

## **Lettre ouverte aux animateurs de la revue *Lexicometrica***

A propos de l'article :

Stephan Vonfelt, *Le graphonaute ou Molière retrouvé.*

(avril 2011)

Avec un post-scriptum de décembre 2011

Nous découvrons sur votre site cet article :

<http://lexicometrica.univ-paris3.fr/numspeciaux/special9/Vanfelt.pdf>

Ce texte (qui serait en circulation depuis plusieurs mois) prétend nous avoir réfutés en « retrouvant » Molière.

**Puisque cet article nous visait directement et explicitement, pourquoi n'avons-nous pas été avertis ?**

M. Vonfelt a travaillé avec des fichiers remis par Cyril Labbé et moi-même. Cela lui faisait obligation de nous communiquer son texte avant publication.

Vous deviez également le faire. Votre abstention est d'autant plus incompréhensible que l'article pose des problèmes évidents.

### **1. Sur la méthode.**

Voici quelques remarques qu'un lecteur attentif ne peut manquer de faire.

- Plusieurs problèmes de calculs ne sont pas éclaircis. Par exemple, comment sont traités l'intervalle entre le début du texte et la première occurrence du caractère étudié et celui séparant la dernière occurrence de ce même caractère de la fin du texte ? Autre exemple : le texte de l'article indique que les apostrophes, traits d'union, tirets, espaces, ponctuations entrent dans le calcul au même titre que les lettres alors que la note 19 suggère que les espaces et les ponctuations sont exclus de ce calcul... Ce flou et l'absence de données chiffrées donnent le sentiment que M. Vonfelt ne souhaite pas que l'on refasse ses expériences...

- Trois "distances" seulement sont publiées (au début § 3.1) : 0,0246, 0,0270 et 0,0276. Ces valeurs sont très petites. Ce serait la 4<sup>e</sup> décimale qui permettrait d'affirmer que Molière est plus proche de Racine que de Corneille ? Comme c'est absurde, aucun autre chiffre n'est donné et les graphiques en p. 7 et 8 sont sans échelle...

De plus,

- Par construction, la "distance Vonfelt" n'est pas indépendante des effectifs. Il suffit de développer la formule pour s'en douter. Quelques tests simples le confirment.

- Cela amène une autre question : comment et sur quels corpus a-t-on testé l'aptitude de cette mesure à discriminer les auteurs ? Il n'y a aucune indication à

ce sujet dans l'article. La thèse ne porte que sur trois romanciers du XXe siècle (Le Clézio, Yourcenar, Tournier) qui ne posent pas de problème d'identification...

- Pourquoi, dans le graphe factoriel de la section 3.1, les points ne sont pas identifiés ? Pourquoi la contribution de chacun des axes n'est pas indiquée comme il est d'usage ? La forme en U suggère l'influence d'une variable dominante. De quelle variable s'agit-il ? Pourquoi ne pas l'avoir recherchée ?

- La classification automatique est un complément indispensable à ce type de démarche. Son absence ne vous a pas alertés ?

- M. Vonfelt tire de cette analyse factorielle que genre et forme sont les facteurs dominants (axe horizontal du graphe) et qu'il faut donc distinguer tragédies et comédies, vers et prose. Surprise : dans les deux sections suivantes, les comédies et les tragédies de Corneille sont mélangées ; les pièces en vers et en prose de Molière également ! Erreur de méthode ? Volonté de dissimulation ?

- les pièces ne sont pas classées de la même manière dans les graphiques des deux sections suivantes. Pourquoi ?

Enfin et surtout,

- Dans le corpus Corneille, une pièce a disparu : *La comédie des Tuileries*. Nous l'avons pourtant remise à M. Vonfelt. Pourquoi l'avoir écartée ?

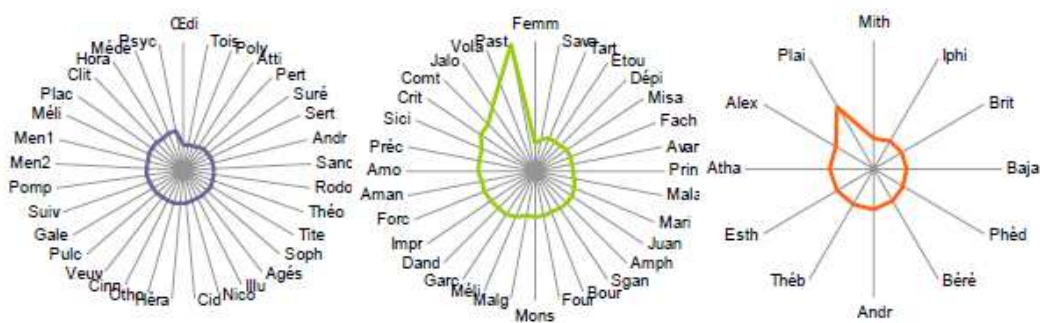
- Corneille et Molière ont officiellement "collaboré" pour une pièce (*Psyché*). Les passages qu'ils sont censés avoir écrit l'un et l'autre sont connus. Molière aurait écrit le premier acte et la première scène des actes 2 et 3 ; Corneille le reste. Nous avons remis ces textes à M. Vonfelt en deux fichiers distincts. Pourtant, il n'utilise que le texte de Corneille. Pourquoi avoir écarté celui de Molière ? Pourquoi le cacher maladroitement dans la note 15 ?

Il y a encore beaucoup d'autres questions, mais celles-ci suffisent pour comprendre combien vous avez eu tort de vous affranchir des règles usuelles en matière de publication scientifique...

## **2. Sur les résultats :**

Premièrement, la "distance Vonfelt" est corrélée avec les effectifs des caractères dont la somme donne la "longueur" des textes comparés... Plus cette longueur est élevée, plus la "distance" est petite.

On le voit dans les diagrammes de la section 3.2 (reproduits ci-dessous)



Ces diagrammes représentent la moyenne des "distances" de chacun des textes à tous les autres composant le corpus considéré (Corneille à gauche, Molière au centre et Racine à droite). Dans ces diagrammes, les textes sont classés dans le sens des aiguilles d'une montre du plus "central" (plus petite "distance" moyenne), situé à 0 h., au plus décalé (plus grande "distance" moyenne). Dans ces trois figures, la courbe s'éloigne du centre au fur et à mesure que l'on tourne et que... la longueur des textes diminue.

Chez Corneille (figure de gauche), d'après la mesure Vonfelt, les pièces les plus "centrales" sont les plus longues : *la Toison d'Or* (20 343 mots), *Œdipe* (18 618 mots) ; les pièces les plus "décalées" sont les plus courtes : *Psyché* (10 067 mots), *Médée* (14 269 mots), *Clitandre* (14 402 mots), *la Place royale* (13 801 mots)... On remarque également que la figure est presque circulaire en dehors d'une très légère pointe pour *Psyché* (nettement plus courte que les autres puisque Corneille est censé en avoir composé seulement les deux tiers). En effet, chez Corneille, la dispersion des longueurs des pièces autour de la moyenne (16 273 mots) est faible (coefficient de variation  $v = 15,9\%$ ). Cela explique aussi le caractère assez régulier de la figure de gauche dans la section 3.3. Enfin, la corrélation entre longueurs des pièces et "distances" moyennes est significative au seuil de 5%.

Chez Racine (figure de droite), d'après la mesure Vonfelt, les pièces les plus centrales sont aussi les plus longues : *Iphigénie* (15 782 mots), *Mithridate* (15 091 mots), *Britannicus* (15 387 mots), etc. Les pièces les plus décalées sont les plus courtes (hormis *Athalie*) : les *Plaideurs* (8 041 mots), *Esther* (11 147 mots) et *la Thébaïde* (13 813). A l'exception des *Plaideurs*, la disposition des points sur le graphique est régulière car la dispersion des longueurs autour de la moyenne (13 885 mots) est faible ( $v = 16,2\%$ ). Cela explique la figure de droite dans la section 3.3. Enfin, la corrélation - entre longueur et "distance" moyenne - est également significative.

Chez Molière (figure du milieu), les pièces les plus "centrales" sont encore les pièces les plus longues, en vers : *l'Etourdi* (18 671 mots), *les Femmes savantes* (16 863 mots), *l'Ecole des femmes* (16 625 mots), *le Dépit amoureux*

(16 242 mots), etc. Les pièces les plus décalées sont les pièces en prose les plus courtes : *la Comédie pastorale* (732 mots), *la Jalousie du barbouillé* (3 501 mots), *le Médecin volant* (3 876 mots), *la Comtesse d'Escarbagnas* (5 564 mots), etc. Ici la figure est plus irrégulière. En effet, la dispersion des longueurs autour de la moyenne (11 405 mots) est beaucoup plus forte ( $v = 49,6\%$ ). Cette dispersion explique le visage particulier de la figure du milieu dans la section 3.3. Enfin, en séparant les pièces en prose et les pièces en vers, on trouve à nouveau une corrélation significative entre longueur et "distance" moyenne.

**M. Vonfelt mesure donc principalement la longueur des pièces et, secondairement, des différences de vocabulaire.**

Si M. Vonfelt avait classé les pièces de la même manière dans les graphiques des sections 3.2 et 3.3, cela aurait été évident...

Reprenons les longueurs moyennes (en mots) : Corneille : 16 273 ; Racine : 13 885 ; Molière : 11 405. **Les longueurs des pièces de Corneille sont plus proches de celles de Racine que de celles de Molière.** Puisque M. Vonfelt prend les longueurs des pièces pour leurs "distances", on devine ses "résultats". **Citons M. Vonfelt : "Corneille et Racine sont les plus proches (...) Molière affiche moins d'affinité avec Corneille qu'avec Racine".**

CQFD !

Deuxièmement, l'"attribution d'auteur" évoquée par M. Vonfelt dans sa section 3.3 est faussée par une erreur de méthode évidente : mélanger tragédies et comédies singularise nécessairement Corneille ; mélanger vers et prose singularise nécessairement Molière.

Malgré tout le résultat est faux pour : *Le menteur* (Corneille 1642), *La Comédie des Tuileries* (Corneille 1634), *Dom Garcie de Navarre* (Molière 1661), *Amphitryon* (Molière 1668), *les Amants magnifiques* (Molière 1670), *Psyché* (les passages attribués à Molière en 1671), *la Thébaïde* (Racine 1664), *Alexandre* (Racine 1665), *les Plaideurs* (Racine 1668)...

**9 pièces ne sont pas attribuées "correctement", soit un taux d'erreur de 11% !** Surtout, la méthode échoue sur *Psyché* : la "mesure Vonfelt" est incapable de "retrouver Molière" dans les passages de cette pièce que celui-ci est censé avoir écrits... M. Vonfelt a donc escamoté ce texte gênant (avec *la Comédies des Tuileries* que sa méthode "attribue" à Molière alors qu'elle est de Corneille !).

Ce résultat médiocre ne surprend pas puisque la "mesure Vonfelt" est d'abord une manière compliquée et imprécise de calculer l'inverse de la longueur des textes.

Au contraire, c'est la "réussite" relative qui surprend. Elle s'explique ainsi : chaque pièce est rapportée à la moyenne des trois corpus qui sont de longueurs très différentes.

- La plupart des pièces officielles de Corneille sont plus longues que celles des autres. M. Vonfelt les "attribue" à l'auteur le plus long ;
- Les pièces de Racine sont pratiquement toutes plus courtes que celles de Corneille ;
- Les pièces parues sous le nom de Molière sont de longueurs plus diverses mais, pour la majorité, plus courtes que celles de Racine...

On comprend maintenant certaines caractéristiques étranges de l'article de M. Vonfelt : explications techniques confuses et lacunaires, absence de tests, de tableaux chiffrés, de classification automatique et d'annexe présentant les pièces, effacement des légendes sur le graphe factoriel, escamotage de certains textes...

On comprend aussi votre volonté de nous dissimuler ce factum aussi longtemps que possible !

Si vous aviez respecté les règles élémentaires en matière de publication scientifique, nous aurions pu vous faire les remarques ci-dessus, sans avoir à les rendre publiques.

En publiant, sans précautions, cet étrange article, vous avez été bien légers.

C'est également le cas de tous ceux qui, comme Wikipedia, ont cité M. Vonfelt sans réfléchir ni vérifier...

Votre manque de sérieux a rendu possible une grossière opération de désinformation contre nous.

Notre travail sur Corneille et Molière en sort renforcé.

Dominique Labbé (11 avril 2011)

[dominique.labbe@iep-grenoble.fr](mailto:dominique.labbe@iep-grenoble.fr)

<http://www.pacte.cnrs.fr/spip.php?article54>

PS : le détail des corpus (genre, dates de création, longueurs et vocabulaires) peut être trouvé en annexe de notre ouvrage : *Si deux et deux sont quatre, Molière n'a pas écrit Don Juan*. Paris : Max Milo, 2009.

## Post-scriptum (décembre 2011)

M. Vonfelt a placé une réponse sur son site, à l'adresse suivante :

<http://graphorythmes.free.fr/corpus/Réponse.pdf>

Ce document sans date confirme toutes nos observations, spécialement à propos des dissimulations et des manipulations auxquelles l'auteur s'est livré. Citons notamment, parmi les aveux implicites ou explicites :

- il a dissimulé la corrélation entre son indice et la longueur des pièces,

- pour maquiller la petitesse extrême de ses résultats, il les a multipliés par 100... C'est bien la 4<sup>e</sup> décimale qui est censée discriminer les auteurs !

- il a enlevé une partie de *Psyché* – et n'en a rien dit dans son article - parce que sa méthode attribue à Corneille les passages de cette pièce que Molière est supposé avoir écrits. La mesure de M. Vonfelt aboutit au constat suivant : le prétendu Molière écrivait comme Corneille. C'est pourquoi, il a maladroitement dissimulé ce résultat. A elle seule, cette dissimulation d'un résultat crucial disqualifie M. Vonfelt,

- surtout, il n'a réalisé aucune des expériences standards nécessaires pour la mise au point d'un outil statistique : son bricolage n'a été appliqué qu'à Corneille, Molière et Racine (avec un taux d'échec considérable)....

Enfin, il reconnaît qu'il n'a pas respecté les règles élémentaires du débat scientifique.

Les Moliéristes ont eu bien tort de mettre en avant ce curieux personnage, mais l'ont-ils seulement lu avant de le publier ?