Le graphonaute ou Molière retrouvé

Stephan Vonfelt

Université de Toulouse Le Mirail - Laboratoire « Lettres, Langages et Arts »
stephan.vonfelt@free.fr

ABSTRACT. *Did Corneille write the plays of Molière? The answers contradict each other, including those refering to the objectivity of figures. In this vein, our study bases on the distribution of characters composing a text. Between two works, the resulting distance renders the contribution of the author, but also the gender, the form and the chronology. The measurement does not incite to melt Molière into Corneille and highlights the variety of his work, probably influenced by several sources.*

KEYWORDS : *author attribution - classification - rhythm - statistics - stylistics – stylometry – text topology*

RESUME. *Corneille a-t-il écrit les pièces de Molière ? Les réponses se contredisent, y compris celles qui se réfèrent à l'objectivité des chiffres. Dans cette veine, notre étude se fonde sur la répartition des caractères composant un texte. Entre deux œuvres, la distance résultante traduit la contribution de l'auteur, mais aussi du genre, de la forme et de la chronologie. Les mesures n'incitent pas à fondre Molière dans Corneille et mettent en lumière la variété de son œuvre, probablement influencée par diverses sources.*

MOTS-CLES : *attribution d'auteur - classification - rythme - statistique - stylistique – stylométrie – topologie textuelle*

1. Introduction

En 1919, le poète Louÿs attribue l'œuvre de Molière à Corneille[1]. Si la thèse est reprise par certains romanciers, elle prend une dimension nouvelle lorsque Labbé la confronte aux statistiques textuelles[2]. Fondée sur la similarité d'un vocabulaire lemmatisé, son étude confirme l'intuition du poète et provoque plusieurs réactions. Brunet confirme les mesures mais conteste leur interprétation[3]. Par ailleurs, Viprey réfute cette conjecture à partir des collocations de mots fréquents[4]. Enfin, Beaudouin & Yvon trouvent dans les séquences de syllabes une proximité

[1] Louÿs P. (1919).
[2] Labbé C. & Labbé D. (2001).
[3] Brunet E. (2004).
[4] Viprey J.M. (2004). A la suite de cette publication, Labbé effectue un calcul analogue qui conforte sa position initiale.

manifeste entre deux pièces des écrivains, mais se gardent de conclure plus généralement[5]. Les divers points de vue sont confrontés à l'occasion d'une table-ronde des *JADT*[6]. Notre étude reprend le principe de la stylistique statistique et cherche à compléter le puzzle, en suivant un sentier différent.

Avant toute analyse se pose le choix des unités linguistiques. Le spectre est large, entre lemmes, catégories grammaticales, sons… Cette première étape franchie, les étiquettes sont-elles justes ? Ces variables de seconde main échappent difficilement à une certaine subjectivité. Cependant, un texte reste une succession de caractères ou graphèmes, d'où la référence à ces éléments tangibles, de surcroît abondants et sources de statistiques significatives. Parmi les travaux qui exploitent cette veine, citons Markov et ses processus aléatoires[7], Shannon et sa théorie de l'information[8], Brunet et son étude du vocabulaire français[9], Khmelev & Tweedie[10] puis Jardino[11] pour l'attribution d'auteur. Dans la même quête, notre thèse constate l'efficacité des caractères par rapport aux catégories grammaticales et aux concepts sémantiques[12]. La piste mérite donc d'être approfondie.

De ces unités, les statistiques textuelles retiennent traditionnellement les fréquences d'apparition : une oeuvre est vue comme un sac de billes dont on compte les éléments, les yeux fermés sur son agencement. La topologie du texte se réduit à une expression élémentaire, sa composition sans son organisation. Une première solution consiste à segmenter un texte en morceaux d'échelle médiane, puis à analyser les évolutions des fréquences. Mais le découpage est souvent délicat, a fortiori la comparaison de deux architectures différentes, comme le soulignent Longrée, Luong & Mellet[13]. Prolongeant la marche vers l'infiniment petit, nous adoptons une division élémentaire qui suit le fil ténu de chaque occurrence. A travers le temps de retour d'une unité, un rythme supposé fondateur est appréhendé.

2. Principes

2.1 Sources textuelles

Le corpus numérisé a été aimablement fourni par Dominique Labbé. Outre Corneille et Molière, il inclut Racine, témoin de la comparaison.

[5] Baudouin V. & Yvon F. (2004).
[6] Table-ronde (2004).
[7] Markov A. (1913).
[8] Shannon C.E. (1951).
[9] Brunet E. (1981).
[10] Khmelev D. & Tweedie F.J. (2001).
[11] Jardino M. (2006).
[12] Vonfelt S. (2008), chapitre 8.
[13] Longrée D, Luong X., Mellet S. (2004).

L'œuvre de Corneille est la plus ancienne. Versifiée et principalement tragique[14], elle rassemble dans l'ordre chronologique : *Mélite, Clitandre, La Veuve, La Galerie du Palais, La Suivante, La Place Royale, Médée, L'illusion comique, Le Cid, Horace, Cinna, Polyeucte, La mort de Pompée, Le Menteur, Rodogune, La Suite du Menteur, Théodore, Héraclius, Andromède, Don Sanche d'Aragon, Nicomède, Pertharite, Œdipe, La Toison d'or, Sertorius, Sophonisbe, Othon, Agésilas, Attila, Tite et Bérénice, Psyché[15], Pulchérie, Suréna.*

Celle de Molière est résolument comique[16] et partagée entre le vers et la prose. Elle comprend *La Jalousie du Barbouillé, Le Médecin volant, L'Etourdi, Le Dépit amoureux, Les Précieuses ridicules, Sganarelle, Dom Garcie de Navarre, L'Ecole des maris, Les Fâcheux, L'Ecole des femmes, La Critique de l'Ecole des femmes, L'Impromptu de Versailles, Le Mariage forcé, La Princesse d'Elide, Le Tartuffe, Dom Juan, L'Amour médecin, Le Misanthrope, Le Médecin malgré lui, Mélicerte, Pastorale comique, Le Sicilien, Amphitryon, George Dandin, L'Avare, Monsieur de Pourceaugnac, Les Amants magnifiques, Le Bourgeois gentilhomme, Les Fourberies de Scapin, La Comtesse d'Escarbagnas, Les Femmes savantes, Le Malade imaginaire.*

L'œuvre de Racine est la plus tardive. Ecrite en vers et essentiellement tragique[17], elle se compose de *La Thébaïde, Alexandre le Grand, Andromaque, Les Plaideurs, Britannicus, Bérénice, Bajazet, Mithridate, Iphigénie, Phèdre, Esther, Athalie.*

2.2 Traitements linguistiques

Par rapport à la version originale de la base *Frantext*, Labbé a réalisé quelques corrections et normalisations orthographiques sans incidence statistique.

Les indications de scènes et de personnages, isolées par des balises, sont passées sous silence pour concentrer l'analyse sur le texte dramatique. Afin de s'affranchir d'une mise en page particulière, les retours à la ligne sont éliminés : chaque texte est ramené à sa plus simple expression, une suite de caractères.

L'intégralité des unités est lue et mémorisée : les espaces, la ponctuation, les lettres en distinguant minuscules, majuscules et caractères accentués, ainsi que les apostrophes et les traits d'union.

2.3 Mesure statistique

2.3.1 Caractérisation d'un texte

En mathématiques, la topologie est souvent présentée comme « une géométrie de la feuille de caoutchouc ». Notre approche se restreint à la géométrie « rigide » d'Euclide, où deux figures sont équivalentes lorsque les distances internes sont conservées. Si l'on isole par la pensée une unité

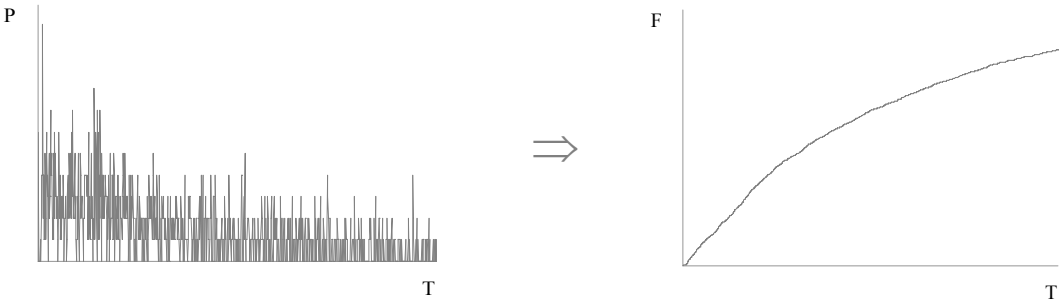
[14] Cependant, Corneille a écrit quelques comédies : *Mélicerte, La Veuve, La Galerie du Palais, La Suivante, La Place Royale, L'illusion comique, Le Menteur, La Suite de Menteur*.
[15] La plupart des vers est de Corneille, alors que scénario est attribué à Molière.
[16] Sauf *Dom Garcie de Navarre* et *Mélicerte*.
[17] A l'exception *des Plaideurs*.

(par exemple un « a »), la structure de ses apparitions dans le texte est donc définie par la succession des temps de retour T_j.

Dans le détail, l'information fournie par ces temps de retour peut se scinder en deux : la valeur moyenne m définit la composition ou le « thème »[18], tandis que les arythmies $T_j/m-1$ traduisent localement l'organisation ou le « style ». Sur la figure, la première séquence est symétrique et sert de référence, la seconde a le même thème et un style différent, tandis que la troisième a un thème différent et le même style.

Dans les statistiques, l'ensemble des valeurs de T est prise en compte. Les corrélations entre des mesures successives sont généralement négligeables[19], si bien que ces variables peuvent être considérées comme indépendantes. Un texte se caractérise alors par la distribution P de chaque unité. Pratiquement, cette courbe est chaotique, d'où le recours à la répartition F qui l'intègre, soit $F(t) = P(T \leq t)$.

Pour une structure symétrique, P se réduit à une raie et F à une marche, définies par la valeur unique du temps de retour. Ailleurs, ces fonctions complètent l'information primaire par la dispersion des mesures.



Cette méthode ignore les interactions spatiales entre les unités. D'un point de vue théorique, il suffit de généraliser ce qui précède et de substituer aux temps de retour d'une unité les temps de transition entre deux unités. En pratique, le coût d'une telle opération est élevé et multiplie la complexité du procédé par le nombre d'unités[20]. Si la voie n'est pas poursuivie ici, la question reste ouverte.

---

[18] Cette moyenne est l'inverse de la fréquence mesurée traditionnellement.

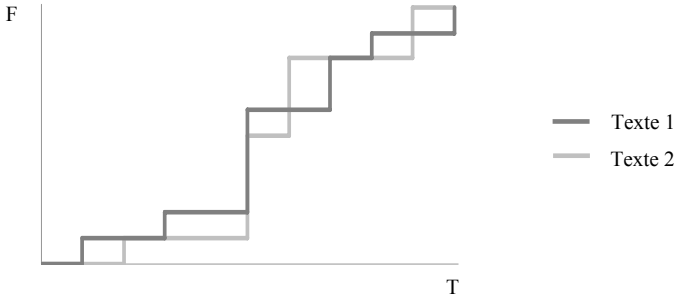
[19] Sur le corpus de notre thèse, les corrélations sont nulles dans le cas des lettres ; légèrement perceptibles pour les espaces et la ponctuation, elles confondent les œuvres et sont donc inadaptées à notre problématique : Vonfelt S. (2008), chapitre7.

[20] Soit un facteur de 79, en incluant espaces, ponctuation, lettres, apostrophes et traits d'union.

2.3.2 Comparaison de deux textes

La mesure porte dans un premier temps sur une unité i : les écarts des répartitions sont moyennés en suivant les valeurs discrètes et croissantes des temps de retour T_{ij}, et en pondérant par leurs effectifs n_{ij} : $D_{12i} = (\Sigma\ (n_{1ij}+n_{2ij})\ (F_1(T_{ij})-F_2(T_{ij}))^2 / (n_{1i}+n_{2i}))^{1/2}$.

Ces écarts locaux sont intégrés sur l'ensemble des unités pour obtenir une distance globale entre deux textes : $D_{12} = (\Sigma\Sigma\ (n_{1ij}+n_{2ij})\ (F_1(T_{ij})-F_2(T_{ij}))^2 / (n_1+n_2))^{1/2}$.

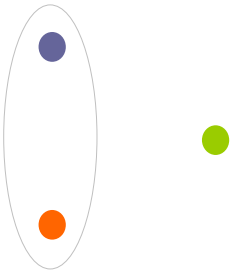


Comprise entre 0 et 1, $D_{12} = <(F_1-F_2)^2>^{1/2}$ est bien une distance : la séparation et la symétrie sont triviales, tandis que l'inégalité triangulaire résulte de celle de Cauchy-Schwarz.

Cette mesure fonde d'ailleurs le test d'Anderson, qui estime la probabilité que deux répartitions empiriques naissent de la même loi théorique[21]. Mais la méthode ne s'applique qu'à des variables continues, d'où notre renoncement à l'inférence pour nous tenir prudemment à la description.

3. Mesures

3.1 Cartes d'ensemble



Une vue générale est donnée par les distances mutuelles entre les œuvres, soit Corneille-Racine=2.46 %, Molière-Racine=2.70 % et Corneille-Molière=2.76 %.

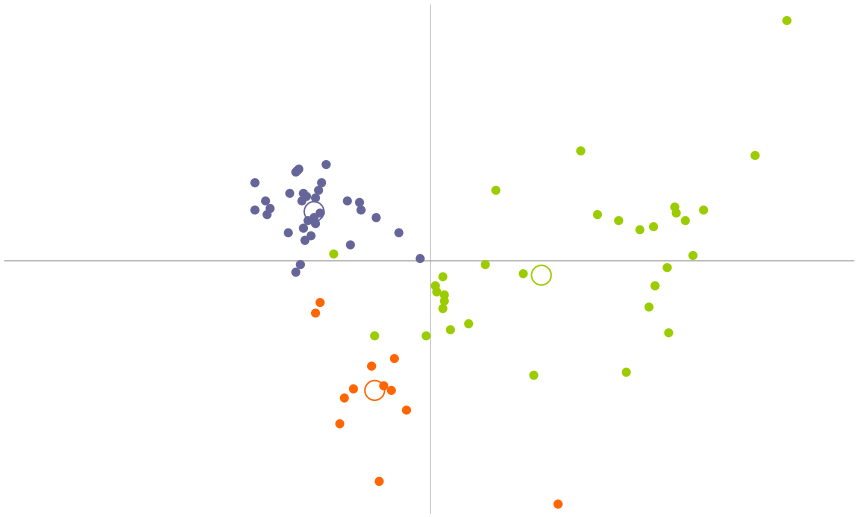
Sur le schéma, la triade formée de Corneille en bleu, Molière en vert et Racine en rouge est équilibrée. Précisément, Corneille et Racine sont les plus proches, naturellement unis par le genre tragique et la forme versifiée. Molière prend ses distances avec ce couple primordial et affiche de surcroît moins d'affinité avec Corneille que Racine : le mariage pressenti par Louÿs se présente sous de sombres auspices.

[21] Anderson T.W. (1962).

Une vue plus détaillée dessine les positions des pièces. Plongées dans les bras multiples d'une géométrie complexe, elles sont projetées vers leur plan principal en diagonalisant la matrice des distances mutuelles[22]. Susceptible de trahir ponctuellement, cette carte traduit en bonne justice les masses en instance.

Les centres d'inertie des trois nuages font retrouver le schéma précédent, avec la même convention de couleurs. Mais la distribution des individus apporte une information supplémentaire : Corneille et Racine administrent sagement des royaumes soudés, tandis qu'un Molière aventureux étend son empire à des marches lointaines.

L'axe horizontal reflète mécaniquement les contributions du genre et de la forme : à gauche, les tragédies en vers, à droite les comédies en prose. Plus diffuse, l'influence de la chronologie semble transparaître sur l'axe vertical, suivant la progression de Corneille à Racine.



3.2 Profils des œuvres

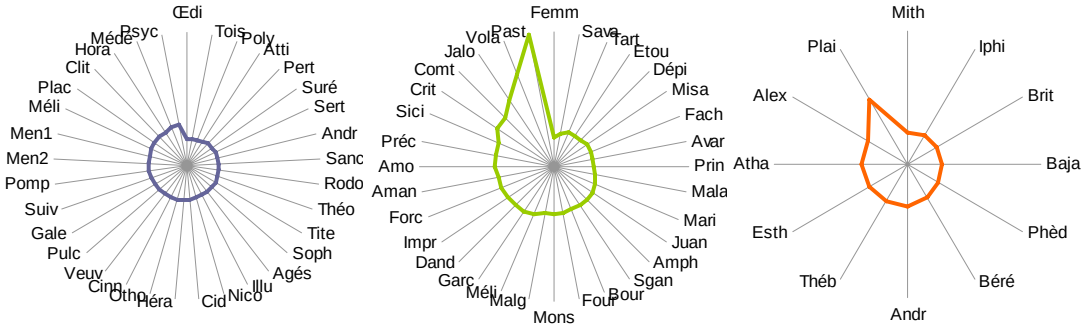
L'attention se focalise ici sur un auteur, précisément sur les distances entre une pièce et l'intégralité de l'œuvre. Représentées par valeurs croissantes sur le cadran, la médiane est lue à six heures, tandis que la plage est délimitée par le saut de la dernière heure.

L'orbite de Corneille reste proche de son foyer, sans grande excentricité. Les tragédies de la maturité (*Œdipe*[23], *La Toison d'or*, *Attila*, *Pertharite* et *Suréna*) sont les plus caractéristiques, à l'inverse de pièces comiques (*Le Menteur* et sa suite, *Mélite*, *La Place Royale*) ou précoces (*Clitandre*, *Horace*, *Médée*). Mais on notera surtout l'atypie de *Psyché*, tragi-comédie écrite avec la main de Molière.

[22] La dimension de cet espace est égale au nombre de pièces diminué de 1. La théorie est donnée dans Benzécri J.P. (1973).
[23] Un temps découragé, Corneille reprend la plume avec cette pièce.

Large et cornue, la trajectoire de Molière est d'une autre nature. Les pièces typiques (*L'Ecole des femmes*[24], *Les Femmes savantes, Le Tartuffe*) sont versifiées, alors que la prose est excentrée (*La Critique de l'Ecole des femmes*, *La Comtesse d'Escarbagnas*, a fortiori les essais de la jeunesse : *La Jalousie du Barbouillé*, *Le médecin volant*). Enfin, l'élément le moins congru reste la burlesque *Pastorale comique.*

Le chemin de Racine se situe entre ces deux auteurs, tant par l'éloignement médian que par l'intervalle de variation. L'unique comédie (*Les Plaideurs*) ressort visiblement, tandis que le reste est homogène. Parmi les pièces caractéristiques, *Mithridate, Iphigénie, Bajazet* et le chef d'œuvre *Phèdre* naissent après la joute théâtrale contre Corneille[25].



3.3 Attributions d'auteurs

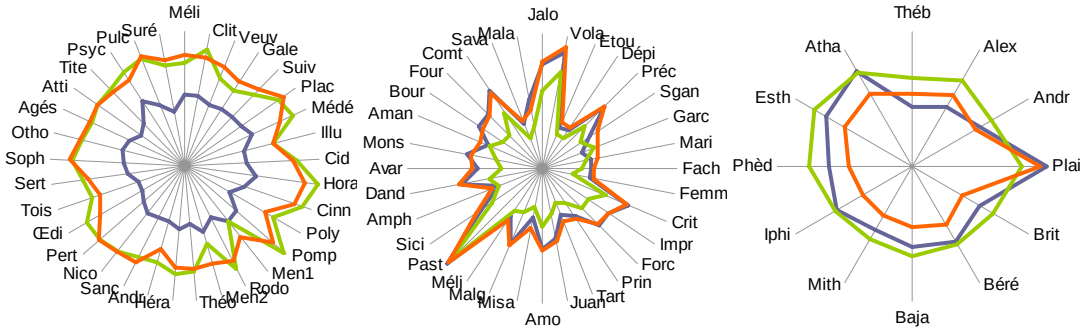
L'attribution d'auteur est un sujet complexe, qui suppose que les écrits d'une personne sont suffisamment homogènes pour être distingués de ceux d'une autre. Sans entrer dans ce débat général, examinons si les trois œuvres du corpus possèdent cette propriété.

Une pièce est attribuée à l'œuvre la plus proche, comparant une distance interne à deux externes. Cette distance interne est légèrement supérieure à celle de la section précédente[26] : pour que le jeu soit équitable, la pièce en question est boutée hors de l'œuvre qui la contient.

Les pièces de Corneille sont affectées sans hésitation, à l'exception du *Menteur*, associé de justesse à Molière : la comédie semble avoir marqué ce dernier[27], d'où le rapprochement effectué. L'opération est plus délicate pour Molière, et échoue avec *Dom Garcie de Navarre*, unique pièce sérieuse du comédien qui répond au *Don Sanche d'Aragon* : personne ne s'étonnera de son attribution à Corneille. La configuration de Racine est plus complexe : les œuvres de la jeunesse (*La Thébaïde*, *Alexandre Le Grand*) restent influencées par le modèle cornélien, tandis que la seule pièce comique (*Les Plaideurs*) est donnée à Molière.

[24] La pièce suit le mariage de Molière et voit l'écrivain trouver sa marque de fabrique.
[25] Avec *Bérénice*, Racine l'emporte sur *Tite et Bérénice* de Corneille.
[26] Les cadrans de la section 3.3 sont tracés à pleine échelle, ce qui facilite la lecture locale mais interdit les comparaisons globales, notamment avec la section 3.2.
[27] « Je dois beaucoup au *Menteur* » : propos rapporté par Boileau et cité dans François de Neufchâteau N. (1819), p. 149.

Finalement, l'expérience réussit dans 94 % des cas, score fort honorable qui laisse peu de doute sur l'authenticité de la tripartition.



4. Conclusion

La topologie textuelle est traduite par la répartition de chacune des unités. On complète ainsi une information traditionnellement limitée à la composition du texte, en négligeant cependant les corrélations temporelles au sein d'une même unité ou les interactions spatiales entre différentes unités. La distance intertextuelle qui en résulte reflète correctement l'auteur, mais aussi le genre, la forme et la chronologie.

Les mesures, loin de fondre Molière dans Corneille, mettent en lumière sa singularité et la variété de ses pièces. Deux explications viennent d'abord à l'esprit : l'écrivain a plusieurs facettes, ou il prête son nom aux plumes de l'ombre. Plus intéressante nous semble une piste exprimée du vivant du comédien, sous le ton du reproche : pressé par les demandes du roi[28], il tisserait avec habileté des éléments rapportés. Pour reprendre les mots de Donneau de Visé, son œuvre serait une vaste « rhapsodie ». A tout le moins, le théâtre de Molière semble charrier les eaux de diverses influences, françaises mais aussi italiennes et espagnoles[29]. Le jeu doit probablement être élargi avant de lever le rideau.

Références

Anderson T.W. (1962), « On the distribution of the two-sample Cramer-Von Mises criterion », *The Annals of Mathematical Statistics*, Beachwood, Institute of Mathematical Statistics, vol. 33, n° 3, p. 1148-1159.

Association Cornélienne de France, *L'affaire Corneille-Molière*, www.corneille-moliere.org.

Baudouin V. & Yvon F. (2004), « Contribution de la métrique à la stylométrie », *Actes des Journées internationales d'Analyse statistique des Données Textuelles,* Louvain, Presses Universitaires de Louvain.

[28] La contrainte est mise en scène dans *L'Impromptu de Versailles*.
[29] Forestier G. (2003).

Benzécri J.P. (1973), *L'analyse des Données*, Paris, Dunod.

Brunet E. (1981), *Le vocabulaire français de 1789 à nos jours*, Paris, Champion.

Brunet E. (2004), « Où l'on mesure la distance entre deux textes », *Texto !*, www.revue-texto.net, rubrique « Dits et inédits ».

Forestier G. (2003), « D'un vrai canular à une fausse découverte scientifique - à propos des travaux de Dominique et Cyril Labbé », *Centre de Recherche sur l'Histoire du Théâtre*, www.crht.org, rubrique « Ressources / Dossiers ».

François de Neufchâteau N. (1819), *L'esprit du grand Corneille*, Paris, Didot.

Jardino M. (2006), « Identification des auteurs de textes courts avec des n-grammes de caractères », *Actes des Journées internationales d'Analyse statistique des Données Textuelles*, Besançon, Presses Universitaires de Franche-Comté.

Khmelev D. & Tweedie F.J. (2001), « Using Markov Chains for Identification of Writers », *Literary and Linguistics Computing*, Oxford, Oxford University Press, vol. 16, n° 4, p. 299-307.

Labbé C. & Labbé D. (2001), « Intertextual Distance and Authorship Attribution : Corneille and Molière », *Journal of Quantitative Linguistics*, London, Routledge, vol. 8, n° 3, p. 213-231.

Longrée D., Luong X., Mellet S. (2004), « Temps verbaux, axe syntagmatique, topologie textuelle : analyse d'un corpus lemmatisé », *Actes des Journées internationales d'Analyse statistique des Données Textuelles,* Louvain, Presses Universitaires de Louvain.

Louÿs P. (1919), « L'Imposteur de Corneille et le Tartuffe de Molière », *Comoedia*, Paris.

Markov A. (1913), « Un exemple de recherche statistique sur le texte d'Eugène Onéguine illustrant la liaison des épreuves en chaînes », *Bulletin de l'Académie Impériale des Sciences*, vol. 7, Saint-Pétersbourg, p. 153-162.

Shannon C.E. (1951), « Prediction and Entropy of Printed English », *Bell Systems Technical Journal*, Hoboken, Wiley, vol. 30, p. 50-64.

Table-ronde (2004), « Corneille et Molière », *Actes des Journées internationales d'Analyse statistique des Données Textuelles*, Louvain, Presses Universitaires de Louvain.

Viprey J.M. (2004), « Analyse séquencée de la micro-distribution lexicale », *Actes des Journées internationales d'Analyse statistique des Données Textuelles*, Louvain, Presses Universitaires de Louvain.

Vonfelt S. (2008), *La musique des lettres : variations sur Yourcenar, Tournier et Le Clézio,* Université de Toulouse, thèse.