Mélanges offerts à Charles Muller pour son centième anniversaire (22 septembre 2009), textes réunis par Christian Delcourt et Marc Hug, Paris, CILF, 2009

Muller le lexicomaître ([1])

ÉTIENNE BRUNET

Le héros de ce jour est le père-fondateur et le maître de notre discipline, dont le nom de baptême est double : tantôt statistique lexicale, tantôt statistique linguistique. Si Pierre Guiraud (1954 et 1959), le parrain, penchait pour la première appellation, Muller a joué sur les deux noms : sa thèse sur Corneille en 1967 était lexicale, mais son premier manuel en 1968 était linguistique. Quand ce même manuel se dédouble en 1973, le premier volume reste linguistique, mais le second redevient lexical. Ce n'est pas de l'hésitation terminologique : lorsque l'exposé décrit la méthode générale, le champ s'élargit aux dimensions de la linguistique. Quand le champ se rétrécit aux résultats, la limite lexicale est précisée puisque la plupart des applications s'arrêtent au lexique, même si le grammairien Muller s'engage parfois dans les exemples grammaticaux et syntaxiques. Mais, dans une discipline naissante, la terminologie fluctue : les affluents qui débouchent dans la rivière peuvent en modifier le cours et même en changer le nom. Et Muller en prend acte dans la préface qu'il donne à la thèse de Pierre Lafon (1984) et où il prend position pour la lemmatisation mais sans prendre parti pour le nom de la discipline qui s'oriente « dans les voies d'une lexicologie quantitative, d'une statistique lexicale, d'une lexicométrie si l'on préfère. » Le développement des outils informatiques a privilégié alors le matériau brut, un texte limité à l'ensemble des graphies, dénué de codes grammaticaux et livré sans préparation au traitement automatique, documentaire et statistique. Et la lexicométrie s'est imposée pour signifier cette démarche.

Mais Muller peut aujourd'hui savourer sa revanche. Les progrès de l'automaticité, après avoir éliminé la grammaire dans un premier temps, l'ont réintroduite dans les traitements actuels. Les lemmatiseurs raffinent le produit textuel et permettent maintenant d'aborder directement la syntaxe, voire même la sémantique. Et le terme de lexicométrie s'en trouve contesté. André Salem propose de lui substituer la textométrie, Damon Mayaffre la logométrie, d'autres la stylométrie.

---

(1) On craint toujours de n'être pas le premier quand on risque une métaphore ou un calembour. « Le premier qui a comparé une femme à une rose était un génie, le second un imbécile » (cette formule de Nerval a souvent été répétée par… les imbéciles). On trouve certes des *géomaîtres* dans la littérature oulipienne et des *chronomaîtres* dans la publicité des marchands de montres. Mais jusqu'à présent le *lexicomaître* ne semble pas avoir de précédent.

Or ce flottement observé pour le nom de la discipline, s'observe aussi dans le détail de ses méthodes. Muller a engagé résolument la démarche du côté du calcul des probabilités. Avec une belle audace il a affronté les grandes lois qui gouvernent l'univers de la statistique : loi normale, loi binomiale ([2]), loi de Poisson et les tests qui combattent l'hypothèse nulle : khi2, écart réduit, coefficient de corrélation. Tout l'appareillage d'une statistique inférentielle a été livré aux mains expertes et à celles qui l'étaient moins. Un mouvement de reflux a suivi, qui a contesté que le domaine du langage puisse être celui du hasard et soutenu que l'hypothèse nulle était malvenue dans un univers surdéterminé où les mots, bien loin d'être sortis d'une urne aléatoire, étaient liés par une logique interne. Les batailles des années 80 ont cessé sans qu'il y ait vainqueur ou vaincu. Les partisans d'une statistique descriptive ont profité des ressources puissantes que l'analyse multidimensionnelle offrait à l'interprétation. De leur côté, les disciples de Muller ont certes reconnu que les écarts étaient plus nombreux et plus forts que ce qu'autoriserait le hasard (et Muller avait expliqué le phénomène de la spécialisation du langage, qui génère ce gauchissement du modèle), mais ils ont maintenu que, faute d'autre référence, le hasard servait à comparer les mesures et les écarts, comme le niveau de la mer sert à mesurer l'altitude. Et ils ont continué à promener leur altimètre dans les paysages textuels.

1. La distance Jaccard. Application à Racine

Dans un domaine cependant la voie tracée par Muller n'a guère été suivie. Il est vrai que la démarche paraissait hardie et, faute de moyens de calcul suffisants, le Maître en avait exposé le principe mais sans en proposer une application en grandeur réelle. Il s'agit du chapitre 6 de sa thèse qui est aussi le dernier chapitre de son manuel et qui porte le même nom : la connexion lexicale. Cette dénomination n'a pas résisté à une appellation plus proche de la tradition mathématique, qui fait appel à la distance. Il s'agit de la même notion, vue sous un angle opposé. La connexion lexicale qui lie deux textes est l'inverse de la distance intertextuelle qui les sépare. En réalité, la relation n'est pas aussi symétrique qu'elle paraît. Si la connexion est bien une relation de proximité, mesurée par l'étendue du vocabulaire commun à deux textes, la distance entre deux textes peut être envisagée de deux façons, selon que la mesure est établie à partir du texte A ou à partir du texte B. Et Muller avec raison distingue deux tests d'indépendance : celui de A vers B, qui est proportionnel au nombre de mots dans A qui ne sont pas dans B, et celui de B vers A, qui compte les mots de B absents dans A. Ces deux indices ne sont pas nécessairement liés l'un à l'autre : si l'un des textes est plus

(2) Le calcul hypergéométrique n'a pas été proposé d'emblée. Il a été généralisé lorsque la technologie a fourni une puissance de calcul suffisante.

riche que l'autre, la part privative de son vocabulaire sera plus étendue et son indépendance plus forte. Mais ils sont liés malencontreusement à la taille des textes, le plus long ayant plus de chances d'absorber la plus grande part du vocabulaire du plus petit et c'est la raison qui a poussé Muller à abandonner les rapports trop simples qui mesurent la part commune et les parts exclusives des deux vocabulaires.

1.1. En réalité il y avait moyen d'échapper à l'influence de l'étendue en combinant les deux indices d'indépendance. Observons en effet que pour le même couple de textes les deux indices évoluent en sens inverse et d'un même pas. Quand croît l'inégalité de taille, l'indépendance du petit texte tend vers 0 quand celle du gros tend vers 1. Ces deux mouvements s'annulent si on les additionne, la résultante tournant autour de 1, entre deux limites 0 et 2, ou mieux entre 0 et 1 si on calcule la moyenne des deux indices. Appliquons le calcul aux pièces de Racine (il s'agit des lemmes, non des graphies). Au croisement de deux textes, l'un en ligne, l'autre en colonne, le premier tableau fait état des vocables communs et le second des vocables exclusifs, qui sont dans un texte sans être dans l'autre. Ainsi les deux premiers de la liste, *La Thébaïde* et *Alexandre,* ont 968 vocables en partage et 558 et 581 respectivement en exclusivité. Rapportés à l'étendue du vocabulaire (indiquée dans la dernière colonne), ces effectifs donnent la mesure de l'indépendance (ou de la distance) de chaque élément à l'égard de l'autre (tableau 3). Comme les deux pièces ont sensiblement la même taille, on serait fondé à dire que la seconde est la plus indépendante des deux (0,375 contre 0,366) et qu'elle a sans doute un vocabulaire ou plus riche ou plus original.

ab (*part commune*)	*a - b* (*part privative de A*)	*b - a* (*part privative de B*)	*A* (*vocab.*)	*B* (*vocab.*)
968	558	581	1526	1549
indépendance de A =	558 / 1526	= 0,366		
indépendance de B =	581 / 1549	= 0,375		

Mais si l'on compare la première pièce à la dernière, on se trouve devant deux indices 0.331 et 0.511 dont l'écart peut être interprété de deux façons suivant qu'on l'attribue à la taille ou à la spécificité lexicale. Or l'effet mécanique de la taille est annulé si l'on fait la moyenne des deux indices, soit 0,421, valeur (multipliée par 1000) que l'on peut lire dans le tableau 4, au croisement de la *Thébaïde* et d'*Athalie* et qui rend compte de la distance mutuelle qui sépare les deux pièces.

	Théb	Alex	Andr	Plai	Brit	Béré	Baja	Mith	Iphi	Phèd	Esth	Atha	**voc.**
Théb	0	968	983	602	1008	943	1006	1022	1017	1040	966	1021	1526
Alex	0	0	1011	608	1057	987	1043	1068	1058	1059	986	1033	1549
Andr	0	0	0	633	1091	997	1074	1099	1122	1105	1013	1059	1595
Plai	0	0	0	0	659	611	665	644	658	677	651	696	1468
Brit	0	0	0	0	0	1111	1156	1161	1196	1192	1103	1150	1849
Béré	0	0	0	0	0	0	1073	1051	1077	1062	990	1051	1556
Baja	0	0	0	0	0	0	0	1143	1161	1159	1048	1133	1715
Mith	0	0	0	0	0	0	0	0	1168	1185	1088	1151	1728
Iphi	0	0	0	0	0	0	0	0	0	1237	1134	1189	1820
Phèd	0	0	0	0	0	0	0	0	0	0	1149	1242	2002
Esth	0	0	0	0	0	0	0	0	0	0	0	1256	1907
Atha	0	0	0	0	0	0	0	0	0	0	0	0	2089

Tableau 1. Effectifs des vocables communs

	Théb	Alex	Andr	Plai	Brit	Béré	Baja	Mith	Iphi	Phèd	Esth	Atha	**voc.**
Théb	0	558	543	924	518	583	520	504	509	486	560	505	1526
Alex	581	0	538	941	492	562	506	481	491	490	563	516	1549
Andr	612	584	0	962	504	598	521	496	473	490	582	536	1595
Plai	866	860	835	0	809	857	803	824	810	791	817	772	1468
Brit	841	792	758	1190	0	738	693	688	653	657	746	699	1849
Béré	613	569	559	945	445	0	483	505	479	494	566	505	1556
Baja	709	672	641	1050	559	642	0	572	554	556	667	582	1715
Mith	706	660	629	1084	567	677	585	0	560	543	640	577	1728
Iphi	803	762	698	1162	624	743	659	652	0	583	686	631	1820
Phèd	962	943	897	1325	810	940	843	817	765	0	853	760	2002
Esth	941	921	894	1256	804	917	859	819	773	758	0	651	1907
Atha	1068	1056	1030	1393	939	1038	956	938	900	847	833	0	2089

Tableau 2. Nombre de vocables privatifs

	Théb	Alex	Andr	Plai	Brit	Béré	Baja	Mith	Iphi	Phèd	Esth	Atha
Théb	0	0,366	0,356	0,606	0,339	0,382	0,341	0,330	0,334	0,318	0,367	0,331
Alex	0,375	0	0,347	0,607	0,318	0,363	0,327	0,311	0,317	0,316	0,363	0,333
Andr	0,384	0,366	0	0,603	0,316	0,375	0,327	0,311	0,297	0,307	0,365	0,336
Plai	0,590	0,586	0,569	0	0,551	0,584	0,547	0,561	0,552	0,539	0,557	0,526
Brit	0,455	0,428	0,410	0,644	0	0,399	0,375	0,372	0,353	0,355	0,403	0,378
Béré	0,394	0,366	0,359	0,607	0,286	0	0,310	0,325	0,308	0,317	0,364	0,325
Baja	0,413	0,392	0,374	0,612	0,326	0,374	0	0,334	0,323	0,324	0,389	0,339
Mith	0,409	0,382	0,364	0,627	0,328	0,392	0,339	0	0,324	0,314	0,370	0,334
Iphi	0,441	0,419	0,384	0,638	0,343	0,408	0,362	0,358	0	0,320	0,377	0,347
Phèd	0,481	0,471	0,448	0,662	0,405	0,470	0,421	0,408	0,382	0	0,426	0,380
Esth	0,493	0,483	0,469	0,659	0,422	0,481	0,450	0,429	0,405	0,397	0	0,341
Atha	0,511	0,506	0,493	0,667	0,449	0,497	0,458	0,449	0,431	0,405	0,399	0

Tableau 3. Indices d'indépendance a > b et b > a

	Théb	Alex	Andr	Plai	Brit	Béré	Baja	Mith	Iphi	Phèd	Esth	Atha
Théb	368	370	369	597	397	388	377	369	387	399	430	421
Alex	370	345	356	596	372	364	359	346	367	393	423	419
Andr	369	356	336	585	362	367	350	337	340	377	416	414
Plai	597	596	585	578	597	595	579	594	595	600	607	596
Brit	397	372	362	597	341	342	350	350	348	379	412	413
Béré	388	364	367	595	342	341	342	358	358	393	422	410
Baja	377	359	350	579	350	342	335	336	342	372	419	398
Mith	369	346	337	594	350	358	336	335	341	361	399	391
Iphi	387	367	340	595	348	358	342	341	339	351	391	388
Phèd	399	393	377	600	379	393	372	361	351	350	411	392
Esth	430	423	416	607	412	422	419	399	391	411	369	370
Atha	421	419	414	596	413	410	398	391	388	392	370	369

Tableau 4. Distances intertextuelles (*méthode Jaccard*)
(distance globale des textes deux à deux, multipliée par 1000)

Isolément, ces mesures laissent l'appréciation dubitative, faute de barème absolu. Car l'échelle des valeurs change selon les options de la lemmatisation. Mais rapprochées des valeurs lues sur la même ligne ou la même colonne, elles dessinent le profil d'un texte et sa place dans l'ensemble du corpus. Et dans l'exemple de la première pièce (de la première ligne), on voit à l'œil nu que de 368 à 421 la distance s'accroît avec le temps. Une courbe fondée sur ces distances montrerait à l'évidence cette évolution. Mais la dérive du temps n'est pas seule en cause : alors que toutes des valeurs lues dans le tableau se situent entre 300 et 500, la ligne consacrée aux *Plaideurs* montre des valeurs nettement supérieures, qui tournent autour de 600 et qui marquent l'exterritorialité d'une comédie isolée au milieu des tragédies. Qu'on n'invoque pas la taille, plus courte, pour expliquer cette différence, car si la taille agissait, elle irait dans le sens contraire. À côté du temps et du genre, le tableau 4 laisse deviner les rapprochements qui tiennent au sujet. Les pièces grecques reconnaissent leur parenté thématique (*Thébaïde*, *Alexandre* et *Andromaque* d'un côté, *Iphigénie* et *Phèdre* de l'autre). De même Rome est l'entremetteuse qui met en rapport *Bérénice* et *Britannicus,* comme la Bible est ce qui unit *Athalie* et *Esther*. Quand l'environnement est partagé, les indices — c'est-à-dire les distances — se réduisent. Pour mettre en évidence le jeu, conjugué ou contrarié, de ces diverses forces, chronologiques, génériques et thématiques, on a besoin d'une vue globale qui synthétise toutes les indications du tableau 4. C'est ce que propose l'analyse arborée dans la figure 5. Les observations précédentes y sont soulignées avec force, aussi bien la chaîne chronologique orientée de bas en haut, que l'exterritorialité générique des *Plaideurs*, isolés à l'extrême droite, et les regroupements qui s'ordonnent autour des sujets romains, grecs ou bibliques.

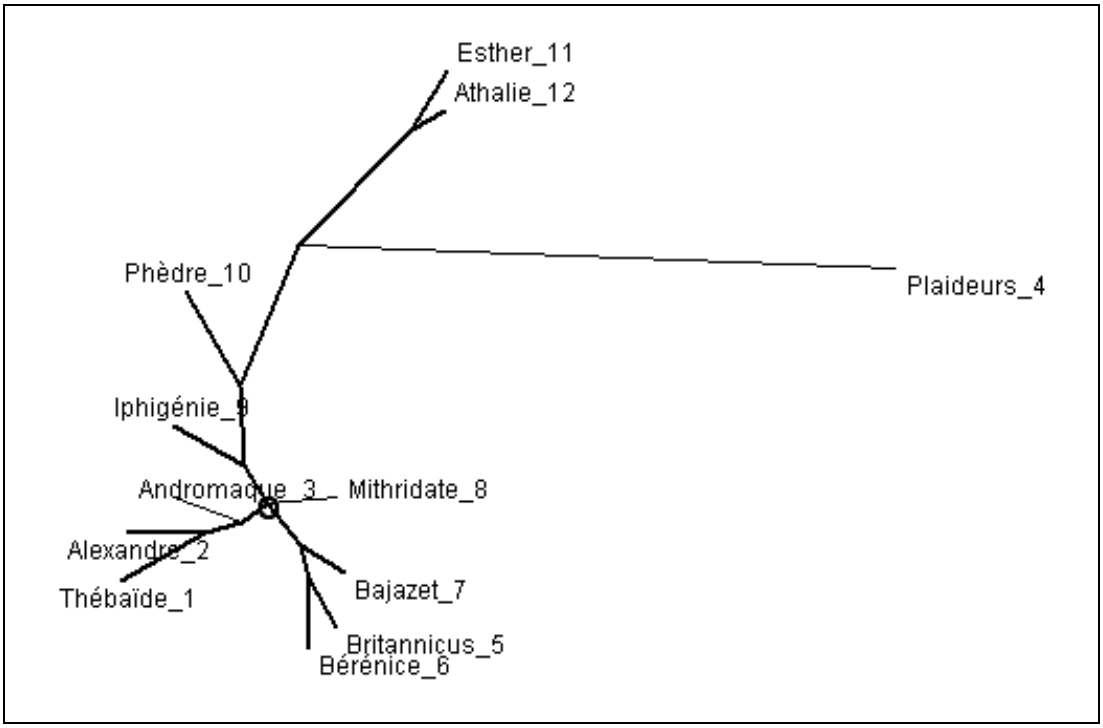


Figure 5. Analyse arborée rendant compte de l'œuvre de Racine
(*méthode Jaccard appliquée aux lemmes*)
(*le rang chronologique est ajouté au nom de la pièce*)

1.2. Muller s'en est tenu à la formule originale de la distance de Jaccard ([3]), sans essayer d'en corriger les défauts. Il recommande seulement « de ne tenir compte de ces indices que pour des textes de longueur égale ou au moins du même ordre » (Muller : 1968, p. 215). Mais il cite les travaux de Étienne Évrard (1966) qui sous le titre de « coefficient de Bernoulli » propose un calcul amélioré de l'indice Jaccard, fondé pareillement sur les éléments communs (*a*), et privatifs (*b* et *c*) mais aussi sur ceux qui sont absents dans les deux ensembles (*d*) :

$$r = ((a * d) - (b * c)) / \sqrt{((a+b)(c+d)(a+c)(b+d))}$$

En un siècle la formule de Jaccard proposée en 1908 a donné naissance à de multiples avatars dont certains ignorent le lien de parenté avec le pionnier. C'est le cas de la formule précédemment exposée, que nous pensions avoir inventée et qui se trouve, parmi une vingtaine d'autres, dont le coefficient de Bernoulli, dans l'inventaire dressé par F.B. Baulieu (1989).

(3) L'indice Jaccard $J(A,B)$ est le rapport entre l'intersection de deux ensembles A et B et l'union de ces deux ensembles. La distance Jaccard en est dérivée : $1 - J(A,B)$. Voir Jaccard (1901).

1.3. Même s'il avait connu toutes ces variantes, Muller les aurait peut-être récusées. Ses réticences supposées tiennent à la nature purement descriptive d'un indice brut qui laisse le jugement statistique en suspens. La démarche qu'il préfère est celle qui s'appuie sur un modèle, calcule un effectif théorique et apprécie les écarts en terme de probabilité. Les réserves tiennent aussi à la fragilité des quotients ou pourcentages quand les données sont de faible étendue. Elles tiennent enfin aux lacunes de l'indice Jaccard appliqué au vocabulaire. L'indice est en effet indifférent à la fréquence des mots. L'indice tient registre des espèces présentes ou absentes, ce qui peut se justifier si l'on fait un recensement de la flore, et qu'on veut circonscrire les limites d'extension de telle ou telle espèce. Cela peut s'appliquer dans les cartes de dialectologie, où l'on recherche les survivances et les témoignages, même uniques. Mais s'il s'agit d'apprécier la distance entre deux textes, l'approche de Jaccard est réductrice, incomplète et même trompeuse. D'une part même si aucun mot n'est exclu de droit, en fait les mots fréquents n'ont aucune place dans le calcul, puisqu'ils se trouvent dans tous les textes et sont automatiquement comptés au nombre des mots communs, si inégal que soit leur emploi dans les textes comparés. D'autre part même les mots de basse fréquence sont couchés sur un lit de Procuste, qui s'ajuste mal à la diversité de leur distribution : supposons en effet qu'un mot soit présent dans les deux textes considérés, mais avec des fréquences très inégales. S'il est vingt fois dans l'un et une fois dans l'autre, Jaccard le verse dans le lot des mots communs et la distance entre les deux textes s'en trouve diminuée. On incline à penser au contraire qu'une répartition aussi déséquilibrée est propre à accroître cette distance, au même titre que la distribution voisine 20 et zéro. Le calcul doit donc faire intervenir la fréquence et prendre en compte non pas seulement les vocables, mais aussi les occurrences.

Muller tente bien de bâtir avec les occurrences un indice *CN* qu'il met en parallèle avec celui des vocables *CV* :

$$CN = \textit{occurrences du vocabulaire commun / occurrences des deux textes} = N(ab) / N(a + b)$$

Mais il y renonce très vite car cet indice *CN* renforce les défauts de *CV*, dans les cas de vocables fortement thématiques. « On retiendra que les valeurs de V […] sont plus sûres que celles de N » ([4]).

(4) Muller : 1967, p. 171. Charles Muller cite le cas du mot *roi* qui a 71 occurrences dans *Pertharite,* une seule dans *Mélite* et aucune dans *Héraclius*. Dans le cas du couple *Pertharite-Mélite* le calcul de *CN* contribuerait à rapprocher les deux textes avec un poids de 72, alors que dans le couple *Pertharite-Héraclius* le même mot pèserait en faveur de l'éloignement avec un poids équivalent. Deux contributions opposées pour une situation quasiment identique.

2. La connexion lexicale de Muller. Application à Corneille

L'abandon de *CN* (et même de *CV*) est d'autant plus facile que Muller imagine un autre moyen de tenir compte de la fréquence dans le calcul de la distance. C'est même par cette méthode qu'il commence son exposé dans le dernier chapitre de son *Initiation.* Et s'il évoque en fin de chapitre les coefficients *CV* et *CN*, c'est à défaut de mieux, quand on n'a pas les moyens de mettre en œuvre le vrai calcul binomial. Rappelons que ces questions théoriques étaient débattues à la fin des années 60, à un moment où les universitaires, surtout les littéraires, n'avaient pas accès aux ordinateurs pour mettre en pratique la théorie Même la recherche d'Étienne Évrard, citée plus haut et publiée par Muller et Pottier en 1966, a été exécutée sur du matériel mécanographique, l'auteur invitant, pour de plus amples travaux, à se tourner vers l'ordinateur ([5]). Quarante ans plus tard, alors que les ordinateurs sont disponibles partout, on doit regretter qu'aucun logiciel n'ait été proposé pour livrer clés en mains le calcul de Muller.

2.1. La méthode, réputée complexe, est pourtant clairement établie dans l'*Initiation.* Nous y renvoyons le lecteur ainsi qu'à un article que nous avons consacré à la connexion lexicale chez Hugo (Brunet : 1988). Le but du présent article n'est pas d'en présenter un nouvel exposé, mais d'en expliciter les vertus et d'en proposer une implémentation dans un logiciel dédié à la statistique linguistique. Il faut pourtant en dire le principe, à partir d'un exemple réel, emprunté précisément à la thèse de Muller. En somme je vous propose de revenir 40 ans en arrière et d'assister à la soutenance de celui qui a présidé ou expertisé la plupart de nos thèses en la matière. Un centenaire à qui on retire quarante ans est encore un homme d'expérience qui approche de la soixantaine et son exposé est magistral (Nous le reproduisons tel quel).

Selon la démarche habituelle chez Muller, la méthode conduit d'abord à un modèle, puis à un relevé des faits dans le texte et enfin à un écart entre le modèle et l'observation. Prenons pour exemple les deux dernières pièces de Corneille, *Pulchérie* et *Suréna*. Leur taille est à peu près la même, soit 19235 et 19148 mots. Les probabilités p et q sont donc très voisines p = 0.5011 et q = 0.4989. On trouvera dans le tableau 7 les effectifs théoriques auxquels conduit la loi binomiale, sachant que les totaux pour chaque classe de fréquence sont les suivants :

fréquence 1 : 827, fréquence 2: 329, fréquence 3 : 206, fréquence 4 : 132, fréquence 5 : 92, fréquence 6 : 82, fréquence 7 : 50, fréquence 8 : 47, fréquence 9 : 45.

(5) « Pour une documentation portant sur quelques centaines d'idiomes, il faudra, me semble-t-il, penser à un traitement par ordinateur » (Évrard : 1966, p. 94). Ajoutons que l'auteur allait sans tarder se lancer dans la programmation et réaliser la chaîne des traitements informatiques qui ont fait le renom du LASLA.

Connaissant l'étendue respective des deux textes A et B, qui est N_a et N_b, on calculera aisément la probabilité pour qu'une occurrence prise au hasard soit dans A ou dans B ; nous appellerons conventionnellement la première p et la seconde q, pour retrouver des notations déjà employées :

$$p = \frac{N_a}{N_a + N_b} \qquad q = \frac{N_b}{N_a + N_b} \qquad p + q = 1 .$$

Il suffira ensuite d'appliquer les développements du binôme $(p + q)^f$ pour construire un modèle ; on appellera f la fréquence du vocable dans l'ensemble, et V_f l'effectif qui lui est associé ; f'_a et f'_b les sous-fréquences dans les deux textes, et V'_{fa}, V'_{fb} leurs effectifs.

La probabilité, pour un vocable de fréquence f, d'avoir les sous-fréquences 0, 1, 2... f dans l'un ou l'autre des textes est alors :

	Probabilité d'une sous-fréquence dans A					id. dans B				
f	0	1	2	3	4...	0	1	2	3	4...
1	q	p	0	0	0	p	q	0	0	0
2	q^2	$2pq$	p^2	0	0	p^2	$2pq$	q^2	0	0
3	q^3	$3pq^2$	$3p^2q$	p^3	0	p^3	$3p^2q$	$3pq^2$	q^3	0
4	q^4	$4pq^3$	$6p^2q^2$	$4p^3q$	p^4	p^4	$4p^3q$	$6p^2q^2$	$4pq^3$	q^4
etc.										

Figure 6. La connexion lexicale selon la loi binomiale
(Muller : 1968, p. 211)

On vérifiera que le total de la ligne 9 des tableaux réel et théorique est bien ce qu'il doit être : 45. Le tableau des écarts confirme la tendance des textes à s'arroger une part privative plus grande que ne le ferait un hasard impartial, ce que Muller appelle la spécialisation lexicale. Dans la zone centrale du tableau, celle où l'on partage, les écarts sont négatifs, tandis que sur les marges, là où règne l'exclusivité, les écarts sont plus souvent positifs. Bien entendu les écarts absolus doivent être convertis en KHI2, et comptabilisés dès que l'effectif théorique atteint ou dépasse la valeur 5. On voit que la dernière ligne qui s'arrête à la fréquence 9 n'épuise pas le calcul puisque pour cette classe l'effectif théorique de certaines cases dépasse encore 10. Il a donc fallu pousser plus loin la chaîne des calculs sans s'effrayer si les probabilités p et q ont des exposants terrifiants (en réalité nul besoin de recourir à l'exponentiation : on passe d'une classe à l'autre en ne recourant qu'à la multiplication). On a donc poursuivi le calcul jusqu'à la classe 50. Lorsque l'effectif théorique n'atteint pas le seuil, on procède au regroupement des effectifs trop minces. Reste à totaliser les KHI2 partiels et à les confronter aux degrés de liberté. Pour l'exemple ci-dessous, pour un ddl = 36, on obtient un KHI2 total de 93 et , lorsque le calcul s'étend jusqu'à la classe 50, le KHI2 s'établit à 360 pour un ddl de 48. Dans les deux cas, que le calcul soit limité ou étendu, la valeur

du KHI2 laisse une chance infinitésimale au hasard: les deux pièces puisent dans des zones distinctes du lexique. Et il en est ainsi de toutes les pièces de Corneille, confrontées deux à deux. Mais ce qui nous importe n'est pas de prouver la spécialisation lexicale, mais de la mesurer et de se servir de cette mesure pour établir une distance entre les textes.

probabilités p : 0,501133314227653 et q : 0,498866685772347

0,00									
414,44	412,56								
82,62	164,50	81,88							
25,93	77,42	77,07	25,58						
8,33	33,15	49,50	32,85	8,18					
2,91	14,47	28,81	28,68	14,28	2,84				
1,30	7,76	19,31	25,62	19,13	7,62	1,26			
0,40	2,77	8,26	13,70	13,64	8,15	2,70	0,38		
0,19	1,49	5,19	10,33	12,85	10,23	5,09	1,45	0,18	
0,09	0,80	3,20	7,43	11,10	11,05	7,33	3,13	0,78	0,09

Tableau théorique

0									
435	392								
86	144	99							
28	68	62	48						
8	39	37	31	17					
7	10	22	23	19	11				
1	9	19	19	18	10	6			
3	4	5	12	9	11	5	1		
2	2	4	12	10	9	4	0	4	
2	1	5	6	4	14	7	2	3	1

Tableau réel

0,00									
20,56	-20,56								
3,38	-20,50	17,12							
2,07	-9,42	-15,07	22,42						
-0,33	5,85	-12,50	-1,85	8,82					
4,09	-4,47	-6,81	-5,68	4,72	8,16				
-0,30	1,24	-0,31	-6,62	-1,13	2,38	4,74			
2,60	1,23	-3,26	-1,70	-4,64	2,85	2,30	0,62		
1,81	0,51	-1,19	1,67	-2,85	-1,23	-1,09	-1,45	3,82	
1,91	0,20	1,80	-1,43	-7,10	2,95	-0,33	-1,13	2,22	0,91

Tableau des écarts

0,00									
1,02	1,02								
0,14	2,55	3,58							
0,17	1,15	2,95	19,66						
0,01	1,03	3,16	0,10	9,53					
0,00	1,38	1,61	1,13	1,56	0,00				
0,00	0,20	0,00	1,71	0,07	0,74	0,00			
0,00	0,00	1,29	0,21	1,58	1,00	0,00	0,00		
0,00	0,00	0,27	0,27	0,63	0,15	0,23	0,00	0,00	
0,00	0,00	0,00	0,28	4,54	0,79	0,02	0,00	0,00	0,00

Tableau des CHI2

Tableaux 7a-d. Calcul de la connexion lexicale pour les basses fréquences

Dans l'application que nous avons faite de ce calcul à l'œuvre de Giraudoux et de Hugo, nous nous étions arrêté là, sans prolonger le calcul dans les classes de fréquence supérieures. Et pourtant la bible du lexicomètre indiquait clairement le chemin : « Ce calcul ne peut être poursuivi que jusqu'au point où les effectifs V_f deviennent trop faibles et où les effectifs théoriques dans les sous-ensembles deviennent inférieurs à l'unité. À partir de ce point, on pourra traiter les vocables non plus par classes de fréquence, mais isolément, et le modèle deviendra très simple : $f_a' = pf$ $f_b' = qf$ » (Muller : 1968, p. 212). Le prolongement du calcul est en effet un jeu d'enfant, chaque mot donnant lieu au calcul d'un KHI2 partiel qui s'ajoute à tous les autres, tandis que les degrés de liberté s'accroissent d'une unité. Pour le couple *Pulchérie-Suréna,* le résultat est un KHI2 de 250 pour un ddl de 97. On l'ajoute au résultat antérieur pour obtenir en fin de compte :

$$CHI2 = 610 \; ddl = 145$$

Mais comme les résultats doivent se lire sur une table à des endroits différents, il est préférable de faire la conversion en un écart réduit dont l'interprétation est directe. Là encore Muller propose la formule de conversion :

$$z = \sqrt{(2 * chi2)} - \sqrt{((2 * ddl) - 1)}$$

L'ensemble des résultats est reproduit dans le tableau 8 qui croise les 34 textes du corpus.

	Mé	Cl	Ve	Ga	Su	Tu	Mé	Pl	Il	Cd	Ci	Ho	Po	Pp	M1	M2	Ro
Mé	36	75	77	80	78	46	82	74	80	94	88	92	89	91	89	90	92
Cl	75	38	85	86	84	48	74	78	79	89	84	89	87	87	96	99	88
Ve	77	85	37	79	77	47	85	73	80	94	91	95	90	95	86	90	93
Ga	80	86	79	36	76	46	85	75	78	93	91	94	88	95	86	87	92
Su	78	84	77	76	35	45	85	73	79	92	90	94	87	94	87	89	92
Tu	46	48	47	46	45	35	63	64	61	71	67	67	65	69	69	71	67
Mé	82	74	85	85	85	63	53	80	76	84	79	83	82	81	98	100	82
Pl	74	78	73	75	73	64	80	54	80	88	85	90	88	95	89	89	90
Il	80	79	80	78	79	61	76	80	51	86	82	85	80	84	84	84	85
Cd	94	89	94	93	92	71	84	88	86	61	85	82	85	86	99	101	88
Ci	88	84	91	91	90	67	79	85	82	85	57	80	79	76	97	98	82
Ho	92	89	95	94	94	67	83	90	85	82	80	57	81	79	99	101	81
Po	89	87	90	88	87	65	82	88	80	85	79	81	55	79	93	93	81
Pp	91	87	95	95	94	69	81	95	84	86	76	79	79	59	99	99	78
M1	89	96	86	86	87	69	98	89	84	99	97	99	93	99	59	75	95
M2	90	99	90	87	89	71	100	89	84	101	98	101	93	99	75	61	93
Ro	92	88	93	92	92	67	82	90	85	88	82	81	81	78	95	93	57
Th	93	90	94	91	89	68	87	90	86	90	84	87	79	83	95	93	81
Hé	94	90	94	96	95	73	89	91	89	91	83	90	85	87	96	95	82
An	90	87	94	91	91	66	82	90	83	87	81	81	78	78	94	94	77
Sa	95	94	97	96	94	71	92	95	90	90	88	87	88	85	96	95	81
Ni	97	94	99	97	95	70	90	96	90	93	85	88	87	81	97	94	82
Pe	91	89	94	95	93	72	87	88	91	89	81	88	86	82	98	97	82
Oe	97	94	102	101	99	70	89	99	91	95	87	86	87	85	101	100	81
To	94	91	101	100	98	70	82	96	91	97	89	90	90	86	105	103	86

Se	95	96	98	97	95	70	92	96	91	96	83	89	87	80	98	95	83
So	93	93	96	94	92	69	89	92	90	93	86	86	85	81	96	93	80
Ot	94	94	96	94	93	67	92	94	89	95	84	87	84	83	95	95	82
Ag	98	101	101	98	95	70	100	96	96	100	94	94	91	91	99	96	88
At	93	92	97	96	94	70	89	93	89	92	83	85	85	81	97	96	78
Ti	93	94	96	92	91	68	92	90	90	96	86	89	84	86	94	92	81
Ps	89	90	95	91	93	63	87	92	86	93	88	87	85	87	97	96	82
Pu	94	97	97	95	93	70	96	94	93	96	88	92	87	89	96	93	85
Su	94	94	96	94	92	68	91	91	92	93	86	88	85	85	96	92	81

	Ro	*Th*	*Hé*	*An*	*Sa*	*Ni*	*Pe*	*Oe*	*To*	*Se*	*So*	*Ot*	*Ag*	*At*	*Ti*	*Ps*	*Pu*
Mé	93	94	90	95	97	91	97	94	95	93	94	98	93	93	89	94	94
Cl	90	90	87	94	94	89	94	91	96	93	94	101	92	94	90	97	94
Ve	94	94	94	97	99	94	102	101	98	96	96	101	97	96	95	97	96
Ga	91	96	91	96	97	95	101	100	97	94	94	98	96	92	91	95	94
Su	89	95	91	94	95	93	99	98	95	92	93	95	94	91	93	93	92
Tu	68	73	66	71	70	72	70	70	70	69	67	70	70	68	63	70	68
Mé	87	89	82	92	90	87	89	82	92	89	92	100	89	92	87	96	91
Pl	90	91	90	95	96	88	99	96	96	92	94	96	93	90	92	94	91
Il	86	89	83	90	90	91	91	91	91	90	89	96	89	90	86	93	92
Cd	90	91	87	90	93	89	95	97	96	93	95	100	92	96	93	96	93
Ci	84	83	81	88	85	81	87	89	83	86	84	94	83	86	88	88	86
Ho	87	90	81	87	88	88	86	90	89	86	87	94	85	89	87	92	88
Po	79	85	78	88	87	86	87	90	87	85	84	91	85	84	85	87	85
Pp	83	87	78	85	81	82	85	86	80	81	83	91	81	86	87	89	85
M1	95	96	94	96	97	98	101	105	98	96	95	99	97	94	97	96	96
M2	93	95	94	95	94	97	100	103	95	93	95	96	96	92	96	93	92
Ro	81	82	77	81	82	82	81	86	83	80	82	88	78	81	82	85	81
Th	58	84	80	87	86	84	86	90	86	85	83	89	83	83	87	87	84
Hé	84	63	85	86	86	81	85	92	88	89	85	92	85	87	92	85	87
An	80	85	56	84	86	86	83	83	86	83	85	90	81	85	79	88	84
Sa	87	86	84	61	82	83	84	90	82	80	83	83	81	82	88	82	79
Ni	86	86	86	82	60	82	83	90	77	78	81	85	79	79	90	83	79
Pe	84	81	86	83	82	62	85	86	84	83	84	87	80	84	91	85	81
Oe	86	85	83	84	83	85	60	87	83	81	82	88	80	82	83	84	78
To	90	92	83	90	90	86	87	60	82	81	84	87	81	82	83	86	83
Se	86	88	86	82	77	84	83	82	60	72	76	79	74	74	86	78	74
So	85	89	83	80	78	83	81	81	72	59	77	80	77	74	85	79	75
Ot	83	85	85	83	81	84	82	84	76	77	57	81	75	73	86	73	76
Ag	89	92	90	83	85	87	88	87	79	80	81	60	79	74	86	76	73
At	83	85	81	81	79	80	80	81	74	77	75	79	60	75	83	77	73
Ti	83	87	85	82	79	84	82	82	74	74	73	74	75	58	83	72	71
Ps	87	92	79	88	90	91	83	83	86	85	86	86	83	83	53	85	82
Pu	7	85	88	82	83	85	84	86	78	79	73	76	77	72	85	60	75
Su	84	87	84	79	79	81	78	83	74	75	76	73	73	71	82	75	58

Tableau 8. Tableau de la connexion lexicale chez Corneille

Mé (*Mélite* : 2956) ; *Cl* (*Clitandre* : 2955) ; *Ve* (*Veuve* : 3030) ; *Ga* (*Galerie* : 2994) ; *Su* (*Suivante* : 2958) ; *Tu* (*Tuileries* : 2185) ; *Mé* (*Médée* : 2907) ; *Pl* (*Place* : 2957) ; *Il* (*Illusion* : 2870) ; *Cd* (*Cid* : 3073) ; *Ci* (*Cinna* : 2876) ; *Ho* (*Horace* : 2952) ; *Po* (*Polyeucte* : 2864) ; *Pp* (*Pompée* : 2885) ; *M1* (*Menteur1* : 3159) ; *M2* (*Menteur2* : 3148) ; *Ro* (*Rodogune* : 2837) ; *Th* (*Théodore* : 2911) ; *Hé* (*Héraclius* : 2981) ; *An* (*Andromède* : 2857) ; *Sa* (*Sanche* : 2940) ; *Ni* (*Nicomède* : 2930) ; *Pe* (*Pertharite* : 2920) ; *Oe* (*Oedipe* : 2967) ; *To* (*Toison* : 3005) ; *Se* (*Sertorius* : 2891) ; *So* (*Sophonisbe* : 2855) ; *Ot* (*Othon* : 2868) ; *Ag* (*Agésilas* : 3006) ; *At* (*Attila* : 2842) ; *Ti* (*Tite* : 2841) ; *Ps* (*Psyché* : 2932) ; *Pu* (*Pulchérie* : 2932) ; *Su* (*Suréna* : 2839)

2.2. Muller n'avait retenu que 32 titres, en éliminant les pièces où la participation de Corneille n'avait été que partielle. Nous avons rétabli les titres délaissés aux lignes 6 (*Tuileries*) et 32 (*Psyché*). La raison de ce sauvetage n'est pas philologique mais méthodologique. Il se trouve que la *Comédie des Tuileries* est nettement plus courte que les autres pièces de Corneille dont la taille tourne autour de 20 000 mots ([6]). Les *Tuileries* n'en ont que 4 245, près de 5 fois moins. Le test de la connexion lexicale résiste-t-il à cette inégalité ? Il semble que non : les valeurs lues dans la ligne 6 (et dans la colonne 6, le tableau étant carré et symétrique) ne dépassent guère 70, quand les autres atteignent ou dépassent 100. D'autres expériences nous amènent à penser que la chute du coefficient est due à une étendue moindre. Les résultats — KHI2 ou écarts réduits — appartiennent en effet à l'univers probabiliste, dans lequel on mesure les chances d'obtenir les faits constatés. La valeur de l'écart observé est prise en compte mais aussi la taille des observations. Pour un même écart en pourcentage, par exemple 10% en plus ou en moins, le KHI2 aura des valeurs très différentes si l'effectif considéré est de l'ordre des centaines, des milliers ou des millions. Il y a donc lieu de procéder à une légère pondération qui fait entrer la taille des deux textes considérés dans le calcul de leur distance. Après de multiples essais, nous n'avons pas trouvé mieux que la racine carrée de leur vocabulaire. Ainsi la valeur 46 qu'on lit au croisement de *Mélite* et des *Tuileries* (1e colonne, 6e ligne) est redresée comme suit :

D = *46 / (√taille du vocabulaire de Mélite * √taille du vocabulaire des Tuileries*)
= *46 / (√2886 *√1100*)
= *46 /(53.72 * 33.16*)
= *0,0256* (*valeur multipliée par 10 000 pour plus de lisibilité*)

Cette valeur 256 est mise en relation avec les autres mesures obtenues pour le même texte, ce que montre le graphique 9. On y constate que les *Tuileries* terminent le cycle des débuts de Corneille dans la comédie, avant la rupture de *Médée.*

Un histogramme peut être appliqué de la même façon à chacun des 34 textes du corpus mais l'intérêt principal est dans la carte globale qu'on peut extraire de ces distances bilatérales, comme celles qu'on peut déduire d'un tableau de distances géographiques établies de ville à ville. Il y a souvent une déformation parce que la distance peut représenter des mesures particulières, par exemple le temps nécessaire pour aller d'une ville à l'autre, qui s'accroît avec les montagnes. La figure 10 montre la distorsion qu'une distance ainsi calculée produit dans le paysage, Paris étant décentré vers le nord parce que les communications y sont plus faciles.

(6) La page 270 de la thèse de Muller propose d'autres chiffres pour l'étendue des pièces. C'est que Muller n'inclut pas les signes de ponctuation dans le décompte.

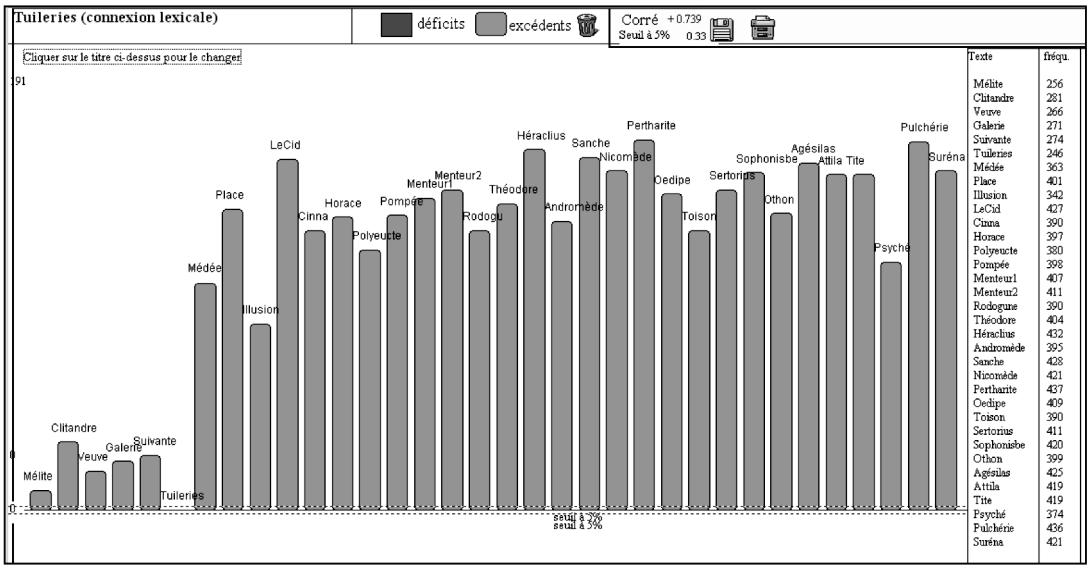


Figure 9. Distance des Tuileries aux autres pièces de Corneille

2.3. La carte 11 dérivée de la connexion lexicale est de lecture facile pour qui a lu la thèse de Muller. Ce qui saute aux yeux c'est d'abord la prédominance du genre. Toutes les pièces qualifiées de comédies se portent à droite, les tragédies s'installant à gauche. Cela ne va pas sans perturber légèrement la chronologie qui elle aussi va de droite à gauche, suivant un arc de cercle caractéristique des données sérielles. Ainsi les deux *Menteur*, quoiqu'ils appartiennent à la seconde période, rejoignent le camp des pièces comiques pour répondre à l'appel plus puissant du genre. De même l'*Illusion* qui suit *Médée* dans le temps de la rédaction rebrousse chemin pour garder le contact avec les comédies. Mais quand le genre n'intervient pas, la chronologie égrène les pièces selon un croissant presque régulier, où les jalons tracés par Muller peuvent être facilement circonscrits, de *Médée* au *Cid,* de *Horace* à *Pertharite* et d'*Oedipe* à *Suréna.*

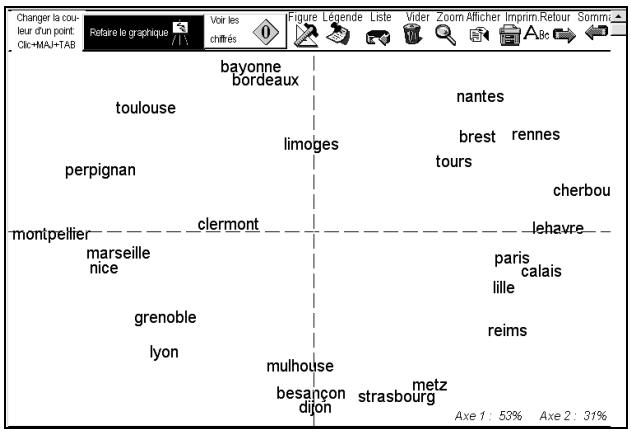


Figure 10. La carte de France établie à partir du temps pour aller d'une ville à l'autre (*analyse factorielle de correspondance*)

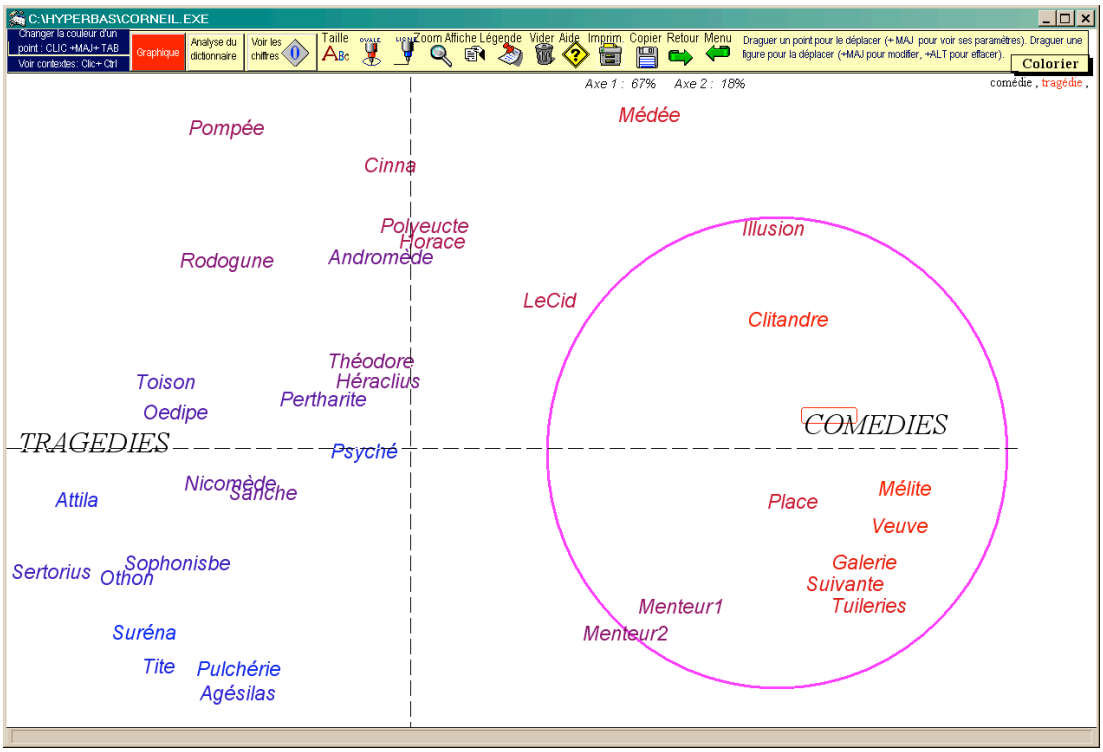


Figure 11. Analyse factorielle de la connexion lexicale chez Corneille

3. Analyse de la connexion. Application à Molière

3.1. Hautes et basses fréquences

Dans la connexion lexicale établie sur les lemmes, on ne sait trop quels facteurs agissent. Le choix des mots est gouverné par diverses influences qui se combattent ou s'appuient : le genre, l'époque, le sujet. Un indice global ne peut que les mêler sans permettre la décantation. Nous avons donc introduit dans le calcul un filtre qui aide à isoler les basses et les hautes fréquences, et peut-être, à travers cette distinction, les facteurs visés. Nous prendrons pour exemple cette fois le corpus de Molière. La connexion lexicale ainsi livrée au spectroscope est décomposée en deux coupes. La première ne tient compte que des fréquences basses (de 1 à 50) ([7]) (figure 12). La seconde ne s'intéresse qu'aux autres fréquences (figure 13). Dans les deux cas l'interprétation est aisée, mais la convergence n'est pas au rendez-vous.

(7) Cette limite n'intéresse pas la fréquence d'un mot dans le corpus, mais celle d'un mot dans l'assemblage de deux textes, comme expliqué précédemment.

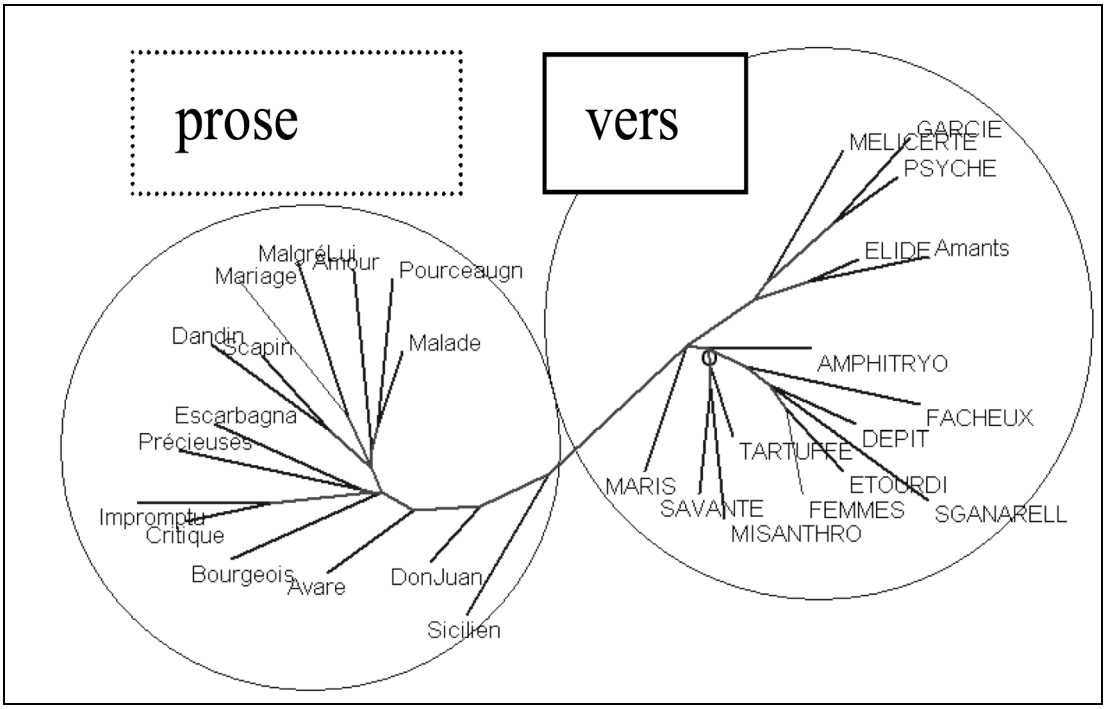


Figure 12. Analyse arborée des basses fréquences isolées dans la connexion lexicale

Dans la figure 12 deux sous-genres se font face. Les pièces en vers (leur nom est en majuscules) ne se compromettent pas avec les pièces en prose (en minuscules). Il n'y a qu'une exception, facilement explicable : si les *Amants magnifiques* se retrouvent dans le camp noble du vers, c'est parce que ce « divertissement royal » est rempli d'intermèdes versifiés. Dans la figure 13 au contraire les majuscules se mêlent aux minuscules et les vers à la prose. Un autre regroupement s'y manifeste qui suit la chronologie : les onze dernières pièces occupent le haut de la figure, les premières se groupent en bas. La conclusion, confirmée par d'autres monographies (par exemple Balzac, Verne, Zola, Proust, Anatole France), est que le choix du genre et du sujet impose davantage sa loi dans les basses fréquences : il y a des mots qui n'ont pas leur place dans une pièce en vers, d'autres qui sont permis dans une comédie mais non dans une tragédie. La poésie a son univers lexical qui n'est pas celui de la correspondance ou du récit, etc. Ces interdits et ces privilèges concernent moins directement les mots fréquents et encore moins les mots-outils parce que leur emploi est inévitable dans tout discours et que l'ostracisme est plus difficile à leur endroit ([8]). Mais les fréquences

(8) Il y a cependant des mots-outils extrêmement sensibles à la situation du discours et aux contraintes ou préférences du genre. Les pronoms personnels ou possessifs sont sujets à de fortes variations de cet ordre, comme les démonstratifs, les subordonnants, les relatifs et certaines prépositions.

hautes n'en sont pas moins animées de mouvements qui paraissent plus lents mais plus profonds et qui décrivent sourdement l'évolution de l'écriture. Ces mouvements de fond sont sans doute moins conscients ou moins volontaires que les choix clairs que l'écrivain fait parmi les genres et les sujets. Plus stylistiques que thématiques, ils sont davantage le reflet de la structure que du contenu.

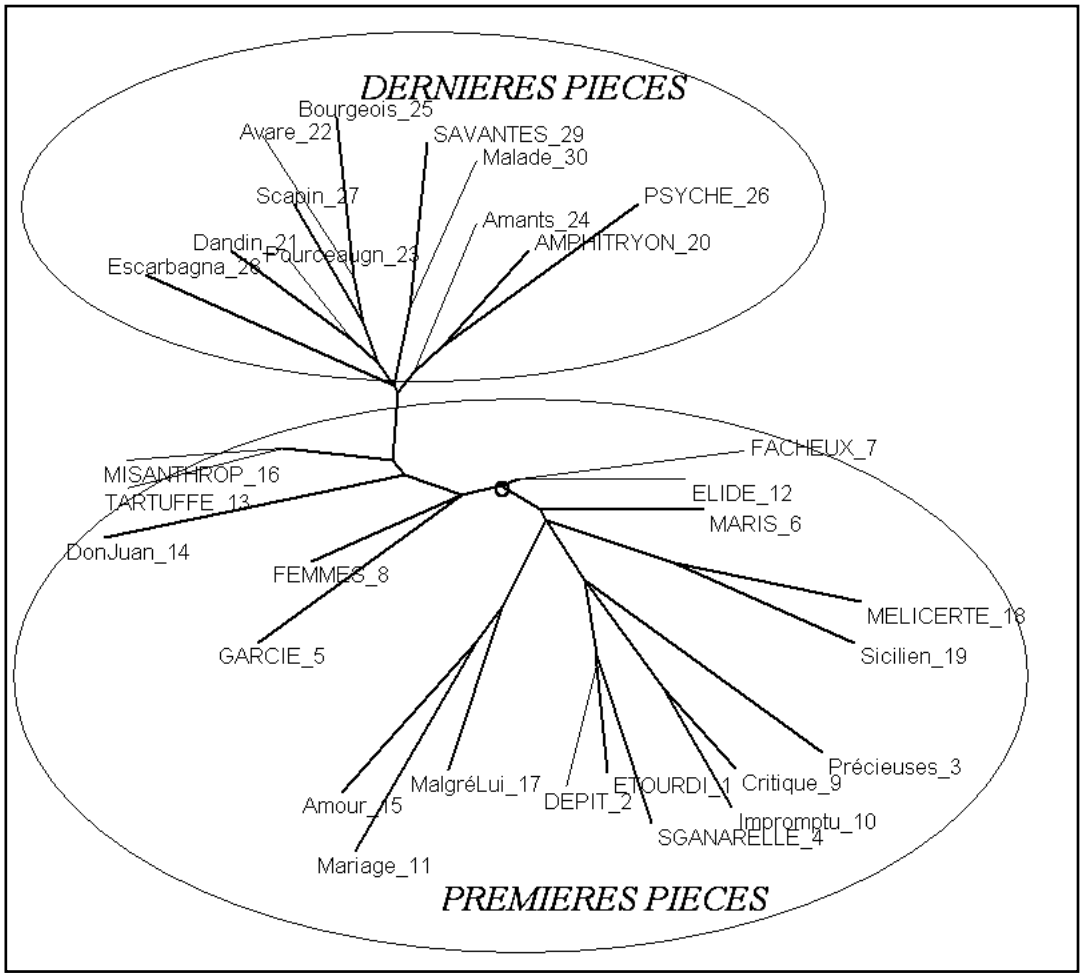
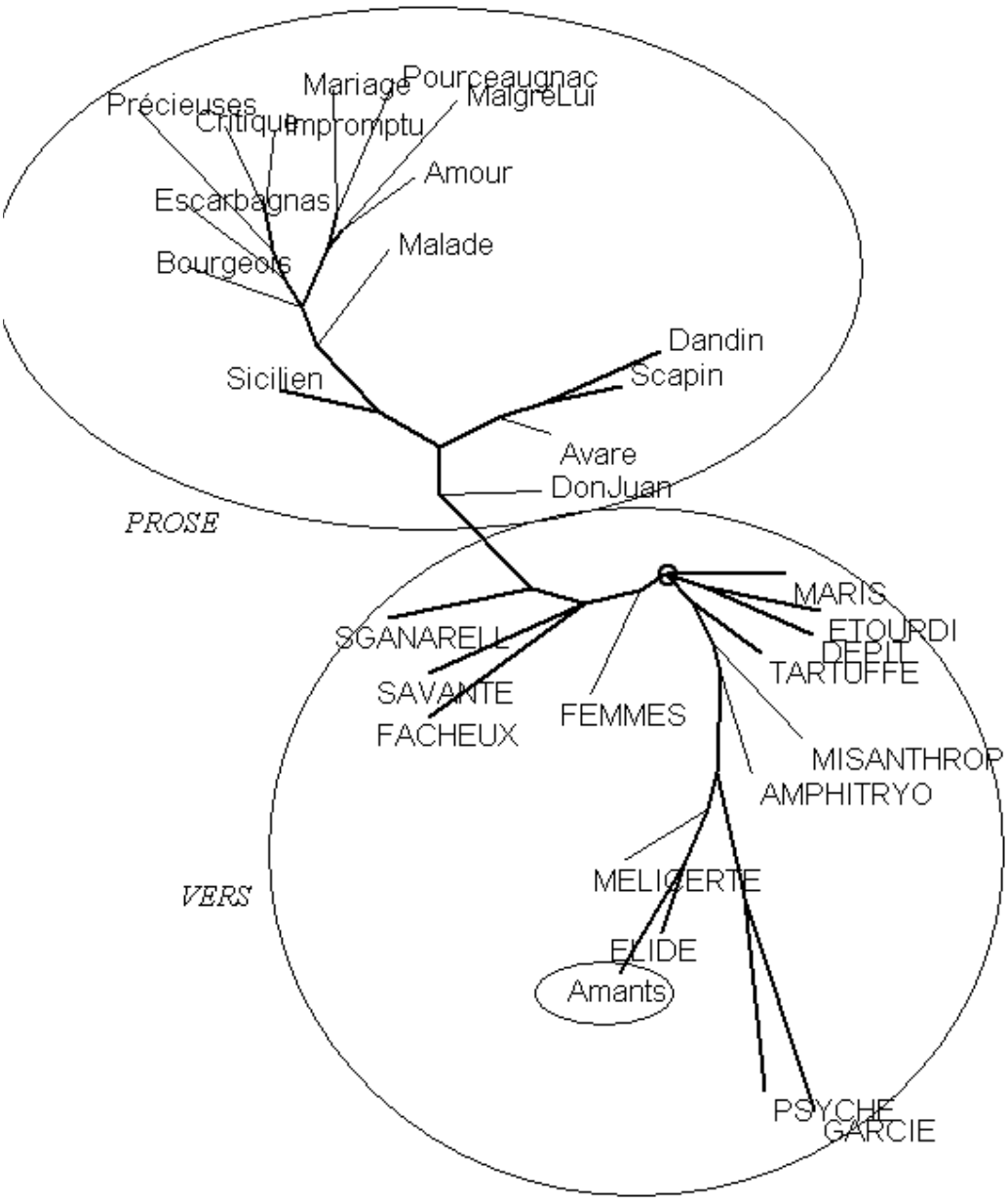


Figure 13. Analyse arborée des hautes fréquences isolées dans la connexion lexicale

De ces deux images contrastées que donne la connexion lexicale, la première, réservée aux basses fréquences, est très proche de la représentation fournie par la distance de Jaccard (graphique 14). Là aussi la dichotomie est absolue qui sépare vers et prose (mis à part l'exception attendue des *Amants magnifiques*). Même schéma d'ensemble. Mêmes détails aussi, comme les paires *Psyché-Garcie, Elide-Amants, Dandin-Scapin, Impromptu-Critique.* On donnera cependant la préférence à la connexion lexicale, parce qu'elle tient compte des

dosages fins de la distribution, alors que la méthode Jaccard, qui s'en tient au jugement expéditif du tout ou rien, brutalise les observations.



Graphique 14. La distance Jaccard appliquée à Molière

3.2. La distance Labbé

Reste à comparer la connexion lexicale à un indice proposé par Dominique Labbé et appliqué avec l'éclat que l'on sait à l'œuvre de Molière et de Corneille. Cette méthode que nous avons intégrée depuis longtemps à notre logiciel

Hyperbase, tient compte, comme la méthode binomiale, des faits de fréquence. Et, comme la distance de Jaccard, elle se réduit à enregistrer des quotients ou des rapports, en dehors des lois probabilistes. Comme le vote de chaque mot est individuel, le poids de chacun tend à être le même, ce qui affaiblit l'influence des mots fréquents et puissants, perdus dans la foule des petits. C'est pourquoi, même si son auteur écarte certains mots rares, et notamment les hapax du texte le plus long, et si d'autres précautions sont prises quand l'écart est trop faible, le test de Labbé privilégie les basses fréquences et dans les faits son témoignage s'accorde très souvent avec la méthode Jaccard, sans permettre d'isoler ce qui tient aux basses fréquences et ce qui tient aux hautes. Nous avons maintenu cet outil dans notre logiciel, car en multipliant les approches, on peut espérer de leur convergence plus de fiabilité dans les résultats. Mais nous ne cachons pas notre préférence pour la connexion de Muller, en regrettant de n'avoir pas introduit plus tôt cette mesure dans notre logiciel. On craignait que la mémoire ne vienne à manquer pour recueillir les observations ou que le temps du traitement soit prohibitif. Il n'en est rien. Certes la distance de Jaccard est plus rapide à calculer mais celle de Labbé est plus lente.

3.3. Le théâtre classique

Et comme nous nous trouvons présentement sur le terrain de Muller, en présence d'une thèse dont le dernier chapitre est à compléter, après quarante ans d'attente, j'aimerais utiliser pour la première fois l'outil du lexicomaître et l'appliquer à son propre domaine, au théâtre classique. Introduisons la clé, la connexion lexicale, et ouvrons la porte. Ce que l'on découvre apparaît dans la figure 15. Comme le bureau du Maître, la salle est bien rangée. Les 75 pièces sont classées par auteurs, Racine à droite, Molière à gauche, et Corneille au centre. Et pour chaque auteur la chronologie et le genre permettent un sous-classement. Chez Racine les premières pièces et les dernières se détachent du reste. Chez Corneille les comédies, à gauche, se distinguent des tragédies à droite. Chez Molière les pièces en vers précèdent les pièces en prose. Si l'on est aveugle aux auteurs, le paysage est encore lisible : à droite c'est la tragédie, à gauche la comédie. À droite c'est le domaine du vers, à droite celui de la prose. Et si l'on croise genre et versification, on a une progression très claire : tragédie en vers, puis comédie en vers, puis comédie en prose.

Il y a pourtant trois ou quatre points où l'harmonie des genres et des signatures ne règne plus. Qu'arrive-t-il lorsqu'il y a conflit ? C'est le genre qui prévaut. Ce qui ne va pas à l'encontre des prérogatives de l'écrivain, puisque c'est lui qui choisit librement le genre, même s'il n'a pas toute liberté pour en modifier les lois. Les points de divergence sont très localisés dans le graphique 15 : à droite au beau milieu des tragédies on trouve une pièce de Molière, *Don Garcie de Navarre ou le Prince jaloux*. Quoique le sous-titre y soit ambigu (« comédie

héroïque »), il s'agit d'une pièce sérieuse où Molière jouait un rôle tragique. Ce fut un échec, pour l'acteur comme pour l'auteur, et Molière se le tint pour dit. À l'opposé, parmi les comédies de Molière, on relève une pièce étrangère, les *Plaideurs,* qui est de Racine, et la seule comédie que Racine ait écrite. Le genre lui a dicté sa place, sans contestation. Enfin reste à considérer le cas des deux *Menteur*. Sans être très éloignés des autres comédies de Corneille, ils prennent place parmi les premières comédies de Molière et surtout celles qui sont écrites en vers. Là encore le genre a parlé. S'y ajoutent les contraintes de la versification, et — plus faiblement — un rapprochement chronologique, les dernières comédies de Corneille n'étant guère antérieures aux premières de Molière.

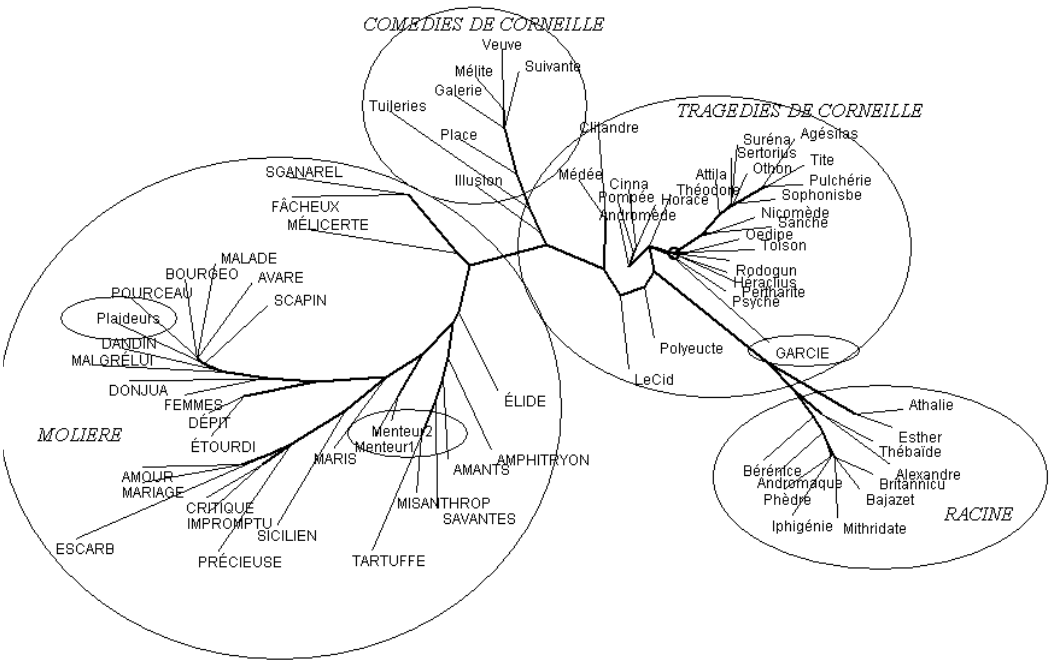


Figure 15. Analyse arborée du théâtre classique, à partir de la connexion lexicale

On restera sur ce constat, qui n'est pas le premier où l'on constate la force du genre. Il y a vingt ans Muller m'avait proposé une expérience de laboratoire qui consistait à mêler les écrivains et les genres. On avait relevé une liste de mots grammaticaux dans des œuvres poétiques, théâtrales et romanesques de trois écrivains de la même école et l'analyse était conviée à un problème d'attribution. Elle en trouva trois : un poète qui avait écrit les *Contemplations,* les *Méditations* et les *Nuits*, et pareillement un romancier et un dramaturge. Le genre s'était interposé devant l'écrivain (Brunet et Muller : 1988).

L'examen statistique, s'il est pratiqué de bonne foi et avec de bons outils, comme la connexion lexicale de Muller, s'inscrit donc en faux contre la thèse de

Pierre Louÿs, bien mal épaulée par Labbé. Au reste dans ces questions historiques, si la statistique peut fournir des indices et même des présomptions, elle ne peut produire des preuves au même titre que la philologie et l'histoire littéraire. Muller il y a quarante ans avait prévu et prévenu ces imprudences dans la conclusion prémonitoire de sa thèse : « Nous avions déclaré d'emblée que cette œuvre ne pose pas de problèmes philologiques importants, et que notre étude ne promettait ni révélations ni solutions inédites. Ne serait-elle pas de nature, plutôt, à mettre en garde ceux qui, en l'absence de renseignements historiques, attendent de la statistique lexicale des certitudes en matière de datation et d'attribution ? » (Muller : 1967, p. 263).

4. Ouvrages cités

BAULIEU (F.B.) : 1989, « A classification of Presence /Absence Based Dissimilarity Coefficients », *Journal of Classification, 6,* pp. 233-246.

BRUNET (Étienne) : 1988, « Une mesure de la distance intertextuelle : la connexion lexicale », *Revue, Informatique et Statistique dans les Sciences humaines, 1-4.*

BRUNET (Étienne) et MULLER (Charles) :1988, « La statistique résout-elle les problèmes d'attrribution ? », *Strumenti critici*, III,3, pp. 367-387.

ÉVRARD (Étienne) : 1966, « Étude statistique sur les affinités de cinquante-huit dialectes bantous », in *Statistique et analyse linguistique* (Paris : PUF), pp. 85-94.

GUIRAUD (Pierre) : 1954, *Bibliographie critique de la statistique linguistique* (Utrecht : Spectrum).

GUIRAUD (Pierre) : 1959, *Problèmes et méthodes de la statistique linguistique* (Dordrecht : D. Reidel Publishing Company).

JACCARD (Paul) : 1901, *Bulletin de la Société Vaudoise des Sciences Naturelles, 37*, pp. 241-272.

LAFON (Pierre) : 1984, *Dépouillements et statistiques en lexicométrie* (Genève-Paris : Slatkine-Champion).

MULLER (Charles) : 1967, *Étude de statistique lexicale. Le vocabulaire de Pierre Corneille* (Paris : Klincksieck).

MULLER (Charles) : 1968, *Initiation à la statistique linguistique* (Paris : Larousse).