# Ce que disent leurs phrases

Cyril Labbé, Dominique Labbé

10th International Conference
Statistical Analysis of Textual Data
SAPIENZA University of Rome
(9-11 june 2010)

Cyril LABBE[1], Dominique LABBE[2]

1. Université Grenoble I (cyril.labbe@imag.fr)

2. Institut d'Etudes Politiques de Grenoble (dominique.labbe@iep.grenoble.fr)

# Ce que disent leurs phrases

Publié dans :

Bolasco Sergio, Chiari Isabella, Giuliano Luca (Eds). *Proceedings of 10th International Conference Statistical Analysis of Textual Data*. Rome : Edizioni Universitarie di Lettere Economia Diritto, 2010, Vol 1, p. 297-307.

**Abstract**

A study of sentence lengths in French 17e century theatre (Corneille, Mairet, Molière, Quinault, Racine). The mid-values (mean, mode, median and medial) and the frequency function highlight some characteristics of each author and the constraints imposed by alexandrine verses and rules of the theatre. Four clusters of sentences are highlighted according to their lengths and their aims: to challenge, to converse, to expound even to soliloquize.

**Keywords :** French theatre ; Sentence lengths ; Alexandrine verses ; Stylometrics

**Résumé**

Etude de la phrase dans le théâtre du XVIIe siècle (Corneille, Mairet, Molière, Quinault, Racine). Les valeurs centrales (moyenne, mode, médiane et médiale) et la distribution des fréquences révèlent des caractéristiques propres à chacun des auteurs et les contraintes très fortes que font peser les vers alexandrins et les règles du théâtre classique. On distingue quatre type de phrases (en fonction de leur longueur) qui remplissent des fonctions différentes : interpeller, dialoguer, exposer voire soliloquer.

**Mots-clefs :** Théâtre français – Corneille – Molière – longueur des phrases – versification alexandrin - stylométrie

## 1. Introduction

Malgré le développement de la statistique textuelle, les études quantitatives portant sur le style sont rares. De son côté, la stylistique s'est détournée des recensements statistiques qui est le "ventre mou de la stylistique française", selon l'expression de G. Molinié (1986 : 54). Pourtant, une stylistique quantitative serait possible, comme le montre le recensement des indices possibles effectué par Molinié (1986 : 146-156).

Cette stylistique quantitative pourrait apprendre beaucoup sur les œuvres et leurs auteurs comme nous allons le montrer à l'aide d'un des indices, connu depuis fort longtemps et que cite Molinié : la longueur des phrases. Le théâtre français du XVIIe offre un terrain d'expérimentation très favorable. Il comporte de nombreux textes contemporains écrits dans un même genre. De plus, certains auteurs – Corneille, Molière, Racine - ont déjà fait l'objet d'analyses approfondies (Muller 1967, Kylander 1995, Bernet 1983). Muller a proposé plusieurs indices stylistiques (Muller 1967, p. 125-128) qui ont été repris par les deux autres, mais la longueur des phrases (et leur construction) ne figure pas parmi ces indices. Cette étude est pourtant instructive comme on l'a déjà montré à plusieurs reprises (par exemple : Monière et Labbé 2002 ; Monière, Labbé & Labbé 2008).

Après avoir présenté le corpus et les conventions utilisées, on étudiera la longueur des phrases dans le théâtre classique du XVIIe. Cette application révèle que des contraintes fortes pesaient sur les auteurs. Il s'agit tout particulièrement de la versification alexandrine.

## 2. Corpus et conventions de dépouillement

On ajoute aux trois corpus de toutes les pièces de Corneille (parues entre 1630 et 1674), de Molière (1659-1673) et de Racine (1664-1691), les trois *Dissertations sur le poème dramatique* (Corneille, 1660 prose) et cinq pièces de deux autres auteurs contemporains. Mairet pour deux tragédies : *Sylvie* (1626) et *Sophonisbe* (1634) ; Quinault pour deux comédies héroïques chantées et mises en musiques - *Atys* (1676) et *Thésée* (1675) - et une comédie : *La mère coquette ou les amants brouillés* (1665). Cela représente un total de 1,151 millions de mots (tableau I).

| | Genre | Nombre de textes | Tragédies | Comédies | Longueur (mots) |
|---|---|---|---|---|---|
| Corneille | Vers | 34 | 25 | 9 | 530 531 |
| | Prose | 3 | - | - | 36 375 |
| Mairet | Vers | 2 | 2 | | 35 979 |
| Molière | Vers | 13 | | 13 | 160 342 |
| | Prose | 18 | | 18 | 187 572 |
| Quinault | Vers | 3 | 2 | 1 | 33 557 |
| Racine | Vers | 12 | 11 | 1 | 166 626 |
| Total | | 85 | 40 | 42 | 1 150 982 |

Tableau 1. Les corpus utilisés pour l'étude des longueurs de phrases dans le théâtre classique

Toutes les pièces en vers sont en alexandrins, sauf cinq composées en "vers libres" (Corneille *Agésilas*, Molière *Amphitryon* et les pièces de Quinault). Nous négligerons le fait, signalé par Bernet, Kylander et Muller, que toutes les pièces en alexandrins contiennent quelques vers n'ayant pas 12 pieds et que *Agésilas* est en majorité composée d'alexandrins.

*Traitements préalables*

Après correction et standardisation des graphies, les textes ont été balisés afin d'exclure de l'analyse les "didascalies" : noms des acteurs, numéro des actes et des scènes, indications scéniques… Ainsi l'analyse ne porte que sur ce qu'entend le spectateur, selon le principe posé par Muller (1967). Chaque mot a reçu une étiquette comportant sa graphie standard (la "forme") et son "lemme" (ou "vocable"), c'est-à-dire l'entrée sous laquelle on trouve ce mot dans un dictionnaire de langue - l'infinitif du verbe, le masculin singulier de l'adjectif, etc. - et sa catégorie grammaticale. Entre autres avantages, cette "lemmatisation" permet de distinguer les "homographes" (une même graphie mais plusieurs entrées de dictionnaire) : par exemple "pouvoir", "devoir", "savoir" qui sont substantifs et verbes à l'infinitif. Ces opérations sont décisives : dans tout texte français, au moins un tiers des mots sont des "homographes".

Dans ce qui suit, on ne compte pas des "formes graphiques" mais des "mots". Par exemple, "aujourd'hui" ou "parce que" comptent deux "formes graphiques" (l'apostrophe ou le blanc étant des "caractères séparateurs"), mais ils ne forment qu'un seul mot puisqu'ils figurent ainsi dans les dictionnaires. En revanche, "du" – une seule forme graphique – est compté pour deux mots (de + le) pour les mêmes raisons (pour une présentation détaillée de ces conventions : Labbé 1990). Dans la suite de cette communication, nous nommerons "mots", ces formes graphiques standardisées, équivalentes aux "tokens" anglo-saxons. Leur recensement figure en dernière colonne du tableau 1 ci-dessus.

*Qu'est-ce qu'une phrase ?*

Une phrase est l'espace de texte compris entre deux ponctuations fortes. Et la longueur de la phrase est mesurée par le nombre de mots compris dans cet espace. Une ponctuation forte est l'un des signes suivants : '.' '…' '?' et '!' quand ils sont suivis d'un mot dont l'initiale est un caractère majuscule. Si un nom propre suit un point, l'opérateur tranche entre deux possibilités : début d'une nouvelle phrase ou abréviation au sein d'un patronyme (par exemple "M. Verdurin"). Le respect de ces conventions est indispensable comme l'a montré la controverse autour de la longueur des phrases chez Proust (Milly 1975 & 1986 : 165-167).

## 3. Les valeurs centrales

Le tableau 2 donne les valeurs centrales pour les différents corpus. Outre le mode principal, les valeurs centrales sont au nombre de trois : moyenne, médiane et médiale. A la moyenne est associé l'écart type qui mesure la dispersion des valeurs composant la série.

| | Mode | Médiane | Médiale | Moyenne | Ecart-type |
|---|---|---|---|---|---|
| Corneille (tragédies en vers) | 9 | 18,7 | 34,7 | 24,2 | 18,2 |
| Corneille (comédies en vers) | 9 | 14,2 | 26,3 | 18,5 | 14,8 |
| Corneille (*Menteurs*, comédies en vers) | 9 | 10,3 | 26,8 | 16,8 | 15,2 |
| Corneille (dissertations en prose) | 27 & 41 | 40,4 | 52,8 | 44,2 | 24,8 |
| Mairet (2 tragédies) | 18 | 18,7 | 36,2 | 26,3 | 22,2 |
| Molière (comédies en vers) | 9 | 10.6 | 28,1 | 16,9 | 17,1 |
| Molière (comédies en prose) | 1 | 8,5 | 21,8 | 13,6 | 14,6 |
| Quinault (2 tragédies en vers) | 9 | 10,6 | 19,3 | 14,4 | 10,4 |
| Quinault (1 comédie en vers) | 9 | 8,1 | 18,2 | 11,6 | 11,3 |
| Racine (tragédies en vers) | 9 | 10,5 | 19,2 | 15,1 | 12,2 |
| Racine (*Plaideurs*, comédie en vers) | 1 & 4 | 5,4 | 10,9 | 8,2 | 8,3 |

Tableau 2. Récapitulatif des valeurs centrales et écart-types dans les corpus étudiés.

*La longueur moyenne* diffère considérablement selon les auteurs et les genres. Il convient d'ajouter trois conclusions. En premier lieu, quel que soit l'auteur (Corneille, Quinault, Racine), la tragédie comporte des phrases en moyenne plus longues que la comédie.

Deuxièmement, dans tous les corpus, l'écart type indique une forte dispersion des longueurs de phrase autour de la moyenne, ce qui invite à la prudence. Enfin, les valeurs moyennes pour chacune des 82 pièces – dont le tableau ne peut être reproduit – font apparaître que, chez un même auteur, ces valeurs peuvent varier considérablement, même pour des pièces appartenant à un même genre, et que, chez Corneille et Racine, il existe peut-être une tendance à la baisse (trait gras sur les figures 1 et 2).

Cependant, dans les deux cas, l'ajustement est médiocre (coefficient de corrélation non significatif au seuil de 10 %). Pour Racine, à partir d'*Andromaque*, la tendance est horizontale. Pour Corneille, à partir de *Rodogune* (1644), la droite d'ajustement est également horizontale. Ces auteurs auraient donc mis un certain temps pour "stabiliser" leur style. En revanche, dans les comédies en vers parues sous le nom de Molière (entre 1659 et 1673), il n'y a aucune variation de la longueur suivant la chronologie.

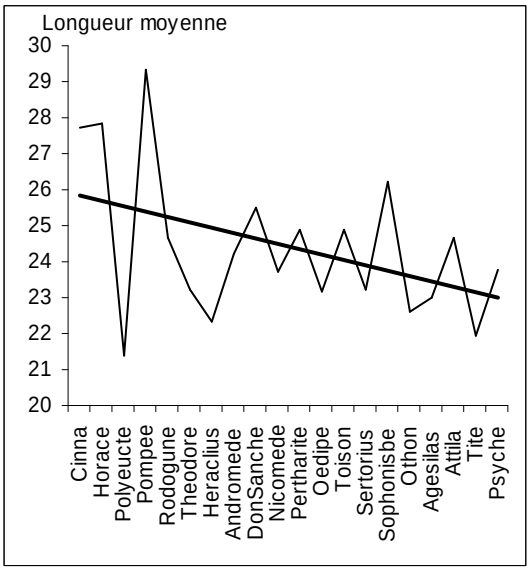


Figure 1. Evolution de la longueur moyenne des phrases dans les tragédies de Corneille (classement chronologique)

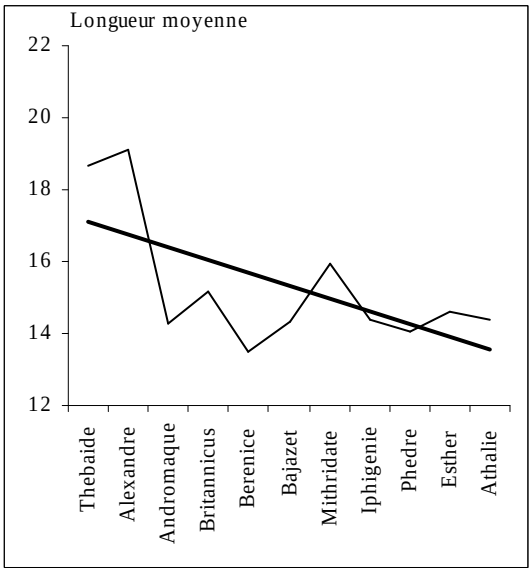


Figure 2. Evolution de la longueur moyenne des phrases dans les tragédies de Racine (classement chronologique)

La *longueur médiane* sépare la population étudiée en deux parties égales. Par exemple, dans les tragédies de P. Corneille, la moitié des phrases comportent 19 mots ou moins, et l'autre moitié 19 et plus. Lorsqu'il passe à la comédie, Corneille augmente le nombre des phrases brèves (leur contenu est examiné plus bas) et il diminue d'autant les phrase longues (d'où une baisse de la longueur moyenne). Ce mouvement est encore accentué dans les deux dernières comédies de Corneille qui occupent une place à part dans son oeuvre : *le Menteur* et la *Suite du Menteur*.

La *longueur médiale* partage la totalité des mots en deux parts égales. Par exemple, dans les tragédies de Corneille, la moitié des mots se trouvent dans des phrases longues de 35 mots et plus. Autrement dit, lorsqu'il assiste à une tragédie de Corneille, le spectateur entend, la moitié du temps, des phrases de longueurs considérables. Des expériences de psychologie ont montré que l'auditeur moyen éprouve des difficultés à décrypter des phrases dont la longueur est supérieure à 18 - 20 mots, selon la complexité du vocabulaire (Richaudeau 1988). Même s'il s'agit d'un simple ordre de grandeur, il est évident que le spectateur des tragédies de Corneille a quelque mal à en suivre certains passages, d'autant plus que l'auteur aimait à mêler plusieurs intrigues secondaires à la principale ! Seuls Racine

et Quinault semblent avoir eu conscience de ce problème de communication, ce qui les a amené à utiliser des phrases beaucoup plus brèves (et un vocabulaire un peu moins étendu).

On observera enfin que les comédies en alexandrins parues sous le nom de Molière (entre 1659 et 1672) présentent les mêmes valeurs centrales que les 2 *Menteurs* (Corneille, 1642-1643). Cela ne surprendra pas : les 2 *Menteurs* sont la matrice d'où sont sorties toutes les comédies en vers écrites par Corneille pour Molière (Labbé & Labbé, 2001 et Labbé, 2009).

**4. Distribution des longueurs de phrase dans le théâtre classique**

Pour chacune des pièces, puis pour chacun des sous-corpus, les longueurs de phrase sont rangées par ordre croissant et représentées dans des histogrammes (figures 3 à 12 ci-dessous). La hauteur de chaque barre donne la densité (relative) dans le corpus de la longueur de phrase correspondante sur l'axe des abscisses. Par exemple, dans les tragédies de Corneille, 5,4% des phrases ont une longueur de 9 mots (comme indiqué dans la figure 3), c'est la taille la plus fréquemment rencontrée ou *longueur modale principale*. On remarque également l'existence de *modes secondaires* à 4-5, 18, 27-28, 36 et 54 mots.

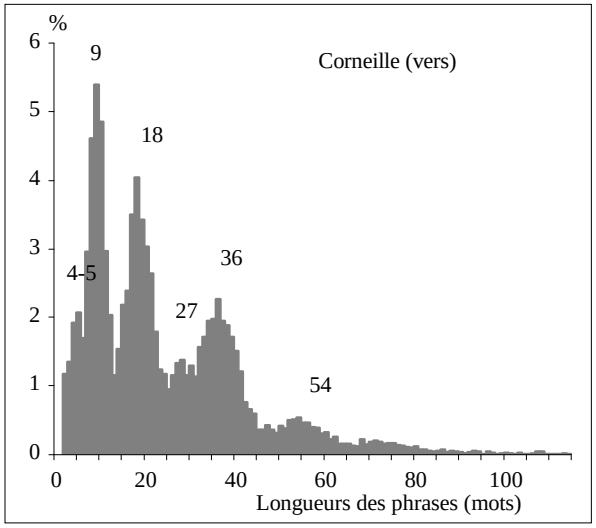


Figure 3. Distribution des longueurs de phrases dans les tragédies de Corneille (vers)

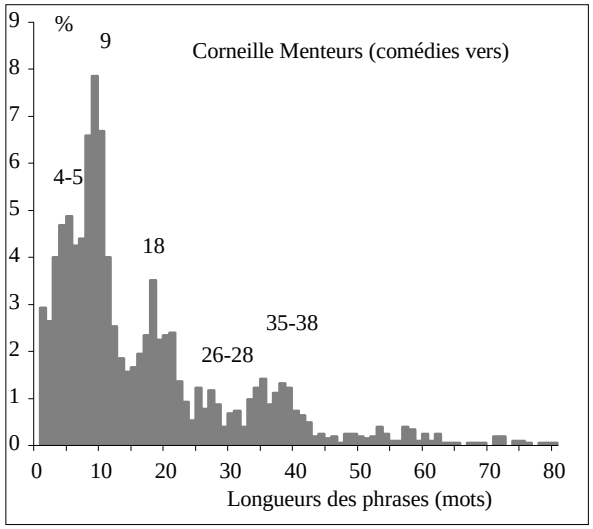


Figure 4. Distribution des longueurs de phrases dans les *2 Menteurs* de Corneille (vers)

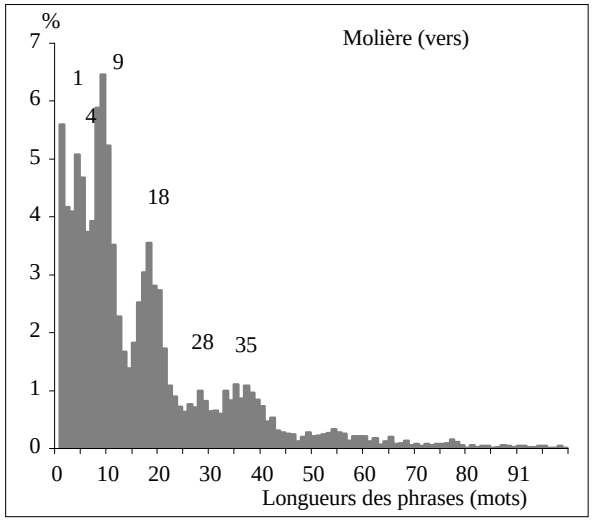


Figure 5. Distribution des longueurs de phrases dans les comédies en vers de Molière

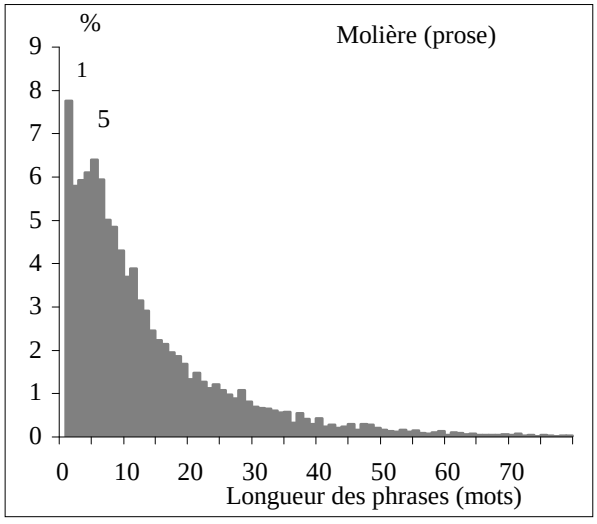


Figure 6. Distribution des longueurs de phrases dans les comédies en prose de Molière

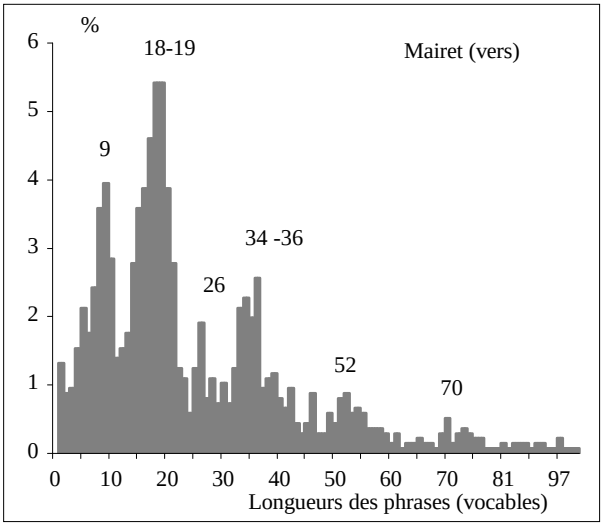


Figure 7. Distribution des longueurs de phrases dans les 2 tragédies de Mairet (vers)

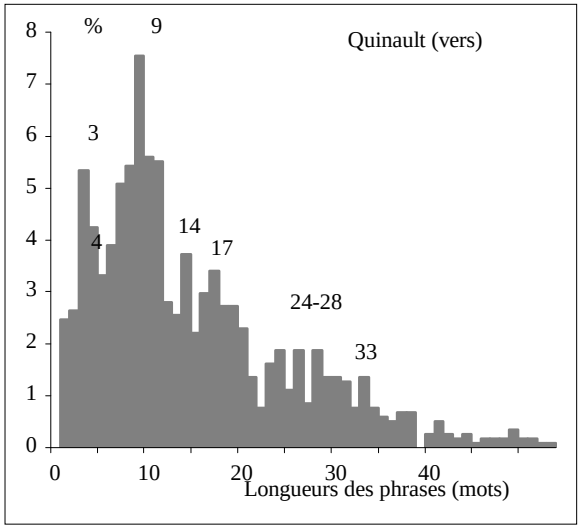


Figure 8. Distribution des longueurs de phrases dans les 2 tragédies de Quinault (vers)

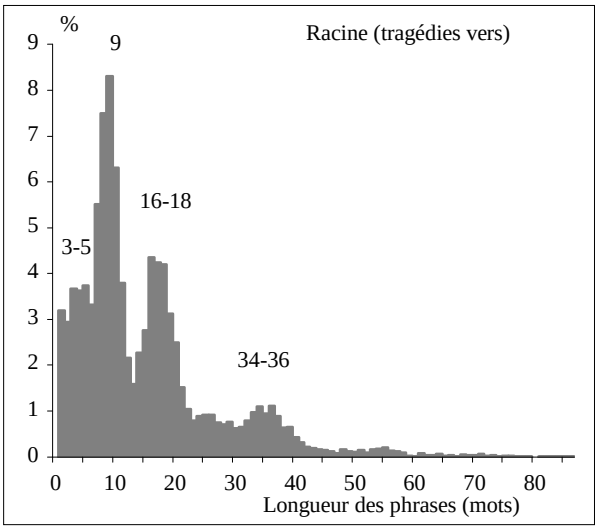


Figure 9. Distribution des longueurs de phrases dans les Tragédies de Racine (vers)

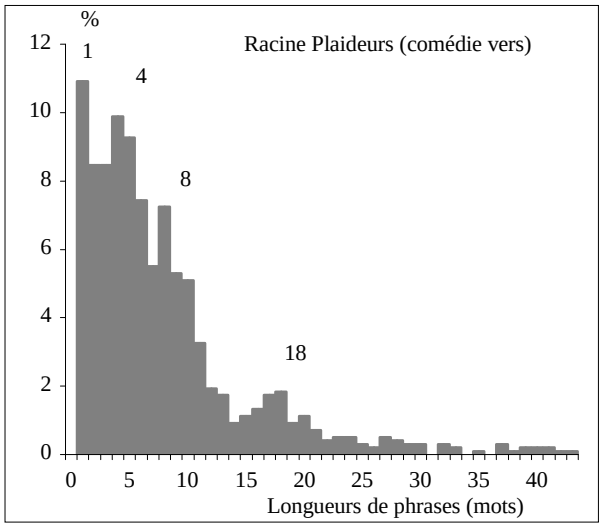


Figure 10. Distribution des longueurs de phrases dans les *Plaideurs* de Racine (vers)

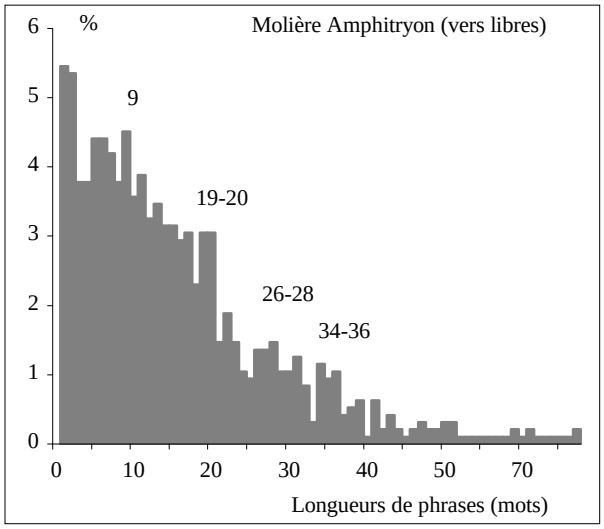


Figure 11. Distribution des phrases dans *Amphitryon* (Vers libres, Corneille sous le nom de Molière)

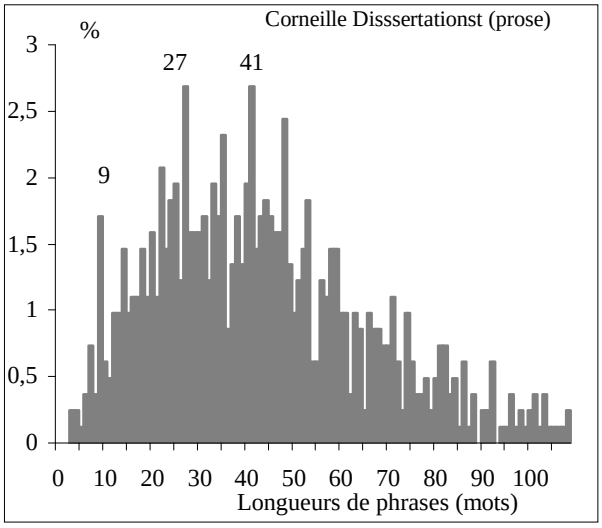


Figure 12. Distribution des longueurs de phrases dans les *Dissertations sur le poème dramatique* (Corneille)

Ces figures "multimodales" sont très éloignées de la forme à laquelle on pourrait s'attendre (courbe en cloche plus ou moins asymétrique). Elles suggèrent que ces phrases ne résultent pas d'un seul patron mais plutôt du mélange de plusieurs populations (groupées autour de chacun des modes). Cette configuration multimodale se retrouve dans pratiquement toutes les œuvres en vers. Le mode de 9 mots semble s'imposer. Toutefois, chez Mairet, il est supplanté par le double : 18 mots. De manière générale, les modes semblent être approximativement des multiples de 4,5 : 9, 18, 27, 36, 54…

On est inévitablement amené à mettre cette caractéristique en relation avec le fait qu'il s'agit de pièces en vers alexandrins (sauf celles de Quinault). Pour des raisons évidentes, une bonne partie des fins de phrase coïncident avec des fins des vers ou sont placées à l'hémistiche. On retrouve alors les proportions établies par C. Muller (Muller 1967 : 49-51) : chez Corneille le nombre moyen de mots par vers oscille entre 9,1 et 9,25 suivant les pièces. Dans les pièces en vers présentées par Molière, cette moyenne est de 9,2 et elle oscille entre 9 pour *l'Etourdi* et 9,3 pour le *Misanthrope* (Kylander 1995: 9-11). Chez Racine, l'alexandrin comporte un peu moins de mots : en moyenne, 8,86 vocables par vers (Bernet 1983 : 33-34, les chiffres de Bernet ne permettent pas de connaître les moyennes par pièce).

Ces caractéristiques ne sont donc pas intrinsèques au style des auteurs, pas plus que la métrique ou la rime des vers. La distribution multimodale – sur un rythme approximatif de 4,5 mots - disparaît quand ces auteurs abandonnent les alexandrins. Le corpus en fournit plusieurs illustrations. Dans les comédies en prose présentées par Molière, les longueurs de phrases se distribuent suivant une courbe presque parfaite, de forme exponentielle dont l'exposant est négatif mais supérieur à -1 (figure 6). Ce profil est à peu près celui que l'on retrouve dans *Amphitryon*, même s'il subsiste des traces des modes "alexandrins" (figure 11). En effet, dans cette pièce, les 12 pieds sont abandonnés et seule l'alternance des rimes est conservée. Dans les *Plaideurs,* en alexandrins, Racine parvient à approcher ce profil qui pourrait être celui de la "prose versifiée" proposée par Corneille comme idéal de la comédie en vers (figure 10).

Dans les *Trois dissertations sur le poème dramatique* de Corneille (figure 12), la distribution des phrases ressemble à une courbe en cloche, mais avec quatre caractéristiques particulières. Premièrement, toutes les valeurs centrales sont beaucoup plus élevées que dans le théâtre. Deuxièmement, il existe deux modes principaux exactement égaux (à 27 et 41 mots comportant chacun 2,7% des phrases), ce qui suggère l'existence d'au moins deux populations mélangées. Troisièmement, un important étalement des longueurs autour de ces modes donne à la courbe un aspect fortement "aplatie". Quatrièmement, la valeur médiane (41) est fort proche de la moyenne (44,4), ce qui signale une asymétrie moins marquée que dans les pièces en vers. On notera enfin dans

cette figure 10, l'existence d'un mode secondaire à 9 mots. Or il n'y aucun vers cité dans ces trois dissertations. P. Corneille conserve une certaine propension à faire des phrases de 12 pieds, même quand il écrit de la prose !

## 5. Caractéristiques des phrases en fonction de leur longueur

Les histogrammes présentés ci-dessus suggèrent que – hormis les *Plaideurs* de Racine – les pièces en vers alexandrins résultent d'un mélange de plusieurs populations de phrases relativement faciles à isoler suivant le schéma de principe ci-dessous (figure 13).

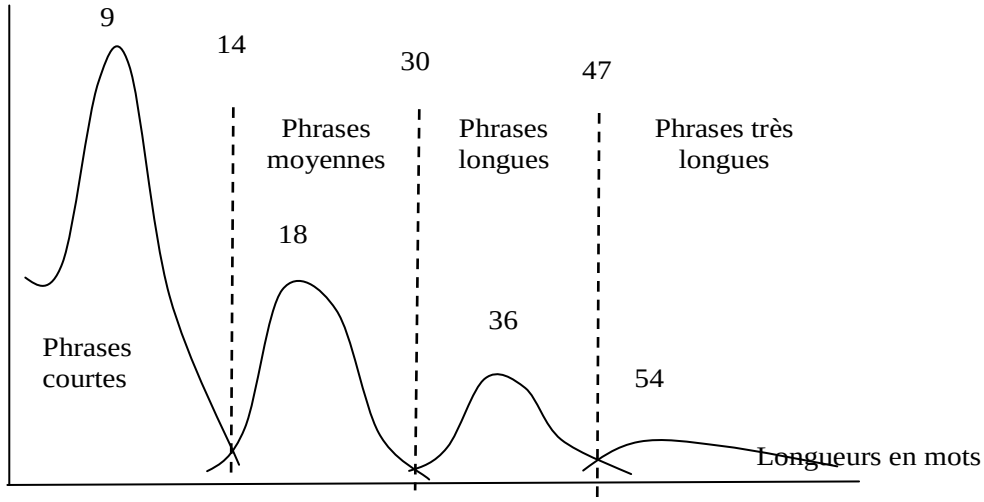


Figure 13. Les quatre groupes de phrases en fonction de leurs longueurs

Ce schéma est établi d'après les *Menteurs* de Corneille (figure 4 ci-dessus). Les coupures (traits pointillés) entre ces populations sont fixées aux points bas de l'histogramme. On constitue quatre sous-corpus avec les phrases dont les longueurs sont comprises entre deux bornes (1-14 ; 15-30 ; etc.) et l'on compare ces sous-ensembles aux trois autres grâce à la méthode du vocabulaire caractéristique (Labbé & Labbé 1994 ; Monière, Labbé & Labbé 2005). Cette comparaison porte sur les catégories grammaticales et sur les vocables. Le tableau 3 ci-dessous présente le poids des catégories grammaticales pour les phrases courtes (colonne B) comparées aux trois autres groupes (colonne A).

Les différences entre A et B sont-elles statistiquement significatives ? Cette question revient à se demander si B peut être considéré comme un échantillon aléatoire tiré dans A. Il y a 8 817 mots dans B ($N_b$). Pour la première ligne, avec un risque d'erreur $\alpha = 5\%$ peut-on considérer que la densité des verbes dans B ($PV_b$ = 222,6‰) ne diffère pas significativement de celle observée dans A ($PV_a$ = 200,1 ‰) ? En utilisant l'approximation normale, il faudrait que:

$$PV_b - 1{,}96\sqrt{\frac{PV_b * (1000 - PV_b)}{N_b}} \leq PV_a \leq PV_b + 1{,}96\sqrt{\frac{PV_b * (1000 - PV_b)}{N_b}}$$

Ce qui donne un intervalle {218,2 – 227,0}. $PV_a$ étant situé largement en dehors de cet intervalle, on conclut que la densité des verbes dans les phrases courtes est significativement supérieure à celle observée dans le reste du corpus. Le même calcul montre que, en dehors des adjectifs démonstratifs, tous les écarts en dernière colonne du tableau 3 sont significatifs.

| Catégories | A (Corpus sans le sous corpus) ‰ | B (Sous corpus phrases courtes) ‰ | (B-A)/A % |
|---|---|---|---|
| Verbes | 200.1 | 222.6 | +11.3 |
| *Formes fléchies* | *131.7* | *160.3* | *+21.7* |
| *Participes passés* | *20.7* | *18.8* | *-9.2* |
| *Participes présents* | *3.4* | *1.4* | *-60.5* |
| *Infinitifs* | *44.2* | *42.2* | *-4.6* |
| Noms propres | 9.6 | 12.7 | +32.8 |
| Noms communs | 148.4 | 134.1 | -9.6 |
| Adjectifs | 40.3 | 36.0 | -10.7 |
| *Adj. participe passé* | *5.9* | *1.9* | *-67.4* |
| Pronoms | 192.4 | 241.1 | +25.3 |
| *Pronoms personnels* | *124.7* | *158.3* | *+27.0* |
| Déterminants | 130.1 | 106.6 | -18.1 |
| *Articles* | *67.5* | *52.4* | *-22.4* |
| *Nombres* | *5.3* | *3.3* | *-38.3* |
| *Possessifs* | *32.7* | *29.0* | *-11.2* |
| *Démonstratifs* | *12.4* | *12.0* | *-2.9* |
| *Indéfinis* | *12.2* | *9.9* | *-19.1* |
| Adverbes | 84.2 | 90.5 | +7.4 |
| Prépositions | 113.9 | 91.0 | -20.1 |
| Conjonctions | 79.5 | 57.7 | -27.4 |
| *Conjonctions de coordination* | *43.4* | *33.7* | *-22.3* |
| *Conjonctions de subordination* | *36.1* | *24.0* | *-33.3* |

Tableau 3. Densité des catégories grammaticales dans les phrases courtes comparées au reste des *Menteurs*.

En dehors des conjonctions de coordination, il est possible de rassembler ces catégories en deux groupes selon qu'elles appartiennent plutôt à l'univers du verbe (pronoms, adverbes, conjonctions de subordination) ou à celui du nom (adjectifs, déterminants, prépositions). Le tableau 4 ci-dessous résume cette information pour les 4 classes de phrases.

| | Courtes | Moyennes | Longues | Très longues |
|---|---|---|---|---|
| Groupes verbaux | +12,8 | +2,6 | -6,1 | -9,8 |
| Groupes nominaux | -14,0 | -2,2 | +7,3 | +12,1 |

Tableau 4. Variation des densités des groupes verbaux et nominaux dans une catégorie de phrases comparée aux 3 autres (en %).

**Les phrases courtes** ont une longueur inférieure à 15 mots. Dans les *Menteurs*, cela représente 6 phrases sur 10 et 27% du texte. Le verbe domine et l'indicatif l'emporte sur les autres temps. Le plus fort écart négatif est enregistré pour les participes présents, ce qui est logique puisqu'il s'agit de la forme la plus nominale du verbe. Les trois vocables les plus caractéristiques de ces phrases sont trois pronoms personnels : *je, vous, tu* suivis de : *monsieur, adieu, madame* et des verbes *aller, dire, écouter, parler*… Ce sont les phrases de l'interpellation et de l'action. Car l'essentiel de l'action – sur une scène de théâtre – consiste à parler et à entrer et sortir de scène ! Dans la tragédie, Racine et Quinault donnent d'avantage de poids à ces brefs échanges verbaux par rapport à Corneille et surtout par rapport à Mairet.

**Les phrases moyennes**, dont la longueur est comprise entre 15 et 30 mots, sont celles de la conversation courante. Dans les *Menteurs*, elles couvrent 29% du texte total. Le groupe verbal continue à l'emporter mais moins nettement que dans les phrases courtes. Les écarts les plus significatifs concernent les participes présents (-26%), les infinitifs (+15%) et les adjectifs (-8%). Une fois établi ce vocabulaire caractéristique, le logiciel relit ces phrases en accordant à

chacune un score en fonction du nombre de vocables spécifiques qu'elle contient. Par exemple, voici la phrase la plus spécifique de ce second groupe :

*"S'il faut ruser ici j'en sais autant que vous et vous serez bien fin si je ne romps vos coups."*

Sur les 22 mots que comporte cette phrase, 9 sont spécifiques de la conversation courante dans les *Menteurs*.

**Les phrases longues** ont une taille comprise entre 31 et 47 mots. On observera que ces deux bornes correspondent – aux arrondis près – à la moyenne augmentée de 1 et de 2 écarts types. Ces phrases représentent 12% du total mais elles couvrent 26% de la surface de ces deux pièces. Autrement dit, pendant environ un quart de la pièce, le spectateur entend des périodes oratoires assez longues et parfois difficiles à comprendre. Dans la tragédie, ces proportions sont nettement plus importantes, surtout chez Corneille et Mairet.

Ces phrases remplissent deux fonctions différentes. La majorité d'entre elles sont des phrases *d'exposition* : récits d'évènements qui se passent en dehors de la scène et sont rapportés au spectateur ; retours en arrière, ou précisions historiques indispensables pour comprendre un incident ou le comportement d'un personnage. En effet, le théâtre classique était enserré dans un très grand nombre de règles. L'action devait se dérouler en moins de 24 heures, on ne devait pas voir sur scène de violences – et encore moins de sang – les personnages devaient se comporter avec bienséance, etc. L'auteur ne pouvait donc pas montrer un combat, un duel ou un meurtre mais les faire raconter par un personnage. On trouve également des phrases longues dans la bouche de quelques personnages – comme les rois, les empereurs, les grands capitaines – lorsqu'ils parlent aux autres dans le cadre de leurs fonctions. Lorsque de tels personnages parlent beaucoup – par exemple : *Cinna* de Corneille ou *Alexandre* de Racine – la moyenne s'en trouve nettement augmentée (Figure 1 & 2 ci-dessus).

**Les phrases très longues** comportent 48 mots et plus (soit la moyenne augmentée de deux écarts-types). Elles ne sont que 99 – soit 4,8% des phrases – mais elles couvrent 18,1% du total du texte. Leur longueur moyenne est considérable : 62,7 mots. On y trouve quelques rappels historiques et, surtout, des phrases qui exposent la pensée d'un personnage clef, soit qu'il parle à un confident, soit qu'il se livre à un monologue (stances). Dans les pièces comiques, ces phrases exposent les ridicules d'un personnage ou d'un caractère. Voici la plus caractéristique de ces phrases très longues dans le *Menteur* (vers 332 à 334) :

*On s'introduit bien mieux à titre de vaillant :*
*Tout le secret ne gît qu'en un peu de grimace,*
*A mentir à propos, jurer de bonne grâce,*
*Etaler force mots qu'elles n'entendent pas,*
*Faire sonner Lamboy, Jean de Vert, et Galas,*
*Nommer quelques châteaux de qui les noms barbares*
*Plus ils blessent l'oreille, et plus leur semblent rares,*
*Avoir toujours en bouche angles, lignes, fossés,*
*Vedette, contrescarpe, et travaux avancés :*
*Sans ordre et sans raison, n'importe, on les étonne ;*
*On leur fait admirer les bayes qu'on leur donne,*
*Et tel, à la faveur d'un semblable débit,*
*Passe pour homme illustre, et se met en crédit.*

Chez Corneille, beaucoup de phrases très longues sont construites par empilement de brèves propositions faiblement coordonnées les unes aux autres et toutes placées sur le même plan. Dans ces vers, on voit que la plupart des ponctuations mineures sont placées en fin de vers ou

à l’hémistiche. Ces procédés rendent aussi les phrases longues plus supportables par l’auditeur. L'accumulation est l'un des ressorts comiques souvent employés par Corneille, procédé que l’on retrouve dans ses pièces présentées par Molière.

## 6. Conclusions

La statistique textuelle a besoin de normes de dépouillement des textes et de vastes corpus étiquetés. C’est parce que nous avons suivi les normes proposées par Muller, Bernet et Kylander que nous avons pu bénéficier des travaux de ces devanciers et rendre comparables nos résultats avec les leurs. C'est parce que chaque mot est étiqueté avec son entrée de dictionnaire et sa catégorie grammaticale que l'on a pu déterminer le caractère plus ou moins nominal ou verbal des phrases en fonction de leur longueur. De même, il faut de vastes corpus. La présence de plusieurs auteurs et d'un grand nombre d'œuvres permet seule de tirer des conclusions assez certaines comme la différence entre les phrases de comédies et de tragédies.

En ce qui concerne le théâtre classique, nous avons montré que le vers alexandrin fait peser de fortes contraintes sur les auteurs et les conduits à mouler la longueur et la structure de leurs phrases sur la longueur du vers et sa césure. Cette conclusion peut paraître banale mais, à notre connaissance, personne n’avait aperçu ce phénomène. Il en est de même pour l’existence de plusieurs groupes de phrases de longueurs et de vocabulaires différents suivant les fonctions qu'elles remplissent.

Au-delà du cas particulier du théâtre du XVIIe, on trouve dans tous les corpus étudiés – spécialement dans les corpus oraux ou ceux destinés à la communication orale -, quelques constantes générales liant le type de phrase et les objectifs de la communication. S'il s'agit d'interpeller quelqu'un, de lui adresser une demande ou de lui donner un ordre, la phrase est brève, avec une prédominance du groupe verbal. La conversation courante utilise également des phrases simples et un excédent de pronoms et de verbes. En revanche, la communication formalisée ou le récit mobilisent des phrases plus longues où le nom l'emporte sur le verbe. Enfin, lorsqu'une personne expose sa pensée ou ses opinions, la phrase peut être très longue et fort complexe.

L'étude de la ponctuation et de la structure des phrases apportent encore beaucoup d'autres renseignements stylistiques intéressants que la dimension limitée de cette communication ne permettait pas d'aborder.

## Références


Bernet C. (1983). *Le vocabulaire des tragédies de Racine (Analyse statistique)*. Genève-Paris: Slatkine-Champion.

Kylander B.-M. (1995). *Le vocabulaire de Molière*. Goteborg : Acta Universitatis, Gothoburgensis.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques*. *Cahier du CERAT n° 7*. Grenoble : CERAT-IEP.

Labbé C. & Labbé D. (1994). *Que mesure la spécificité du vocabulaire ?* Grenoble : CERAT. Décembre 1994 & juin 1997. Repris dans : *Lexicometrica*. 3-2001.

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

Labbé C., Labbé D. & Monière D. (2008). "Les styles discursifs des premiers ministres québécois de Jean Lesage à Jean Charest". *Revue canadienne de science politique*. 41-1, p 43-69.

Labbé D. (2009). *Si deux et deux sont quatre, Molière n'a pas écrit Dom Juan*. Paris : Max Milo.

Milly J. (1975). *La Phrase de Proust*. Paris : Larousse (Réédition Paris : Champion, 1983).

Milly J. (1986). *La longueur des phrases dans "Combray"*. Paris-Genève : Champion-Slatkine.

Molinié G. (1986). *Eléments de stylistique française*. Paris : PUF.

Monière D. & Labbé D. (2002). "Essai de stylistique quantitative. Duplessis, Bourassa et Lévesque". In Morin A. & Sébillot P. (eds). *VIe Journées Internationales d'Analyse des Données Textuelles (Saint-Malo 13-15 mars 2002*). Rennes: IRISA-INRIA, 2002, vol. 2, p 561-569.

Monière D., Labbé C. & Labbé D. (2005). "Les particularités d'un discours politique : les gouvernements minoritaires de Pierre Trudeau et de Paul Martin au Canada". *Corpus*, 4, p 79-104.

Monière D., Labbé C. & Labbé D. (2008). "Les styles discursifs des premiers ministres québécois de Jean Lesage à Jean Charest". *Revue canadienne de science politique*. 41:1, p.43-69.

Muller C. (1967). *Etude de statistique lexicale. Le vocabulaire du théâtre de Pierre Corneille*. Paris : Larousse.

Richaudeau F. (1988). *Ce que révèlent leurs phrases*. Paris : Retz.