

RÉPONSES À DOMINIQUE LABBÉ
SUR NOTRE ARTICLE « LE GRAPHONAUTE OU MOLIÈRE RETROUVÉ »
- STEPHAN VONFELT -

DÉCEMBRE 2011.....	2
FÉVRIER 2012.....	13

DÉCEMBRE 2011

Dans une lettre ouverte¹, M. Labbé s'émeut de notre article paru dans la revue *Lexicometrica*, à l'occasion d'un numéro consacré à la topographie et à la topologie textuelles².

Que notre lecteur se rassure : nulle flèche ne le « vise directement », nulle animosité ne se nourrit de sa personne. Ainsi disposée, notre parole est sereine.

Remercions à nouveau ce chercheur, qui a nous transmis le corpus établi par l'*Institut National de la Langue Française*. Les informations qu'il a élaborées, les lemmes distingués par leurs catégories grammaticales, n'ont d'ailleurs pas servi à notre étude, fondée sur les caractères du texte.

Discutons d'abord l'axe central – notre distance ne mesurerait que les effectifs – avant de répondre point par point.

1 DISTANCE ET EFFECTIFS³

1.1 CENTRALITÉ

M. Labbé remarque que la centralité est liée aux effectifs. Faut-il s'en étonner ? Distance entre un texte et le corpus qui le contient, cette grandeur décroît fatalement avec la taille de l'échantillon, et s'annule lorsque les deux témoins se confondent. Le phénomène est indépendant de la distance

¹ Labbé, 2011.

² Vonfelt, 2009.

³ Les valeurs sont données en annexe.

employée.

1.2 ATTRIBUTION

Mineur en première instance, cet effet est éliminé lors du jugement de l'attribution. Ainsi pour affecter *Mélite*, une distance interne mesure l'écart entre la pièce et l'œuvre de Corneille, extirpée de la souche commune ; cette grandeur est alors comparée aux distances externes vers les œuvres de Molière et Racine.

Une distance impliquant deux effectifs, évaluons en première approche leur influence mutuelle par $N^{-1} = n_1^{-1} + n_2^{-1}$. Sur l'ensemble des expériences, la corrélation entre la distance et cette grandeur est de 61 %. On soutiendra difficilement que cette mesure ne compte que les effectifs. Si ces derniers font entendre leur voix et modèrent les écarts, faut-il pour autant s'alarmer ?

Revenons à la définition et considérons pour simplifier un caractère, par exemple « a » : la distance est la moyenne quadratique des écarts des fonctions de répartition sur les temps de retour de ce caractère dans chaque texte, notée $D = \|F_1 - F_2\|$ par la norme L_2 . Formellement, cette distance est bien indépendante de la taille des textes ; cependant, elle s'incline vers 0 si les répartitions particulières se rendent à la langue commune, sous le joug des effectifs.

Mathématiquement, comment traduire cette tendance ? Le cadre est ici aléatoire : dans le champ d'un texte, les caractères sont égrenés librement, balancés par l'inspiration de l'écrivain ; l'hypothèse « nulle » dispose qu'un même semeur œuvre sur les deux champs, en d'autres termes que les lois de

distribution sont identiques. Le théorème de Glivenko-Cantelli⁴ assure alors la convergence uniforme de F_1 et F_2 vers la fonction commune F , l'inégalité triangulaire entraînant D vers 0. Dans le cas continu, $N^{1/2} D$ tend vers une distribution indépendante de F^5 ; dans le cas discontinu qui nous concerne, la question est plus complexe et sort de cet exposé. On retiendra en première approche que $N^{-1/2}$ semble un bon marqueur de la décroissance de D .

Pour clore cette section, notons que la loi des grands nombres s'applique dans ces conditions au traditionnel comptage des unités linguistiques. La distance classique, fondée sur les écarts de fréquences, est donc également vouée à plier sous la charge des effectifs. La remarque vaut pour la distance utilisée par M. Labbé au niveau sémantique, mais les effectifs de chaque classe diffèrent évidemment : il y a plus de caractères que de mots dans un texte, et surtout moins de signes que de vocables.

2 VARIA

Nous répondons ici aux remarques en suivant leur chronologie.

La mesure porte sur les temps de retour d'un caractère. La marge entre le début du texte et la première occurrence est donc muette, comme entre la dernière occurrence et la fin du texte. Les hapax sont par définition ignorés.

Les distances mutuelles entre les trois auteurs sont effectivement voisines, et mettent précisément à mal un rapprochement entre Corneille et Molière.

⁴ Glivenko-Cantelli, 1933.

⁵ Anderson, 1962.

Les espaces et la ponctuation sont intégrés dans les statistiques. La note 19 précise que sur le corpus de notre thèse, les temps de retour de ces caractères sont légèrement corrélés, et que cette information ne permet pas de discriminer les textes⁶.

A dessein, le graphe factoriel n'identifie pas les points : il ne s'agit pas de détailler illusoirement une carte gauchie par les projections, mais de cerner les ensembles. Les axes de la synthèse, combinaisons des contributions de chaque caractère, n'ont pas plus d'étiquette fallacieuse.

Si la classification automatique a ses vertus, elle sort de notre propos : loin de fondre des ensembles, nous affectons chaque pièce à des corpus prédéfinis, avant d'évaluer la partition. En d'autres termes, nous catégorisons.

Pour tester loyalement la méthode, nous ignorons candidement le genre, la forme, et les étiquettes culturelles : tragédies ou comédies, en vers ou en prose, sont mécaniquement alignées sur un auteur, au fil des caractères qui les composent.

Psyché a été initialement traitée d'un bloc ; dans le détail, les passages de Corneille et Molière sont identiquement donnés au tragédien, qui semble avoir imprimé sa marque sur cette œuvre. Écrite à cinq mains⁷, *La Comédie des Tuileries*, a priori écartée d'un corpus limité à trois auteurs, est logiquement attribuée à Corneille⁸.

In fine, l'attribution « échoue » sur 6 pièces : *Le menteur*, un passage de

⁶ Vonfelt, 2008, p. 232-233.

⁷ Outre Corneille : Boisrobert, Colletet, L'Estoile, Rotrou.

⁸ Les mesures complémentaires sont données dans l'annexe 3.4.

Psyché, Dom Garcie de Navarre, La Thébaïde, Alexandre le Grand, Les Plaideurs. Mais elle « réussit » sur 71 pièces, soit 92 % des cas : le taux paraît très honorable.

Encore faut-il s'entendre sur les prémisses de cette expérience, qui scrute les écrits sans sonder les âmes. Fût-ce à l'aide de statistiques, un texte ne saurait à coup sûr assigner son auteur. Il suffit d'évoquer la falsification ou l'imitation pour s'en convaincre⁹. Au plus, on espère des présomptions.

RÉFÉRENCES

ANDERSON T.W., 1962, « On the distribution of the two-sample Cramer-Von Mises criterion », *The annals of Mathematical Statistics*, vol. 33, n° 3, p. 1148-1159.

BRENNAN M. & GREENSTADT R., 2009, « Practical Attacks Against Authorship Recognition Techniques », *Innovative Applications of Artificial Intelligence Conference*, Pasadena.

GLIVENKO V. & CANTELLI F.P., 1933, « Sulla determinazione empirica della legge di probabilita », *Giornale dell'Istituto Italiano degli Attuari*, n° 4, p. 92-99 et 421-424.

LABBÉ D., 2011, « Lettre ouverte aux animateurs de la revue *Lexicometrica* », www.pacte.cnrs.fr.

VONFELT S., 2008, *La musique des lettres*, Université de Toulouse.

VONFELT S., 2009, « Le graphonaute ou Molière retrouvé », *Lexicometrica*.

⁹ Brennan & Greenstadt, 2009.

ANNEXES

0 CORPUS

	Corneille	Molière	Racine
1	<i>Mélite</i>	<i>La Jalousie du Barbouillé</i>	<i>La Thébaïde</i>
2	<i>Clitandre</i>	<i>Le Médecin volant</i>	<i>Alexandre le Grand</i>
3	<i>La Veuve</i>	<i>L'Étourdi</i>	<i>Andromaque</i>
4	<i>La Galerie du Palais</i>	<i>Le dépit amoureux</i>	<i>Les Plaideurs</i>
5	<i>La Suivante</i>	<i>Les Précieuses ridicules</i>	<i>Britannicus</i>
6	<i>La Place Royale</i>	<i>Sganarelle</i>	<i>Bérénice</i>
7	<i>Médée</i>	<i>Dom Garcie de Navarre</i>	<i>Bajazet</i>
8	<i>L'illusion comique</i>	<i>L'École des maris</i>	<i>Mithridate</i>
9	<i>Le Cid</i>	<i>Les Fâcheux</i>	<i>Iphigénie</i>
10	<i>Horace</i>	<i>L'École des femmes</i>	<i>Phèdre</i>
11	<i>Cinna</i>	<i>La Critique de l'École des femmes</i>	<i>Esther</i>
12	<i>Polyeucte</i>	<i>L'Impromptu de Versailles</i>	<i>Athalie</i>
13	<i>La mort de Pompée</i>	<i>Le Mariage forcé</i>	
14	<i>Le menteur</i>	<i>La Princesse d'Élide</i>	
15	<i>Rodogune</i>	<i>Le Tartuffe</i>	
16	<i>La Suite du menteur</i>	<i>Dom Juan</i>	
17	<i>Théodore</i>	<i>L'Amour médecin</i>	
18	<i>Héraclius</i>	<i>Le Misanthrope</i>	
19	<i>Andromède</i>	<i>Le Médecin malgré lui</i>	
20	<i>Don Sanche d'Aragon</i>	<i>Mélicerte</i>	
21	<i>Nicomède</i>	<i>Pastorale comique</i>	
22	<i>Pertharite</i>	<i>Le Sicilien ou L'Amour peintre</i>	
23	<i>Œdipe</i>	<i>Amphitryon</i>	
24	<i>La Toison d'or</i>	<i>George Dandin</i>	
25	<i>Sertorius</i>	<i>L'Avare</i>	
26	<i>Sophonisbe</i>	<i>Monsieur de Pourceaugnac</i>	
27	<i>Othon</i>	<i>Les Amants magnifiques</i>	
28	<i>Agésilas</i>	<i>Le Bourgeois gentilhomme</i>	
29	<i>Attila</i>	<i>Les Fourberies de Scapin</i>	
30	<i>Tite et Bérénice</i>	<i>La Comtesse d'Escarbagnas</i>	
31	<i>Psyché</i>	<i>Les Femmes savantes</i>	
32	<i>Pulchérie</i>	<i>Le Malade imaginaire</i>	
33	<i>Suréna</i>		

1 EFFECTIFS¹⁰

	Corneille	Molière	Racine
1	86 183	18 625	71 361
2	75 294	20 338	73 580
3	91 424	79 555	78 842
4	83 993	84 920	42 843
5	78 477	34 709	82 829
6	70 944	31 210	71 178
7	75 041	88 416	81 654
8	79 814	54 213	80 246
9	85 902	41 040	84 613
10	84 823	85 182	77 392
11	83 146	44 833	59 621
12	85 300	37 994	82 455
13	84 954	31 899	
14	85 285	58 580	
15	86 885	93 440	
16	90 726	88 872	
17	88 311	31 883	
18	89 134	88 048	
19	80 517	48 026	
20	86 910	28 678	
21	87 137	4 003	
22	87 060	27 763	
23	94 366	78 131	
24	104 620	56 088	
25	90 975	107 472	
26	85 689	61 935	
27	86 477	62 831	
28	92 997	89 045	
29	85 251	73 169	
30	84 078	29 392	
31	83 389	86 737	
32	83 609	106 142	
33	83 164		

¹⁰ Nombre de caractères, espaces compris.

2 CENTRALITÉ¹¹

	Corneille	Molière	Racine
1	2,33	4,70	2,49
2	2,35	5,90	2,79
3	2,16	2,18	2,48
4	2,21	2,18	4,46
5	2,25	3,54	1,94
6	2,33	2,75	2,32
7	2,44	3,20	2,05
8	1,99	2,62	1,91
9	2,06	2,33	1,93
10	2,38	1,73	2,14
11	2,10	4,04	2,69
12	1,68	3,27	2,76
13	2,26	3,32	
14	2,30	2,37	
15	1,91	2,17	
16	2,27	2,69	
17	1,91	3,47	
18	2,07	2,32	
19	1,90	2,84	
20	1,90	3,03	
21	2,00	7,94	
22	1,78	3,60	
23	1,59	2,72	
24	1,61	3,22	
25	1,88	2,36	
26	1,98	2,81	
27	2,09	3,36	
28	1,98	2,75	
29	1,76	2,78	
30	1,95	4,04	
31	2,52	1,93	
32	2,18	2,47	
33	1,82		

¹¹ Les distances sont multipliées par 100.

3 ATTRIBUTION¹²

3.1 CORNEILLE

	Corneille	Molière	Racine
1	2,40	3,47	3,77
2	2,41	4,01	3,76
3	2,22	3,17	3,62
4	2,27	3,03	3,40
5	2,31	3,36	3,63
6	2,39	3,91	4,06
7	2,50	4,05	3,54
8	2,04	3,07	3,09
9	2,12	3,79	3,68
10	2,45	4,55	4,12
11	2,16	4,18	3,93
12	1,73	3,45	3,12
13	2,33	4,45	4,00
14	2,37	2,36	3,10
15	1,97	3,87	3,63
16	2,34	2,72	3,45
17	1,96	3,56	3,50
18	2,14	3,67	3,46
19	1,95	3,41	2,93
20	1,95	3,44	3,68
21	2,06	3,65	3,73
22	1,84	3,81	3,85
23	1,65	3,83	3,32
24	1,67	3,30	3,06
25	1,94	3,49	3,32
26	2,04	3,82	3,90
27	2,15	3,56	3,63
28	2,04	3,43	3,52
29	1,81	3,64	3,59
30	2,01	3,59	3,49
31	2,59	3,79	3,49
32	2,24	3,87	3,99
33	1,87	3,45	3,64

¹² Les distances sont multipliées par 100.

3.2 MOLIÈRE

	Corneille	Molière	Racine
1	6,34	4,74	6,46
2	7,21	5,95	7,53
3	2,75	2,28	3,03
4	2,76	2,28	2,94
5	5,29	3,60	5,36
6	3,54	2,79	4,06
7	2,57	3,35	3,59
8	3,00	2,68	3,42
9	2,94	2,38	3,13
10	2,94	1,81	3,16
11	5,65	4,13	5,55
12	4,79	3,34	4,83
13	4,90	3,38	4,72
14	3,51	2,44	3,71
15	3,05	2,28	3,47
16	4,29	2,80	4,50
17	4,87	3,52	4,94
18	3,03	2,42	3,66
19	4,95	2,91	5,03
20	3,40	3,07	3,76
21	8,15	7,95	8,02
22	4,42	3,64	4,81
23	2,85	2,84	3,28
24	4,94	3,31	5,08
25	3,89	2,48	4,20
26	4,66	2,90	4,37
27	3,62	3,46	4,06
28	4,67	2,88	4,33
29	4,66	2,88	4,47
30	5,65	4,10	5,55
31	2,92	2,02	3,09
32	4,20	2,63	3,62

3.3 RACINE

	Corneille	Molière	Racine
1	2,21	3,27	2,70
2	2,54	3,66	3,01
3	2,87	3,39	2,71
4	4,97	4,07	4,74
5	2,90	3,47	2,14
6	3,20	3,30	2,51
7	2,98	3,34	2,26
8	2,72	3,06	2,12
9	3,22	3,25	2,11
10	3,09	3,81	2,32
11	3,65	4,15	2,87
12	4,06	4,01	3,06

3.4 ADDENDA

	Corneille	Molière	Racine
<i>Psyché-C</i>	2,98	4,20	3,94
<i>Psyché-M</i>	3,41	4,04	4,01
<i>Tuileries</i>	3,04	4,04	4,54

FÉVRIER 2012

Poursuivant un dialogue qui semble le passionner, M. Labbé a publié en décembre un post-scriptum nous qualifiant de « curieux personnage ». Nous accueillons avec bonheur de n'être pas coulé dans la masse. Nous ne prendrons cependant pas ce sentier glissant, peu conforme aux « règles du débat » réclamées, et préférons la voie sûre des arguments.

Recommandons d'abord à M. Labbé de lire correctement, sans prêter des intentions imaginaires : notre distance n'est pas un « indice », et n'a jamais prétendu s'abstraire des effectifs. Mieux : aux yeux d'une statistique descriptive, la taille est une caractéristique essentielle d'un texte, qu'il ne convient pas de voiler pudiquement. On voit donc mal l'objet d'une « dissimulation ».

Une distance non indicielle procédant par comparaison, les valeurs absolues sont insignifiantes, seules comptent les valeurs relatives. Avec un écart-type relatif de 0.33, les mesures sont aisément séparées¹³.

Nous rejoignons volontiers M. Labbé sur la proximité de *Psyché* et de *Dom Garcie de Navarre* avec Corneille, ainsi que sur celle du *Menteur* avec Molière. Quant aux autres pièces, nos mesures confirment les attributions traditionnelles. L'hypothèse de notre lecteur a été testée de surcroît, ébauchant de nouvelles frontières : soit X le sous-ensemble de Molière qui serait à attribuer à Corneille (*Dom Garcie de Navarre*, *L'Étourdi*, *Le dépit amoureux*, *L'École des maris*, *Les Fâcheux*, *L'École des femmes*, *La Princesse d'Élide*, *Le Tartuffe*, *Dom Juan*, *Le*

¹³ Sur les 77 distances, la moyenne est de $3.48 \cdot 10^{-2}$, l'écart-type de $1,13 \cdot 10^{-2}$.

Misanthrope, Mélicerte, L'Avare, Les Femmes savantes), M* le corpus de Molière amputé de X, et C celui de Corneille. A l'aune de notre distance, X est plus proche de M* que de C, invalidant ce partage inédit¹⁴. Sauf erreur, M. Labbé n'a pas fait cette simple évaluation, et nous l'invitons à compléter ses expériences.

Nous suggérons enfin à notre lecteur de consulter d'autres « bricolages » portant sur Yourcenar, Tournier, Le Clézio, Homère et Hésiode¹⁵. Il pourra se convaincre que les mesures sont droites et fermes.

¹⁴ $XC = 2,37 \cdot 10^{-2}$, $XM^* = 2,03 \cdot 10^{-2}$.

¹⁵ *La musique des lettres* (2008) et *Archéologie numérique de la poésie grecque* (2010), publiées à l'Université de Toulouse et sur graphorythmes.free.fr.