence de manuscrits de Molière, de multiples indices historiques concordants justifient l'examen, par les méthodes statistiques, de la paternité de certaines pièces représentées sous son nom.

Et les conclusions de cette étude sont solides, plus solides que l'expérience "sans biais" de *JQL 06* qui va être examinée d'un peu plus près.

Chapitre 8
Sans biais ?

VALERE
Sans biais ?
HARPAGON
Oui.
VALERE
Ah ! je ne dis plus rien. Voyez-vous ? voilà une raison
tout à fait convaincante ; il se faut rendre à cela. (…)
HARPAGON
Sans biais.
VALERE
Vous avez raison : voilà qui décide tout, cela s'entend.
HARPAGON.
Sans biais.
VALERE
Ah ! il n'y a pas de réplique à cela : on le sait bien ;
qui diantre peut aller là contre ? (…)
HARPAGON.
Sans biais.
VALERE.
Il est vrai : cela ferme la bouche à tout, "sans biais".
Le moyen de résister à une raison comme celle-là ?

(D'après P. Corneille, l'Avare)

L'analyse présentée dans les § 8.10 à 8.12 de *JQL 06* est tronquée. Il manque les contrôles indispensables dans le cas d'une expérience cruciale comme celle-ci. Voici ces contrôles[63]. Cela permettra de vérifier si cette analyse est réellement "sans aucun biais" et de savoir précisément quelles conclusions on peut en tirer.

Les distances du Chi2 et leur représentation graphique

En retirant le prologue de Psyché par Quinault, il reste 67 pièces (et non pas 65). Ces pièces sont rangées selon l'ordre de l'annexe 1. Elles constituent les 67 colonnes du "tableau lexical" ("Lexical Table" : Lebart, Salem & Berry 1998, p. 35). Les 6 879 vocables de fréquence supérieures à 1 sont rangés par ordre alphabétique et constituent

[63] Nous utilisons principalement l'ouvrage de Lebart, Salem & Berry (1998) puisque c'est le seul cité par *JQL 06*. On peut aussi se reporter à Lebart, Morineau & Tabart (1977) ou Lebart, Morineau & Fénelon (1982) dans lesquels se trouvent les formules et les algorithmes de calcul utilisés dans ce chapitre. Enfin, l'ouvrage d'Escoffier & Pagès (1988) donne une autre présentation commode de ces techniques.

les lignes du tableau. Celui-ci comporte donc 460 893 cellules (6 879 * 67) à l'intersection des lignes (vocables) et des colonnes (textes). Dans chaque cellule : la fréquence d'un vocable (ligne) dans le texte considéré (colonne).

Pour chacune des 67 colonnes (c'est-à-dire les pièces), le programme[64] calcule sa distance du Chi2" à chacune des 66 autres colonnes. La même opération peut être effectuée sur les lignes (distances entre chaque paires de lignes) et donne alors la distance entre les vocables.

Pour les colonnes (les pièces), le calcul aboutit donc à un tableau des distances du Chi2 organisé comme celui qui a été présenté dans le chapitre II (p. 30 de ce dossier).

La figure 6 (p 88 de ce dossier) et le graphique du tableau VII.1 (p. 93 ci-dessus) sont l'une des nombreuses représentations possibles de ce tableau des distances du chi2 entre les pièces (les colonnes du tableau lexical). Pour obtenir cette figure, le tableau est d'abord converti en un espace à 65 dimensions (67 - 2) dans lequel chacun des textes est représenté par un point dont les coordonnées sont déterminées par les 66 distances du Chi2 mesurées sur toutes les paires que l'on peut constituer entre ce texte et chacun des 66 autres. Le calcul détermine alors le plan passant au plus près de tous ces points, puis les points sont projetés orthogonalement sur ce plan, ce qui donne la figure 6.

Une information importante est fournie par les % inscrits sur les 2 axes de la figure 6 : 89% de l'information est "perdue" (ou "bruitée"). Cela signifie que certains points (textes) peuvent être plus ou moins "mal" représentés sur le graphique. Dans ce cas, rien ne remplace un examen direct du tableau des distances du Chi2, comme on vient de le faire dans le chapitre II (avec la distance intertextuelle).

On va lire ci-dessous cet examen que *JQL 06* aurait dû faire.

Quelques nuages dans un ciel sans biais

Pour permettre la comparaison - entre la méthode employée *JQL 06* et la distance intertextuelle dont les résultats sont discutés dans le chapitre II - la démarche suivie sera la même : examen du texte le plus court - il s'agit ici de la Comédie pastorale – puis du

[64] Nous utilisons à nouveau R (version 2.4) et nos propres programmes (formules et algorithmes dans : Lebart, Morineau & Tabart 1977 et Lebart, Morineau & Fénelon 1982). Le script de l'expérience est à la disposition des chercheurs qui le désireraient. Les formules et les explications peuvent aussi être trouvées dans Lebart, Salem & Berry 1998, p. 52-54 ou Escoffier & Pagès 1988, p. 226-230.

plus long (l'Avare). Les graphiques ci-dessous sont construits selon la procédure suivie au chapitre II (avec la distance intertextuelle). Par exemple, sur le tableau VIII.1, l'abscisse (position sur l'axe horizontal) de chaque point est déterminée par la longueur du texte dont on a mesuré la distance du Chi2 à la Comédie Pastorale. L'ordonnée (axe vertical) du même point correspond à la distance du Chi2 de la Comédie pastorale au texte considéré. Le nuage s'étend donc de la Jalousie du barbouillé (3 501 mots) jusqu'à l'Avare (21 033 mots) et comporte 66 points.

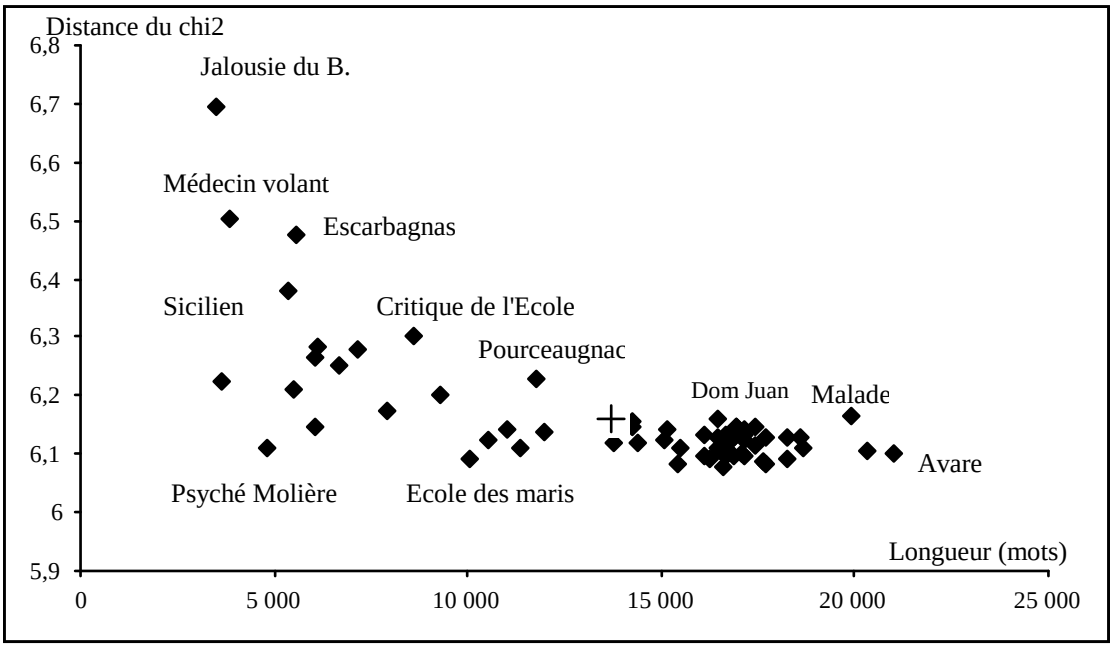


Tableau VIII.1 Valeurs des distances du Chi2 séparant la Comédie pastorale des 66 autres pièces du corpus Corneille-Molière (classées par longueurs croissantes).

Si les deux variables étaient indépendantes, le nuage s'étirerait à l'horizontale (à titre de comparaison, voir les tableaux II.4 et II.5, p. 34 - 35 de ce dossier).

Les valeurs de la distance du Chi2 évoluent clairement en raison inverse de l'allongement des textes. Le coefficient de corrélation linéaire est égal à -0,51[65]. La droite d'ajustement du nuage passe par le point moyen indiqué par une croix (correspondant à la longueur moyenne, soit 13 684 mots). La qualité de cet ajustement est bonne. La corrélation est donc avérée.

[65] Avec 65 degrés de liberté, la limite d'acceptation du coefficient de corrélation linéaire est de ±0,32 (avec un risque de première espèce inférieur à 1% de chances de se tromper en acceptant l'existence de cette corrélation).

On note également, dans la partie inférieure droite, une convergence très nette des distances – conforme au modèle de dépendance qu'on a cherché en vain à propos de la distance intertextuelle – mais aussi que, tout au long, les pièces en prose se situent dans la partie supérieure et les pièces en vers en dessous (sans considération d'auteurs).

Les conséquences de cette liaison avérée sont faciles à imaginer : sur la figure 6 (p. 88 ci-dessus), la position du point 56 est principalement déterminée par le rapport de la longueur de cette pièce à celles des 64 autres.

Le tableau VIII.2 présente le nuage obtenu avec l'autre extrême du corpus, la plus longue pièce (l'Avare). Le point correspondant à la distance entre l'Avare et la Pastorale est trop haut pour figurer sur le graphique.

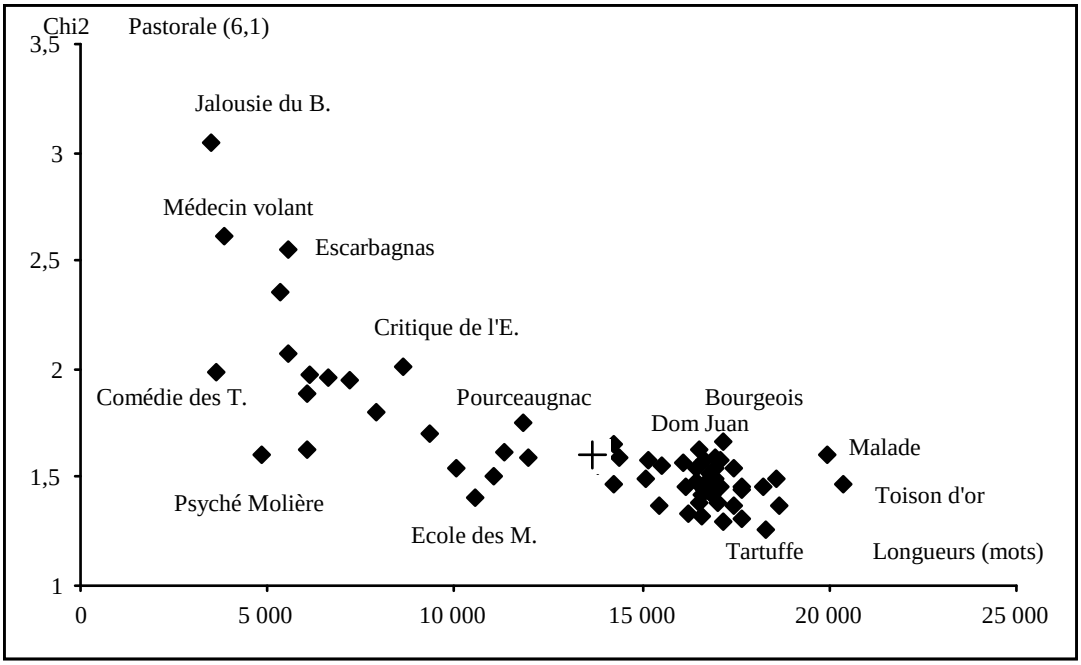


Tableau VIII.2 Distances du Chi2 séparant l'Avare des 66 autres pièces du corpus Corneille-Molière (classées par longueurs croissantes).

Le coefficient de corrélation est égal à -0,701. La qualité de l'ajustement est excellente. La droite d'ajustement suit la seconde diagonale du tableau. Elle est plus inclinée que la droite d'ajustement du tableau VIII.2. Les distances du Chi2 entre l'Avare et les autres pièces obéissent donc à la loi suivante : plus la pièce est courte, plus la distance du Chi2 avec l'Avare est forte. C'est le facteur essentiel déterminant ces distances. Secondairement, on retrouve les mêmes mécanismes que ci-dessus : convergence des distances au fur et à mesure de l'allongement des textes ; les pièces en

prose dans la partie supérieure du nuage, les pièces en vers dans la partie inférieure (quel que soit l'auteur).

Les conséquences de cette liaison avérée sont faciles à imaginer : sur la figure 6 (p. 88 de ce dossier), la position du point 67 est principalement déterminée par le rapport de la longueur de l'Avare à celles des 66 autres pièces et secondairement par le fait qu'elle est en prose…

Le tableau VIII.3 ci-dessous superpose les deux séries précédentes. Le nuage supérieur correspond à la petite pièce et celui en bas à la plus longue (l'Avare). Ce second nuage commence au niveau du premier (le triangle le plus à gauche correspond à la distance du Chi2 entre l'Avare et la Comédie Pastorale, c'est aussi le losange le plus à droite).

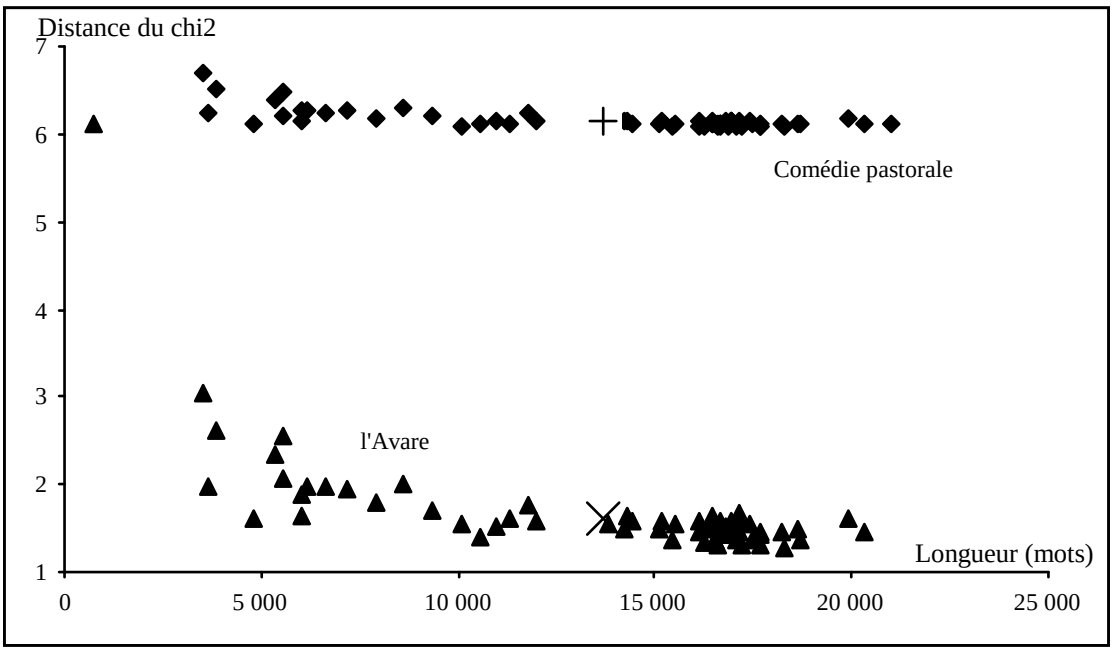


Tableau VIII.3 Distances du Chi2 de la Comédie pastorale (ligne supérieure) et de l'Avare (ligne inférieure) aux 66 autres pièces classées par longueurs croissantes (avec les points moyens).

La confrontation de ces tableaux amène les conclusions suivantes. Plus les textes comparés sont brefs, plus les distances du Chi2 sont élevées (mais aussi un peu moins sensibles à l'allongement des textes) et inversement. De ce fait, les distances du Chi2 générées par le petit texte sont beaucoup plus élevées que celles générées par le texte le plus long. Comme l'indique la position des points moyens, le nuage supérieur,

correspondant aux distances engendrées par la Comédie pastorale (732 mots), est 3.7 fois plus élevé que celui correspondant aux distances générées par l'Avare (21 000 mots). Autrement dit, dans la détermination des coordonnées du graphique 6 (p. 88 de ce dossier), chaque mot de la Comédie pastorale pèse en moyenne 108 fois plus lourd que chacun de ceux de l'Avare…

Pour donner une vision synthétique du phénomène, le total des distances du Chi2 générées par chaque texte (colonnes du tableau lexical) est rapporté au total général de ces distances (total de la marge en colonne du tableau lexical). Cela permet de constater que les distances du Chi2 générées par la Comédie pastorale représentent 4,7% du total général alors que l'Avare pèse 1,29% de ce total. Le tableau VIII.4 récapitule les résultats pour les 67 pièces (à comparer avec le tableau II.11, p. 42 ci-dessus).

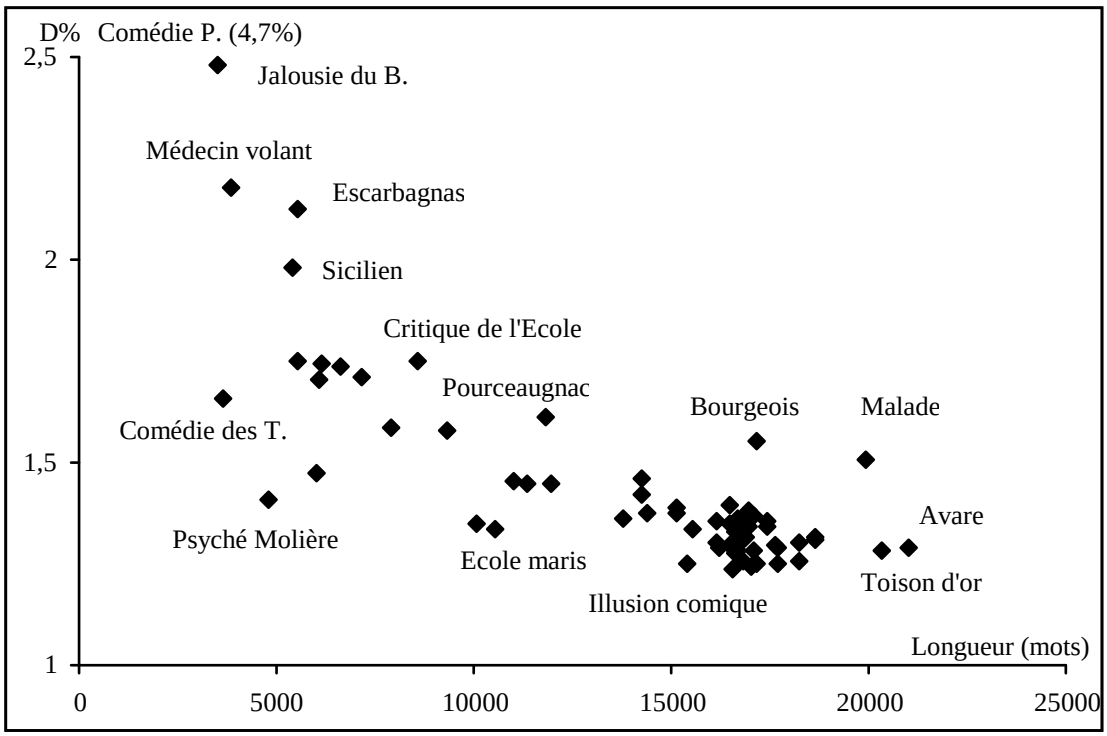


Tableau VIII.4 Poids de chacune des 67 pièces dans la distance du Chi2 totale.

On retrouve la même liaison linéaire selon la seconde diagonale et la convergence des points vers la droite. A longueur égale, les pièces en prose sont toujours au-dessus des pièces en vers (quel que soit l'auteur). L'écart est beaucoup plus fort pour les pièces les plus courtes. Par exemple, à longueur égale, la Jalousie du Barbouillé (farce en prose, Molière) pèse 1,7 fois plus lourd que la Comédie des Tuileries (Corneille, vers).

Enfin, la longueur d'un grand nombre de pièces étant peu différente (la majorité des pièces a une longueur comprise entre 13 000 et 18 000 mots), leurs distances du Chi2 mutuelles sont moins influencées par les différences de taille (et l'on retrouve alors des relations qui ressemblent apparemment à celles que l'on obtient avec la distance intertextuelle). Cependant, certaines pièces sont manifestement "déviantes" (par exemple, le Bourgeois gentilhomme, le Malade imaginaire, Dom Juan) pour des raisons que l'on va découvrir dans un moment…

Ce schéma – rapporté à nos tableaux II.11 et II.12 (p. 42-43 de ce dossier) – permet de comprendre pourquoi *JQL 06* tombe "grosso modo" sur le "bon" résultat. D'une part, la majorité des pièces ont des longueurs assez proches et, à l'exception de Mélicerte et de la Comédie des Tuileries, les plus courtes sont des comédies en prose. D'autre part, les pièces en vers, de longueurs assez proches (sauf les 2 mentionnées ci-dessus) sont moins affectées par le "facteur longueur". Dès lors, sur la figure 6, les trois groupes de pièces correspondent au classement suivant : les tragédies – dont les longueurs sont les plus fortes et qui sont peu différentes entre elles - (en bas à droite), les comédies en vers – dont les longueurs moyennes sont moins élevées et qui sont à peine plus différentes - (dans la partie supérieure) et les comédies en prose – en moyenne plus courtes et plus diverses - en bas à gauche.

Dans le détail, la figure 6 (p. 88 de ce dossier) comporte plusieurs "aberrations". Ainsi, la forme en fer à cheval du nuage de points indique l'existence d'une gradation selon une dimension dominante[66]. Surtout, les tentacules dessinées sur la figure 6 posent une question évidente. De deux choses l'une : il y a eu une erreur sur le genre de 4 pièces ou l'analyse est erronée. En tous cas, il est absurde de présenter un tel tableau sans réfléchir aux raisons pour lesquelles de pareilles aberrations peuvent se produire.

Une bonne méthode mal utilisée

Toutes ces aberrations ne sont pas des "biais" mais des conséquences normales des propriétés de la distance du Chi2, propriétés que l'auteur de *JQL 06* ignore

[66] Lorsque les points épousent à peu près la forme d'un fer à cheval, on parle d' "effet Guttman" (Lebart, Salem & Berry 1998, p. 149-150). Une telle configuration indique que les données sont organisées selon une "gradation" unidimensionnelle.

probablement. Les anomalies ne viennent pas tant de la longueur des textes que du choix de conserver les mots "rares".

Le calcul de la distance du Chi2 ne peut porter que sur des mots fréquents et présents dans un nombre suffisant de textes. Lebart, Salem et Berry écrivent à ce sujet que : "les calculs ne sont significatifs que si les mots [utilisés dans ces calculs] apparaissent avec une fréquence minimale : les hapax et les mots rares sont éliminés" (op.cit, p. 104). Dans l'exemple donné par ces auteurs, plus de 90% des vocables sont éliminés (ibid, p. 106).

L'auteur de *JQL 06* a-t-il lu le manuel qu'il cite ?

Les mots qui pèsent le plus lourd dans son expérience sont ceux qui ont les fréquences les plus faibles, tout en étant présents dans un seul texte de petite taille. Par exemple, les trois mots les plus "lourds" sont : *guéret, jouvenceau* et *glacer*, car ils sont utilisés uniquement dans… la Comédie pastorale. A eux seuls cinq vers - répétés deux fois à la fin de cette pièce - pèsent autant dans le calcul de la distance du Chi2 que tout le vocabulaire de l'Illusion comique :

> *Quand l'hiver a glacé nos guérets,*
> *Le printemps vient reprendre sa place,*
> *Et ramène à nos champs leurs attraits ;*
> *Mais, hélas ! quand l'âge nous glace,*
> *Nos beaux jours ne reviennent jamais.*

La critique littéraire était passée à côté de ces vers de Molière…

A l'inverse, les mots qui pèsent le moins lourd dans le calcul sont les vocables les plus fréquents et qui figurent dans toutes les pièces. Cela donne aux 2 202 occurrences de "aimer" ou aux 2 361 de "amour" - qui sont présents dans 64 des 65 pièces - le privilège de peser moins lourd, dans la distance du Chi2 totale, que les 6 apparitions de "pisser" jointes aux 3 occurrences de "urine" parce que ces occurrences se produisent toutes dans le seul Médecin Volant de Molière (3 876 mots)…

Voilà pourquoi le Malade imaginaire, le Bourgeois gentilhomme ou Dom Juan se singularisent sur le tableau VIII.4 : le faux turc, le latin de cuisine et le faux Picard !

Plus généralement, le tableau lexical, utilisé dans cette expérience, présente les caractéristiques suivantes :

- sur les 460 893 cellules, 82,3% contiennent un zéro ;

- dans les 67 colonnes, la proportion de cases nulles oscille entre 75,3% (l'Etourdi) et 95,4% (la Comédie pastorale).

Un tableau "creux" comme celui-ci ne se prête pas à une analyse "classique" des correspondances (et à aucune des analyses issues de la statistique du χ^2). C'est autant un problème de répartition – la répartition mesure le nombre de textes où apparaît le vocable considéré - qu'un problème de fréquence. Par exemple, Chimène (54 occurrences) ou Rodrigue (67 occurrences) peuvent être considérés comme des mots fréquents, mais leur présence dans une seule pièce (le Cid) pose problème.

Dans le tableau lexical du corpus Corneille-Molière, 85 vocables seulement - moins de 1% du vocabulaire - sont présents au moins une fois dans les 67 pièces. Il s'agit essentiellement de "mots outils" (les auxiliaires être et avoir, les articles, pronoms, adverbes, conjonctions et prépositions usuelles). Il ne reste aucun nom propre ; seulement 13 verbes, 4 adjectifs et substantifs (*beau, cœur, coup, jour*)… Certes, l'ensemble couvre un peu plus de la moitié de la surface des textes mais peut-on prétendre rendre compte d'un corpus pareil avec si peu de mots ?

De ce fait, *JQL 06* donne un poids exorbitant à quelques pièces marginales : Comédie Pastorale, Jalousie du barbouillé, Médecin volant, Sicilien, Escarbagnas, etc. En suivant les recommandations de Lebart, Salem & Berry (1998, p. 78), ces pièces devaient être sorties du calcul, leur position étant indiquée sur le graphique ou dans la classification comme "éléments illustratifs".

Enfin, *JQL 06* n'a pas réalisé les contrôles de routine indispensables quand on effectue une expérience cruciale. Les tableaux VIII.1 à VIII.4 ont déjà présenté l'un de ces contrôles simples. Il faut également examiner les plans factoriels suivants (ici on aurait pu se rendre compte qu'un "effet Guttman" perturbe l'analyse) et, enfin, il était indispensable de réaliser au moins une classification automatique comme celle qui a été présentée dans le chapitre précédent.

Toute distance présente des propriétés qui doivent être connues et comprises par ceux qui les utilisent.

JQL 06 ne présente pas une "analyse des correspondances classique" ; il ne respecte même pas les principes de base et les précautions d'usage pour une telle analyse.

A l’inverse, nos tests sont de véritables expériences de laboratoire. Les textes sont choisis avec soin. Leur orthographe est entièrement révisée et leurs graphies sont standardisées (normalisation). Les fichiers sont balisés et lemmatisés, en suivant toujours les mêmes conventions. Nous écrivons tous les programmes informatiques, de telle sorte que nous savons précisément ce qu'ils font et ce qu'ils ne peuvent pas faire. Enfin, lors des tests – effectués dans les règles de l’art -, toutes les variables et tous les paramètres sont contrôlés. Cela exclut évidemment les traitements à l'aveuglette sur n'importe quel corpus, avec des outils que l'on ne comprend pas, dans le seul but d'impressionner la galerie, en alignant les centaines de millions de mots qui ne sont que la mesure du néant.

Conclusions de la 2e partie

La discussion ci-dessus permet de comprendre l'intérêt de la standardisation des graphies et de la lemmatisation. Outre la précision des calculs, ces opérations entraînent une réduction considérable du nombre de lignes dans le tableau lexical (Labbé 2000). Dans le corpus Corneille-Molière, il y a 28 982 formes "brutes" - en conservant les majuscules de début de vers et les graphies multiples pour un même mot - mais seulement 9 995 vocables. La lemmatisation divise donc le nombre de lignes par 2,9 tout en respectant exactement le texte. Cette réduction porte sur les plus basses fréquences, ce qui facilite les calculs de distance, quel que soit le procédé employé. Il est donc absurde de la rejeter *a priori* comme "exorbitante".

La discussion permet également de comprendre pourquoi il a été proposé - c'est le mot employé dans notre article *JQL01* - d'éliminer du calcul certains vocables rares. En effet, dans un grand tableau – spécialement si l'on utilise un dérivé du χ^2 pour calculer les distances entre lignes ou entre colonnes -, toutes les cases devraient contenir des valeurs positives et, si possible, au moins égales à l'unité. Cette proposition n'appartient pas au cœur du raisonnement et elle a - sur le calcul de la distance intertextuelle - une portée pratique faible (qui ne justifiait pas l'enflure du ton et la longueur des § 4.4 à 4.9 de *JQL 06*). En revanche, il aurait fallu réfléchir sérieusement à ce problème avant de conserver les vocables rares pour le calcul de la distance du Chi2 !

Ce qui précède permet enfin de comprendre les raisons qui ont conduits à élaborer la distance intertextuelle.

Réduire autant que possible la sensibilité de la mesure par rapport à la longueur des textes comparés.

Ne pas hypertrophier le poids des petits ou des grands textes, tout en essayant de les maintenir dans le calcul (ou du moins en sachant précisément à quel moment ce calcul n'est plus raisonnable).

Conserver une proportion aussi importante que possible du vocabulaire des textes à comparer, tout en évitant que les tableaux de calcul soient encombrés par un grand nombre de cases nulles et en limitant l'influence des décimales dans le calcul.

Donner à chaque mot un poids dans le calcul équivalent à sa fréquence d'utilisation dans les textes comparés.

Rechercher un indice dont la signification soit aisément accessible, etc…

…tout en présentant effectivement les propriétés d’une distance (identité, symétrie, inégalité triangulaire, robustesse, etc.) afin de pouvoir réaliser des classifications fiables sur de grandes collections de textes.

Au passage, cette méthode s’est révélée un excellent outil d’attribution d’auteur.

Conclusions

9. Conclusions

[9.1] 1. The original objective of DCL was to submit a new measurement of distance between texts. The result is disappointing since the inter-textual distance *has two biases which made it unusable, even to compare contrastive pairings (text A is nearer to B than to C…). Moreover, the idea of a rigid threshold-based interpretation scale in itself contradicts any operational use, to say nothing of the unsolved problems raised by normalization and lemmatization of data which perturb the* ID.

[*9.2] 2. With regard to authorship attribution, it is certainly tempting for some people to have at their disposal a unique measurement, of explicit appearance, in order to automatically determine uncertain attributions. But that would amount to a denial (beyond all scientific caution) of the extent to which the constitution and determination of discourse are complex and diverse, be it literary discourse or not. Some elementary tests, made upon prominently distinctive authors, easily show the disturbances to which adopting this method could lead.*

[9.3] 3. This last point is already illustrated by the application to Corneille's and Molière's cases. We certainly do not intend to grant unlimited credit to the 150 last years of international academic research. But it would be even less serious to settle, by a single measurement, a question of such an importance.

[9.4] 4. DCL's paper and its widespread repercussions are likely to seriously weaken the credibility of statistical methods in the humanities, and particularly in literature. It is worth noting, furthermore, that all the authors referred to by DCL's paper in the field of lexical statistics have expressed themselves against DCL's proposition: Etienne Brunet (Brunet, 2004), Charles Muller (Le Point *11.04.2003), Jean-Pierre Barthélémy* (Le Monde *11.06.2003). Meanwhile DCL have not received any significant approbation during the last three years.*

Appendix

Table of the 101 Maupassant's Contes selected for study mentioned in Sections 4 and 6.

1	*Sur l'Eau*	*36*	*Normand (Un)*	*71*	*Bonheur (Le)*
2	*Maison Tellier (La)*	*37*	*Parricide (Un)*	*72*	*Aveu (L')*
3	*Aventure parisienne (Une)*	*38*	*Réveillon (Un)*	*73*	*Coco*
4	*Partie de campagne (Une)*	*39*	*Ruse (Une)*	*74*	*Crime au Père Boniface (Le)*
5	*Aux Champs*	*40*	*Veillée (La)*		
6	*Aveugle (L')*	*41*	*Vieux Objets*	*75*	*Gueux (Le)*
7	*Bécasse (La)*	*42*	*Voleur (Le)*	*76*	*Ivrogne (L')*
8	*Bûche (La)*	*43*	*Yveline Samoris*	*77*	*Lettre trouvée sur un Noyé*
9	*Ce cochon de Morin*	*44*	*A cheval*		
10	*Clair de Lune*	*45*	*Ami Joseph (L')*	*78*	*Mère Sauvage (La)*
11	*Confessions d'une femme*	*46*	*Auprès d'un Mort*	*79*	*Notes d'un voyageur*
12	*Correspondance*	*47*	*Aventure de Walter Schnaffs (L')*	*80*	*Parure (La)*
13	*Farce normande*			*81*	*Petit Fût (Le)*
14	*Folle (La)*	*48*	*Confession (La)*	*82*	*Rose*
15	*Fou ?*	*49*	*Denis*	*83*	*Souvenir*
16	*Gâteau (Le)*	*50*	*Deux Amis*	*84*	*Tombe (La)*
17	*Histoire vraie*	*51*	*En Mer*	*85*	*Lâche (Un)*
18	*Lit (Le)*	*52*	*Farce (La)*	*86*	*Vieux (Le)*
19	*Loup (Le)*	*53*	*Ficelle (La)*	*87*	*Bête à Maît' Belhomme (La)*
20	*Madame Baptiste*	*54*	*Humble Drame*		
21	*Mademoiselle Fifi*	*55*	*Main (La)*	*88*	*Mes Vingt-cinq jours*
22	*Marroca*	*56*	*Mon oncle Jules*	*89*	*Cri d'alarme*
23	*Menuet*	*57*	*M. Jocaste*	*90*	*Epave (L')*
24	*Mots d'amour*	*58*	*Orphelin (L')*	*91*	*Fermier (Le)*
25	*Nuit de Noël*	*59*	*Père Milon (Le)*	*92*	*Mademoiselle Perle*
26	*Peur (La)*	*60*	*Petit (Le)*	*93*	*Etrennes*
27	*Pierrot*	*61*	*Première Neige*	*94*	*Allouma*
28	*Relique (La)*	*62*	*Remplaçant (Le)*	*95*	*Hautot père et fils*
29	*Rempailleuse (La)*	*63*	*Réveil*	*96*	*Soir (Un)*
30	*Roche aux Guillemots (La)*	*64*	*Sabots (Les)*	*97*	*Champ d'oliviers (Le)*
		65	*Saint-Antoine*	*98*	*Mouche*
31	*Rouerie*	*66*	*Serre (La)*	*99*	*Après*
32	*Saut du Berger (Le)*	*67*	*Tombouctou*	*100*	*Colporteur (Le)*
33	*Testament (Le)*	*68*	*Duel (Un)*	*101*	*Père (Le)*
34	*Coq chanta (Un)*	*69*	*Vendetta (Une)*		
35	*Fils (Un)*	*70*	*Vengeur (Le)*		

References

[1] Adam, J. -M. (1999) *Linguistique textuelle: des genres de discours aux textes*, Paris: Nathan.

[2] Bakhtine, M. (1977) *Marxisme et philosophie du langage*, Paris: Minuit.

[3] Barthélémy, J. -P. and Guénoche, A. (1991) *Trees and Proximity Representations*, New York: John Wiley & Sons.

[4] Brunet, E. (2004) Où l'on mesure la distance entre les distances, *Texto!*, [en ligne], mars 2004. Rubrique Dits et inédits (http://www.revuetexto.net/Inedits/Brunet/Brunet_Distance.html).

[5] Habert, B., Nazarenko, A. and Salem, A. (1997) *Les linguistiques de corpus*, Paris: Colin.

[6] Harris, Z. S. (1969) Analyse du discours, *Langages*, 13, pp. 11–65.

[7] Kendall, M. G. (1962) *Rank Correlation Methods*, London: Griffin.

[8] Lebart, L, Salem, A. and Berry, L. (1998) *Exploring Textual Data*, Boston: Kluwer Academic Publisher.

[9] Labbé, D. (2003) *Corneille dans l'ombre de Molière*, Paris, Bruxelles: Les Impressions nouvelles.

[10] Labbé, D. and Labbé, C. (2001) Inter-textual distance an authorship attribution, *Journal of Quantitative Linguistics*, 8(3), pp. 213–228.

[11] Luong, X. (1988) Using a tree-model in textual analysis, *Computers and the Humanities*, 23, pp. 397–402.

[12] Muller, C. (1992a) *Initiation aux méthodes de la statistique linguistique*, Paris: Champion.

[13] Muller, C. (1992b) *Principes et méthodes de statistique lexicale*, Paris: Champion.

[14] Muller, C. (1993) *Langue française: débats et bilans*, Paris: Champion.

[15] Viprey, J. -M. (1997) *Dynamique du vocabulaire des Fleurs du mal*, Paris: Champion.

[16] Viprey, J. -M. (1998) Une norme endogène pour le calcul stylistique du vocabulaire, *JADT 1998, 4èmes Journées internationales d'Analyse statistique des Données Textuelles*. Nice: CNRS-UNSA.

[17] Viprey, J. -M. (2002) *Analyses textuelles et hypertextuelles des Fleurs du mal*, Paris: Champion.

Nos conclusions

Toutes les expériences présentées dans *JQL 06* confirment la solidité de nos méthodes et valident l'attribution à Corneille de 16 pièces représentées sous le nom de Molière.

La distance intertextuelle - combinée avec les méthodes modernes de classification - vient une nouvelle fois de démontrer sa solidité et sa fiabilité.

Le mérite en revient aux nombreuses personnes qui ont aidé à sa mise au point depuis près de 10 ans. Grâce à ces personnes, cette méthode a subi avec succès un très grand nombre de tests sévères, organisés dans les règles de l'art, dont plusieurs ont été réalisés en "aveugle". Les comptes rendus de trois de ces expériences ont été publiés (Labbé 2002a ; Monière & Labbé 2006 ; Labbé 2007). Ces épreuves ont toutes été couronnées d'un plein succès : tous les textes attribués à un même auteur se sont révélés avoir été effectivement écrits par la même personne ; tous les pièges imaginés, pour faire échouer la méthode, ont été déjoués.

Ces outils désignent clairement P. Corneille comme l'auteur d'au moins 16 pièces représentées sous le nom de Molière (toutes les comédies en vers plus le Dom Juan et l'Avare).

De plus, cette attribution ne repose pas que sur la distance intertextuelle combinée avec plusieurs classifications (dont l'efficacité vient d'être démontrée une nouvelle fois). Elle repose sur un faisceau d'indices – lexicaux, stylistiques et historiques – concordants et solides dont aucun n'a été remis en cause depuis décembre 2001.

Remerciements

Institutions et personnes anonymes

Les réviseurs de l'article de 2001 qui l'ont jugé digne d'être publié et qui ont communiqué quelques remarques fort utiles.

Les réviseurs de nos travaux ultérieurs qui les ont examinés sans se laisser intimider par les ragots et les calomnies.

Les responsables des revues scientifiques qui ont bien voulu publier nos travaux sur la distance intertextuelle sans prêter attention aux rumeurs (Journal of Quantitative Linguistics, Literary and Linguistic Computing, Computers and the Humanities, Corpus…)

Le Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, (LIMSI-Paris VI-Orsay) et le département de Français de l'Université de Dublin (Trinity College) qui nous ont donné la parole et ont accepté que ces communications soient mises en ligne.

Les organisateurs des Journées d'Analyse des Données Textuelles de Louvain-la-Neuve (2004) qui ont bien voulu nous inviter à une table ronde sur cette question.

Les chercheurs qui ont contribué à nos travaux

André Pibarot, décédé en 1998, et le général Picard sans qui le développement des programmes d'analyse statistique du langage auraient été abandonnés au milieu des années 1990.

Pierre Hubert (Ecole des Mines de Paris) qui travaille avec nous depuis plus de 30 ans et qui nous a signalé l'énigme Corneille-Molière et le livre d'H. Wouters.

Charles Bernet (Ecole normale supérieure de Lyon) qui a fourni les textes de Corneille, Molière et Racine. Il a bien voulu apporter de très utiles conseils lors de la mise au point des outils informatiques pour le traitement des textes ;

Jean-Guy Bergeron (Université de Montréal), Mathieu Brugidou (EDF-GREST), Pierre Hubert (Ecole des Mines de Paris), Jean et Nelly Leselbaum (Université de Paris-

X), Denis Monière (Université de Montréal) qui ont entrepris avec nous les premières expériences sur la distance intertextuelle.

Xuan Luong (Université de Nice) qui nous a initié à l'analyse arborée et a bien voulu nous conseiller dans la réalisation des programmes.

Gaétan Paéquin et Mathieu Ruhlmann (Polytech' Grenoble) qui ont aidé à réaliser les programmes de classification arborée et de segmentation automatique des corpus.

Edward Arnold (University of Dublin), Gérard Ledger et Tom Merriam pour leurs aides précieuses dans l'application de la distance intertextuelle à l'attribution d'auteur en langue anglaise.

Francis Lapierre pour nos travaux communs sur le nouveau testament.

Benoît Peeters et Jan Baetens (Les impressions nouvelles) sans qui Dominique Labbé n'aurait pas écrit son essai Corneille dans l'ombre de Molière et grâce à qui nos résultats sur le cas "Gary-Ajar" ont été rendus publics.

Dominique Andolfatto (CNAM-Paris), Edward Arnold (Trinity College – Dublin), Guy Bensimon (Institut d'Etudes Politiques de Grenoble), Tom Merriam et Denis Monière (Université de Montréal) qui ont bien voulu relire la première version de ce dossier et faire de très utiles commentaires.

Hippolyte Wouters pour ses encouragements et son humour.

Ils sont trop nombreux pour tous les nommer :

Les journalistes qui ont rendu compte honnêtement de ce dossier sans se laisser impressionner par les ragots, les insultes et les calomnies.

Les très nombreux érudits, internautes et simples curieux qui ont écrit si nombreux pour nous encourager et signaler tel ouvrage ou article utiles pour nos recherches…

Bibliographie

Bergeron Jean-Guy & Labbé Dominique (2000). "L'évaluation de la négociation raisonnée par les acteurs. Une analyse lexicométrique". (Communication au XVI^{e} Congrès international de l'Association internationale des sociologues de langue française, Québec, juillet 2000). Reproduit dans Bernier Colette & Al. Formation, relations professionnelles à l'heure de la société-monde. Paris-Québec : L'Harmattan - Les Presses de l'Université Laval, 2002, p. 239-252.

Bourqui Claude (1999). Les sources de Molière. Répertoire critique des sources littéraires et dramatiques. Paris: SEDES.

Brugidou Mathieu & Labbé Dominique (1999). Le discours syndical français contemporain (CGT, CGT, FO en 1996-98). Grenoble-Paris, CERAT-EDF(GRETS).

Brunet Etienne & Muller Charles (1988). "La statistique résout-elle les problèmes d'attribution ?". Strumenti Critici, septembre 1988, p. 367-387.

Dodge Yadolah (1993). Statistique. Dictionnaire encyclopédique. Paris: Dunod.

Escoffier Brigitte & Pagès Jérôme. Analyses factorielles simples et multiples. Paris : Dunod.

Forestier Georges (1990). Molière en toutes lettres. Paris: Bordas.

Grimarest (1675). La vie de M. de Molière. Genève : Slatkine Reprints, 1973.

Guiraud Pierre (1955-1964). Index du vocabulaire du théâtre classique. Paris: Klincksieck (Oeuvres de P. Corneille et de J. Racine).

Harris John W. & Stocker Horst (1998). Handbook of Mathematics and Computational Science. New York-Berlin: Springer-Verlag.

Hatzfeld Adolphe, Darmeisteter Arsène & Thomas Antoine (1898 env). Dictionnaire général de la langue française du commencement du XVIIe siècle jusqu'à nos jours. Paris : Delagrave.

Hubert Pierre & Labbé Dominique (1998). "La connexion des vocabulaires" in Mellet Sylvie. IVe journées internationales d'analyse statistique des données textuelles. Nice : Université de Nice-Sophia Antipolis, février 1998, p. 361-369.

Juilliard Michel & Luong Xuan (2001). "On consensus between Tree-Representation of Linguistic Data". Literary and Linguistic computing, 16-1, p. 59-76.

Katsberg Söjblom Margareta (2006). L'écriture de J.M.G. Le Clézio : des mots aux thèmes. Paris: Honoré Champion.

Labbé Cyril & Labbé Dominique (2001). "Inter-Textual Distance and Authorship Attribution Corneille and Molière". Journal of Quantitative Linguistics. 8-3, December 2001, p 213-231.

Version française : htpp://halshs.archives-ouvertes.fr/halshs-00137675.

Labbé Cyril & Labbé Dominique (2003). "La distance intertextuelle". Corpus, 2, p 95-118.

http://revel.unice.fr/corpus/document.html?id=31

Labbé Cyril, Labbé Dominique & Hubert Pierre (2004). "Automatic Segmentation of Texts and Corpora". Journal of Quantitative Linguistics, December 2004, 11-3. p 193-213.

Labbé Cyril & Labbé Dominique (2005). "How to measure the meanings of words ? Amour in Corneille's work", Langage Resources Evaluation, 2005, 39, p 335-351.

Labbé Cyril & Labbé Dominique (2006). "A Tool for Literary Studies: Intertextual Distance and Tree Classification". Literary and Linguistic Computing. 21-3, 2006, p 311-326.

Labbé Dominique (1990). Normes de saisie et de dépouillement des textes politiques. Grenoble: Cahier du CERAT.

http://web.upmf-grenoble.fr/cerat/Recherche/PagesPerso/LabbeNormes

Labbé Dominique (2000). "Analyse des données textuelles et Statistique lexicale (Textual Data Analysis and Lexical Statistics)". Conférence introductive aux 5e journées internationales d'analyse des données textuelles. Lausanne : Ecole polytechnique fédérale, 2000. Reproduite dans Lexicometrica, 4, 2002.

http://web.

Labbé Dominique (2002a). Qui a écrit quoi ? Grenoble: CERAT.

http://www.upmf-grenoble.fr/cerat/Recherche/PagesPerso/LabbeExperience

Labbé Dominique (2002b). "La lemmatisation des grandes bases de textes. Un exemple : Corneille, Molière et Racine", Communication au colloque *L'édition électronique en littérature et dictionnairique, évaluation et bilan*, Rouen, 17-21 juin 2002.

http://web.upmf-grenoble.fr/cerat/Recherche/PagesPerso/LabbeRouen

Labbé Dominique (2003). Corneille dans l'ombre de Molière. Histoire d'une recherche. Bruxelles : Les impressions nouvelles.

Labbé Dominique (2004a). "Interventions lors de la Table ronde "Corneille et Molière"". 7e Journées d'Analyse des Données Textuelles (Louvain-la-Neuve, 11 mars 2004).

http://www.upmf-grenoble.fr/cerat/Recherche/PagesPerso/LabbeLouvain

Labbé Dominique (2004b). "Corneille in the shadow of Molière". Dublin : University of Dublin (Trinity College), French Department Research Seminar, April 6 2004. http://web.upmf-grenoble.fr/cerat/Recherche/PagesPerso/LabbeDublin

Labbé Dominique (2004c). "Corneille et Molière". Séminaire du Groupe Langues Informations Représentations - Université de Paris VI-Orsay : Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), 13 janvier 2004.

http://web.upmf-grenoble.fr/cerat/Recherche/PagesPerso/LabbeCorneilleMoliere

Labbé Dominique (2007). "Experiments on Authorship Attribution by Intertextual Distance in English". Journal of Quantitative Linguistics, April 2007, 14-1. p 33-80.

Labbé Dominique & Monière Denis (2000), "La connexion intertextuelle. Application au discours gouvernemental québécois", Martin RAJMAN et Jean-Cédric

CHAPPELIER (eds), *Actes des 5e journées internationales d'analyse des données textuelles*, Lausanne, Ecole polytechnique fédérale, vol 1, p 85-94.

Labbé Dominique & Monière Denis (2003). Le vocabulaire gouvernemental. Canada, Québec, France (1945-2000). Paris: Champion.

Lafon Michel & Peeters Benoît (2006). Nous est un autre. Paris, Flammarion.

Lebart Ludovic, Morineau Alain & Fénelon Jean-Pierre (1982). Traitement des données statistiques. Méthodes et programmes. Paris: Dunod.

Lebart Ludovic, Morineau Alain & Tabard N. (1977). Techniques de la description statistique. Méthodes et logiciels pour la description des grands tableaux. Paris: Dunod.

Lebart Ludovic & Salem André (1994). Statistique textuelle. Paris: Dunod.

Lebart Ludovic, Salem André & Berry Lisette (1998). Exploring Textual Data. Dordrecht: Kluwer. Traduction anglaise de Lebart Ludovic & Salem André (1994).

Love Harold (2002). Attributing Authorship: An Introduction. Cambridge: Cambridge University Press.

Luong Xuan (1988). Méthodes d'analyse arborée. Algorithmes, applications. Thèse pour le doctorat ès sciences. Université de Paris V.

Luong Xuan dir. (2003). "La distance intertextuelle". Corpus. 2.

Mayaffre Damon (2004). Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la Ve République. Paris: Champion.

Merriam Thomas (2002). "Intertextual Distances between Shakespeare Plays, with Special Reference to *Henry V* (verse)". Journal of Quantitative Linguistics. 9-3, December 2002, p. 260-273.

Merriam Thomas (2003a). "An Application of Authorship Attribution by Intertextual Distance in English". Corpus, 2, 2003, p. 167-182.

Merriam Thomas (2003b). "Intertextual Distance, Three Authors". Literary and Linguistic Computing. 18-4. November 2003. p. 379-388.

Monière Denis & Labbé Dominique (2006). "L'influence des plumes de l'ombre sur les discours des politiciens". In Condé Claude & Viprey Jean-Marie (dir.). Actes des 8e Journées internationales d'Analyse des données textuelles. Besançon : 19-21 avril 2006, II, p 687-696.

http://halshs.archives-ouvertes.fr/halshs-00010477

Mongrédien Georges (1971). La querelle de l'Ecole des femmes. Paris: Didier.

Wouters Hippolyte & Ville de Goyer Christine de (1990). Molière ou l'auteur imaginaire ? Bruxelles : Complexe.

Annexe I. Les pièces de Corneille et de Molière

(N° Labbé & Labbé : JQL 2001)

N° *JQL 01*	N° *JQL 06*		Année probable de création	Genre*	Taille en mots
Corneille					
1	1	Mélite	1630	Comédie vers	16 690
2	2	Clitandre	1631	Tragi-comédie vers	14 402
3	3	La Veuve	1631	Comédie vers	17 661
4	4	La Galerie du Palais	1632	Comédie vers	16 140
5	5	La Suivante	1633	Comédie vers	15 160
6	6	Comédie des Tuileries	1634	Comédie vers	3 627
7	7	Médée	1635	Tragédie vers	14 269
8	8	La Place Royale	1634	Comédie vers	13 801
9	9	L'illusion comique	1636	Comédie vers	15 428
10	10	Le Cid	1636	Tragi-comédie vers	16 677
11	11	Cinna	1641	Tragédie vers	16 126
12	12	Horace	1640	Tragédie vers	16 482
13	13	Polyeucte	1641	Tragédie vers	16 472
14	14	Pompée	1642	Tragédie vers	16 492
15	15	Le Menteur 1	1642	Comédie vers	16 653
16	16	Le Menteur 2	1643	Comédie vers	17 675
17	17	Rodogune	1644	Tragédie vers	16 842
18	18	Théodore	1645	Tragédie vers	17 121
19	19	Héraclius	1647	Tragédie vers	17 433
20	20	Andromède	1650	Tragédie vers	15 514
21	21	Don Sanche	1650	Comédie héroïque vers	16 947
22	22	Nicomède	1651	Tragédie vers	16 923
23	23	Pertharite	1651	Tragédie vers	17 121
24	24	Oedipe	1659	Tragédie vers	18 618
25	25	Toison d'Or	1661	Tragédie vers	20 343
26	26	Sertorius	1662	Tragédie vers	17 675
27	27	Sophonisbe	1663	Tragédie vers	16 858
28	28	Othon	1664	Tragédie vers	16 971
29	29	Agésilas	1666	Tragédie vers	18 227
30	30	Atilla	1667	Tragédie vers	16 788
31	31	Tite et Bérénice	1670	Comédie héroïque vers	16 697
32	33**	Pulchérie	1672	Comédie héroïque vers	16 630
33	34**	Suréna	1674	Tragédie vers	16 545
Psyché					
34	***	*Psyché Corneille*	*1671*	*Tragi-comédie-ballet vers*	*10 067*
35	***	*Psyché Molière*	*1671*	*Tragi-comédie-ballet vers*	*4 816*
36	***	*Psyché Quinault*	*1671*	*Tragi-comédie-ballet vers*	*1 299*
Molière					
37	36	La jalousie	1660	Comédie prose	3 501
38	37	Médecin volant	1660	Comédie prose	3 876
39	38	L'étourdi	1660	Comédie vers	18 671
40	39	Dépit amoureux	1660	Comédie vers	16 242
41	40	Précieuses ridicules	1660	Comédie prose	6 648

42	41	Sganarelle	1660	Comédie vers	6 042
43	42	Dom Garcie	1661	Comédie héroïque vers	17 049
44	43	L'école des maris	1661	Comédie vers	10 536
45	44	Les fâcheux	1661	Comédie vers	7 922
46	45	L'école des femmes	1662	Comédie vers	16 625
47	46	Critique de l'école	1663	Comédie prose	8 610
48	47	L'impromptu	1663	Comédie prose	7 168
49	48	Mariage forcé	1664	Comédie prose	6 058
50	49	Princesse d'Elide	1664	Comédie vers et prose	11 333
51	50	Le Tartuffe	1664	Comédie vers	18 271
52	51	Dom Juan	1665	Comédie prose	17 452
53	52	L'amour médecin	1665	Comédie prose	6 147
54	53	Le Misanthrope	1666	Comédie vers	17 180
55	54	Médecin malgré lui	1666	Comédie prose	9 317
56	55	Mélicerte	1666	Comédie vers	5 540
****	56	Comédie pastorale	1667	Comédie vers libres	732
57	57	Le sicilien	1667	Comédie prose	5 375
58	58	Amphytrion	1668	Comédie vers libres	15 117
59	59	Georges Dandin	1668	Comédie prose	11 009
60	60	L'avare	1668	Comédie prose	21 033
61	61	M. de Pourceaugnac	1669	Comédie prose	11 803
62	62	Amants magnifiques	1670	Comédie vers libres & prose	11 983
63	63	Bourgeois gentilhomme	1670	Comédie prose	17 132
64	64	Fourberies de Scapin	1671	Comédie prose	14 245
65	65	Comtesse d'Escarbagnas	1671	Comédie prose	5 564
66	66	Femmes savantes	1672	Comédie vers	16 863
67	67	Malade imaginaire	1673	Comédie prose	19 919

* Selon l'indication portée sur la première édition ou sur l'édition de référence.
** Il n'y a pas de point n° 32 sur la figure 6 de *JQL 06*.
*** *JQL 06* aurait retiré Psyché de son expérience.
**** Pièce retirée dans l'expérience de 2001 (à cause de sa petite taille) mais maintenue dans l'expérience *JQL 06*.

Sources :

Corneille : Charles Marty-Laveaux. Œuvres complètes de P. Corneille. Paris : Hachette 1862. Collection Les Grands écrivains de la France.

Molière : Eugène Despois. Œuvres complètes de Molière. Paris : Hachette, 1876. Collection Les Grands écrivains de la France.

Le textes électroniques ont été remis par l'INaLF et aimablement transmis par M. Charles Bernet ; ces textes (ou leurs versions numériques ultérieures) sont consultables sur la base Frantext de l'ATILF.

Dominique Labbé a réalisé quelques ultimes corrections, la normalisation orthographique, le balisage et la lemmatisation. Ces données sont à la disposition des chercheurs qui en feront la demande et qui s'engageront à respecter les règles traditionnelles de la recherche, notamment en matière de publication.

Annexe 2
La distance intertextuelle

Etant donné la confusion des § 2.3-2.5 de *JQL 06* la présentation de la distance intertextuelle est résumée ci-dessous. Puis on relève certaines erreurs contenues dans la section 2 de cet article (reproduite en p. 14-15 de ce dossier).

Le calcul

Given 2 texts A and B:

V_a and V_b: number of types in A and B

F_{ia}: absolute frequency of the *ith* type in A

F_{ib}: absolute frequency of the *ith* type in B

N_a and N_b: number of tokens in A and B with $N_a = \sum F_{ia}$, $N_b = \sum F_{ib}$ and $N_a \leq N_b$

The mathematical expectancy of a type *i* in the text A, given his frequency in B (F_{ib}) is:

$$E_{ia(u)} = F_{ib} * U \quad with \ U = \frac{N_a}{N_b}$$

The total of the mathematical expectancies of the B types in A:

$$N'b = \sum_{1}^{V_b} E_{ia(u)}$$

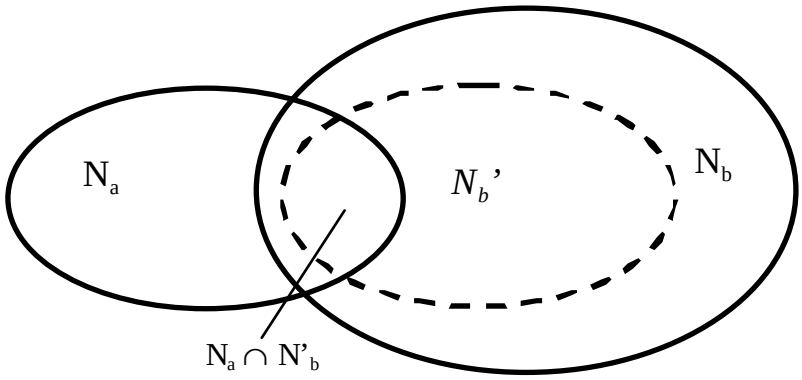
The index of the relative distance between texts A and B is :

$$(1) \quad Drel_{(A,B)} = \frac{\sum_{i \in (A,B)} \left| F_{ia} - E_{ia(u)} \right|}{\sum_{i \in A} F_{ia} + \sum_{i \in B} E_{ia(u)}} = \frac{\sum_{i \in (A,B)} \left| F_{ia} - E_{ia(u)} \right|}{N_a + N'_b}$$

The maximum is 1 (the two texts share no type), the minimum is 0 (same types used in A and B with the same frequencies).

The formula (1) simulates the reduction of the larger text (B) to the size of the smaller (A).

To make clear this reasonning, we added the following figure :



In our 2001 article, we comment these formulae:
"It is worth noting that the metric accuracy is slightly reduced by rounding. In fact, the observed frequencies are always integers whereas mathematical expectations include decimals which will contribute to the distance. This drawback will increase when low frequency types are an important part of the texts, that occurs in the case of small texts. To overcome this, it is convinient not to apply the calculation to too small texts —we never applied this calculation under the limit of 1 000 tokens (so that the small excerpt of the *Comédie pastorale* cannot be examined) — and to avoid a too large scale of sizes (around 1/10)[67]. In the application above, the shortest text counts 3 500 tokens — it is Molière's first comedy (see the appendix) and the largest counts 20 300 (Corneille's *Toison d'or*)[68]. For the same reasons, all results under .50 are eliminated from the numerator ($|F_{ia} - E_{ia(u)}| < .5$)."

Les erreurs de *JQL 06* (réponses à la section 2, p 14-15 de ce dossier)

La formule présentée dans cette section est fausse et ne correspond pas à la "notation ensembliste". Il faudrait au moins une intersection au "numérateur" (et non un delta) et, au "dénominateur", une union et non pas une intersection… Mais cette notation n'a pas le moindre intérêt pour la discussion qui suit.

[67] 2007 The correct expression is : "less than 1/10".
[68] The longest play is l'Avare (21 033 tokens).

Le schéma présenté en dessous de la formule est faux (voir le schéma exact en page précédente) :

- on confond B et B', c'est-à-dire le plus long des deux textes (toujours désigné par la lettre B) avec sa réduction (N'_b) à la longueur de A (N_a). Cette réduction est un point essentiel de notre raisonnement comme l'indique le texte ci-dessus ;

- la formule ajoutée à gauche du cadre ne signifie rien.

La fin du § 2.3 ajoute une nouvelle erreur : en cas de textes de même taille ($N_a = N_b$), le dénominateur de la formule est égal à $N_a + N_b$ (distance maximale possible entre deux textes n'ayant aucun mot en commun) et non pas à : $(N_a + N_b) - N_{a \cup b}$. Cette erreur rend caduque l'expérience présentée dans la section 4 (§ 4.6 à 4.8) qui est censée porter sur deux textes de taille égale.

Les paragraphes suivants contiennent deux nouvelles erreurs.

Le calcul de la distance intertextuelle porte sur tout le vocabulaire du petit texte (A). Donc, quand un vocable est présent dans A, il est toujours pris en compte dans ce calcul, quelle que soit sa fréquence dans le grand texte (B). Les seuls vocables de B qui sont exclus[69] du calcul sont ceux qui, absents de A, ont, du fait de leur faible fréquence en B, une espérance mathématique d'occurrence dans A inférieure à 1.

On verra, en lisant les § 4.4 à 4.9 puis 7.2 que cette première "erreur" n'en est peut-être pas une… En tous cas, notre article *JQL 01* et le programme de calcul sont clairs sur ce point.

Enfin, l'exclusion des différences inférieures à .50 n'affecte en rien le texte A ni le dénominateur de la formule de l'indice. L'objection présentée à la fin du § 2.5 n'a donc aucun sens.

[69] Dans notre texte, nous utilisons "we propose"… La justification de cette exclusion apparaît clairement dans le chapitre 8 (p. 106 de ce dossier).

Annexe 3.
Corneille – Molière et Racine

1. Distances intertextuelles entre les deux Menteurs (Corneille, comédies en vers), les Plaideurs (Racine, comédie en vers et les pièces représentées sous le nom de Molière.

N°	Pièces	Genre	Le Menteur (Corneille 1642)	Suite du Menteur (Corneille 1643)	Les plaideurs (Racine : 1668)
15	Le Menteur (1642)	Vers	0,000	0,180	0,296
16	La suite du Menteur (1643)	Vers	0,180	0,000	0,293
34	Psyché Corneille (1671)	Vers	0,288	0,273	0,348
35	Psyché Molière (1671)	Vers	0,329	0,325	0,354
37	La jalousie du barbouillé (avant 1660)	Prose	0,341	0,331	0,327
38	Médecin volant (avant 1660)	Prose	0,310	0,293	0,302
39	**L'étourdi (1658)**	Vers	**0,205**	**0,206**	0,269
40	**Dépit amoureux (1658)**	Vers	**0,215**	**0,212**	0,270
41	Précieuses ridicules (1660)	Prose	0,315	0,314	0,314
42	Sganarelle ou le cocu imagin. (1660)	Vers	0,259	0,253	0,293
43	Dom Garcie de Navarre (1661)	Vers	0,280	0,273	0,359
44	**L'école des maris (1661)**	Vers	**0,223**	**0,217**	0,279
45	**Les fâcheux (1661)**	Vers	**0,248**	**0,248**	0,306
46	**L'école des femmes (1662)**	Vers	**0,226**	**0,217**	0,261
47	Critique de l'école des femmes (1663)	Prose	0,323	0,319	0,340
48	L'impromptu de Versailles (1663)	Prose	0,321	0,316	0,323
49	Mariage forcé (1664)	Prose	0,322	0,302	0,320
50	**Princesse d'Elide (1664)**	Vers Prose	0,251	**0,243**	0,314
51	**Le Tartuffe (1664)**	Vers	**0,242**	**0,228**	0,275
52	**Dom Juan (1665)**	Prose	0,259	**0,248**	0,281
53	L'amour médecin (1665)	Prose	0,292	0,289	0,287
54	**Le Misanthrope (1666)**	Vers	0,252	**0,234**	0,283
55	Médecin malgré lui (1666)	Prose	0,298	0,289	0,296
56	**Mélicerte (1666)**	Vers	0,257	**0,250**	0,322
57	Le sicilien ou l'amour peintre (1667)	Prose	0,277	0,260	0,301
58	Amphytrion (1668)	Vers libres	0,253	0,256	0,297
59	Georges Dandin (1668)	Prose	0,292	0,279	0,292
60	**L'Avare (1668)**	Prose	0,256	**0,244**	0,270
61	M. de Pourceaugnac (1669)	Prose	0,292	0,283	0,285
62	Amants magnifiques (1670)	Prose	0,282	0,279	0,329
63	Bourgeois gentilhomme (1670)	Prose	0,294	0,280	0,286
64	Fourberies de Scapin (1671)	Prose	0,269	0,263	0,281
65	Comtesse d'Escarbagnas (1671)	Prose	0,311	0,300	0,305
66	**Femmes savantes (1672)**	Vers	0,260	**0,248**	0,283
67	Malade imaginaire (1672)	Prose	0,282	0,270	0,278
	Distance mean with Molière's work		*0,275*	*0,266*	*0,299*
	Mean with Molière's plays in verses		***0,241***	***0,234***	*0,290*
	Distance mean with Corneille's work		*0,252*	*0,249*	*0,347*
	Distance mean with Racine's work		*0,314*	*0,311*	*0,376*

Bien qu'écrits 16 ans avant la première pièce représentée sous le nom de Molière, les 2 Menteurs (Corneille) sont plus proches des pièces de Molière que de celles de Corneille (même contemporaines des Menteurs). Le très grand nombre de distances inférieures à .250 (en gras dans le tableau ci-dessus) ne se rencontre pas chez deux auteurs différents, même lorsqu'ils travaillent, en même temps, dans le même genre, sur les mêmes thèmes.

A titre de comparaison, la dernière colonne du tableau ci-dessus donne les distances entre les Plaideurs (comédie en alexandrins de Racine, contemporaine de Molière) et les œuvres représentées sous le nom de Molière : ces distances sont toujours supérieures à 0.260 et sont également très élevées avec les 2 Menteurs de Corneille.

Le même raisonnement s'applique à Dom Garcie et à Psyché (Molière), vis-à-vis des œuvres contemporaines de Corneille (tableau ci-dessous).

2. Distances entre Dom Garcie (Molière), Psyché (Corneille et Molière) et les pièces contemporaines de Corneille.

Last plays of Corneille	Dom Garcie (Molière,1661)	Psyché (Corneille & Molière, 1671)
Rodogune (1644)	0,245	0,231
Théodore (1645)	0,234	0,245
Héraclius (1647)	0,248	0,273
Andromède (1650)	0,241	**0,218**
DonSanche (1650)	**0,224**	0,251
Nicomède (1651)	0,244	0,264
Pertharite (1651)	0,235	0,263
Œdipe (1659)	**0,223**	**0,226**
Toison d'or (1661)	**0,221**	**0,220**
Sertorius (1662)	0,230	0,238
Sophonisbe (1663)	**0,228**	0,236
Othon (1664)	0,235	0,240
Agésilas (1666)	0,234	0,233
Attila (1667)	0,235	**0,227**
Tite et Bérénice (1670)	**0,227**	0,235
Psyché (1671)	***0,230***	—
Pulcherie (1672)	0,230	0,226
Surena (1674)	**0,216**	0,224
Mean Corneille	*0,243*	*0,244*
Mean Molière	*0,286*	*0,297*

3. Distances les plus faibles entre les pièces de Corneille et de Racine à l'époque de Tite et Bérénice.

	Tite et Bérénice (Corneille, 1670)	Bérénice (Racine, 1670)
CORNEILLE :		
Agésilas (1666)	0.159	0.278
Attila (1667)	0.180	0.289
Tite et Bérénice (1670)	0	**0.256**
Pulchérie (1672)	0.155	0.271
Suréna (1672)	0.156	0.264
RACINE :		
Andromaque (1667)	0.259	0.225
Britannicus (1669)	0.251	0.209
Bérénice (1670)	**0.256**	-
Bazajet (1672)	0.262	0.220
Mithridate (1673)	0.249	0.206

Le tableau ci-dessus donne les distances les plus faibles existant entre Corneille et Racine. Bien que travaillant tous deux, les mêmes années, dans le genre le plus contraignant (la tragédie en alexandrins) et parfois sur les mêmes thèmes (Bérénice), Corneille et Racine se distinguent parfaitement. Le "facteur genre" n'efface donc pas le "facteur auteur".

4. Les syntagmes verbaux "verbes modaux + infinitifs" (fréquence pour 100000 mots)

Corneille		Molière		Racine	
Combinations	Frequency	Combinations	Frequency	Combinations	Frequency
faire voir	***33,8***	***faire voir***	***31,5***	aller voir	12,0
pouvoir être	***18,8***	**pouvoir être**	***25,5***	**pouvoir voir**	**9,6**
pouvoir faire	***18,4***	**pouvoir faire**	***25,5***	faire entendre	9,0
faire naître	13,9	vouloir dire	24,9	**pouvoir faire**	8,4
pouvoir voir	**13,4**	***vouloir faire***	***19,5***	aller chercher	7,8
devoir être	12,7	pouvoir dire	14,5	faire parler	7,8
pouvoir souffrir	10,8	pouvoir avoir	13,7	**pouvoir être**	7,8
vouloir faire	***9,9***	aller faire	13,2	venir chercher	7,2
faire connaître	9,6	avoir faire	13,2	faire éclater	6,6
devoir faire	8,7	**pouvoir voir**	**12,3**	falloir partir	6,6

Racine partage seulement pouvoir (être, faire, voir) avec les deux autres, mais avec un classement et des densités fort différentes. En revanche, Molière et Corneille partagent 5 combinaisons sur 10 dont les trois premières dans le même ordre et avec des densités très proches. Etant donné le très grand nombre de combinaisons possibles, une telle caractéristique ne peut être le fait du hasard.

La situation devrait d'ailleurs être inverse puisque les œuvres de Corneille et de Racine sont toutes deux dominées par les tragédies en alexandrins alors qu'il n'y a que des comédies dans les pièces représentées sous le nom de Molière.

Annexe 4.
Un test simple
Relation entre la distance intertextuelle et la longueur des pièces de Corneille et Molière

JQL 06 affirme que l'indice de la distance intertextuelle - qui mesure la similarité de deux textes (A et B) - souffre d'une multitude de "biais" qui infirment l'attribution à P. Corneille des pièces en vers de Molière et de deux de ses pièces en prose (Dom Juan et L'Avare). Les deux principaux seraient que l'indice varie d'une part en raison inverse des longueurs des deux textes étudiés (notées Na, N_b) et d'autre part, dans le même sens que le rapport entre ces tailles (Na/Nb). Enfin, la section 6 introduit également le total des longueurs des deux textes ($N_a + N_b$).

Depuis 2001, il a été expliqué à de nombreuses reprises que la baisse de l'indice de la distance intertextuelle en fonction de l'allongement des textes est suffisamment lente pour pouvoir être négligée sur des corpus comme ceux des œuvres de Corneille et Molière et qu'elle n'a eu aucune incidence sur les résultats de nos recherches.

Puisque l'on feint d'ignorer ces explications, voici les tableaux de calcul.

1. L'indépendance de l'indice par rapport à la taille des textes

Dans notre article *JQL 01*, le point essentiel porte sur la faible distance entre les deux comédies de P. Corneille (Le Menteur et La Suite du Menteur) avec 10 pièces en vers de Molière et 2 en prose. On a donc deux tableaux de 12 lignes chacun (les pièces de Molière qui sont attribuées à Corneille) et de deux colonnes : la longueur en mots de chaque pièce et la valeur de l'indice de la distance entre la pièce considérée et l'un des deux Menteurs (ces valeurs sont données dans l'annexe 3 ci-dessus ; elles sont rappelées dans les 4 tableaux ci-dessous).

Il s'agit de tester l'hypothèse selon laquelle l'indice de la distance de l'un quelconque de ces 12 textes (numéroté i, i variant de 1 à 12) avec les Menteurs (distance notée y_i dans l'équation standard ci-dessous) est une variable inversement dépendante de la longueur (en mots) de ce texte i (longueur notée x_i dans l'équation standard ci-dessous).

Pour les textes étudiés (i variant de 1 à 12), cette dépendance supposée est vérifiée grâce au coefficient de corrélation linéaire (dit de Bravais-Pearson) :

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

Avec : $\bar{x}$ moyenne arithmétique simple des longueurs des 12 textes et $\bar{y}$ moyenne arithmétique simple des distances de ces douze textes au Menteur puis à la Suite du Menteur. Dans les tableaux ci-dessous, les expressions (x_i - $\bar{x}$) et (y_i - $\bar{y}$) sont notées X et Y.

Pièces de Molière attribuées à Corneille	Longueurs (mots)	Distance au Menteur	X^2	Y^2	XY
Etourdi	18 671	0,2048	14 940 802	0,00132	-140,57
Dépit Amoureux	16 242	0,2153	2 063 053	0,00067	-37,06
Ecole des Maris	10 536	0,2231	18 230 053	0,00033	77,15
Fâcheux	7 922	0,2477	47 384 867	0,00004	-44,79
Ecole des Femmes	16 625	0,2256	3 309 974	0,00024	-28,31
Elide	11 333	0,2514	12 059 414	0,00011	-35,60
Tartuffe	18 271	0,2416	12 008 535	0,00000	1,55
Dom Juan	17 452	0,2592	7 003 080	0,00033	47,76
Misanthrope	17 180	0,2524	5 637 459	0,00013	26,80
Mélicerte	5 540	0,2569	85 852 579	0,00025	-145,97
Avare	21 033	0,2560	38 779 680	0,00022	92,46
Femmes Savantes	16 863	0,2598	4 232 620	0,00035	38,39
Moyennes et sommes	14 806	0,2412	251502117	0,00397	-148,21

Tableau AIII.1 Tableau de calcul d'une éventuelle corrélation linéaire entre les distances avec le Menteur et la longueur (en mots) des pièces (de Molière) attribuées à Corneille

$$r = \frac{-148,21}{\sqrt{251\ 502\ 117 * 0,0397}} = -0,148$$

Le coefficient de corrélation mesure la force de la liaison existant entre deux variables. Ce coefficient varie entre 0 (absence de liaison) et 1 (liaison rigide et variation dans le même sens) ou –1 (liaison ridide mais variation inverse). Plus on s'approche de ±1, plus cette liaison est forte. Cependant, l'appréciation du coefficient dépend du nombre d'observations et, plus précisément, du nombre de "degrés de liberté" (effectifs du nombre d'observations moins deux).

Avec dix degrés de liberté (12-2), la limite d'acceptation de l'hypothèse est de ±0,708 (en acceptant un risque de première espèce de 1% de chances de se tromper en rejetant cette hypothèse), de ±0,567 (avec un risque de 5%) et de ±0,497 (avec un risque de 10%).

L'hypothèse d'une liaison linéaire entre les deux variables doit donc être rejetée sans aucune discussion possible.

Effectuons maintenant le même test avec la Suite du Menteur.

Pièces de Molière attribuées à Corneille	Longueurs (mots)	Distance à la Suite du Menteur	X^2	Y^2	XY
Etourdi	18 671	0,2060	14 940 802	0,00072	-103,54
Dépit Amoureux	16 242	0,2120	2 063 053	0,00043	-29,83
Ecole des Maris	10 536	0,2168	18 230 053	0,00026	68,25
Fâcheux	7 922	0,2477	47 384 867	0,00022	-102,52
Ecole des Femmes	16 625	0,2167	3 309 974	0,00026	-29,26
Elide	11 333	0,2425	12 059 414	0,00010	-33,93
Tartuffe	18 271	0,2284	12 008 535	0,00002	-15,13
Dom Juan	17 452	0,2471	7 003 080	0,00021	37,92
Misanthrope	17 180	0,2332	5 637 459	0,00000	0,95
Mélicerte	5 540	0,2499	85 852 579	0,00029	-158,90
Avare	21 033	0,2439	38 779 680	0,00012	69,56
Femmes Savantes	16 863	0,2490	4 232 620	0,00026	33,48
Moyennes et sommes	14 806	0,2328	251 502 117	0,00289	-262,96

Tableau AIII.2 Tableau de calcul d'une éventuelle corrélation linéaire entre les distances avec la Suite du Menteur et la longueur (en mots) des pièces (de Molière) attribuées à Corneille

$$r = -0,309$$

Pour les textes considérés, on peut conclure sans aucun risque d'erreur significatif à une absence de relation linéaire entre les valeurs de l'indice de la distance et la longueur de ces textes.

On note au passage que le test porte sur l'une des pièces les plus courte (Mélicerte) et sur la plus longue (L'Avare). Il couvre donc pratiquement toute l'amplitude des tailles du corpus Corneille-Molière…

Il reste toutefois à examiner la possibilité d'une liaison curviligne pour laquelle le modèle linéaire serait inadapté (Tableau AIII.3). Sur l'abscisse du graphique, la taille des textes dont on mesure la distance aux Menteurs, sur l'ordonnée, leurs distances intertextuelles avec le Menteur (point rond) et avec la Suite du Menteur (losanges).

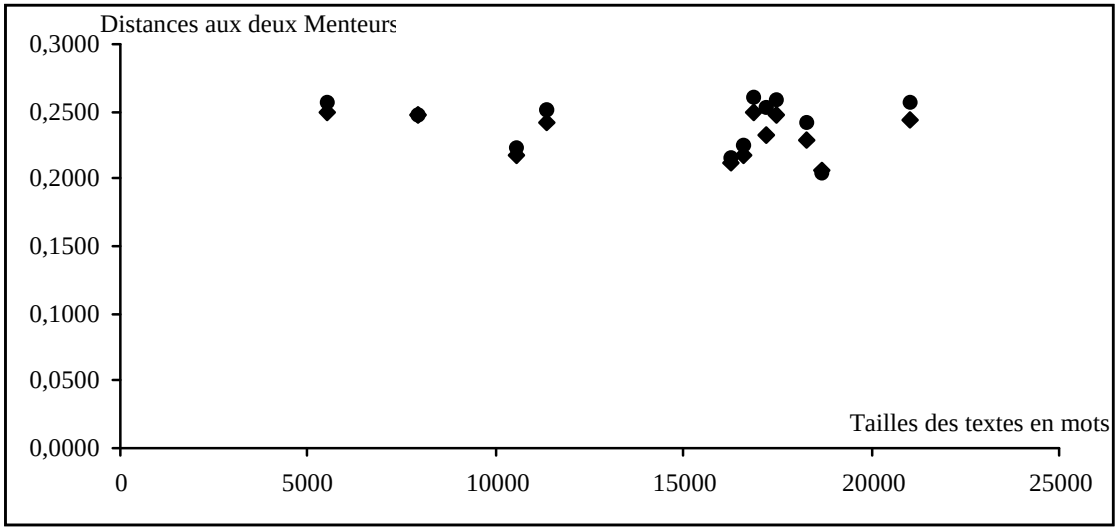


Tableau AIII.3 Examen graphique du nuage des points (valeurs de l'indice de la distance intertextuelle avec les deux Menteurs en fonction de la taille des textes comparés).

Le nuage de points est horizontal avec une légère dilatation dans la zone centrale où se trouvent la majorité des textes considérés. Cet examen suffit pour rejeter l'hypothèse d'une liaison curviligne.

Il n'y a aucune relation statistique avérée entre la longueur des pièces de Molière attribuées à Corneille et leurs distances intertextuelles aux deux Menteurs… Les longueurs de ces pièces ne jouent aucun rôle dans leur très grande parenté avec ces deux comédies de Corneille.

L'affirmation de *JQL 06* peut être rejetée sans aucun risque d'erreur significatif.

2. L'indépendance de l'indice de la distance avec le rapport Na/Nb

Examinons maintenant la seconde affirmation : l'indice de la distance diminuerait au fur et à mesure qu'augmenteraient les différences de longueurs entre les textes comparés et cette propriété expliquerait la parenté observée entre les deux Menteurs et les pièces attribuées à Corneille.

JQL 06 mesure ces différences de longueurs grâce au ratio noté "*Na/Nb*" (avec N_a taille du petit texte et N_b taille du plus grand). Il s'agirait d'un second "biais" différent de celui qui vient d'être examiné ci-dessus :

DCL do not explicitly mention this bias related to the quotient Na/Nb. *They allude to it (as well as to the bias related to* Na *and* Nb) (§ 4.3, dans ce dossier, p. 28).

Le Menteur compte 16 653 mots et La Suite du Menteur : 17 675. Ces longueurs sont rapportées à celles des pièces attribuées et ce rapport devient la variable indépendante ou "explicative" (seconde colonne des deux tableaux ci-dessous).

Pièces de Molière attribuées à Corneille	ratio N_a/N_b	Distance au Menteur	X^2	Y^2	XY
Etourdi	1,1212	0,2048	0,0539	0,00132	-0,0084
Dépit Amoureux	0,9753	0,2153	0,0074	0,00067	-0,0022
Ecole des Maris	0,6327	0,2231	0,0657	0,00033	0,0046
Fâcheux	0,4757	0,2477	0,1709	0,00004	-0,0027
Ecole des Femmes	0,9983	0,2256	0,0119	0,00024	-0,0017
Elide	0,6805	0,2514	0,0435	0,00011	-0,0021
Tartuffe	1,0972	0,2416	0,0433	0,00000	0,0001
Dom Juan	1,0480	0,2592	0,0253	0,00033	0,0029
Misanthrope	1,0316	0,2524	0,0203	0,00013	0,0016
Mélicerte	0,3327	0,2569	0,3096	0,00025	-0,0088
Avare	1,2630	0,2560	0,1398	0,00022	0,0056
Femmes Savantes	1,0126	0,2598	0,0153	0,00035	0,0023
Moyennes et sommes	0,8891	0,2412	0,9069	0,00397	-0,0089

Tableau AIII.4 Tableau de calcul d'une éventuelle corrélation linéaire entre les distances au Menteur (16 653 mots) et le ratio Na/Nb pour les pièces (de Molière) attribuées à Corneille.

$$r = - 0,148$$

Le résultat est exactement le même que celui obtenu dans le premier calcul ci-dessus.

Pièces de Molière attribuées à Corneille	ratio N_a/N_b	Distance à la Suite du Menteur	X^2	Y^2	XY
Etourdi	1,056	0,2060	0,0478	0,00072	-0,0059
Dépit Amoureux	0,919	0,2120	0,0066	0,00043	-0,0017
Ecole des Maris	0,596	0,2168	0,0584	0,00026	0,0039
Fâcheux	0,448	0,2477	0,1517	0,00022	-0,0058
Ecole des Femmes	0,941	0,2167	0,0106	0,00026	-0,0017
Elide	0,641	0,2425	0,0386	0,00010	-0,0019
Tartuffe	1,034	0,2284	0,0384	0,00002	-0,0009
Dom Juan	0,987	0,2471	0,0224	0,00021	0,0021
Misanthrope	0,972	0,2332	0,0180	0,00000	0,0001
Mélicerte	0,313	0,2499	0,2748	0,00029	-0,0090
Avare	1,190	0,2439	0,1241	0,00012	0,0039
Femmes Savantes	0,954	0,2490	0,0135	0,00026	0,0019
Moyennes et sommes	0,838	0,2328	0,8050	0,00289	-0,0149

Tableau III.5 Tableau de calcul d'une éventuelle corrélation linéaire entre les distances avec la Suite du Menteur (17 675 mots) et le ratio Na/Nb pour les pièces (de Molière) attribuées à Corneille.

$$r = -0,309$$

Même calcul que ci-dessus, même résultat et même conclusion : pour ce corpus, il n'y a aucune liaison entre les valeurs de l'indice de la distance intertextuelle (avec la Suite du Menteur) et le rapport entre la longueur du texte comparé avec celle de la Suite du Menteur.

Dans *JQL 06*, spécialement le § 6.8, la somme Na + Nb est présentée comme une nouvelle dimension du problème.

Les résultats des calculs avec ces sommes sont respectivement : - 0,148 et -0,309…

Le lecteur l'avait deviné : il s'agit de la même relation exprimée de trois manières différentes !

JQL 06 n'apporte pas le moindre démenti à la conclusion principale de notre article de 2001 : la proximité remarquable de 12 pièces de Molière avec les deux Menteurs de Corneille…

Annexe 5
Deux nouvelles de G. de Maupassant

La tombe

Le dix-sept juillet mil huit cent quatre-vingt-trois, à deux heures et demie du matin, le gardien du cimetière de Béziers, qui habitait un petit pavillon au bout du champ des morts, fut réveillé par les jappements de son chien enfermé dans la cuisine.

Il descendit aussitôt et vit que l'animal flairait sous la porte en aboyant avec fureur, comme si quelque vagabond eût rôdé autour de la maison. Le gardien Vincent prit alors son fusil et sortit avec précaution.

Son chien partit en courant dans la direction de l'allée du général Bonnet et s'arrêta net auprès du monument de MadameTomoiseau.

Le gardien, avançant alors avec précaution, aperçut bientôt une petite lumière du côté de l'allée Malenvers. Il se glissa entre les tombes et fut témoin d'un acte horrible de profanation.

Un homme avait déterré le cadavre d'une jeune femme ensevelie la veille, et il le tirait hors de la tombe.

Une petite lanterne sourde, posée sur un tas de terre, éclairait cette scène hideuse.

Le gardien Vincent, s'étant élancé sur ce misérable, le terrassa, lui lia les mains et le conduisit au poste de police.

C'était un jeune avocat de la ville, riche, bien vu, du nom de Courbataille.

Il fut jugé. Le ministère public rappela les actes monstrueux du sergent Bertrand et souleva l'auditoire.

Des frissons d'indignation passaient dans la foule. Quand le magistrat s'assit, des cris éclatèrent: "A mort! A mort!" Le président eut grand'peine à faire rétablir le silence.

Puis il prononça d'un ton grave:

"Prévenu qu'avez-vous à dire pour votre défense?"

Courbataille, qui n'avait point voulu d'avocat, se leva. C'était un beau garçon, grand, brun, avec un visage ouvert, des traits énergiques, un œil hardi.

Des sifflets jaillirent du public.

Il ne se troubla pas, et se mit à parler d'une voix un peu voilée, un peu basse d'abord, mais qui s'affermit peu à peu.

"Monsieur le président,

"Messieurs les jurés,

"J'ai très peu de choses à dire. La femme dont j'ai violé la tombe avait été ma maîtresse. Je l'aimais.

"Je l'aimais, non point d'un amour sensuel, non point d'une simple tendresse d'âme et de cœur, mais d'un amour absolu, complet, d'une passion éperdue.

"Ecoutez-moi:

"Quand je l'ai rencontrée pour la première fois, j'ai ressenti, en la voyant, une étrange sensation. Ce ne fut point de l'étonnement, ni de l'admiration, ce ne fut point ce qu'on appelle le coup de foudre, mais un sentiment de bien-être délicieux, comme si on m'eût plongé dans un bain tiède. Ses gestes me séduisaient, sa voix me ravissait, toute sa personne me faisait un plaisir infini à regarder. Il me semblait aussi que je la connaissais depuis longtemps, que je l'avais vue déjà. Elle portait en elle quelque chose de mon esprit.

"Elle m'apparaissait comme une réponse à un appel jeté par mon âme, à cet appel vague et continu que nous poussons vers l'Espérance durant tout le cours de notre vie.

"Quand je la connus un peu plus, la seule pensée de la revoir m'agitait d'un trouble exquis et profond; le contact de sa main dans ma main était pour moi un tel délice que je n'en avais point imaginé de semblable auparavant, son sourire me versait dans les yeux une allégresse folle, me donnait envie de courir, de danser, de me rouler par terre.

"Elle devint donc ma maîtresse.

"Elle fut plus que cela, elle fut ma vie même. Je n'attendais plus rien sur la terre, je ne désirais rien, plus rien. Je n'enviais plus rien.

"Or, un soir, comme nous étions allés nous promener un peu plus loin le long de la rivière, la pluie nous surprit. Elle eut froid.

"Le lendemain une fluxion de poitrine se déclara. Huit jours plus tard elle expirait.

"Pendant les heures d'agonie, l'étonnement, l'effarement m'empêchèrent de bien comprendre, de bien réfléchir.

"Quand elle fut morte, le désespoir brutal m'étourdit tellement que je n'avais plus de pensée. Je pleurais.

"Pendant toutes les horribles phases de l'ensevelissement ma douleur aiguë, furieuse, était encore une douleur de fou, une sorte de douleur sensuelle, physique.

"Puis quand elle fut partie, quand elle fut en terre, mon esprit redevint net tout d'un coup et je passai par toute une suite de souffrances morales si épouvantables que l'amour même qu'elle m'avait donné était cher à ce prix-là.

"Alors entra en moi cette idée fixe:

"Je ne la reverrai plus."

Quand on réfléchit à cela pendant un jour tout entier, une démence vous emporte! Songez! Un être est là, que vous adorez, un être unique car dans toute l'étendue de la terre il n'en existe pas un second qui lui ressemble. Cet être s'est donné à vous, il crée avec vous cette union mystérieuse qu'on nomme l'Amour. Son œil vous semble plus vaste que l'espace, plus charmant que le monde, son œil clair où sourit la tendresse. Cet être vous aime. Quand il vous parle, sa voix vous verse un flot de bonheur.

"Et tout d'un coup il disparaît! Songez! Il disparaît non pas seulement pour vous, mais pour toujours. Il est mort. Comprenez-vous ce mot? jamais, jamais, jamais, nulle part, cet être n'existera plus. Jamais cet œil ne regardera plus rien; jamais cette voix, jamais une voix pareille, parmi toutes les voix humaines, ne prononcera de la même façon un des mots que prononçait la sienne.

"Jamais aucun visage ne renaîtra semblable au sien. Jamais, jamais! On garde les moules des statues; on conserve des empreintes qui refont des objets avec les mêmes contours et les mêmes couleurs. Mais ce corps et ce visage, jamais ils ne reparaîtront sur la terre. Et pourtant il en naîtra des milliers de créatures, des millions, des milliards, et bien plus encore, et parmi toutes les femmes futures, jamais celle-là ne se retrouvera. Est-ce possible? On devient fou en y songeant!

"Elle a existé vingt ans, pas plus, et elle a disparu pour toujours, pour toujours, pour toujours! Elle pensait, elle souriait, elle m'aimait. Plus rien. Les mouches qui meurent à l'automne sont autant que nous dans la création. Plus rien! Et je pensais que son corps, son corps frais, chaud, si doux, si blanc, si beau, s'en allait en pourriture dans le fond d'une boîte sous la terre. Et son âme, sa pensée, son amour, où?

"Ne plus la revoir! Ne plus la revoir! L'idée me hantait de ce corps décomposé, que je pourrais peut-être reconnaître pourtant. Et je voulus la regarder encore une fois!

"Je partis avec une bêche, une lanterne, un marteau. Je sautai par-dessus le mur du cimetière. Je retrouvai le trou de sa tombe; on ne l'avait pas encore tout à fait rebouché.

"Je mis le cercueil à nu. Et je soulevai une planche. Une odeur abominable, le souffle infâme des putréfactions me monta dans la figure. Oh! son lit, parfumé d'iris!

"J'ouvris la bière cependant, et je plongeai dedans ma lanterne allumée, et je la vis. Sa figure était bleue, bouffie, épouvantable! Un liquide noir avait coulé de sa bouche.

"Elle! c'était elle! Une horreur me saisit. Mais j'allongeai le bras et je pris ses cheveux pour attirer à moi cette face monstrueuse!

"C'est alors qu'on m'arrêta.

"Toute la nuit j'ai gardé, comme on garde le parfum d'une femme après une étreinte d'amour, l'odeur immonde de cette pourriture, l'odeur de ma bien-aimée!

"Faites de moi ce que vous voudrez."

Un étrange silence paraissait peser sur la salle. On semblait attendre quelque chose encore. Les jurés se retirèrent pour délibérer.

Quand ils rentrèrent au bout de quelques minutes, l'accusé semblait sans craintes, et même sans pensée.

Le président, avec les formules d'usage, lui annonça que les juges le déclaraient innocent.

Il ne fit pas un geste, et le public applaudit.

(Texte publié dans Gil Blas du 29 juillet 1884, sous la signature de Maufrigneuse)

Le remplaçant

"Madame Bonderoi ?

- Oui, Madame Bonderoi.

- Pas possible ?

- Je-vous-le-dis.

- Madame Bonderoi, la vieille dame à bonnets de dentelle, la dévote, la sainte, l'honorable Madame Bonderoi dont les petits cheveux follets et faux ont l'air collé, autour du crâne ?

- Elle-même.

- Oh ! voyons, vous êtes fou ?

- Je-vous-le-jure.

- Alors, dites-moi tous les détails ?

- Les voici. Du temps de Monsieur Bonderoi, l'ancien notaire, Madame Bonderoi utilisait, dit-on, les clercs pour son service particulier. C'est une de ces respectables bourgeoises à vices secrets et à principes inflexibles, comme il en est beaucoup. Elle aimait les beaux garçons ; quoi de plus naturel ? N'aimons-nous pas les belles filles ?

Une fois que le père Bonderoi fut mort, la veuve se mit à vivre en rentière paisible et irréprochable. Elle fréquentait assidûment l'église, parlait dédaigneusement du prochain, et ne laissait rien à dire sur elle.

Puis elle vieillit, elle devint la petite bonne femme que vous connaissez, pincée, sûrie, mauvaise.

Or, voici l'aventure invraisemblable arrivée jeudi dernier :

Mon ami Jean d'Anglemare est, vous le savez, capitaine aux dragons, caserné dans le faubourg de la Rivette.

En arrivant au quartier, l'autre matin, il apprit que deux hommes de sa compagnie s'étaient flanqué une abominable tripotée. L'honneur militaire a des lois sévères. Un duel eut lieu. Après l'affaire, les soldats se réconcilièrent, et interrogés par leur officier, lui racontèrent le sujet de la querelle. Ils s'étaient battus pour Madame Bonderoi.

- Oh !

- Oui, mon ami, pour Madame Bonderoi !"

Mais je laisse la parole au cavalier Siballe :

"Voilà l'affaire, mon capitaine. Y a z'environ dix-huit mois, je me promenais sur le cours, entre six et sept heures du soir, quand une particulière m'aborda.

Elle me dit, comme elle m'avait demandé son chemin : "Militaire, voulez-vous gagner honnêtement dix francs par semaine ?"

Je lui répondis sincèrement : "A vot' service, madame."

Alors ell' me dit : "Venez me trouver demain, à midi. Je suis Madame Bonderoi, 6, rue de la Tranchée.

- J' n'y manquerai pas, madame, soyez tranquille."

Puis, ell' me quitta d'un air content en ajoutant : "Je vous remercie bien, militaire.

- C'est moi qui vous remercie, madame."

Ça ne laissa pas que d'me taquiner jusqu'au lendemain.

A midi, je sonnais chez elle.

Ell' vint m'ouvrir elle-même. Elle avait un tas de petits rubans sur la tête.

"Dépêchons-nous, dit-elle, parce que ma bonne pourrait rentrer."

Je répondis : "Je veux bien me dépêcher. Qu'est-ce qu'il faut faire ?"

Alors, elle se mit à rire et riposta : "Tu ne comprends pas, gros malin ?"

Je n'y étais plus, mon capitaine, parole d'honneur.

Ell' vint s'asseoir tout près de moi, et me dit : "Si tu répètes un mot de tout ça, je te ferai mettre en prison. Jure que tu seras muet."

Je lui jurai ce qu'ell' voulut. Mais je ne comprenais toujours pas. J'en avais la sueur au front. Alors je retirai mon casque oùsqu'était mon mouchoir. Elle le prit, mon mouchoir, et m'essuya les cheveux des tempes. Puis v'là qu'ell' m'embrasse et qu'ell' me souffle dans l'oreille :

"Alors, tu veux bien ?"

Je répondis : "Je veux bien ce que vous voudrez, madame, puisque je suis venu pour ça."

Alors ell' se fit comprendre ouvertement par des manifestations. Quand j'vis de quoi il s'agissait, je posai mon casque sur une chaise ; et je lui montrai que dans les dragons on ne recule jamais, mon capitaine.

Ce n'est pas que ça me disait beaucoup, car la particulière n'était pas dans sa primeur. Mais y ne faut pas se montrer trop regardant dans le métier, vu que les picaillons sont rares. Et puis on a de la famille qu'il faut soutenir. Je me disais : "Y aura cent sous pour le père, là-dessus."

Quand la corvée a été faite, mon capitaine, je me suis mis en position de me retirer. Elle aurait bien voulu que je ne parte pas sitôt. Mais je lui dis : "Chacun son dû, madame. Un p'tit verre ça coûte deux sous, et deux p'tits verres, ça coûte quatre sous."

Ell' comprit bien le raisonnement et me mit un p'tit napoléon de dix balles au fond de la main. Ça ne m'allait guère, c'te monnaie-là, parce que ça vous coule dans la poche, et quand les pantalons ne sont pas bien cousus, on la retrouve dans ses bottes, ou bien on ne la retrouve pas.

Alors que je regardais ce pain à cacheter jaune en me disant ça, ell' me contemple ; et puis ell' devient rouge, et ell' se trompe sur ma physionomie, et ell' me demande :

"Est-ce que tu trouves que c'est pas assez ?" Je lui réponds :

"Ce n'est pas précisément ça, madame, mais, si ça ne vous faisait rien, j'aimerais mieux deux pièces de cent sous."

Ell' me les donna et je m'éloîgnai.

Or, voilà dix-huit mois que ça dure, mon capitaine. J'y vas tous les mardis, le soir, quand vous consentez à me donner permission. Elle aime mieux ça, parce que sa bonne est couchée.

Or donc, la semaine dernière, je me trouvai indisposé ; et il me fallut tâter de l'infirmerie. Le mardi arrive, pas moyen de sortir ; et je me mangeais les sangs par rapport aux dix balles dont je me trouve accoutumé.

Je me dis : "Si personne y va, je suis rasé ; qu'elle prendra pour sûr un artilleur." Et ça me révolutionnait.

Alors, je fais demander Paumelle, que nous sommes pays ; et je lui dis la chose : "Y aura cent sous pour toi, cent sous pour moi, c'est convenu."

Y consent, et le v'là parti. J'y avais donné les renseignements. Y frappe ; ell' ouvre ; ell' le fait entrer ; ell' l'y regarde pas la tête et s'aperçoit point qu'c'est pas le même.

Vous comprenez, mon capitaine, un dragon et un dragon, quand ils ont le casque, ça se ressemble.

Mais soudain, elle découvre la transformation, et ell' demande d'un air de colère :

"Qu'est-ce que vous êtes ? Qu'est-ce que vous voulez ? Je ne vous connais pas, moi ?"

Alors Paumelle s'explique. Il démontre que je suis indisposé et il expose que je l'ai envoyé pour remplaçant.

Elle le regarde, lui fait aussi jurer le secret, et puis elle l'accepte, comme bien vous pensez, vu que Paumelle n'est pas mal aussi de sa personne.

Mais quand ce limier-là fut revenu, mon capitaine, il ne voulait plus me donner mes cent sous. Si ça avait été pour moi, j'aurais rien dit, mais c'était pour le père ; et là-dessus, pas de blague.

Je lui dis :

"T'es pas délicat dans tes procédés, pour un dragon, que tu déconsidères l'uniforme."

Il a levé la main, mon capitaine, en disant que c'te corvée-là, ça valait plus du double.

Chacun son jugement, pas vrai ? Fallait point qu'il accepte. J'y ai mis mon poing dans le nez. Vous avez connaissance du reste.

Le capitaine d'Anglemare riait aux larmes en me disant l'histoire. Mais il m'a fait aussi jurer le secret qu'il avait garanti aux deux soldats.

"Surtout, n'allez pas me trahir, gardez ça pour vous, vous me le promettez ?

- Oh ! ne craignez rien. Mais comment tout cela s'est-il arrangé en définitive ?

- Comment ? Je vous le donne en mille ! ... La mère Bonderoi garde ses deux dragons, en leur réservant chacun leur jour. De cette façon, tout le monde est content.

- Oh ! elle est bien bonne, bien bonne !

- Et les vieux parents ont du pain sur la planche. La morale est satisfaite."

(Texte publié dans Gil Blas du 2 janvier 1883 sous le titre "Les remplaçants" et signé Maufrigneuse, puis dans le recueil Mademoiselle Fifi.)

Annexe 6
Les Nouvelles de Maupassant sélectionnées par *JQL 06**
(classement des titres par ordre alphabétique)

N° Viprey	Titre	Longueur	Vocables	Ponctuations
44	A Cheval**	2 387	746	486
94	Allouma	8 133	1 585	1 437
45	Ami Joseph (L')	1 855	629	329
99	Après	2 204	678	333
46	Auprès d'un mort	1 387	484	267
5	Aux Champs**	2 091	597	436
47	Aventure de W. Schnaffs	2 927	893	481
3	Aventure parisienne	2 344	767	518
72	Aveu (L')**	1 774	559	370
6	Aveugle	1 298	486	195
87	Bête à Maît' Belhomme (La)**	2 733	759	599
71	Bonheur (Le)	2 191	704	419
8	Bûche (la)	1 861	612	413
9	Ce cochon de Morin	4 115	960	928
97	Champ d'oliviers (Le)	9 346	1 795	1 780
10	Clair de lune	1 392	468	270
10bis	Clair de Lune	1 954	654	296
73	Coco	1 488	517	263
100	Colporteur (Le)	2 524	740	470
48	Confession (La)	1 789	473	418
48bis	Confession (La)	2 641	723	536
48ter	Confession (La)	2 143	666	481
11	Confessions d'une femme	1 643	578	299
34	Coq chanta (Un)	1 669	581	332
12	Correspondance	1 921	632	311
89	Cri d'alarme	2 365	624	587
74	Crime au père Boniface (Le)**	1 936	626	357
49	Denis	2 362	699	435
50	Deux Amis	2 253	711	458
68	Duel**	1 576	517	322
51	En Mer**	1 885	613	336
90	Epave (L')**	4 182	1 065	768
93	Etrennes	1 989	556	454
52	Farce (La)	2 021	638	352
13	Farce normande**	1 718	605	320
91	Fermier (Le)**	2 517	700	551
53	Ficelle (La)**	2 386	760	455
35	Fils (un)	3 619	1 000	689
14	Folle (La)**	1 137	422	185
15	Fou	1 555	510	295

16	Gâteau	1 596	528	280
75	Gueux (Le)**	1 960	618	284
95	Hautot	4 549	1 028	982
17	Histoire vraie**	1 900	565	379
54	Humble Drame	1 958	600	349
76	Ivrogne (L')**	1 841	577	351
85	Lâche (un)	2 913	777	539
77	Lettre trouvée sur un noyé (La)	2 007	640	362
18	Lit (Le)	1 295	470	218
19	Loup (Le)	2 035	661	325
20	Madame Baptiste	2 186	683	373
21	Mademoiselle Fifi	4 666	1 342	840
92	Mademoiselle Perle	5 980	1 252	1 139
55	Main (La)**	2 170	705	398
2	MaisonTellier	10 455	2 211	1 781
22	Marroca	3 441	1 003	651
23	Menuet	1 628	588	266
78	Mère Sauvage (La)	2 651	784	447
88	Mes 25 Jours	2 842	835	437
56	Mon Oncle Jules	2 716	764	488
57	Monsieur Jocaste	1 607	487	283
24	Mots d'amour	1 297	436	228
98	Mouche	3 330	991	591
71	Normand (Un)**	2 231	693	441
79	Notes d'un voyageur	1 713	638	304
25	Nuit de Noël	1 496	527	333
58	Orphelin (L')	2 357	707	402
72	Parricide (Un)	2 206	621	434
4	Partie de campagne	4 385	1 163	764
80	Parure (La)	2 929	811	577
101	Père (Le)	1 586	547	316
101bis	Père (Le)	3 055	794	636
59	Père Milon (Le)**	2 296	689	442
60	Petit (Le)	1 982	613	435
81	Petit Fût (Le)**	1 981	565	393
26	Peur (La)	2 328	725	420
26bis	Peur (La)	2 605	755	445
27	Pierrot	1 883	599	383
61	Première Neige	2 898	799	551
28	Relique (La)	1 725	581	319
29	Rempailleuse (La)	2 421	676	446
62	Remplacant (Le)**	1 316	395	311
63	Réveil	1 862	604	309
38	Réveillon (Un)	1 978	669	405

30	Roche aux Guillemots (La)**	1 449	533	292
82	Rose	2 054	676	396
31	Rouerie	1 988	565	416
39	Ruse (Une)	1 991	614	471
64	Sabots (Les)**	1 879	542	404
65	Saint Antoine**	3 059	863	569
32	Saut du Berger (Le)**	1 463	605	202
66	Serre (La)	1 905	608	423
96	Soir (Un)	6 234	1 455	1 225
83	Souvenir	1 474	530	322
83bis	Souvenir	2 343	716	444
1	Sur l'Eau	2 407	688	356
33	Testament (Le)	1 720	569	292
84	Tombe	1 393	489	257
67	Tombouctou**	2 482	800	464
40	Veillée (La)	1 413	510	304
69	Vendetta (Une)**	1 665	553	342
70	Vengeur (Le)	1 706	495	389
86	Vieux (Le)**	2 520	725	491
41	Vieux Objets	1 386	432	229
42	Voleur (Le)	1 597	557	337
43	Yveline Samoris	1 293	463	262
Total		258 991	10 537	49 175

Notes

* Au catalogue de la Bibliothèque nationale, il ne figure pas de recueil des nouvelles de Maupassant, aux éditions Conard pour l'année 1929. Il est donc impossible de savoir laquelle de ces nouvelles ont été choisies :

- Clair de Lune. Deux nouvelles de Maupassant portent ce titre. La première a été publiée dans Le Gaulois du 1er juillet 1882, la seconde dans Gil Blas le 19 octobre 1882 sous la signature de Maufrigneuse.
- La confession. Trois nouvelles portent ce titre. La première a été publiée dans Le Gaulois du 21 octobre 1883 sous le titre L'aveu, puis sous le titre La confession dans le recueil Contes du jour et de la nuit. La seconde dans Gil Blas du 12 août 1884. La troisième dans Le Figaro du 10 novembre 1884.
- Le Père. Deux nouvelles portent ce titre. La première a été publiée dans Gil Blas du 20 novembre 1883, sous la signature de Maufrigneuse, la seconde dans Gil Blas du 26 juillet 1887.
- La Peur. Deux nouvelles portent ce titre. La première a été publiée dans Le Gaulois du 23 octobre 1882, la seconde dans Le Figaro du 25 juillet 1884.
- Souvenir. Deux nouvelles portent ce titre. La première a été publiée dans Gil Blas du 16 février 1882, sous la signature de Maufrigneuse, la seconde dans Gil Blas du 20 mai 1884, sous la signature de Maufrigneuse.

Elles ont donc toutes été traitées dans notre expérience.

** Les nouvelles dont les titres sont suivies d'un astérisque contiennent une proportion telle de "jargon" qu'elle est susceptible d'augmenter significativement leur distance par rapport aux autres.

Les nouvelles de G. de Maupassant ont été déchargées sur le site de Thierry Selva. Les transcriptions ont été contrôlées sur l’édition de la Pléiade des Œuvres complètes de G. de Maupassant.

Ces 106 textes normalisés et lemmatisés sont à la disposition des chercheurs.

Annexe 7
Flaubert et Maupassant

Pourquoi Gustave Flaubert (1821-1880) et Guy de Maupassant (1850-1893) et plus spécialement Madame Bovary (1857), l'Education sentimentale (1869) et Une vie (1883) ? L'histoire littéraire a enregistré l'influence que le premier a exercé sur le second. Nous avons signalé, dans l'essai de 2003, l'étrange proximité de ces trois romans pourtant séparés par un laps de temps considérable pour l'histoire littéraire (une génération pendant laquelle on passe du romantisme au naturalisme).

Outre les dates de publication, les tableaux ci-dessous indiquent également les tailles de ces textes qui sont oubliées une fois de plus par *JQL 06*...

Les textes ont été standardisés et lemmatisés selon les normes utilisées pour tous nos corpus (Labbé, 1990).

La distance intertextuelle est calculée grâce à la méthode de la "fenêtre glissante" (présentée en conclusion de Labbé & Labbé 2003 et exposée exhaustivement dans : Labbé 2007). Cette fenêtre est ici à la taille du plus petit texte (Hérodias, 10 599 mots) et glisse le long des textes selon un pas de 2.000 mots. Cette taille ramène les distances dans l'intervalle de validité de l'échelle de la distance intertextuelle. La matrice de 12x12 ne pouvant être reproduite, elle a été découpée en trois tableaux : les distances internes aux œuvres de chacun des auteurs puis celles séparant les deux auteurs.

Tableau AVI.3 Distances internes à l'œuvre de G. Flaubert

	Bovary	Salambo	Education sentimentale	Cœur simple	Hérodias	Bouvard & Pécuchet
Date (première pub.)	1857	1862	1869	1877	1877	1881
Longueurs (mots)	122 660	109 378	152 890	12 161	10 599	95 238
Bovary	-	0,328	**0,260**	0,293	0,349	0,308
Salambo	0,328	-	0,338	0,322	0,296	0,323
Education sentimentale	0,260	0,338	-	0,294	0,344	0,293
Cœur simple	0,293	0,322	0,294	-	0,334	0,310
Hérodias	0,349	0,296	0,344	0,334	-	0,342
Bouvard & Pécuchet	0,308	0,323	0,293	0,310	0,342	-
Moyenne	0,308	0,321	0,306	0,310	0,333	0,315

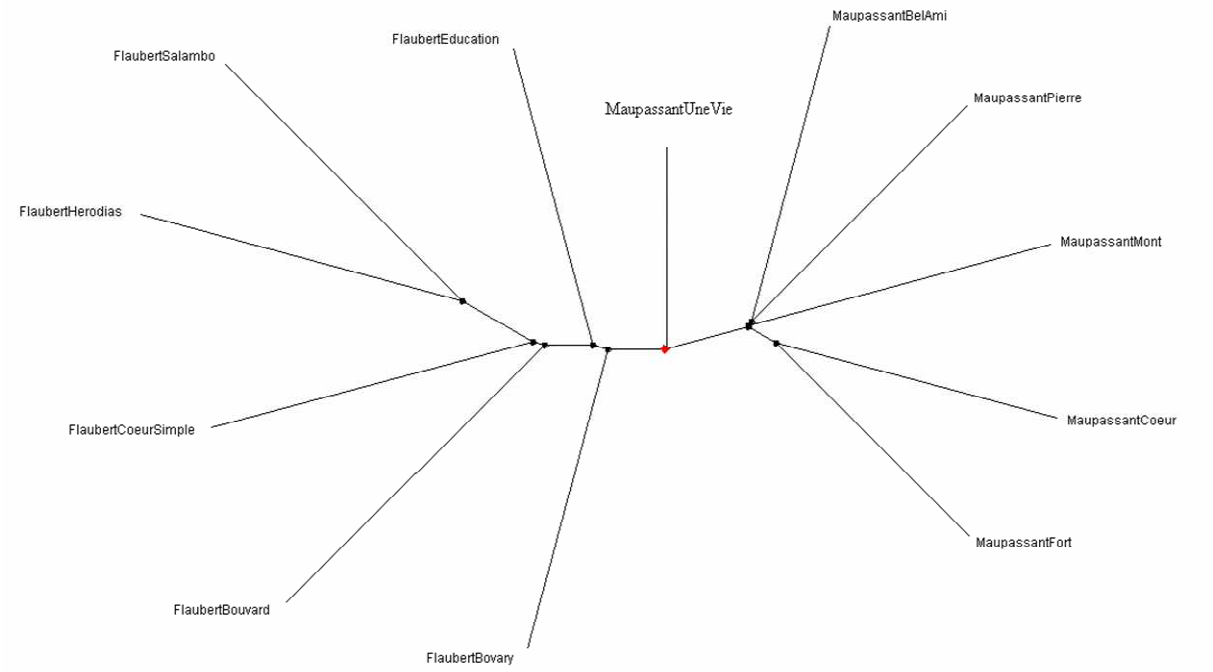
Tableau AVI.2 Distances internes à l'œuvre de G. de Maupassant

	Une vie	Bel Ami	Mont-Oriol	Pierre & Jean	Fort comme la mort	Notre coeur
Date	1883	1885	1887	1888	1889	1890
Longueurs (mots)	78 671	112 729	85 905	45 560	76 869	60 589
Une vie	-	0,269	0,266	0,255	0,264	0,280
Bel ami	0,269	-	0,254	**0,246**	0,252	0,258
Mont-Oriol	0,266	0,254	-	0,250	0,252	0,262
Pierre & Jean	0,255	0,246	0,250	-	**0,247**	0,254
Fort comme la mort	0,264	0,252	0,252	0,247	-	**0,225**
Notre coeur	0,280	0,258	0,262	0,254	0,225	-
Moyenne	0,267	0,255	0,257	0,251	0,248	0,256

Tableau AVI.3 Distances entre les œuvres de G. Flaubert et de G. de Maupassant

Flaubert / Maupassant	Bovary	Salambo	Education sentimentale	Cœur simple	Hérodias	Bouvard & Pécuchet
Une vie	**0,279**	0,292	0,295	0,294	0,288	0,298
Bel ami	0,354	0,386	0,374	0,387	0,378	0,388
Mont-Oriol	0,293	0,289	0,297	0,298	0,290	0,297
Pierre & Jean	0,317	0,353	0,351	0,350	0,345	0,353
Fort comme la mort	0,368	0,394	0,384	0,395	0,385	0,395
Notre coeur	0,340	0,348	0,338	0,353	0,345	0,348
Moyenne	0,325	0,344	0,340	0,346	0,339	0,346

Classification arborée sur le corpus Flaubert-Maupassant



Pour la méthode, voir : Labbé & Labbé, 2006.

Rappel sur les règles de lecture des arbres : la distance entre les textes est mesurée par le chemin à parcourir le long de l'arbre pour joindre les deux feuilles les symbolisant.

La qualité globale de cet arbre est de 98,7%, c'est-à-dire que la quasi-totalité de l'information contenue dans la matrice des distances est fidèlement restituée par ce graphe. Le chemin le plus "problématique" unit le roman de G. de Maupassant Une vie à la nouvelle de G. Flaubert Un cœur simple (indice de qualité : 95,4%). Ce sont aussi les deux textes qui ont l'indice de qualité le moins bon (98,1%).

Au total cette expérience distingue bien les deux auteurs et elle leur "attribue" correctement tous leurs romans mais elle montre aussi la grande diversité de l'œuvre de G. Flaubert et, surtout, combien a été grande l'influence de deux de ses romans (l'Education sentimentale et surtout Madame Bovary) sur l'œuvre romanesque de G. de Maupassant. La distance intertextuelle combinée avec la classification arborée donne une mesure exacte de cette influence. C'est pourquoi elle est un outil précieux pour les études littéraires.

Annexe 8
Romain Gary et Emile Ajar

(Lettre à Benoît Peeters et Michel lafon pour leur livre Nous est un autre)

Grenoble 5 juin 2004

Vous avez bien voulu m'interroger sur le cas Gary-Ajar qui semble intéresser encore le public comme en témoigne la parution récente d'une nouvelle biographie de cet auteur (Bona 1987).

Est-ce que la statistique permet d'identifier Gary dans l'ombre de Ajar, avec quels outils et avec quel degré de certitude ?

Comme indiqué dans mes précédents courriers, et comme je l'explique dans l'essai sur *Corneille dans l'ombre de Molière*, la principale difficulté réside dans l'impossibilité d'accéder aux fichiers électroniques de la quasi-totalité des romans français du XXe siècle. Certes, ces fichiers existent à l'Institut de la Langue Française, mais cet organisme refuse de communiquer les œuvres couvertes par des copyrights, même aux fins de recherche. Il fallait donc passer ces livres au scanner, ce qui est un travail très long du moins si l'on veut un traitement sans faute…

Pour cette raison, la note ci-dessous se limite à une comparaison des 4 romans signés Ajar avec quatre romans contemporains de Gary. Pour Ajar, les 4 romans ont été publiés au Mercure de France : *Gros-Câlin* (1974), *La vie devant soi* (1975), *Pseudo* (1976) et *L'angoisse du roi Salomon* (1979). Les 4 romans contemporains signés Gary, ont tous paru chez Gallimard : *Chien blanc* (1970), *Au-delà de cette limite, votre ticket n'est plus valable* (1974), *Charge d'âme* (1977), *Clair de femme* (1977). Au total ce corpus compte 396 944 mots (soit environ 200.000 mots pour chacune des "deux" auteurs) et 14 126 vocables différents.

Après passage au scanner, les textes ont été normalisés et lemmatisés selon des procédés éprouvés sur lesquels je ne reviens pas ici (Labbé, 1990) avant de subir les traitements statistiques standards ;

Parmi les traitements intéressants pour l'attribution d'auteur, cette note présente les résultats du calcul de la distance intertextuelle et l'observation de certaines combinaisons caractéristiques de mots (verbe + verbe).

I. La distance intertextuelle

L'ensemble du raisonnement est présenté dans les articles parus en décembre 2001 dans le *Journal of Quantitative Linguistics* et, en français, dans le numéro 2 de la revue *Corpus* (2003). Il s'agit de mesurer la proximité ou l'éloignement de deux ou plusieurs textes, les uns par rapport aux autres, en **considérant tous leurs mots**. On calcule ainsi un indice de la distance qui varie uniformément entre zéro (exactement les mêmes mots avec les mêmes fréquences) et un (aucun mot en commun).

Quatre facteurs influencent cette distance :

— le **genre**. On ne parle pas comme on écrit ; la fiction romanesque a ses codes qui ne sont pas ceux du théâtre, etc. Le carcan imposé par le genre est plus ou moins rigide.

Celui de la fiction romanesque à l'époque contemporaine étant très lâche, la distance aura tendance à être plus forte que dans d'autres genres plus codifiés ;

— le vocabulaire de **l'époque**. L'œuvre de Gary s'étend sur plus de 35 ans. Sur de telles étendues de temps, le style et le vocabulaire d'un auteur évoluent nécessairement. D'ailleurs, il est lui-même pris dans le flux qui change lentement le lexique de la langue. Il est donc indispensable de comparer des œuvres contemporaines, ce qui est fait ici.

— le **thème** traité. Chaque thème a un vocabulaire propre. Ce sont d'abord des noms de lieux, de personnes et une série de substantifs particuliers…

— enfin et surtout : **l'auteur**…

Pour rechercher la paternité d'un texte anonyme ou dont l'auteur est contesté, il faut donc le comparer à d'autres dont la signature n'est pas contestée, ayant été écrits à la même époque et traitant de thèmes voisins, dans un même genre (poésie, prose, roman, théâtre…). Ce point est important. En l'état actuel de la technique, on ne peut comparer que ce qui est comparable : un roman avec un autre roman, le théâtre avec le théâtre, une lettre avec de la correspondance…

Ce calcul appliqué à plusieurs milliers de textes en français de toutes origines (romans, pièces de théâtre, poésie, articles de presse, discours politiques, entretiens…) depuis le XVIIe siècle a permis de confirmer la validité du raisonnement et d'étalonner une échelle de la distance. Les valeurs ci-dessous s'appliquent aux textes dont les longueurs sont comprises entre 5 000 et 20 000 mots (pour les textes plus longs, on utilise des extraits) :

— une valeur inférieure ou égale à 0.20 ne se rencontre jamais chez des auteurs différents ;

— entre 0.20 et 0.25, il est pratiquement certain que l'auteur est le même. Sinon, les deux textes ont été écrits à la même époque, dans un même genre, sur un sujet et avec des arguments semblables. Ce cas se rencontre parfois dans les articles de presse, à propos d'un événement, parce que les journalistes travaillent à partir des mêmes sources et citent les mêmes noms de lieux et de personnes... Dans le cas d'œuvres littéraires appartenant à deux auteurs différents, il est très probable que le second s'est plus qu'"inspiré" du premier (dans l'ordre chronologique). En tous cas, ce genre de "collision" peut difficilement se produire plusieurs fois entre deux auteurs distincts.

— au-dessus 0.25, on entre dans une zone "grise" où deux hypothèses sont envisageables : un même auteur et des thèmes très différents ou deux auteurs contemporains traitant, dans un même genre, un thème proche avec leur style propre… De telle sorte que, plus on s'élève au-dessus de ce seuil, plus il sera difficile d'attribuer la paternité d'un texte anonyme à l'auteur considéré sans que, pour autant, cette paternité puisse être rejetée ;

— au-dessus de 0.40 les auteurs sont très probablement différents ou bien, pour un même auteur, les textes sont de genres très éloignés, par exemple : oral et écrit.

Le tableau I présente les résultats détaillés obtenus sur l'ensemble du corpus (les résultats significatifs sont placés en gras).

Tableau I. Les distances intertextuelles entre les huit romans de Gary-Ajar*

	Ame	ChienB.	Clair	Ticket	Gros-Câlin	Pseudo	Salomon	Vie devant
Ame		0,263	0,314	0,280	0,311	0,323	0,360	0,383
ChienBlanc	0,263		0,267	**0,227**	0,255	0,266	0,305	0,327
Clair	0,314	0,267		**0,229**	**0,244**	**0,238**	**0,226**	0,275
Ticket	0,280	0,227	0,229		**0,238**	**0,247**	0,263	0,310
Gros-Câlin	0,311	0,255	0,244	0,238		**0,224**	**0,247**	0,273
Pseudo	0,324	0,267	0,238	0,247	0,225		**0,246**	0,256
Salomon	0,360	0,305	0,226	0,263	0,247	0,246		**0,176**
Vie devant	0,384	0,328	0,276	0,311	0,274	0,257	0,177	

* Attention, cette matrice est symétrique selon la diagonale. On ne considère que la moitié nord-est.

Pour rendre les résultats plus intelligibles, les plus faibles distances sont classées par ordre croissant (tableau II). Une seule de ces distances se situe au-dessous du seuil qui désigne avec certitude un même auteur. Elle est apparemment sans intérêt puisqu'elle concerne deux œuvres écrites sous le nom de Ajar (*La vie devant soi* et *L'angoisse du roi Salomon*). En fait, cette distance très faible concerne le dernier Ajar paru et montre que Gary, du moins après *La vie devant soi*, avait "codifié" le style et le vocabulaire propre à "Ajar" et qu'il était capable de le reproduire assez fidèlement…

Tableau II. Les couples les plus proches
(en gras les distances entre des œuvres signées sous deux noms différents)

1	Ajar Salomon	Ajar Vie devant soi	0.176
2	Ajar Gros-Câlin	Ajar Pseudo	0.224
3	**Gary Clair de femme**	**Ajar Salomon**	**0.226**
4	Gary Chien Blanc	Gary Au-delà de cette limite	0.227
5	Gary Clair de femme	Gary Au-delà de cette limite	0.229
6	**Gary Au-delà de cette limite**	**Ajar Gros-Câlin**	**0.238**
7	**Gary Clair de femme**	**Ajar Pseudo**	**0.238**
8.	Ajar Pseudo	Ajar Salomon	0.246
9.	Ajar Pseudo	Ajar Gros-Câlin	0.247

A l'opposé, les 4 œuvres signées Gary sont au-dessus de ce seuil, ce qui montre une certaine diversité (ou un certain renouvellement). Toutefois, deux romans de Gary (*Clair de femme et Au-delà de cette limite.*…), très proches entre eux, sont également très proches des premier et troisième livres de Ajar (*Gros-Câlin* et *Pseudo*). Un tel classement croisé entre quatre romans publiés à quelques mois d'intervalle avec des distances aussi faibles ne laisse aucun doute. Ce n'est pas un "accident" ponctuel qui serait nécessairement limité à deux ouvrages. L'auteur est donc le même…

Il faut également considérer l'autre bout de la distribution. N'existe-t-il pas entre ces œuvres des distances telles qu'elles pourraient faire douter de l'existence d'un auteur unique ? Le tableau III apporte la réponse.

Tableau III. Les distances les plus grandes
(en gras les distances les plus grandes entre des œuvres signées du même nom)

21	Gary ChienBlanc	Ajar Salomon	0.305
22	Gary Ticket	Ajar Vie devant soi	0.310
23	Gary Charge d'âme	Ajar Gros-Câlin	0.311
24	**Gary Charge d'âme**	**Gary Clair de femme**	**0.314**
25	Gary Charge d'âme	Ajar Pseudo	0.323
26	Gary Chien blanc	Ajar Vie devant soi	0.327
27	Gary Charge d'âme	Ajar Salomon	0.360
28	Gary Charge d'âme	Ajar Vie devant soi	0.383

Sans atteindre le seuil de 0,40 au-dessus duquel on doit prendre en considération l'hypothèse de deux auteurs différents (pour des textes contemporains appartenant à un même genre), on note les distances très élevées séparant deux Ajar (*La vie devant soi* et *L'angoisse du roi Salomon*) d'un roman signé Gary dont l'écriture est pourtant contemporaine (*Charge d'âme*).

Généralisons le raisonnement. Voici ci-dessous les moyennes des distances séparant les 4 romans parus sous son nom (Gary-Gary), celles entre les 4 publiés sous le nom de Ajar (Ajar-Ajar) et enfin celles séparant ces deux ensembles (Gary-Ajar).

Gary-Gary : 0,264
Ajar-Ajar :0,237
Gary-Ajar : 0,285

Sans pouvoir parler, à propos de Gary, d'un "Caméléon" comme le fait l'une de ses récentes biographes, cette dernière mesure souligne malgré tout, une forte diversité dans l'écriture. En fait, dans les dix dernières années de sa vie, il y a bien deux Gary mais la coupure ne passe pas seulement entre Gary et Ajar mais au sein même de Gary.

La coupure au sein de l'œuvre "officielle" séparerait :

— un auteur de facture relativement "classique" dont les œuvres sont écrites à la troisième personne et qui comportent une intrigue, une progression dramatique, des rebondissements… Dans l'ensemble de l'œuvre, ce premier ensemble domine nettement. Dans le corpus, seul *Charge d'âme* se rattache vraiment à ce premier ensemble. Pour les seules années 1970, il faudrait aussi considérer la dernière œuvre de Gary qui appartient fort probablement à ce genre "classique" : *Les cerfs-volants* (Gallimard, 1980).

— une écriture relativement "nouvelle" où la narration est faite à la première personne, dans un ton plus personnel, avec une forme imitant l'oralité. L'auteur y cultive le saugrenu, l'humour noir, le non-sens, le coq-à-l'âne… *La danse de Gengis Cohn* (1966) inaugure probablement cette veine particulière. Dans le corpus, *Chien blanc* (1970) représente ce second versant avant "l'invention de Ajar". Au cours de sa période "ajarienne", Gary signe sous son nom deux romans qui appartiennent manifestement à ce second "genre" : *Au-delà de cette limite…* (1974) et surtout : *Clair de femme* (1977).

La faiblesse des distances séparant ces deux romans de Gary avec les œuvres signées "Ajar" ne laisse aucun doute sur le fait que l'auteur est le même. Surtout cette proximité souligne que Ajar n'est pas une pure création "ex-nihilo". Il existait potentiellement dans ce second versant de l'œuvre de Gary.

Pour conclure définitivement dans ce sens — dualité de l'œuvre et origine de "Ajar" dans l'une des deux parties de celle-ci —, il faudrait naturellement traiter l'ensemble de

l'œuvre de Gary — quel que soit le nom sous lequel les textes ont paru depuis la fin des années 1940.

Si le calcul montre la diversité de l'œuvre signée Gary, il souligne également l'homogénéité de l'œuvre "Ajar". Autrement dit, Gary avait "codifié" l'écriture particulière de celui-ci et il a su s'y tenir dans les quatre livres écrits sous ce nom au cours des cinq années qu'a duré la supercherie. Il est également probable qu'il a essayé de "singulariser" autant que possible cette écriture de sa manière propre d'écrire. Mais il n'y est pas vraiment parvenu. En annexe à cette note, le résultat de la classification arborée qui isole bien 2 des 4 œuvres de Gary (à gauche de l'arbre) et 2 des romans de Ajar (*La vie devant soi* et *L'angoisse du roi Salomon* à droite de l'arbre). En revanche, 3 autres se rejoignent ensemble au milieu du graphe (*Clair de femme, Gros-Câlin et Pseudo*)… Cela confirme que *le Ticket* et surtout *Clair de femme* — qui ont été écrits alors que Gary produisait aussi ses Ajar — ont été fortement "contaminés" par ce style si particulier et que l'auteur n'est pas parvenu à maintenir la "cloison" entre ses deux œuvres.

Il est amusant de constater que ce n'est pas Gary qui "contamine" Ajar mais l'inverse puisque *Clair de femme* est postérieur à trois des Ajar (*Gros-Câlin, La vie devant soi* et *Pseudo)* dont il est fort proche…

L'auteur a imité son double ! Je crois me souvenir que D. Bonna signale, dans sa biographie de Gary, que celui-ci aurait été conscient de cet étrange phénomène.

Cette "attribution d'auteur" est-elle confirmée par d'autres indices lexicaux ou stylistiques ?

II. Les syntagmes répétés

Le calcul qui vient d'être présenté considère l'ensemble du vocabulaire. D'autres indices prennent en compte certains aspects plus particuliers. Du point de vue des questions d'attribution d'auteur, l'analyse des "syntagmes répétés" les plus fréquents représente l'une des voies les plus prometteuses.

Ce sont des combinaisons stables de plusieurs vocables qui conservent leur indépendance contrairement au mot composé et à la locution où les composants perdent cette indépendance. Puisqu'on dispose des catégories grammaticales, grâce à la lemmatisation, on peut même s'intéresser à certaines combinaisons particulières comme les combinaisons "verbe +verbe".

Dans la langue française, les "pseudo-auxiliaires", dits encore "verbes modaux", suivis d'un infinitif permettent de réaliser une infinité de groupes verbaux complexes (sur le modèle de "pouvoir faire"). Chaque individu trahit, dans les combinaisons qu'il privilégie, un certain rapport à soi-même, aux autres et aux choses. Le sujet peut envisager ces rapports sous l'angle de l'action ("faire", "aller", "dire"), du possible ("pouvoir"), de la volonté ("vouloir") de l'obligation morale ou légale ("devoir") de l'impératif ("falloir"), de la connaissance ("savoir"). Dans chacune de ces catégories, il existe de multiples nuances (souhaiter, prétendre, entendre, vouloir, exiger…) Ces pseudo-auxiliaires sont suivis d'un verbe à l'infinitif : verbe d'action (faire, aller ou dire), de la pensée (penser, connaître), d'état (être) ou de la possession (avoir).

On demande donc à l'ordinateur de retrouver toutes ces combinaisons en tenant compte de ce que des adverbes ou des locutions adverbiales peuvent se glisser dans le couple recherché (par exemple : "faire (très bien) voir", etc).

Cette méthode a été élaborée avec A. Pibarot en 1997. Elle a été appliquée à plusieurs milliers de textes différents et s'est montrée apte à discriminer les thèmes (grâce aux groupes nominaux) et les auteurs (grâce aux groupes verbaux). Le tableau IV ci-dessous présente les résultats obtenus sur Gary et Ajar.

Tableau IV. Les principaux groupes verbaux complexes chez Ajar et chez Gary
(Classement en fonction de la fréquence exprimée pour 10.000 mots)

Ajar Syntagmes	**‰***	**Gary Syntagmes**	**‰***
vouloir dire	**7,20**	**vouloir dire**	**7,40**
pouvoir être	**4,90**	**pouvoir être**	**3,70**
pouvoir faire	**4,40**	**pouvoir faire**	**3,40**
devoir être	**3,10**	**devoir avoir**	**2,40**
aller faire	**2,90**	**aller faire**	**2,10**
pouvoir vivre	2,70	**devoir être**	**2,10**
laisser tomber	**2,30**	**laisser tomber**	**2,00**
devoir avoir	**2,10**	**pouvoir dire**	**2,00**
aller voir	**2,00**	mettre rire	1,80
devoir faire	**1,80**	entendre parler	1,60
pouvoir avoir	1,80	**devoir faire**	**1,30**
aller chercher	1,70	**faire parler**	**1,30**
falloir faire	1,60	laisser aller	1,30
falloir être	**1,60**	entendre dire	1,20
vouloir savoir	**1,60**	pouvoir comprendre	1,20
aller mourir	1,50	**aller voir**	**1,10**
faire chier	1,50	faire penser	1,10
faire parler	**1,50**	**falloir être**	**1,10**
falloir savoir	1,50	**vouloir savoir**	**1,10**
pouvoir dire	**1,40**	faire tuer	1,00

* pour 10.000 mots

En ne considérant que les 20 premières combinaisons, Gary et Ajar en partagent 13, notamment les trois premières dans le même ordre et pratiquement avec les mêmes fréquences relatives. Etant donné l'extraordinaire diversité des combinaisons possibles, il est impossible de rencontrer deux individus présentant les mêmes préférences pour plusieurs combinaisons spécifiques, à peu près dans le même ordre et avec des densités voisines. Certes, on pourrait objecter que, par exemple, "vouloir", "pouvoir", "dire", "être" et "faire" sont des verbes très fréquents et que leurs combinaisons sont assez banales — en fait, même à ce niveau-là, les préférences individuelles se distinguent très bien — mais alors que dire de "laisser tomber" par exemple ? Cette combinaison se trouve en septième position dans les deux œuvres avec quasiment le même nombre d'emplois. Il en est de même pour une autre combinaison très rare : "faire parler". Elle est totalement spécifique à Gary ! Ce qui caractérise cet auteur, sous son nom comme sous celui de Ajar, c'est l'insistance sur le "dire" et sur "parler", dimension qui se retrouve par exemple dans l'importance que les dialogues prennent dans l'œuvre de Gary dès l'origine, importance que l'on retrouve également chez Ajar...

Naturellement, certaines de ces combinaisons appartiennent au thème traité dans l'un ou l'autre des romans. Par exemple, le vingtième syntagme chez Gary ("faire tuer") appartient en quasi-totalité à *Charge d'âme*. D'autres, comme "faire chier" ont un poids très marginal chez Gary : ce n'est qu'à l'abri du nom d'un autre, qu'il peut donner libre cours à un vocabulaire "scatologique" dont il limite assez strictement l'emploi dans son œuvre "officielle"…

En conclusion

Les quatre romans signés Ajar sont en décalage assez net par rapport à l'œuvre "classique" de Gary, celle pour laquelle il était surtout connu lorsque Ajar paraît sur la scène littéraire. Cela peut expliquer que l'on n'ait pas songé à lui avant que son lien de parenté avec Pavlowitch n'ait été révélé. Si Gary n'avait pas donné *Chien blanc* et surtout *: Au-delà de cette limite…* et *Clair de femme*, le calcul statistique aurait difficilement pu identifier la "plume de l'ombre" cachée derrière le "pseudo" de Ajar. A condition aussi que Gary puis Pavlowitch n'aient pas dévoilé la supercherie, il aurait été très difficile d'écarter l'idée selon laquelle ce dernier était bien l'auteur des œuvres qu'on lui prêtait. Mais, après la parution de *Clair de femme* (1977), le doute n'est plus permis. Les distances intertextuelles désignent un auteur unique : Gary…

A la fin des années 1970, les ordinateurs disposaient déjà d'une puissance respectable et les outils intellectuels nécessaires pour "démasquer" Ajar, existaient potentiellement. Pourtant, personne, à notre connaissance, n'a songé à utiliser la statistique lexicale pour résoudre cette énigme, y compris parmi les chercheurs. L'explication essentielle réside dans la croyance, encore largement partagée aujourd'hui, selon laquelle "l'auteur est mort". En fait, non seulement l'auteur est bien présent dans les œuvres qu'il publie sous son nom mais de plus — même lorsqu'il s'agit d'un vieux routier comme Gary — il lui est très difficile de dissimuler longtemps les traces qu'il laisse malgré lui dans les textes qu'il écrit pour d'autres (réels ou imaginaires !)

Dominique Labbé

Références bibliographiques

Bona Dominique (1987), *Romain Gary,* Paris, Mercure de France.

Labbé Dominique, (1990), *Normes de saisie et de dépouillement des textes politiques*, Grenoble, Cahier du CERAT.

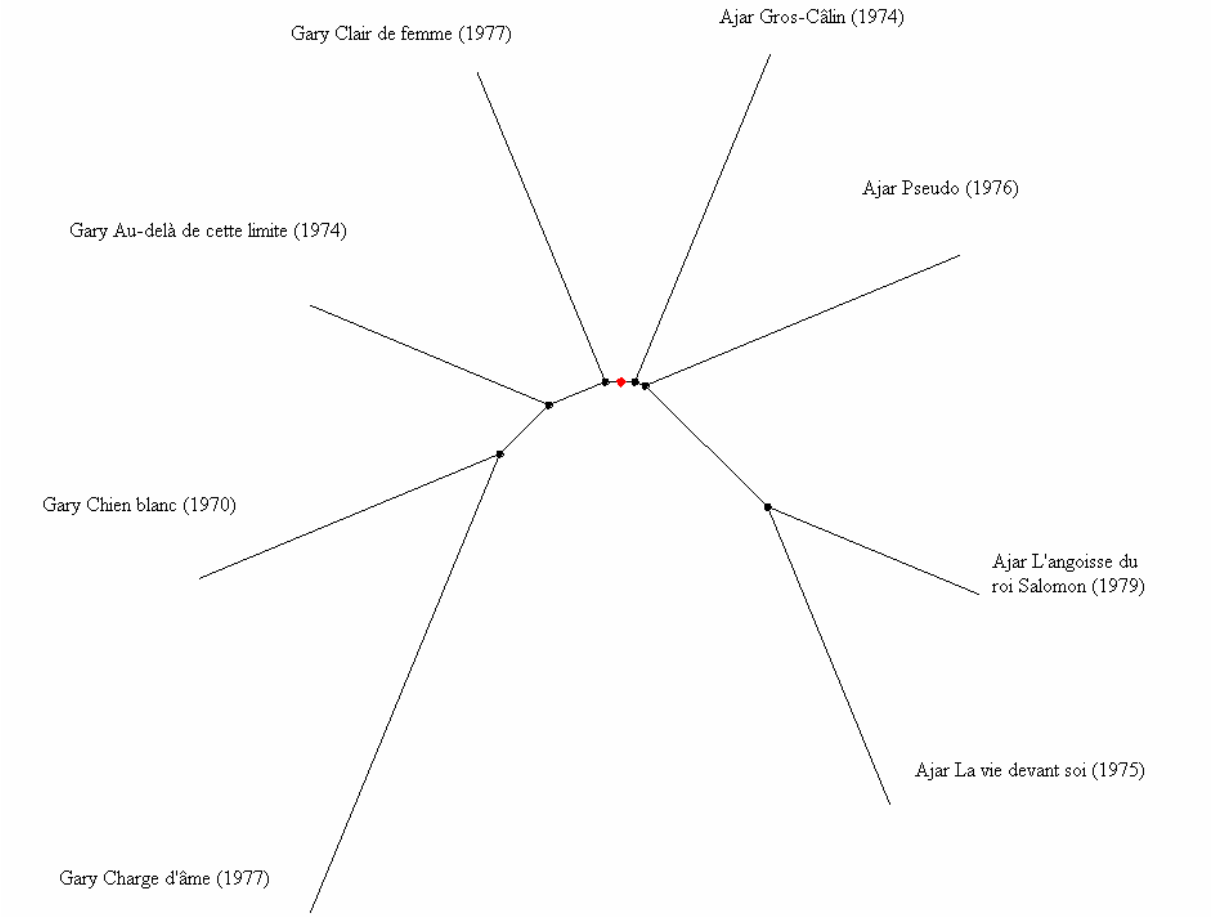
Labbé Cyril, Labbé Dominique (2001). "Inter-Textual Distance and Authorship Attribution Corneille and Molière". *Journal of Quantitative Linguistics*. 8-3. December 2001.

Labbé Cyril, Labbé Dominique (2003). "La distance intertextuelle". *Corpus*. 2-2003.

Labbé Dominique (2003). *Corneille dans l'ombre de Molière. Histoire d'une recherche*. Bruxelles, Les impressions nouvelles.

Pibarot André, Labbé Dominique (1998) "Les syntagmes répétés dans l'analyse des commentaires libres", in Sylvie Mellet (dir), *4e Journées d'analyse des données textuelles*, Nice, p 507-516.

Annexe Classification arborée des huit romans de Gary-Ajar



Rappel : pour lire ce graphe, il faut se souvenir que la position absolue dans l'espace des différentes feuilles n'a pas de signification et que seul compte le chemin à parcourir, en suivant les arêtes du graphe, pour unir les feuilles terminales entre elles.

Annexe 9
Un étrange procès

JQL 06 instruit un étrange procès en dissimulant un certain nombre de pièces et en lançant de nombreuses accusations erronées. Voici l'essentiel.

Remarques préalables

JQL 06 ne se contente pas de citer un seul article et un seul ouvrage - et d'en travestir le contenu - il dissimule au lecteur la plupart des informations dont il dispose.

En avril 2003, il a été remis à l'auteur de *JQL 06* le "corpus Corneille-Molière" (décrit en annexe 1) avec :

- le code source du programme de calcul de la distance intertextuelle ;

- une bibliographie, un rapport (Labbé 1990) et une série d'articles - qui présentent nos méthodes de travail, la manière dont les calculs ont été élaborés et les propriétés de la distance - comportant les textes suivants : Bergeron & Labbé (2000), Brugidou & Labbé (1999), Labbé & Labbé (2003), Labbé (1990), Labbé (2002a et 2002b), Labbé & Monière (2000). Ces textes présentent les propriétés de la distance intertextuelle - propriétés que *JQL 06* prétend avoir "découvertes" et qu'il caricature - ainsi que les réponses aux questions qu'il feint de poser. Certains sont cités dans notre article de 2001. Tous ces documents sont en ligne sur notre page personnelle.

L'auteur de *JQL 06* a participé avec D. Labbé à une table ronde à Louvain-la-Neuve, le 11 mars 2004. Au cours de cette table ronde, il a été de nouveau répondu à ses objections (Labbé 2004). Dans son article, il dissimule l'existence de cette rencontre alors qu'il en a lui-même rédigé un compte rendu fort éclairant[70]…

JQL 06 fait comme si ces informations n'existent pas. Par exemple, le § 5.1 affirme que nous n'avons pas précisé le sens de l'expression "normalisation" des graphies. Notre article de 2001 renvoie à ce sujet à : Labbé 1990 (dont le titre est Normes de saisie…) Ce document a été remis à l'auteur de *JQL 06*. De plus, comme rappelé ci-dessous, il a assisté à une conférence (et a reçu le texte de celle-ci) au cours de laquelle ces méthodes ont été présentées longuement avec le corpus Corneille-Molière-Racine.

Faute de rien pouvoir démontrer contre les méthodes et les résultats, on distille donc des insinuations et des soupçons contre les personnes. Comment le lecteur de bonne foi pourrait-il se douter que toutes ces insinuations sont fausses ?

Ce procès est d'autant plus ridicule que tout a été mis dans le domaine public - fichiers, méthodes, programmes - et que l'auteur de *JQL 06* a pu faire toutes les expériences qu'il souhaitait avec le résultat que l'on voit.

Le texte ci-dessous se contente de répondre aux attaques principales ce qui permettra au lecteur de comprendre que toutes les insinuations de *JQL 06* sont dérisoires et délibérément FAUSSES.

[70] http://laseldi.univ-fcomte.fr/Archives/affaireMorneilleColiere/morneille6.htm

Réponses à la section 1

1. A propos de la distance intertextuelle

Nous ne sommes pas les "inventeurs" du terme "distance intertextuelle"[71].

Depuis la première présentation (Monière et Labbé, 2000), la méthode a été utilisée dans une multitude de publications. Pour n'en citer que quelques-unes :

- Le Journal of Quantitative Linguistics a publié une expérience d'attribution d'auteur utilisant ces techniques (Merriam 2002). T. Merriam est également l'auteur de deux autres expériences utilisant la distance intertextuelle (Merriam 2003a et Merriam 2003b). Le même journal vient de publier un nouvel article qui présente une expérience en "double aveugle", entièrement couronnée de succès (Labbé 2007) .

- L'auteur de *JQL 06* connaît bien une collection ("Lettres numériques" aux éditions Champion) qui a publié récemment au moins trois ouvrages dans lesquels nos concepts - spécialement la distance intertextuelle - et nos algorithmes sont largement utilisés : Mayaffre, 2003 (la distance intertextuelle combinée à la classification arborée est utilisée à 7 reprises), Katsberg 2006 (notamment p. 73 sq.) et Labbé & Monière, 2003.

- La distance intertextuelle est au centre d'un numéro spécial de la revue Corpus (Luong 2003).

- En 2006, l'auteur de *JQL 06* était membre du comité d'organisation d'une rencontre internationale durant laquelle a été présentée une communication utilisant la distance intertextuelle et la classification arborée pour une expérience d'identification d'auteur en aveugle qui a été couronnée d'un total succès (Monière & Labbé, 2006).

Cette notion fait donc l'objet d'un large consensus terminologique et d'un vif intérêt qu'il est absurde de remettre en cause…

2. A propos du texte

Le § 1.2, reproche de considérer le "texte" comme un sorte de sac de mots et d'ignorer sa "structure". Ce reproche s'adresse à toute la statistique appliquée aux textes et spécialement au calcul présenté par *JQL 06*, dans sa dernière section (voir chapitres 7 et 8).

De plus, nous avons apporté des réponses à ces questions avec plusieurs instruments statistiques efficaces pour :

• localiser les ruptures thématiques et stylistiques dans un texte ou un corpus (par exemple : Labbé, Labbé & Hubert 2004) ;

• étudier les réseaux de collocations (combinaisons stables et répétées de deux ou plusieurs vocables)[72]. Ainsi, avons-nous démontré, à Louvain-la-Neuve le 11 mars 2004, en présence de l'auteur de *JQL 06*, que les collocations des mots les plus fréquents dans les pièces de Molière et Corneille indiquent clairement un auteur unique (Labbé 2004a).

[71] E. Brunet (1988) aurait utilisé ce terme le premier. Des attestations antérieures sont possibles.
[72] Voir par exemple : Labbé & Labbé 2005.

3. A propos du genre et de l'auteur

Dans le § 1.4, *JQL 06* réclame une définition des notions de "genre", "thème", "auteur", etc… mais n'apporte aucune précision à ce sujet tout en utilisant constamment ces notions par la suite (spécialement : § 8.2 et 8.6). On aurait aimé découvrir le "concept d'auteur" !

L'auteur et le genre sont portés sur la couverture des premières éditions (voir annexe 1)[73]. Naturellement, on peut mettre en doute ce qui est écrit sur la couverture, spécialement… le nom de l' "auteur" !

Réponses à la section 3

Les affirmations contenues dans les § 3.1 à 3.3 (et dans la note 3) sont fausses.

Les textes qui ont servi à expérimenter la distance intertextuelle ont été choisis avec soin et avec l'aide des spécialistes des domaines concernés car nous n'avons pas travaillé seuls (contrairement à un sous-entendu constant de *JQL 06* qui éclate dans la conclusion).

Les corpus ayant servi aux tests ont été soigneusement sélectionnés[74]. Dans ces corpus, il y a un grand nombre de poésies et de pièces de théâtre, à commencer par toute l'œuvre théâtrale de J. Racine, contemporain de P. Corneille et de J.B. Poquelin-Molière (voir en annexe 3 les distances les plus faibles entre Racine et Corneille-Molière).

L'auteur de *JQL 06* le sait très bien. Par exemple, voici un extrait du programme du congrès international sur l'Edition électronique en littérature et dictionnairique qui s'est tenu à Rouen du 17 au 21 juin 2002 :

> Séance du 18 juin 2002 :
> 15h00 « Corpus littéraires : pourquoi il faut les étiqueter et comment » (J.-M. Viprey).
> 16h00 « La lemmatisation des grandes bases de textes : conventions retenues, procédures et logiciels » (D. Labbé).
> 17h10 « Table ronde »

L'auteur de *JQL 06* venait de parler, et il était à la tribune, aux côtés de D. Labbé, quand celui-ci a présenté nos propres méthodes, en s'appuyant sur le corpus du théâtre classique (Corneille Molière et Racine : Labbé 2002b). Il a largement participé à la discussion, comme le prouvent les archives électroniques de cette séance.

Plusieurs de ces tests ont été publiés. Par exemple : Monière & Labbé 2000, Bergeron & Labbé 2000, Labbé 2002b, Labbé & Labbé 2003[75]. On verra comment a été étalonnée l'échelle des distances, en suivant scrupuleusement les méthodes usuelles en la matière.

Ces articles ont été remis à l'auteur de *JQL 06* et sont consultables en ligne.

---

[73] Voir aussi le fac-similé de la couverture de l'édition originale du Dépit amoureux (annexe 10, p. 161 de ce dossier) qui porte "comédie" juste en dessous du titre…

[74] Sur ces corpus, voir notamment Labbé 2003 (p. 64-66). Les indications citées par *JQL 06* sont situées à plusieurs années de distance. L'expérience sur Corneille et Molière a été réalisée en 1999. A l'époque, les corpus destinés aux tests comportaient 11 276 000 mots. Début 2003, lorsque D. Labbé a rédigé son essai, ces corpus avaient doublé. Il est exact que, dans ces corpus destinés à la mise au point des logiciels, la taille moyenne des textes est inférieure à 10.000 mots et que, pour les longs romans, des extraits ont été utilisés et non les œuvres complètes.

[75] Voir également : Labbé 2007 (pour l'anglais).

L'auteur de *JQL 06* connaît bien nos méthodes et leur sérieux.

Il sait qu'il y a d'autres pièces de théâtre dans nos corpus.

Dans les § 3.1 à 3.3 et dans la note 3, il a délibérément imprimé de fausses informations pour jeter des soupçons sur notre sérieux et notre intégrité.

La manière dont est reproduite l'échelle de la distance est de la pure casuistique. On sort les citations de leur contexte et on leur fait dire le contraire de ce qu'elles disent. Notre article de 2001 ne présente pas une échelle valable quelle que soit la dimension des textes. Il affirme exactement le contraire[76]. Sinon pourquoi est-il souligné à plusieurs reprises que le calcul ne peut porter que sur des textes de longueurs pas trop différentes ?

Dans notre article de 2001, juste au-dessus de l'échelle, il est écrit : "In this application, the shortest text has 3 500 tokens - it is Molière's first comedy (see the Appendix) and the largest is 20 300 (Corneille's Toison d'or)"[77].

Tout lecteur de bonne foi comprend que l'échelle concerne les textes dont les longueurs sont comprises dans cet intervalle et qu'elle a été étalonnée sur des textes de ces tailles[78].

Cette plage de validité a été précisée publiquement à plusieurs reprises, y compris à Louvain-la-Neuve en présence de l'auteur de *JQL 06*. Au cours de son exposé, le modérateur, L. Lebart, l'a personnellement interrompu pour le lui rappeler. Voici cet échange d'après le compte rendu que l'auteur de *JQL 06* a mis en ligne après cette table ronde :

> *"Mes observations, faites à partir de l'application de l'indice de D.Labbé, me conduisent à supposer un biais dans cet indice : il croît tendanciellement à mesure que les textes analysés s'allongent. N'étant pas moi-même statisticien mais utilisateur expert, je soumets cette question aux statisticiens tout en maintenant pour l'instant ma position. (Ludovic Lebart souligne à ce moment que, dans une publication plus récente, D. Labbé a précisé que l'échelle n'est valide que pour des textes de 10 000 mots)"*[79].

Pourquoi, l'auteur de *JQL 06* feint-il d'ignorer cette précision, alors qu'il confirme lui-même en avoir été informé devant plus d'une centaine de chercheurs ? C'est que, dans la suite de son article, il utilise des textes - dont les longueurs se situent très en dehors de cette plage de validité - pour prétendre avoir mis en défaut cette échelle.

Les § 3.7 et 3.8 (p. 18 de ce dossier) nous font dire exactement le contraire de ce nous avons écrit. Les graduations sur l'échelle de la distance sont comparables à celles

---

[76] Voir également Labbé 2003, spécialement p. 54 : "il serait vain de chercher un outil universel".

[77] Labbé 2001, p 218. Au passage : nous avons commis une erreur : le plus long texte du corpus est L'Avare (Molière) : 21 033 tokens.

[78] Précision rappelée notamment lors de nos conférences d'Orsay et de Dublin (Labbé 2004b et 2004c).

[79] Nous soulignons (http://laseldi.univ-fcomte.fr/Archives/affaireMorneilleColiere/morneille6.htm). En fait, L. Lebart a rappelé la dépendance entre la distance intertextuelle et la longueur des textes ainsi que la plage de validité de l'échelle. Il a ajouté que, pour plus de clarté, le raisonnement pourrait se faire en "pour 10 000 mots" (X. Luong a également émis cette suggestion).

d'une éprouvette. De plus, il a été précisé que : "L'indice de la distance varie uniformément - entre 0 (même vocabulaire et fréquence identique de chaque vocable dans les deux textes) et 1 (aucun vocable en commun) – sans saut, ni effet de seuil entre certaines valeurs"[80]. Un simple regard à la figure 3 de *JQL 06* (p. 54) – permet de constater que cela est exact.

Réponses à la section 8

DCL begin with a selection of eight plays, which they present as Molière's **best-known** *plays. This cannot fail to astonish a scientific mind, since no notoriety criterion is made explicit. (Cyril et Dominique Labbé commencent avec une sélection de 8 pièces, présentées comme les pièces* **les plus connues** *de Molière. Cela ne peut manquer d'étonner un esprit scientifique puisqu'aucun critère de notoriété n'est indiqué"* (§ 8.2, p. 77 de ce dossier).

Voici le titre du tableau original de 2001 : "Distances Between Molière's **Well-Known** Works" ("des oeuvres **bien connues** de Molière").

JQL 06 affirme également que nous avons "commis l'erreur" (*DCL fail to mention that*…) de ne pas indiquer le genre de ces 8 pièces "les plus connues" (*ibid*). Voici le long passage de notre article *JQL 01* placé immédiatement sous le tableau original.

"The calculation shows an important similarity between all these plays, although their topics are very different. The smallest (.167) is between *Tartuffe* and *le Misanthrope*, two plays in Alexandrines in which Molière does not use farce nor colloquial language, nor jargon. The greatest (.239) is between *le Misanthrope* and *le Bourgeois gentilhomme* or *le Malade imaginaire*. The first one is in verse, the two others in prose and they contain a lot of inventions in «turkish» or in «latin». More generally, distances greater than .20 separate *l'Ecole des femmes*, *Tartuffe*, *le Misanthrope* and *les Femmes savantes* — written in verse — and *Dom Juan*, *l'Avare*, *le Bourgeois gentilhomme* and *le Malade imaginaire*, written in prose. Considering these differences, it is obvious that all these masterpieces are from the same author. Some cases seem particularly clear : *Tartuffe* and *Dom Juan* —two plays which caused scandal and were withdrawn — are written with, the first in verse and the second in prose. In spite of a lot of «patois» in the second one, which increases their distance, they remain very close (.199): this confirms that they have only one author and that they were written during the same period (the same comment can be said for *l'Avare* and *Tartuffe*)[81]"

A propos du § 8.7 et de la figure 5 (p. 79 de ce dossier), l'auteur de *JQL 06* trouve normal de trafiquer les graphiques des autres et ne se gêne pas pour modifier les informations qui le gênent. En effet, voici la légende de sa Fig. 5 (p. 79 de ce dossier) :

"DCL's tree-analysis"

[80] The Intertextual Distance index "varies in the same way – between 0 (the same vocabulary and similar frequency of each type in the 2 texts) and 1 (no common type) – without jump, nor threshold effect around some values" (Labbé & Labbé 2001, p. 214).
[81] Labbé & Labbé 2001, p. 221.

et voici la légende qui figurait sous la figure originale :

"We thank M. Xuan Luong (Nice University) for this graph."

Puisque le graphe original a été réalisé par X. Luong (et non par nous !), avant de publier la figure 5, il fallait vérifier que la modification est acceptable et que les commentaires sont corrects. En effet, d'après le code source du programme de X. Luong (voir aussi Luong 1988), ce programme trace les arbres en deux temps. Premièrement, sur le tronc central, il rassemble les nœuds les plus proches pour former un petit nombre de groupes ("clusters"). Puis il met en place les feuilles autour de ces nouveaux nœuds selon le sens inverse des aiguilles d'une montre, en commençant par les textes ou groupes de textes les plus proches de ce nœud central et en terminant par les plus éloignés.

Dans le graphique de gauche, la "feuille" des deux Menteurs (de Corneille) est attachée en dernier au groupe des pièces en vers de Molière. C'est donc la plus décalée (ce qui justifiait l'adverbe "quite"). Dans le graphique de droite - celui que l'auteur de *JQL 06* se vante d'avoir trafiqué – cette feuille est attachée en second (et non plus en dernier comme dans l'analyse originale). Bref, grâce à sa rectification bienvenue :

The Corneille's Menteurs stand no longer quite in the centre of Molière's works (in verses), but exactly in this centre…

Quite a conjuring trick, is'nt it ?

Le § 8.9 (p. 80 de ce dossier) prétend que nous n'avons pas "confronté" G. Flaubert et G. de Maupassant, A. Dumas, H. de Balzac et beaucoup d'autres. Le lecteur peut se reporter à : Labbé 2003 (p. 79-81). C'est dans ces pages que l'auteur de *JQL 06* a trouvé l'exemple de G. Flaubert et de G. de Maupassant. Comme il cite lui-même cet essai, il peut difficilement prétendre ne pas avoir remarqué ce long passage.

Tous ces auteurs ont évidemment été traités avec beaucoup d'autres. Ils ont toujours été clairement identifiés, sauf quand ils n'en faisaient qu'un – comme dans les cas de Gary-Ajar et de… Corneille-Molière.

Le § 8.14 (p. 85 de dossier) prétend que l'essai (Labbé 2003) "évoque un complot du silence organisé par les Moliéristes et/ou les Corneillistes (sic)". Evidemment, aucune page n'est citée : nous n'avons jamais rien écrit de tel !

En conclusion, une question simple se pose : depuis quand mène-t-on ainsi le procès des gens, avec un dossier grossièrement falsifié, sans leur donner la possibilité de se défendre ?

Les notes 8 et 9 (p. 74-75 de ce dossier) - à propos de l'adverbe "quite" que nous avons employé trois fois ! - éclairent cet article sans équivalent dans la science moderne. L'expression "ter relaps" appartient au vocabulaire de l'inquisition. Comme on pourra le constater en consultant tout bon dictionnaire : il signifie que nous sommes retombés trois fois dans l'hérésie…

Annexe 10
Les avertissements de trois premiers éditeurs

Psyché (1671)

Le libraire au lecteur

Cet ouvrage n'est pas tout d'une main. M. Quinault a fait les paroles qui s'y chantent en musique, à la réserve de la plainte italienne. M. de Molière a dressé le plan de la pièce, et réglé la disposition, où il s'est plus attaché aux beautés et à la pompe du spectacle qu'à l'exacte régularité. Quant à la versification, il n'a pas eu le loisir de la faire entière. Le carnaval approchait, et les ordres pressants du Roi, qui se voulait donner ce magnifique divertissement plusieurs fois avant le carême, l'ont mis dans la nécessité de souffrir un peu de secours. Ainsi, il n'y a que le prologue, le premier acte, la première scène du second et la première du troisième dont les vers soient de lui. M. Corneille a employé une quinzaine au reste ; et, par ce moyen, Sa Majesté s'est trouvée servie dans le temps qu'elle avait ordonné.

Le Dom Juan versifié par Thomas Corneille (1683)

Le libraire au lecteur

Cette pièce, dont les comédiens donnent tous les ans plusieurs représentations, est la même que feu M. de Molière fit jouer en prose peu avant sa mort. Celui qui l'a mise en vers a pris soin d'adoucir certaines expressions qui avaient blessé les scrupuleux, et il a suivi la prose dans tout le reste, à l'exception des scènes du troisième et cinquième actes, où il fait parler des femmes. Ce sont scènes ajoutées à cet excellent original, et dont les défauts ne doivent point être imputés au célèbre auteur sous le nom duquel cette comédie est toujours représentée.

(In : Thomas Corneille. Le festin de Pierre (Edition critique par Alain Niderst. Paris : Champion, 2000, p 34).

Dépit amoureux (1662)

(Dédicace du libraire – fac simile page suivante)

Il y a longtemps que j'avais résolu de vous présenter quelque chose qui vous marqua mes respects ; mais ne trouvant rien qui fut digne de vous être offert, et qui fut proportionné à vos mérites, j'avais toujours différé le juste et respectueux hommage que je m'étais proposé de vous rendre ; et j'eusse encore tardé longtemps à le faire, si le Dépit Amoureux de l'auteur le plus approuvé de ce siècle[82] ne me fut tombé entre les mains.

[82] D'après G. Forestier, en 1662, Pierre Corneille dominait la scène théâtrale "de la tête et des épaules" depuis un quart de siècle. Nous ajoutons que, de son vivant, Pierre Corneille était couramment désigné comme "l'auteur le plus approuvé du siècle". Cela n'a jamais été le cas pour Molière.

DE'PIT AMOVREVX

COMEDIE,

REPRESENTE'E SVR LE Theatre du Palais Royal.

DE I. B. P. MOLIERE.

A PARIS,
Chez CLAVDE BARBIN, au Palais, sur le Degré deuant la Sainte Chapelle, au Signe de la Croix.

M. DC. LXIII.

AVEC PRIVILEGE DV ROY.

A MONSIEVR

MONSIEVR

HOVRLIER,

ESCVYER SIEVR DE Mericourt, Conſeiller du Roy, Lieutenant General Ciuil & Criminel au Baillage du Palais à Paris.

MONSIEVR,

Si cette Piece, n'avoit receu les applaudiſſemens de toute la France,

Annexe 11

From the very beginning…

Le § 8.14 de *JQL 06* (p. 85 de ce dossier) conteste cette phrase de l'article de 2001 : "From the very beginning, it was rumoured that Molière was not the writer of his plays" (Depuis le tout début, le bruit a couru que Molière n'était pas l'auteur de ses pièces).

Cette phrase n'a pas été écrite à la légère.

Outre les trois avertissements des éditeurs, cités dans l'annexe précédente, voici quelques exemples, choisis au tout début des succès de Molière.

• En 1660, dans la Préface à sa comédie Les véritables précieuses, Baudeau de Somaise écrit que Molière est "l'auteur prétendu des Précieuses ridicules" et il conclut :

"Qu'attendre d'un homme qui tire toute sa gloire des Mémoires de Guillot-Gorju qu'il a achetées à sa veuve et dont il s'adopte tous les ouvrages ?" (Mongrédien 1986, p. 36).

• En 1663, dans sa Lettre sur les Affaires du Théâtre, Donneau de Visé[83] écrit que, pour l'Ecole des femmes, Molière s'est servi de "mémoires" que lui ont remis des "personnes de qualité" et que ce n'est pas Molière qui est visé par ses critiques contre cette pièce (Mongrédien, 1971, I, p. LIV).

• En 1663, dans sa comédie Zélinde, le même Donneau de Visé s'adresse ainsi à Molière :

"Je ne crois pas que cette pièce [Critique de l'Ecole des femmes], qui n'est en beaucoup d'endroits qu'une imitation de celles que vous avez déjà fait voir[84], eût pu réussir sous le nom d'un autre". "Cela n'empêche pas que vous n'ayez de grandes obligations au chevalier Doriste [l'Abbé du Buisson] dont vous avez si bien tourné les vers en prose" (Mongrédien 1971, p. 40 et 43).

Plus loin dans la même comédie, il fait dire à un auteur dramatique (Aristide) : "Je crois bien que s'il [Molière] était obligé de faire une entière restitution, qu'il resterait, non seulement nu, mais que ses ouvrages seraient dépouillés de ce qu'ils ont de plus beau" (Mongrédien 1971, p. 72).

• La même année, dans sa Réponse à l'Impromptu de Versailles, Donneau de Visé met les propos suivants dans la bouche de P. Corneille (sous le nom d'Ariste[85]) :

---

[83] Fondateur avec Thomas Corneille, du Mercure galant, la première gazette littéraire, Donneau de Visé est, selon G. Mongrédien, parfaitement "informé des dessous de la vie littéraire". Les extraits suivants sont tirés de sa pièce Zélinde qui est une attaque contre l'Ecole des femmes.

[84] "fait voir" (représentées) et non pas "écrites".

[85] Dans cette pièce à clefs, le personnage d'Ariste est P. Corneille. Dans son Panégyrique de l'Ecole des femmes, C. Robinet l'identifie sans doute possible : "Le grand Ariste, dont je ne prends que Le Menteur pour l'opposer à tout le misérable comique de Zoïle" (Mongresdien 1970, I, p. 204-205).

"Les enfants [de Molière] ont plus d'un père […] Personne n'ignore qu'il sût bien retourner des vers en prose en faisant la Critique [de l'Ecole des femmes], et que plusieurs de ses amis ont fait des scènes aux Fâcheux ; c'est pourquoi, si Monsieur Boursault lui répond, il pourra lui dire plus justement que tout le Parnasse s'assemble, lorsqu'il veut faire quelque chose." (Ibid, p. 271).

• Egalement en 1663, Charles Robinet, un autre satiriste bien informé des dessous de la vie littéraire, dans Le panégyrique de l'Ecole des femmes. Conversation comique sur les œuvres de M. de Molière :

"On ne peut pas dire que Zoïle [Molière] soit une source vive, mais seulement un bassin qui reçoit ses eaux d'ailleurs, pour ne point le traiter plus mal, en le comprenant dans la comparaison que quelques-uns ont faite des compileurs de passages à des ânes, seulement capables de porter de grands fardeaux" (Mongrédien 1971, I, p 210).

Voici quelques exemples, pour le "very beginning".

Mais le meilleur témoin est Molière lui-même, s'adressant à son ami Chapelle : "Vous qui faites si fort l'habile homme, et qui passez, à cause de votre bel esprit, pour avoir beaucoup de part dans mes pièces" (Grimarest 1675, p. 103).

Molière confirme ainsi que l'on jasait en ville et qu'on ne le jugeait pas capable d'avoir écrit seul ses pièces. Qu'avons-nous écrit d'autre ?

Enfin, cette idée des "rumeurs" figure à la fin de l'avant-propos de l'ouvrage d'un "Moliériste" convaincu (Forestier 1990).

Comme on le voit, la petite phrase - "From the very beginning, it was rumoured that Molière was not the writer of his plays" – n'a pas été écrite à la légère.

Les écrits des contemporains de Molière ont été dépouillés avec soin. Rien dans ces écrits ne va contre l'attribution à Corneille de toutes les pièces en vers représentées sous le nom de Molière ainsi que du Dom Juan et de l'Avare. Bien au contraire, les allusions comme celles que l'on vient de lire sont légion et, avec les commentaires nécessaires, elles pourraient remplir un volume comparable à ce dossier.

Annexe 12

Lettre au rédacteur en chef (Editor) du Journal of Quantitative Linguistics

Mr. Koehler
Journal of Quantitative Linguistics
Department of Computational Linguistics
University of Trier
TRIER
Germany

Grenoble, 20th December 2006,

Dear Sir,

As you know, a few days ago, we discovered in the last issue of the JQL a review article ("About Labbé's Inter-Textual Distance") on our December 2001 paper ("Inter-Textual Distance and Authorship Attribution") authored by Mr. Viprey.

It is unusual for a scientific journal to publish such an attack on another paper which has previously appeared in its own columns. When this event does occur, normal procedure and common courtesy requires the author of the attack to provide a copy to those who are the object of his criticism. This procedure prevents avoidable errors and misunderstandings before publication. In this particular case, it seems it was all the more necessary as our critic used files, data and software that we provided him without restrictions or conditions (which he has failed to acknowledge as basic courtesy and intellectual honesty demand).

Your readers may well assume that, in accordance with normal practice, we were alerted in advance of this attack on our work, and that our ensuing silence is a tacit admission of our inability to reply.

The readers of a scientific journal also assume that, in a case of such importance, the formulae, main reasoning, quotations, tables and references have been checked for accuracy (a first reading shows that it is not the case as you can see below).

If common practice had been respected, and we had been informed and allowed to answer within a reasonable time - because "About Labbé's Intertextual Distance" is

sometimes very obscure and because the matter deserves extended study -, little of the objections to our work would have remained.

Some examples spring to mind in reading this article for the first time (see below).

1. Discussion of the Inter-Textual Distance (p. 265-276).

Below are a few examples of the important questions that are raised:

- why was there no mention at the opening of his text of the fact that the idea, and indeed the term "Inter-Textual Distance" itself, can be attributed to E. Brunet (we quoted his seminal article in our 2001 paper)?

- in May 2003, we provided our critic with a copy of the proofs of an article (Labbé & Labbé 2003) where the properties of Inter-Textual Distance are carefully set out and where the effects of differences of sizes on its values are made clear. Why did he not quote this paper?

- at the beginning of 2002, an important blind test was carried out with the assistance of Mr. Brunet. In his conclusions, our critic quotes Brunet's record of this blind test (Brunet 2004 in his references). Consequently, he clearly was aware of the existence of this test: why did he not quote our own record (Labbé 2002)?

- he asks for information about the corpora used for the calibration of the Inter-Textual Distance (p. 268): why did he not pose this technical question directly to us instead of doing so in the form of an aside to the readership?

- why did he not check with us the results of his experiments on Balzac, Flaubert, and Maupassant? Why did he not place his files and records in the public domain, as we have done with our own data on the 17th century theatre? This would have allowed interested researchers to check these results independently.

- as to the matter of our "lemmatisation" (p. 273), he has in his possession the book in which are detailled all the standards we have used (Labbé 1990): why did he not refer to it? Can he quote the French dictionary where he found specific entries for: "afin de", "afin que", "bien que", "sans cesse"… and where "cesse" is not a feminine noun? Are all French lexicographers "mistaken" and "naïve"?

2. The discussion of our "application to Corneille and Molière" (p. 278-279)

Again, there seem to be several inconsistences. Even limiting ourselves to a few important questions we see:

- he maintained that we "fail to mention that [four Molière plays] are written in verse and four are not" (p. 277). Did he not read our JQL December 2001 article following page 220?

- why did he write that there are "two versions" of Psyché (p. 278 and 279)? Up now, only one version of this play is known to exist (with three identified authors whose contributions are precisely delimited). Why did he not give the references of this second version? And why did he not mention the names of the "several hands" that wrote these two versions?

- are the two graphs in Fig. 5 (p. 279) actually “equivalent”? Was this question actually adressed to X. Luong, the eminent mathematician who did the tree-analysis and drew the original figure published on page 227 in our JQL December 2001 article? Why did he remove the caption under this diagram?

- can he quote sources precisely (books, pages and sentences) of the 17th century specialists who are supposed to have written, before December 2001, that "Molière's Dom Garcie is related to Corneille's heroic comedies" (p. 278)? Can he also provide precise references for the specialists who are supposed to have written that "Corneille's two Menteurs are more related to Molière's comedies (especially to his versified ones) than to Corneille's early comedies" (idem)? He asserts that this was already known by "relevant specialists counted by (*sic*, “in their”) hundreds, all around the world" (p. 281). At least some of them could have been quoted…

3. Correspondence Analysis (p. 279-281)

Again, some obvious questions arise:

- why did he not present the reasons for withdrawing the three parts of Psyché from the experiment?

- without Psyché, there are actually 33 plays by Corneille and 32 by Molière = 65 plays. On Fig. 6 (p. 280), there are 67 plots (n° 67 is the lowest one on the left) and the Corneille "cluster" contains 34 plots. Are there two unknown plays by Corneille and/or Molière?

- why did he not provide the list of the plays corresponding to the plots? What are the lengths of these plays and when were they written?

- is he sure that a Correspondence Analysis applied to "all the lemmas" (see the title of Fig. 6) is "without any bias" (as written p. 280)?

- what are the inferences to be drawn from the specific horseshoe shape of this Fig. 6?

- are the lines drawn between the four supposed clusters the result of a classification? Can he provide the results of this classification?

4. About the conclusions

Is he sure that the problem of the authorship of the Molière plays "was raised three times in total" and not more frequently?

Can he quote the pages from our papers or from our book in which it is allegedly written that:

- the Molieristes and/or the Corneillists (sic) have organised a "silent plot"?

-we have provided a "unique measurement, of explicit appearance, in order to automatically determine uncertain attributions"?

-we settle our attribution by using a single measurement?

Finally, it is written that "ALL the authors referred to by [Cyril and Dominique Labbé's] paper in the field of lexical statistics have expressed themselves against" our conclusions. In our paper we referred to many individuals. Can he demonstrate that all these people have indeed expressed themselves about our work?

These are some of the more obvious questions, among many others, that should have been addressed to the author of "About Labbé's Inter-Textual Distance" before publication. Of course, you can guess the answers and appreciate more fully that he would have been well-advised to follow common, established practice.

All these unfounded comments have attacked the honour and reputation of two honest researchers who regularly read your journal, contribute to it, and quote it with confidence.

May we propose some solutions to resolve this situation?

We would appreciate a short foreword in the next JQL available issue (and, as soon as possible, on the JQL web site) telling your readers that, contrary to common practices, Mr. Viprey did not inform Cyril and Dominique Labbé of the publication of his article - "About Labbe's Inter-Textual Distance" in the JQL June-December 2006 issue – neither he did offer them the opportunity to answer. Mr. Viprey also failed to acknowledge that he used files, data and software provided by C. & D. Labbé without restrictions or conditions.

We intend to write a complete answer as soon as possible. After its acceptance by the JQL reviewers and the Editor - and after communication to Mr. Viprey - should not this answer be placed on the web site of the review?

In the event of you acceding to this request, we would welcome the possibility of publishing a non-commercial French version of this response on our website.

With our best regards,

Dominique Labbé
CERAT-IEP
BP 48 - F 38040 Grenoble Cedex 9

References quoted in this letter, and missing in "About Labbé's Inter-Textual Distance", are available online.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques*. Cahier du CERAT, n° 7, Grenoble: CERAT-IEP, avril 1990

- http://web.upmf-grenoble.fr/cerat/Recherche/PagesPerso/LabbeNormes.pdf

Labbé D. (2002). *Qui a écrit quoi? Compte-rendu d'une expérience en double aveugle*.

- http://web.upmf-grenoble.fr/cerat/Recherche/PagesPerso/LabbeExperience.pdf

Labbé C. & Labbé D. (2003). "La distance intertextuelle". *Corpus*. 2, p. 95-118.

- http://revel.unice.fr/corpus/document.html?id=31

Annexe 13

STATEMENT
by Cyril and Dominique LABBE
(2007, January, 12th)

In its December 2006 issue (Vol. 13, Number 2-3, pp. 265-283), the Journal of Quantitative Linguistics published a paper by Mr. Jean-Marie Viprey, entitled :

"About Labbé's “Intertextual Distance”"

This paper discusses our article published in the same Journal in December 2001 (Vol. 8, n°3, pp. 213-231) :

"Inter-Textual Distance and Authorship Attribution. Corneille and Molière".

We are happy to see that our methods and findings stimulate interest, discussion and controversies that are a prerequisite for the advancement of science. However, we have very serious reservations concerning Mr. Viprey's article:

1. A first reading shows that many things, presented as novel in this article, were already known and have been caricatured. Many of the attacks are unjustifiable and, in some case, entirely false.

2. Contrary to common practice, Mr. Viprey did not provide us with a copy of his article before its publication. In this particular case, it was all the more necessary as our critic used files, data and software that we provided him with (which he has equally failed to acknowledge). Consequently, we did not have the possibility of responding to this attack.

3. The polemic tone of Mr. Viprey’s prose does not follow the conventions of scientific debate nor the constraints of common courtesy.

We will be providing an extended and in-depth study of this article and we shall respond to it appropriately. This response will respect the accepted procedures of scientific publication which entail several months.

Until such time as our response is available, the reader is advised to reserve judgement on Mr. Viprey's article. To obtain a reliable and detailed presentation of our works:

• On the Inter-Textual Distance:

- Labbé Cyril & Labbé Dominique (2003). "La distance intertextuelle". Corpus, 2, p 95-118.
http://revel.unice.fr/corpus/document.html?id=31

- Labbé Cyril & Labbé Dominique. "A Tool for Literary Studies: Intertextual Distance and Tree Classification". *Literary and Linguistic Computing*. 21-3, 2006, p 311-326.

• On trials and calibration:

Labbé Dominique (2002). Qui a écrit quoi ? Grenoble: CERAT.
http://web.upmf-grenoble.fr/cerat/Recherche/PagesPerso/LabbeExperience.pdf

• On Corneille and Molière:

- "Corneille in the shadow of Molière". Dublin : University of Dublin (Trinity College), French Department Research Seminar, April 6 2004.
http://web.upmf-grenoble.fr/cerat/Recherche/PagesPerso/LabbeDublin.pdf

- French version:
http://web.upmf-grenoble.fr/cerat/Recherche/PagesPerso/LabbeCorneilleMoliere.pdf

Our data and software are at the disposal of researchers.

Cyril Labbé - Université Joseph Fourrier Grenoble II - cyril.labbe@imag.fr
Dominique Labbé - Institut d'Etudes Politiques de Grenoble –
dominique.labbe@iep.upmf-grenoble.fr