Où l'on mesure la distance entre les distances (à propos de l'affaire Corneille-Molière)

Étienne Brunet

Où l'on mesure la distance entre les distances
(à propos de l'affaire Corneille-Molière)

Etienne Brunet

L'exposé qui va suivre reprend et prolonge le thème d'une conférence prononcée en avril à la Sorbonne dans le cadre d'un cycle intitulé, non sans humour, « Tous ceux qui comptent ». Un journaliste, qui se trouvait dans la salle, s'est fait l'écho, plus sonore que fidèle, des propos que j'ai cru devoir tenir dans l'affaire Corneille-Molière qui s'étalait alors sur la place publique et où mon nom avait été imprudemment cité. L'article paru le 11 avril 2003 dans l'hebdomadaire *Le Point* passait sous silence les longs développements que j'avais consacrés à la méthode prônée par Dominique Labbé, pour n'en retenir que la conclusion, laquelle contestait l'interprétation donnée aux faits observés, mais non pas leur mesure. L'affaire s'est envenimée dans les médias et sur Internet, au point que le modérateur du Forum spécialisé LITOR a dû suspendre un débat que la suspicion, la violence et la mauvaise foi avaient dénaturé. Si pour la première fois nous confions à la publication plutôt qu'au silence notre idée sur cette affaire, longtemps après avoir été mis en cause, c'est pour garder et défendre la mesure, pour empêcher qu'on ne profite de cet échec pour condamner sans appel la lexicométrie, et même, pour défendre Labbé et son œuvre contre ses propres excès.

Sans être un spécialiste du XVII^{e} siècle, il se trouve que j'ai été amené à m'intéresser, bien avant que Labbé s'en préoccupe, des rapports entre Molière et Corneille. Un de mes collègues à l'Université, s'était laissé convaincre par la thèse de Pierre Louÿs en y ajoutant un argumentaire de son propre cru. Un autre collègue de la même université, spécialiste incontesté de la comédie au XVII^{e} siècle, opposait son scepticisme à cette thèse, et l'ordinateur était sollicité de part et d'autre pour une expertise objective. Je fus donc conduit à consulter et à traiter

les données du théâtre classique, qui étaient disponibles depuis vingt ans au *Trésor de la Langue Française* et que Labbé allait reprendre quelques années plus tard, en les complétant. Je m'en suis tenu en effet aux pièces classiques les plus célèbres dont 13 de Molière, 8 de Corneille et 10 de Racine. Or les trois auteurs dramatiques, soumis à un calcul de distance lexicale (le calcul de Jaccard) et à l'analyse factorielle, se détachaient fort bien les uns des autres. Le commentaire de cette expérience se trouve encore aux pages 102-103 du manuel de notre logiciel Hyperbase : « La spécificité des trois écrivains y est excellemment soulignée puisque chacun occupe un coin du graphique. Mais la loi suprême du genre est respectée : le *Menteur* et les *Plaideurs*, tout en s'écartant le moins possible de leur auteur, passent dans le camp de la comédie. »

Cette expérience, déjà ancienne, semblait confirmer les leçons d'une recherche, plus ancienne encore, réalisée avec Charles Muller. Ce spécialiste de Corneille – qui n'a jamais ajouté foi à la thèse de Pierre Louÿs – m'avait proposé un exercice de laboratoire en isolant trois écrivains de la même période et de la même école romantique. En fournissant à l'ordinateur une liste de soixante éléments choisis parmi les mots grammaticaux (on pensait écarter ainsi les aléas thématiques pour mieux cerner les faits stylistiques), nous voulions savoir si les mesures lexicométriques permettraient de reconnaître la griffe de Hugo, de Lamartine et de Musset dans les textes poétiques, romanesques ou dramaturgiques où les relevés avaient été faits. La machine eut beau jeu de reconnaître trois écrivains : un poète, qui avait écrit les *Méditations*, les *Contemplations* et les *Nuits*, un dramaturge qui avait écrit *Lucrèce Borgia* et *Il ne faut jurer de rien* et un prosateur qui était l'auteur de *Raphaël*, de *Notre-Dame de Paris* et des *Confessions d'un enfant du siècle*. Le genre avait malencontreusement recouvert les vraies signatures.

1. Une expérience de laboratoire

Mais les machines et même les hommes ont fait des progrès et la conclusion négative et presque désabusée des tentatives précédentes n'est peut-être plus de saison. Des outils et des traitements nouveaux sont maintenant disponibles, en particulier ceux que propose Dominique Labbé. D'où l'idée d'une collaboration avec ce chercheur.

1- Cependant, pour éviter à la machine une autre humiliation, j'ai cette fois neutralisé le genre. Les textes que la nouvelle expérience met en jeu relèvent tous du genre narratif. En revanche, la variable chronologique, ignorée précédemment, entre en ligne de compte, puisque

deux siècles s'interposent entre le texte le plus ancien (*La Vie de Marianne*, Marivaux, 1731) et le texte le plus récent (*Le Temps retrouvé*, Proust, 1927). L'objectif proposé au programme étant de reconnaître la paternité des textes, il suffit, pour chaque auteur, de traiter deux textes qui lui appartiennent et de vérifier si l'algorithme les attribue à la même plume. Pour corser la difficulté, on a choisi pour chaque écrivain d'associer deux œuvres situées aux deux extrémités de sa carrière, pourvu qu'elles partagent le même genre narratif. Il y a ainsi dix-huit ans entre le premier grand succès de Balzac (*Les Chouans*, 1829) et le dernier roman publié de son vivant *(Le Cousin Pons*, 1847). Un laps de temps plus grand encore sépare le premier roman naturaliste de Zola (*Thérèse Raquin*, 1867) et l'un des derniers titres des *Rougon-Macquart* (*La Bête humaine*, 1890). Entre l'un des tout premiers titres de Jules Verne (*De la Terre à la lune*, 1865) et le dernier manuscrit qu'il ait remis à son éditeur Hertzel, quelques jours avant sa mort (*Le secret de Wilhelm Storitz*, 1905), c'est une carrière de quarante ans qui s'est déroulée, modifiant l'inspiration et l'écriture. Cet écart systématique recherché entre les deux spécimens des onze écrivains retenus tendait à dilater au maximum, dans les limites du genre, les différences internes, afin de voir si elles résisteraient aux oppositions externes qui s'exercent entre les écrivains et empêcheraient l'attribution correcte des textes. En somme, nous voulions comparer les distances *intra* (entre les textes d'un même écrivain) et les distances *inter* (entre les écrivains).

En réalité le *nous* collectif que je viens d'utiliser est un abus de langage. Car j'ai été le seul responsable des conditions de l'expérience et du choix des textes. Dominique Labbé voulait en effet participer à l'expérience en ignorant tout des données, afin qu'aucun préjugé subjectif ne puisse pervertir le traitement. Dans beaucoup de disciplines, l'ignorance est ainsi la garantie de la connaissance, et notre modèle a été le protocole en aveugle que la recherche médicale applique au traitement des malades et au test des médicaments. Or il y a plusieurs distances possibles, selon qu'il s'agit de deux écrivains différents, ou de deux textes du même écrivain, ou de deux extraits du même texte. On a donc dédoublé tous les textes retenus, afin qu'il y ait pour chaque texte deux extraits différents, mais aussi proches que possible, puisqu'on les a choisis contigus, l'un suivant l'autre. Il y a ainsi pour chacun des onze écrivains quatre extraits qui lui sont attribués, soit 44 au total. Naturellement Dominique Labbé n'a eu droit qu'à des numéros anonymes (pour Proust c'était 21 et 43, 22 et 44 respectivement). Il n'a pas cherché à les identifier, même si des indices assez clairs – surtout les

noms propres – pouvaient aider au décryptage[1]. Pour déjouer toute tentative de cet ordre, un piège avait été tendu dans les six derniers extraits (numérotés de 45 à 50). Car ce ne sont pas des textes suivis, mais des agrégats constitués de pages empruntées aux 44 textes du corpus, à raison d'une page par texte. Le texte 45 réunit la première page de chaque texte, le texte 46 la dixième, etc. Cela donne des clones qui ne se distinguent pas les uns des autres, mais aussi des portraits robots qui font la synthèse de tous les textes du corpus et en constituent une sorte de moyenne ou d'échantillonnage raisonné. Ce piège a fortement intrigué Labbé, sans l'égarer, et nous invitons le lecteur à lire son commentaire, qui ne manque pas de perspicacité[2].

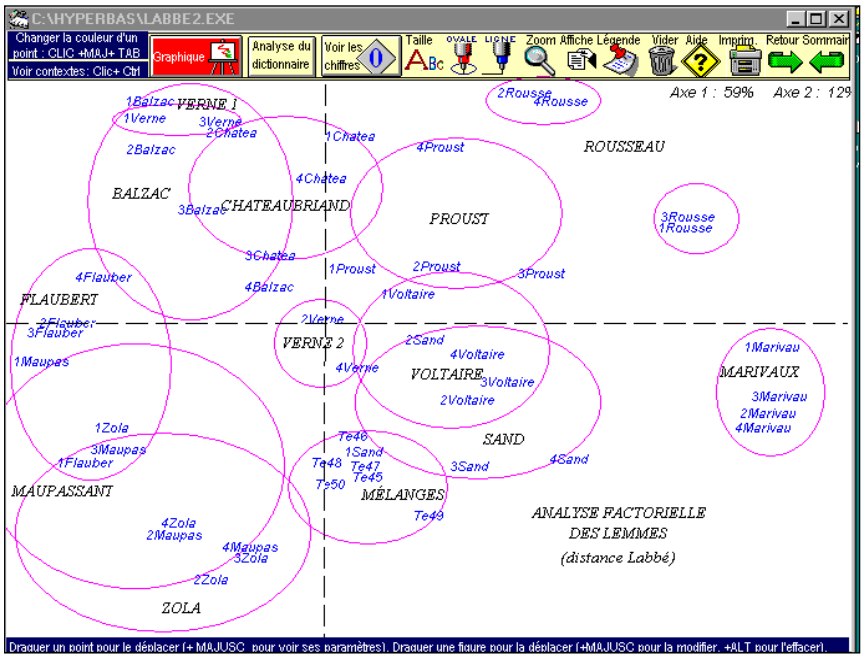


Figure 1. Analyse factorielle de la distance lexicale (formule de Labbé, appliquée aux lemmes)

1 Il ne serait pas sans intérêt de confronter la sagacité humaine à l'expertise de la machine. La lecture humaine, armée de connaissances externes et attentive aux indications du texte, devrait pouvoir reconnaître les extraits qui vont ensemble. Mais cela dépend de la culture du lecteur. Un ignorant fera des erreurs. On verra dans l'analyse de Dominique Labbé que l'ordinateur, appuyé sur les seuls comptages, et dénué de toute culture, ne se trompera pas une seule fois dans l'identification des couples.

2 Cet article se trouve sur le site de l'auteur à l'adresse : http://www.pacte.cnrs.fr/spip.php?article54

2- Avec les mêmes données et des méthodes semblables aux siennes, nous obtenons les mêmes résultats. Dans la panoplie des outils d'analyse multidimensionnelle, à côté de la classification automatique et de l'analyse arborée, dont D. Labbé a fait usage, on dispose de l'analyse factorielle, qui est illustrée dans la figure 1. Confirmation est donnée du lien très fort qui unit les couples : tout extrait portant l'indice 1 se trouve à proximité immédiate de l'extrait correspondant qui en est la suite et qui est numéroté 3, et il en est ainsi des extraits pourvus des indices 2 et 4. Mais encore les deux couples qui se rattachent au même écrivain ne sont jamais très éloignés, en sorte qu'il est facile de circonscrire dans un cercle plus ou moins étroit les quatre extraits qui relèvent de la même plume. Les concentrations les plus fortes sont le fait des extraits que D. Labbé a désignés comme étant sûrement de la même source : à l'extrême droite les textes de Marivaux (codés 1, 23, 2 et 24) et au centre ce que nous appelons les « mélanges » et qui concerne les extraits de 45 à 50.

Il n'en reste pas moins que l'analyse factorielle, c'est l'aire du soupçon. Elle fournit des présomptions sur une échelle continue qui ne rejoint la certitude que de façon asymptotique. Les certitudes sont parfois positives (par exemple le doute n'est guère permis pour Marivaux), mais plus souvent négatives : il est très peu probable que des points diamétralement sur le graphique soient de la même source. Entre ces deux extrêmes on trouve des situations relativement claires et d'autres plus troubles. Parmi les premières, on citera les configurations qui tournent autour de Rousseau, de Voltaire, Chateaubriand, Balzac et Proust. Mais le troupeau des textes réalistes et naturalistes, à gauche et en bas de la figure, est plus indistinct, comme si les bergers avaient mêlé leurs bêtes. Si Flaubert se distingue assez nettement de Zola, Maupassant évolue librement de l'un à l'autre, plus proche de Flaubert dans *Une Vie*, et de Zola dans *Pierre et Jean*. Le désaccord le plus criant est relatif à Jules Verne : si l'excentricité du premier texte (*De la terre à la lune*, extraits 19 et 41), soulignée par D. Labbé, est bien confirmée par la position extrême (en haut à gauche) des points *1Verne* et *3Verne*, la liaison est rompue avec l'autre texte de Verne qui se situe au centre du graphique (points *2Verne* et *4Verne* recouvrant le *Secret de Wilhelm Storitz*). Ces deux textes de Verne se trouvaient aussi très distants dans l'analyse de D. Labbé. L'explication tient non seulement à la distance chronologique qui sépare les deux textes (40 ans), mais aussi à l'évolution d'un écrivain qui commence par écrire des romans d'aventure pour enfants et qui finit candidat à l'Académie française avec des récits

fantastiques et psychologiques écrits à la manière du *Horla* de Maupassant.

En présence du graphique 1, l'œil peut être sensible en outre au mouvement d'ensemble qui, de la droite à la gauche, semble soumettre les textes et les auteurs à la dérive du temps. On observe une sorte de croissant, caractéristique des données sérielles, où prennent place, successivement et dans l'ordre chronologique, Marivaux, Rousseau, Chateaubriand, Balzac, Flaubert, Maupassant et Zola. Cette décantation du temps est pareillement observable dans les deux graphiques de D. Labbé. Rien de très surprenant : en deux siècles la langue a évolué, le mouvement des idées et des sensibilités s'est précipité, et le progrès technique a changé le monde. Pourtant le courant n'emporte pas tous les écrivains à vitesse constante : il peut se rencontrer des obstacles, des résistances et des remous et certains écrivains semblent remonter le courant. C'est le cas de G. Sand et, plus nettement encore, de Proust qui sur le graphique s'éloigne autant que possible du naturalisme et préfère en haut et à droite la compagnie de Rousseau et Chateaubriand. S'agit-il des thèmes proustiens ou de la phrase proustienne ? Les effets sont mêlés car la mesure proposée par D. Labbé tient compte de la fréquence de tous vocables et est sensible aux faits stylistiques autant que thématiques.

3- Pour y voir plus clair et distinguer le thème de la syntaxe, nous avons entrepris d'autres investigations, en poursuivant l'enquête en deçà ou au-delà du lemme. On gardait la mesure de la distance telle que la propose D. Labbé, mais en l'appliquant à d'autres objets isolés dans le même corpus : des graphies, des codes grammaticaux, des structures syntaxiques ou des étiquettes sémantiques. Mais d'autres mesures de la distance étaient aussi proposées et comparées à celle de Labbé. Comme on a rendu compte de cette expérience dans une autre publication[3], nous nous bornerons à reproduire la carte des distances établie sur les graphies. Cette fois nous utiliserons le programme d'analyse arborée que nous avons incorporé à notre logiciel HYPERBASE, parallèlement à l'analyse factorielle de correspondance. La méthode arborée en effet est particulièrement adéquate lorsque le tableau à analyser est une matrice carrée, où lignes et colonnes désignent les mêmes objets et où sont identiques les valeurs lues symétriquement de chaque côté de la diagonale principale (la distance de *A* à *B* est la même que de *B* à *A*). Les

[3] Actes des Troisièmes Journées de la linguistique de corpus, Lorient, 2003, disponible sur http://www.revue-texto.net/Inedits/Brunet/Brunet_Distance.html.

données de la figure 2 sont relatives aux simples graphies, avant toute lemmatisation.

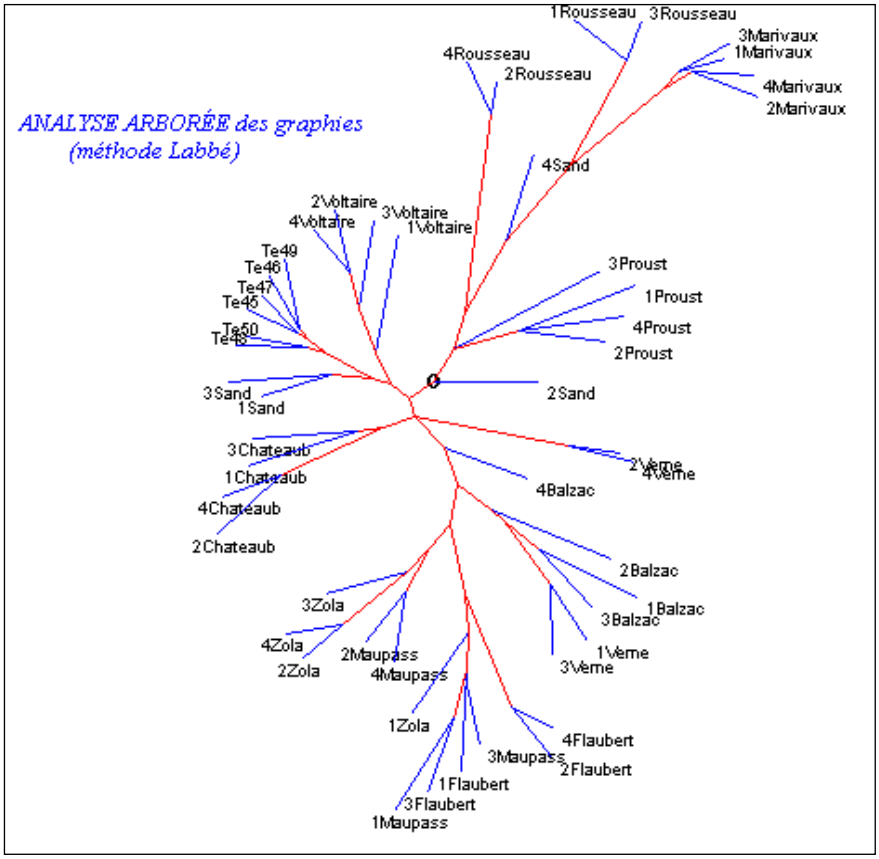


Figure 2. Analyse arborée
(distance établie sur les graphies)

L'interprétation de tels graphes est aisée dans son principe. La distance d'un texte à un autre est directement proportionnelle à la longueur des segments qu'il faut parcourir pour relier les deux points. L'angle, la direction, les tournants et les carrefours n'importent pas, seule compte la longueur du parcours dans un relief tourmenté où les routes empruntent les vallées et les cols.

En partant du haut du graphique, on rencontre d'abord Marivaux dont les quatre extraits sont serrés les uns contre les autres, puis le chemin conduit à Rousseau (mais les deux textes de Rousseau, s'ils

débouchent sur la même voie, sont assez distants l'un de l'autre, car il y a loin entre le récit des amours romantiques et l'essai sur l'éducation des enfants). Ensuite la rencontre de Proust serait inattendue, si nous ne l'avions déjà croisé à cet endroit dans l'analyse des lemmes. Puis la route hésite ; des voyageurs en retard (Sand, deuxième Verne) ou en avance (Voltaire), ou bien des collectivités indifférenciées (les « mélanges » 45 à 50), ou bien encore un isolé que la naissance a placé au croisement des deux siècles (Chateaubriand), encombrent le carrefour qui conduit à la vallée opposée. Balzac attend là[4], qui passe le relais à Flaubert, puis à Maupassant et enfin à Zola. C'est à peu de choses près le chemin qu'a emprunté D. Labbé, les yeux bandés, en suivant les lemmes.

4- Reste à écarter un dernier doute, l'expérience ayant été menée à travers des textes tronqués. De plus, même si le corpus a une taille suffisante, à cause da la multiplication des textes (10 000 x 50 = 500 000 occurrences), chacun des textes traités reste relativement étroit. Les conclusions ne seraient-elles pas plus claires et plus sûres avec des textes complets et une étendue élargie ? La figure 3 répond à cette question en proposant un corpus quatre fois plus vaste (2 millions d'occurrences), constitué des mêmes textes, cette fois sans extraction ni troncature. Reprenons le problème initial et la méthode de Labbé et voyons si la distance lexicale, établie sur les lemmes, pourrait apparier les textes deux à deux et reconnaître une signature commune. Rappelons que les deux textes d'un même auteur ont été choisis à des moments fort différents de la carrière et que rien ne garantit que les thèmes et l'écriture y soient constants. Ils sont pourtant plus proches l'un de l'autre que de tout autre texte. Et cet air de famille est reconnu par l'analyse arborée qui distribue les couples tout au long de la chaîne. Si les liens familiaux sont prépondérants, l'appartenance à la même époque crée des liens secondaires, de sorte que la procession des couples se fait grossièrement par rang d'âge. Mais l'ordre chronologique est bousculé à certains endroits, Voltaire se rapprochant de l'époque moderne, tandis que Proust rompant avec le naturalisme semble appartenir au siècle précédent. Ces remous dans le fleuve chronologique montrent que le tempérament propre d'un écrivain peut résister au courant et que les procédures d'attribution que nous venons de mettre en œuvre sont plus efficaces contre le temps qu'elles ne le sont contre le genre.

[4] Le premier Verne, celui de la lune, aussi.

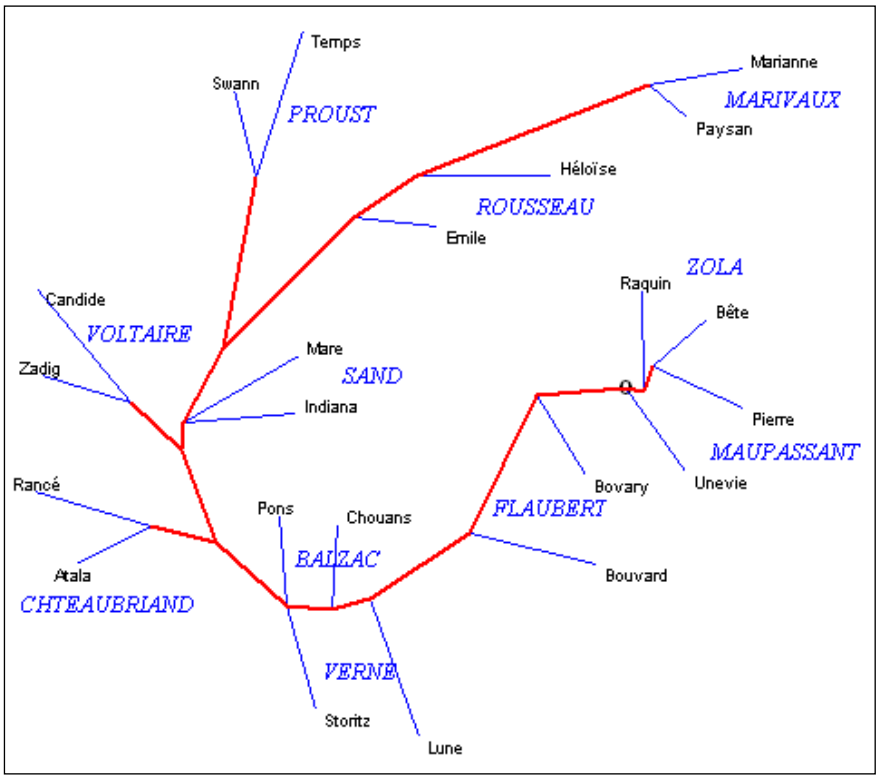


Figure 3. Analyse de la distance lexicale dans les textes complets
(les distances sont établies sur les lemmes, selon la méthode Labbé)

2. Limites de la formule de Labbé

Fort de cette expérience, D. Labbé a cru que la clé pouvait ouvrir d'autres portes et résoudre des problèmes d'attribution plus difficiles que l'exercice d'école qui précède. L'intention est louable car on ne peut réduire toujours la statistique linguistique à un rôle subalterne et ne solliciter son témoignage que lorsque l'affaire est déjà jugée. L'affaire Corneille-Molière – tardivement suscitée par Pierre Louÿs, trois siècles après la mort des intéressés – n'avait pas fait long feu et semblait classée depuis longtemps par les historiens de la littérature. Mais Labbé a fait appel de ce jugement, en invoquant non pas des faits nouveaux mais une méthode d'expertise nouvelle, celle que nous venons de mettre à l'épreuve. Depuis que le recours à l'ADN est autorisé devant les tribunaux, bien des affaires ont été éclaircies que les témoignages et les

autres indices n'auraient pu élucider. ADN, empreintes digitales, carbone 14, ces techniques de dépistage scientifique ont été évoquées dans le procès littéraire où Labbé s'est engagé. Mais la mesure de distance qu'il propose peut-elle jouer ce rôle ? C'est ce que nous nous proposons d'examiner.

1- Observons tout d'abord, d'un point de vue théorique, que la statistique peut emprunter deux voies : l'une est inférentielle, l'autre descriptive. La première s'appuie sur les lois probabilistes et permet, à partir d'observations réalisées sur un échantillon, de confirmer ou d'infirmer des hypothèses et de projeter des conclusions sur la population dont l'échantillon est extrait, tout en mesurant la précision et la sûreté de cette projection. La seconde est plus modeste, comme le note le mathématicien Barthélémy, auquel on doit l'analyse arborée et qui s'indigne de l'usage qui en est fait : « Cette utilisation des méthodes que j'ai contribué à mettre au point est un non-sens. On ne peut faire passer pour des statistiques inférentielles, avec lesquelles on peut éprouver des hypothèses, des statistiques descriptives, d'abord destinées à faire réfléchir des spécialistes[5] ». Or les techniques multidimensionnelles dont on fait usage en lexicométrie, qu'il s'agisse d'analyse factorielle, d'analyse arborée ou de classification hiérarchique, ne sont que des représentations analogiques, qui peuvent fournir des indices, des présomptions, mais non des preuves. Tout est affaire d'interprétation et la nôtre, avec les mêmes données et les mêmes résultats, est assez différente de celle de Labbé, ce que nous montrerons plus loin. L'essentiel du débat – dans la presse comme dans la discussion technique engagée sur la liste LITOR – a porté sur cette prétention de prouver, jugée imprudente et abusive.

Labbé n'a certes jamais caché que sa démarche est empirique, comme celle de tous les chercheurs qui s'adonnent à la lexicométrie, et l'empirisme ne se justifie que par la qualité, l'ampleur et la représentativité des observations. Quoique l'expérience de Labbé soit très large et solide, elle s'est surtout exercée jusqu'ici sur des textes modernes, en relation avec la politique, la sociologie et l'économie. Dans les discours ou entretiens qu'il a étudiés, même parfois sous la plume de de Gaulle ou de Mitterrand, la notion d'auteur a des aspects flous, parce qu'une équipe a souvent préparé ou même rédigé partiellement le texte. Dans le domaine littéraire, la paternité est plus chatouilleuse. On y est

[5] Cité dans un article du journal *Le Monde*, du 10 juin 2003, sous la signature de Fabienne Dumontet, (*Molière et Corneille confondus)*.

sensible aux sources, aux emprunts, aux plagiats, aux querelles d'école, aux contraintes du genre et aux propriétés de l'écriture. Or Labbé a jusqu'ici rarement exploré ce domaine particulier, sinon dans l'expérience que nous venons de relater. Est-ce assez pour affirmer la valeur universelle d'un test ? Si l'étalonnage de ce test est réellement fondé, comme on nous l'affirme, sur des « milliers de textes », encore faut-il que la représentativité de ces textes soit assurée. Combien de ces textes appartiennent à la littérature, combien au théâtre, combien à la comédie, combien à la tragédie, combien au genre versifié, combien au XVII[e] siècle ? Labbé se déclare prêt à mener des enquêtes dans ces directions. Que n'a-t-il commencé par là, avant de proposer imprudemment une échelle absolue.

2- Nous ne contestons pas l'intérêt de la mesure de Labbé, sans quoi nous ne nous serions pas prêté à l'expérience précédente. Mais, faute d'essais suffisants, nous refusons l'idée d'une échelle fixe, d'un barème arbitraire, attaché à une seule mesure, globale et indifférenciée, appliquée, qui plus est, à un seul aspect – lexical – du langage. Nous croyons même que la formule de Labbé vaut mieux que l'usage qu'il en fait, et nous nous sommes attaché dans les pages qui précèdent à diversifier son emploi, en l'appliquant à d'autres objets linguistiques que le lemme : aux graphies, aux codes grammaticaux, aux structures syntaxiques ou aux réseaux sémantiques. Naturellement l'échelle des valeurs obtenues varie selon l'objet étudié et le barème pour les lemmes ne vaudrait plus pour les graphies (J.M. Viprey a fort bien observé un décalage approximatif de 4 points[6]). Bien entendu pour les codes et les structures – on pourrait songer aussi aux mesures rythmiques ou prosodiques – l'échelle exigerait des accommodements plus importants. Mais même dans les conditions précises où se place Labbé, une échelle absolue est impraticable. Elle dépend en effet de certaines options – toutes pareillement justifiables – qui commandent le toilettage du texte, le comptage des mots, et la lemmatisation. Labbé a des exigences particulières quant à la présentation des textes (les hors-texte et didascalies par exemple sont écartés), quant au traitement des mots composés (il en relève un minimum dans les textes classiques), quant à la prise en compte des ponctuations dans le dénombrement des occurrences et surtout quant aux principes de lemmatisation. Le logiciel dont il est l'auteur – et qui a quelque mérite, ayant été construit par un homme seul

[6] Dans le cas du corpus Molière-Corneille-Racine, le décalage moyen est de 0,031, pour 75 textes et 2 775 mesures.

– se contente d'un codage minimum, qui n'envisage pas la fonction des mots et n'approfondit guère leur nature (ni le temps, ni le mode, ni la personne des verbes ne sont repérés). Il permet cependant de réduire les homographies, à condition que des retouches manuelles viennent suppléer aux embarras de la machine. Ces retouches évitent certes bien des erreurs grossières qu'on constate dans les résultats des lemmatiseurs automatiques, comme ceux de Cordial. Mais le prix à payer en temps est élevé, sans garantir la constance des décisions, qui varient d'un chercheur à l'autre et parfois même d'un moment à l'autre. Sauf à confier à Labbé le traitement de tout texte que l'on veut soumettre à son calcul de distance (d'autant que son lemmatiseur n'est pas commercialisé), on voit mal comment on pourrait appliquer son échelle, si les conditions de mesure ne sont pas semblables. Tous les linguistes appellent de leurs vœux une standardisation minimale dans la saisie, le codage, la lemmatisation et le traitement des textes mais cela ne peut résulter que d'un consensus international fixant des normes précises (ce que l'entreprise de *Texte Encodage Initiative* s'emploie à réaliser, d'autant que le codage XML en donne les moyens), ou à tout le moins sur une tradition nationale – qui en France est représentée majoritairement par *Frantext* et l'*Institut de Linguistique Française*. Toute tentative individuelle, même excellente, est vouée à l'échec.

La conséquence de cette situation est que les méthodes et les résultats de Labbé sont infalsifiables, puisqu'on doit passer par lui pour les approuver ou les combattre. Il est certes facile de trouver des contre-exemples où le barème invite à considérer deux textes comme appartenant à la même plume, alors qu'on sait de façon sûre qu'il n'en est rien. Mais Labbé peut toujours les récuser, en prétendant que les conditions du calcul n'ont pas été remplies, puisqu'il est le seul à pouvoir les remplir. Dans une base publiée il y a cinq ans, et distribuée par l'Éducation nationale sous le nom de *Batelier*, nous avions appliqué le calcul de Labbé à une soixantaine de textes, dont le *Menteur* et une trentaine de pièces classiques. Certes la proximité du *Menteur* (et aussi de l'*Illusion Comique*) avec les pièces en vers de Molière y avait été observée, mais aussi celle des *Fleurs du mal* et des *Poésies* de Rimbaud. À l'époque la formule de Labbé n'avait pas les correctifs qu'elle a reçus depuis et nous lui en avions ajouté un (en refusant les hapax non seulement du texte le plus long, ce que recommande Labbé, mais aussi du plus court). Et bien entendu nous ne disposions pas de la lemmatisation Labbé. Le résultat (0,182) n'a donc pas à être confronté à l'échelle établie depuis lors mais aux autres résultats obtenus dans le même corpus, avec

les mêmes options et les mêmes conditions. Or cette proximité entre les recueils de Baudelaire et de Rimbaud est aussi étroite que celle qui lie au *Menteur Don Juan* (0,180), le *Misanthrope* (0,173), l'*Avare* (0,177), les *Femmes savantes* (0,173), le *Bourgeois gentilhomme* (0,222) et le *Malade imaginaire* (0,207). Si donc on conclut que l'auteur du *Menteur* est le même que celui des pièces citées, on doit pareillement conclure qu'il n'y a qu'un auteur pour les *Poésies* rimbaldiennes et les *Fleurs du mal*[7]. Ce contre-exemple n'est d'ailleurs pas le seul que nous ayons relevé : en réunissant dans une même base l'œuvre de Molière et celle de Marivaux, les calculs de distance montrent bien une séparation nette entre les deux dramaturges, à l'exception de la première pièce de Marivaux, qui, il est vrai, est fort courte et la seule qu'il ait écrite en vers. Sans doute aussi s'inspire-t-elle du grand devancier mais elle n'est pas de Molière malgré les indications du barème. Nous avons pareillement réuni l'œuvre de Flaubert et celle de Maupassant et là encore le seuil de fusion est atteint pour *Madame Bovary* et *Une vie*. Il l'est aussi si l'on compare les quatre évangiles dans trois traductions françaises qui en ont été faites, soit douze versions différentes. Le calcul semble indiquer un auteur unique, qu'il s'agisse ou non du Saint-Esprit.

3- Une autre raison invite à renoncer au barème proposé par Labbé, c'est l'obscurité qui s'attache à une mesure unique et globale. Ce que l'on gagne en synthèse est perdu en analyse. Comment en effet interpréter une mesure de proximité quand plusieurs facteurs sont en cause. Labbé est sensible à cette difficulté et il détaille les influences qui entrent en ligne de compte : l'auteur, le genre, le sujet, l'époque. Mais dans une mesure donnée, rien ne permet de distinguer ces influences variables, dont le dosage échappe au calcul. Dès lors ce qu'un chercheur interprète comme caractéristique d'un écrivain, un autre critique peut l'attribuer aux contraintes exercées par le genre, voire aux lieux communs que le sujet entraîne. Les cas où le calcul est opérant sont ceux où les variables indésirables sont neutralisées. Dans une émission sur France-Culture, Labbé opposait à ses contradicteurs le cas de *Tite et Bérénice* où son calcul fait merveille pour distinguer la pièce de Racine et celle de Corneille. On aurait pu lui répondre que les conditions idéales étaient réunies (même sujet, même année et même genre) pour rendre le calcul efficace et explicite, mais qu'elles ne l'étaient plus dans le cas Corneille-

[7] En reprenant les mêmes données avec la formule exacte, les conclusions sont les mêmes : le coefficient pour Baudelaire-Rimbaud (0,296) est du même ordre que les autres (respectivement 0,289 0,271 0,289 0,277 0,332 0,323).

Molière qui faisait l'objet du débat. Quoi de plus attendu que la proximité du *Menteur* et des pièces de Molière ? Ce sont des comédies et celles qui sont les plus proches sont celles qui, comme le *Menteur*, sont écrites en vers. La seule comédie que Racine ait écrite, les *Plaideurs*, est également plus proche de Molière que de Racine. Pourquoi ne pas se contenter de ces remarques de bon sens. Pourquoi s'ingénier à chercher une explication hypothétique du côté de l'auteur, en refusant le facteur le plus évident, c'est à dire le genre[8] (d'autant que le genre est très contraignant à l'époque classique où de surcroît la versification impose des exigences supplémentaires). Devant l'impossibilité de démêler des facteurs entrecroisés et indissociables, le principe de précaution est de ne pas parler de preuve et de laisser à Pierre Louÿs le soin de défendre sa rêverie et ses intuitions.

4- Reste à apprécier en elle-même la formule par laquelle Labbé mesure la proximité entre deux textes. Nous préférons le terme de proximité à celui de distance. Car la distance, notion familière et abstraite dans l'esprit des mathématiciens, peut prêter, dans d'autres esprits, à des confusions engendrées par la métaphore géographique. Et il arrive à Labbé de tomber dans ce piège : « La distance est une mesure physique. Par exemple, St-Germain-en-Laye […] et Paris forment aujourd'hui une seule agglomération alors que Rouen est suffisamment éloignée pour être considérée comme une entité urbaine distincte […] Il est absurde d'objecter à cela que nous devons d'abord mesurer Paris-Lyon, Paris-

[8] Labbé me prête des propos que je n'ai jamais tenus (« le genre est tout, l'auteur n'est rien »). Il s'obstine aussi à m'attribuer, pour aussitôt la contester, une formule qui m'est parfaitement étrangère et qui ne se trouve nullement dans l'article qu'il cite. La formule que je propose depuis dix ans pour mesurer la distance lexicale est dérivée de celle de Jaccard. Elle s'écrit comme suit :

d = ((a-ab)/a) + ((b-ab)/b)

où ab désigne la partie commune aux vocabulaires *a* et *b* (*a-ab* et *b-ab* recouvrant les parties privatives). Nulle part on n'y fait intervenir la fréquence des mots (*Fia* et *Fib*) et la taille des textes *Na* et *Nb*, ingrédients de la formule qu'on m'attribue et qui se trouve répétée une fois de plus dans l'article de Labbé « Inter-textual distance and authorship attribution » (*Journal of Quantitative Linguistics*,2001, vol 8, n° 3, p 215). Enfin pour en finir avec les allégations inexactes, aucun logiciel digne de ce nom ne traite différemment les minuscules et les majuscules qu'on trouve en tête de vers (ou en tête de phrase). Labbé croit voir là un défaut qui discrédite les travaux lexicométriques portant sur les vers. Tous les utilisateurs d'HYPERBASE savent que la majuscule est neutralisée dans l'indexation et les traitements. Ils y trouveront aussi (p. 58 du manuel) la formule de Jaccard que le logiciel exploite et que je viens d'expliciter.

Lille… et pourquoi pas : Paris-Oulan Bator [9]? » Or la distance entre deux textes, c'est comme la proximité entre deux êtres ou deux cultures : elle suppose d'autres textes, plus ou moins proches, un espace où les accointances ou répulsions réciproques puissent se déployer. La distance intertextuelle est relative et n'a pas de sens si manquent les points de repères.

Et surtout la distance est multiple. Il y a bien des façons de rapprocher deux textes ou deux objets. Les mathématiciens en ont inventé des centaines. Et il y a chance que celle qu'on croit trouver a déjà été imaginée par quelqu'un autre. Ainsi nous avons eu la surprise de retrouver récemment dans une revue datant de 1989[10], la formule de Jaccard que nous avions aménagée à notre façon pour la rendre indépendante de l'étendue. Cette formule figure avec vingt autres, pareillement justifiées, et toutes établies, non sur la fréquence, mais sur la présence/absence. Nous en avons profité pour modifier notre calcul en empruntant à cette source un quatrième ingrédient jusqu'ici négligé : le nombre de mots qui ne figurent dans aucun des deux textes comparés. Car la proximité peut résulter non seulement de goûts communs, mais aussi de dégoûts partagés. En taxinomie, s'il est utile de connaître les propriétés qui appartiennent aux deux éléments comparés, et celles qu'on ne trouve que dans un seul, il n'est pas sans intérêt de savoir celles qui sont exclues de part et d'autre. En lexicométrie, le calcul exige évidemment qu'on soit enfermé dans un corpus fini afin qu'on puisse dénombrer les mots qui manquent dans la confrontation de deux textes mais qu'on rencontre dans les autres.

5- Les calculs de proximité qui font intervenir la fréquence sont beaucoup moins nombreux. Et c'est pourquoi nous avons porté intérêt à la formule de Labbé. On dispose certes de deux procédures dont l'une remonte à Muller. Dès 1968 dans son *Initiation à la statistique linguistique*[11], Charles Muller proposait l'application de la loi binomiale au calcul de ce qu'il appelait la connexion lexicale. Ce calcul reposait sur les classes de fréquence, et donc éliminait complètement la composante sémantique et thématique des textes. Nous renvoyons le lecteur aux deux

---

[9] D. Labbé, *Réponse à mes contradicteurs*, à l'adresse Internet : http://www.pacte.cnrs.fr/IMG/pdf_LabbeReponse-2.pdf

[10] F.B. Baulieu, "A classification of Presence/Absence Based Dissimilarity Coefficients", *Journal of Classification* 6:233-246 (1989).

[11] Ce manuel, publié d'abord chez Larousse, puis, en deux volumes, chez Hatier, est maintenant au catalogue de Champion.

applications que nous en avons faites, à propos de Giraudoux et de Hugo, et qui sont, à notre connaissance, sans autre exemple. La chaîne des calculs y est en effet fort longue et, si elle aboutit à un Chi2 synthétique qui évalue la proximité des deux textes comparés, elle nécessite une pondération qui amortisse l'effet des grands nombres, et donc de l'étendue des textes, sur toute mesure probabiliste[12]. La seconde méthode est en revanche très connue, très classique et très rapide. Elle est recommandée par A. Salem et J.M. Viprey, au moins pour une première approche. Il s'agit tout bonnement de l'analyse factorielle appliquée au TLE (tableau lexical entier), c'est-à-dire au dictionnaire des fréquences et sous-fréquences, que les logiciels d'indexation construisent tous à un moment ou à l'autre du traitement. En réalité, le TLE est rarement proposé en entier, car les calculs, peu légitimes dans les basses fréquences, allongeraient exagérément le nombre de lignes du tableau. Mais l'algorithme étant très rapide, des tableaux de quelques milliers de lignes (c'est-à-dire de mots différents) sont traités en quelques secondes.

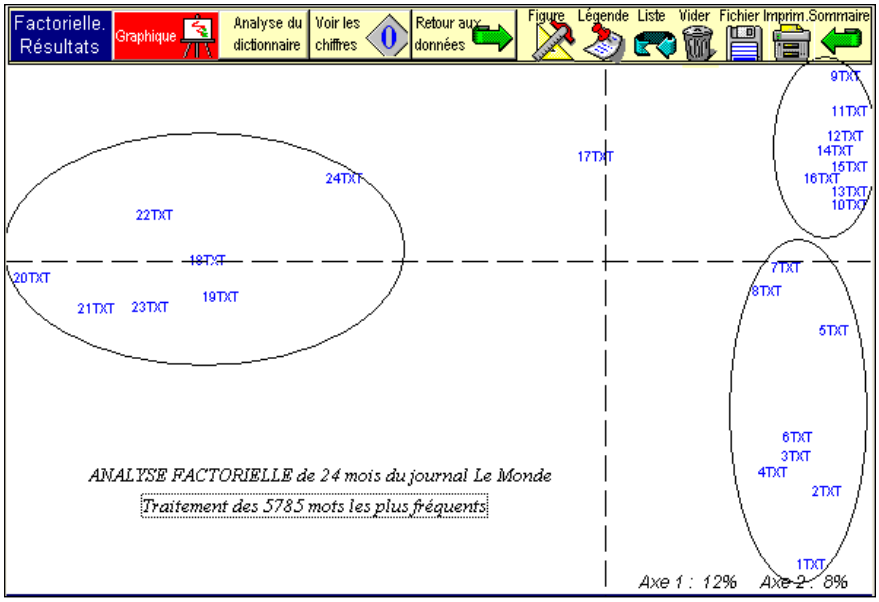


Figure 4. La distance lexicale mesurée par l'analyse des hautes fréquences
Les 5 785 mots analysés représentent plus de 5 millions d'occurrences (sur 6 millions)

[12] *Le Vocabulaire de Giraudoux. Structure et évolution*, Slatkine, 1978, p. 369-396. *Le Vocabulaire de Victor Hugo*, Slatkine, tome 1, p.277-305.

On en donne un exemple dans la figure 4 qui résume l'évolution du vocabulaire de deux années du journal *Le Monde*. Trois étapes de huit mois régulièrement échelonnés structurent le graphique. Les mois successifs qui se sont donné la main dans le temps de la réalité se donnent la main dans l'espace du graphique.

6- La formule de Labbé apporte un heureux complément à la méthode précédente. Car elle est plus sensible aux fréquences basses qu'aux mots fréquents, les premières accaparant 40% de la distance totale quand les seconds, pour une surface avoisinante, ne rendent compte que de 5% de la variance. On trouvera sur ce point, dans la revue *Corpus* (n°2, *La distance intertextuelle*, Nice, décembre 2003), la mesure détaillée que Labbé fait de la contribution des différentes classes de fréquence (et aussi des parties du discours). L'explication qui en est donnée ne nous convainc qu'à moitié : les hautes fréquences seraient plus régulièrement distribuées que les basses, mis à part quelques mots très sensibles à la situation du discours comme les pronoms personnels. En réalité l'influence prépondérante des basses fréquences vient de leur nombre. Comme il y a un vote par mot, rare ou fréquent, pauvre ou riche, la voix des puissants se perd dans la rumeur du peuple. La démocratie égalitaire y a pourtant ses limites. Labbé recommande d'éliminer les hapax et plus précisément les mots rares qu'on rencontre dans le texte le plus long et dont la fréquence théorique dans le plus court serait inférieure à 1. Il invite aussi à ne pas tenir compte des écarts inférieurs à 0,5. Ces retouches sont probablement fondées en pratique, mais elles affaiblissent la pureté de la formule et, en limitant la population appelée à voter, elles diminuent un peu le crédit de la consultation. La formule de Jaccard au contraire est dénuée de rustines et d'emplâtres. Tous les mots, hapax compris, sont invités aux urnes, même si le vote de certains est connu d'avance : les mots très fréquents ne peuvent éviter de se trouver dans la zone commune.

On comprend mieux maintenant la convergence, très souvent observée, des mesures de Jaccard et de Labbé. Quoique l'une s'attache à la simple présence et l'autre à la fréquence, toutes les deux rendent compte en priorité des basses fréquences. Et toutes les deux ont à lutter contre les perturbations que l'étendue inégale des textes peut provoquer. Elles y réussissent certes, mais imparfaitement. Considérons en effet les 75 textes du corpus classique réunissant les pièces de Corneille, Molière et Racine. Cela fait 2 775 distances à calculer, soit n*(n-1)/2. Un tri sur la longueur des textes (ou plus exactement sur le rapport d'étendue des textes deux à deux) met en relief une légère distorsion, de même nature,

dans les deux procédures. Dans les deux mesures, la distance est plus faible, et plus fiable, quand les textes sont de longueur voisine et que le rapport d'étendue s'approche de 1 (de 100 sur le graphique 5).

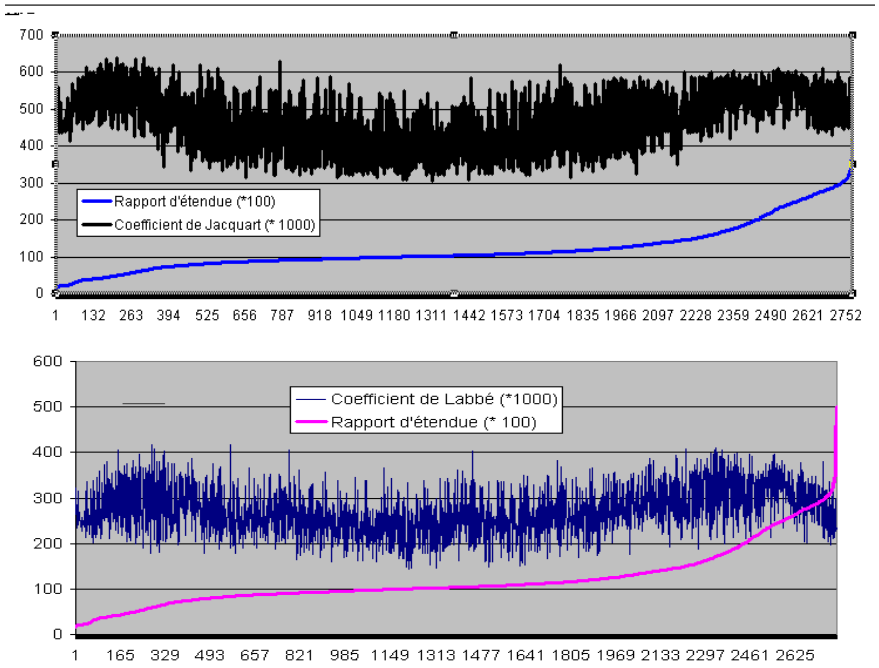


Figure 5. L'influence de l'étendue sur les distances de Jaccard et de Labbé

Cette convergence des deux approches – jusque dans les défauts – est rassurante et c'est pourquoi nous avons tenu à mettre en parallèle les deux programmes de distance dans notre logiciel. Le programme de Jaccard est d'une simplicité et d'une rapidité extrême quand on dispose du TLE. L'algorithme de Labbé est pareillement simple à mettre en œuvre mais le temps du calcul est nettement plus long et augmente exponentiellement avec le nombre de textes du corpus. Aussi avons-nous rendu cette fonction facultative. Quelques lignes de code suffisent à la traduire : on les déchiffrera aisément, si l'on sait que pour un mot donné les sous-fréquences sont cataloguées dans le tableau TABLE, *nb* étant le nombre de textes du corpus, tandis que *dista* et *disma* reçoivent la sommation du numérateur et du dénominateur de la formule. La boucle est à répéter pour chaque lemme.

Tableau TABLE

```
step k from 1 to nb-1
step l from k+1 to nb
if table[k] = 0 and table[l] = 0
continue step
end
if taille [k] < taille [l]
coef = taille [k]/taille[l]
theo = table [l] * coef
if table [k] = 0 and theo < 1
continue step
end
ecart = abs (table [k] - theo)
if ecart < 0.5
continue step
end
```

```
else
coef = taille [l]/taille[k]
theo = table [k] * coef
if table [l] = 0 and theo < 1
continue step
end
ecart = abs (table [l] - theo)
if ecart < 0.5
continue step
end
end
dista[k][l]= dista[k][l]+ecart
disma[k][l]=disma[k][l]+(table[l]+ theo)
end
end
```

3. Le cas Corneille-Molière. Un problème d'interprétation

Après avoir tenté d'évaluer et de comparer les mérites et les limites des mesures de distance, reste à interpréter leurs indications. Et si jusqu'ici nous avons suivi et assez souvent approuvé la démarche de Labbé, tout en refusant son barème, nous nous en séparons radicalement au moment crucial de l'interprétation. Dans le cas du théâtre classique, les résultats pour qui sait les lire sans idée préconçue n'invitent nullement à conclure que Corneille aurait écrit les chefs-d'œuvre de Molière. Bien au contraire, la mesure de Labbé tendrait plutôt à distinguer les deux écrivains. Considérons en effet la carte des proximités, telle que la dessine l'analyse arborée (figure 6).

Il est facile d'en détacher la branche Racine (si on peut dire) qui se dégage mollement d'abord de l'influence de Corneille (la *Thébaïde* et *Alexandre* sont proches du grand devancier) puis affirme son indépendance. On y distingue même la rupture qui à partir d'*Iphigénie* conduit l'auteur à *Athalie*. Une telle finesse dans le détail a tout pour plaire aux exégètes les plus exigeants. Mais ne cherchons pas là la comédie des *Plaideurs*. Personne n'a mis en doute son authenticité. Mais comme elle relève d'un autre genre, le calcul a déplacé cette pièce très

loin sur la gauche, au beau milieu des comédies de Molière. On chercherait vainement une autre explication : le genre est ici prédominant.

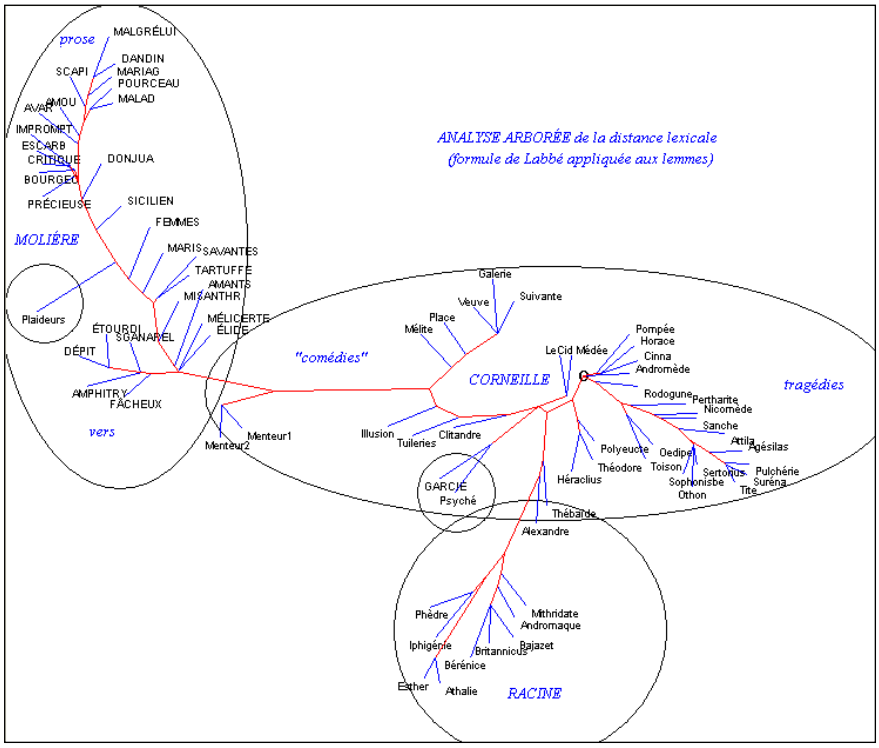


Figure 6. Analyse arborée de la distance lexicale (méthode Labbé appliquée aux lemmes)

De la même façon, la seule pièce sérieuse qu'ait écrite Molière, *Dom Garcie de Navarre*, a déserté la moitié gauche, où toutes les comédies de Molière sont rassemblées, pour se fixer dans le camp opposé, parmi les tragédies. Est-ce suffisant pour prétendre que Corneille (ou Racine) ait écrit cette pièce ? Le genre suffit à expliquer ce déplacement, comme celui de *Psyché*, qui se situe au même endroit, et dont le genre hybride (tragédie-ballet) est également éloigné de la comédie[13]. Comme Molière

[13] Dans le cas de *Psyché*, une raison supplémentaire s'ajoute à l'influence du genre : si la pièce figure bien parmi les œuvres de Molière qui en a créé et développé le canevas en prose, la versification en a été faite, en grande partie, par Pierre Corneille, comme la version versifiée de *Don Juan*, réalisée après la mort de Molière, est due à Thomas Corneille.

et Racine n'ont guère exploité qu'un seul genre, mis à part ces trois exceptions, leur individualité est fort bien circonscrite par le calcul et toutes les comédies de Molière campent à gauche, les pièces en prose en haut et les pièces en vers en bas[14], tandis que les tragédies de Racine sont serrées les unes contre les autres dans le quadrant inférieur droit. Il n'en est pas de même avec Corneille dont l'œuvre est plus diversifiée, plus étalée dans le temps et qui s'est illustré dans plusieurs genres. La surface que le calcul lui attribue est plus large, plus aplatie, et répartie en deux zones : celle des tragédies à droite et celle des pièces comiques ou assimilées à gauche. Il n'en reste pas moins que l'originalité des trois auteurs est préservée, malgré la polarisation du genre. Même les pièces de Corneille forment un bloc, dans lequel entrent les deux *Menteurs*. Les *Menteurs* se rapprochent certes des pièces en vers de Molière, près de la frontière. Mais ce sont des frontaliers, non des transfuges. Quant aux pièces de Molière, aucune ne se compromet avec les pièces de Corneille. Et l'on comprend mal que Labbé, au vu d'un tel graphique, ait pu les attribuer à Molière. En réalité au lieu de considérer le jeu d'en haut, d'un regard impartial et neutre, Labbé, barème et baromètre en mains, s'est introduit dans la partie, en privilégiant un ou deux joueurs parmi les 75 en jeu. En focalisant son attention sur les *Menteurs*, qui se situent à la frontière, il a rassemblé sous le même drapeau tous ceux qui se trouvaient dans le voisinage, et les a soumis au même suzerain (il a choisi Corneille, mais Molière aurait pu tout aussi bien revendiquer la conquête en annexant à son territoire les comédies de Corneille, de *Mélite* à l'*Illusion comique).* L'erreur d'interprétation réside dans ce parti pris que rien ne justifie. Quand on a 2 775 mesures de proximité à synthétiser, cela ne peut se faire qu'en prenant du recul, pour les embrasser du regard sans en fixer aucune en particulier. Les méthodes multidimensionnelles (l'analyse factorielle des mêmes données est aussi claire) servent précisément à élargir le champ de la vision en évitant la myopie et à faire apparaître dans le paysage les massifs et les lignes de partage.

Au besoin, avant ou après cette synthèse, rien n'interdit de concentrer son attention sur une ligne ou une colonne du tableau, par exemple celle qui correspond au *Menteur*, comme dans la figure 7. On

[14] L'influence du genre peut être complexe, car la notion de genre, comme l'a bien montré Rastier, admet des sous-catégories. À un certain niveau le choix se fait entre comédies et tragédies. Au niveau supérieur, on devrait choisir entre théâtre, roman, correspondance, essai, etc…Au niveau inférieur deux options se présentent, vers ou prose, au moins pour la comédie (car il y a peu d'exemples de tragédies en prose au XVII^{e} siècle).

constate en effet que cette comédie a des accointances fortes non seulement avec les autres comédies de Corneille, mais aussi avec celles de Molière, pourvu qu'elles soient en vers. Et, comme on l'a vu avec les deux premières pièces de Racine, l'influence de Corneille est la plus forte au début de la carrière, dans les premiers essais de Molière, l'*Étourdi* et le *Dépit amoureux*, ce qui n'en fait pas nécessairement des chefs-d'œuvre[15]. Ce gros plan sur une pièce est certes riche d'informations, mais les 74 autres contiennent autant de renseignements, parfois concordants, parfois divergents. La difficulté des taxinomies et des calculs de proximité vient de l'absence de transitivité. Si A ressemble à B et à C, il ne s'ensuit pas que B ressemble à C. C'est le nœud gordien des 2 775 coefficients entrelacés qu'il faut dénouer et il ne suffit pas de tirer sur un fil.

Beaucoup d'autres analyses viennent renforcer l'interprétation qui s'impose dans la figure 6[16]. Celle qui suit (figure 8) reprend le même corpus en lui appliquant un calcul de distance différent, expliqué précédemment sous le nom de Jaccard. Il faut bien se persuader que le programme d'analyse arborée place automatiquement tous les textes, en s'arrangeant pour que s'assemblent ceux qui se ressemblent, comme ferait avec ses invités une maîtresse de maison avisée. Les routes et les chemins sont également tracés, de sorte que le travail d'interprétation ne consiste guère qu'à reconnaître, circonscrire et désigner les agglomérations. Elles sont trois, là encore, et faciles à nommer : la première s'appelle Racine (en haut), la seconde Corneille (au centre) et la troisième Molière (en bas). Impossible de répartir autrement la population. Les trois circonscriptions sont indépendantes et franchement séparées. Si le résultat avait ressemblé à la carte des Balkans, avec des ethnies dispersées et entremêlées, le regroupement aurait pu se justifier.

[15] Les distances, multipliées par 1000, servent d'ordonnées à la représentation graphique. Elles sont lisibles dans les deux colonnes de droite. Celles que Labbé a publiées partiellement sont dans la dernière. On les comparera aux nôtres qui apparaissent dans l'avant-dernière et qui ont été calculées avec le même algorithme mais en tenant compte des ponctuations et des hors-texte et en les soumettant à la lemmatisation de *Cordial*. Nos chiffres sont légèrement et constamment inférieurs, de 1% en moyenne, ce qui n'a aucune influence sur l'analyse.

[16] Elles portent sur les graphies, les parties du discours, les structures syntaxiques, la segmentation de la phrase, la longueur des mots, les classes de fréquence, etc. La convergence est au rendez-vous mais la place nous manque pour développer ces points de vue. On est loin d'avoir tout dit sur un texte quand on a fait le relevé des lemmes. Bien d'autres aspects doivent être envisagés, qui font intervenir la syntaxe, la thématique, la métrique.

Mais ici tout est en ordre et les trois écrivains règnent sur des terres que nul ne conteste (mis à part les trois exceptions qu'on a relevées précédemment et qui jouissent de l'exterritorialité du genre littéraire). Ce n'est pas que le genre s'efface. On voit bien qu'il suggère une bipartition : toutes les tragédies sont en haut, et toutes les comédies en bas, et cela sans aucune exception. On voit aussi qu'une décantation se fait qui, chez Corneille, ne mêle pas les comédies et les tragédies et, chez Molière, les vers et la prose. On voit enfin que d'un bout du graphique à l'autre une hiérarchie s'établit entre les pièces : le théâtre d'en bas, c'est la comédie en prose, celui d'en haut, c'est la tragédie en vers, et entre les deux c'est la comédie en vers, que Corneille et Molière se partagent.

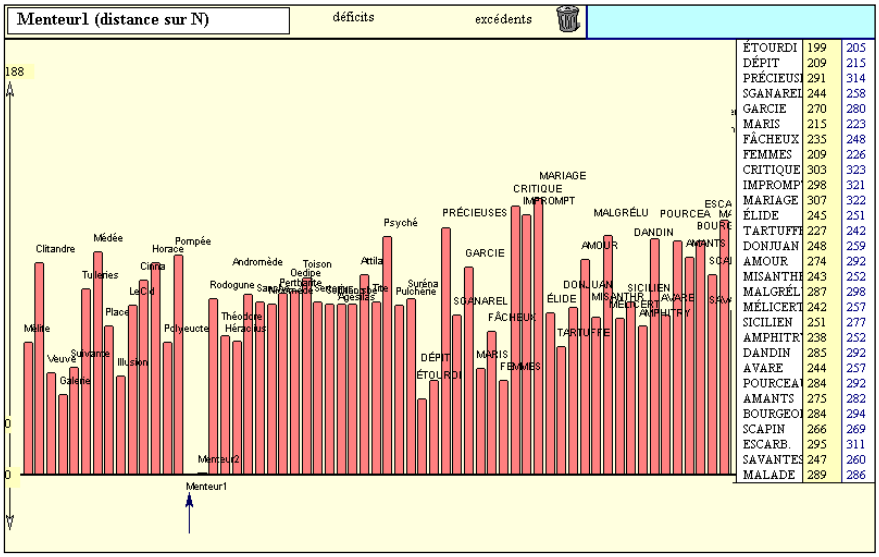
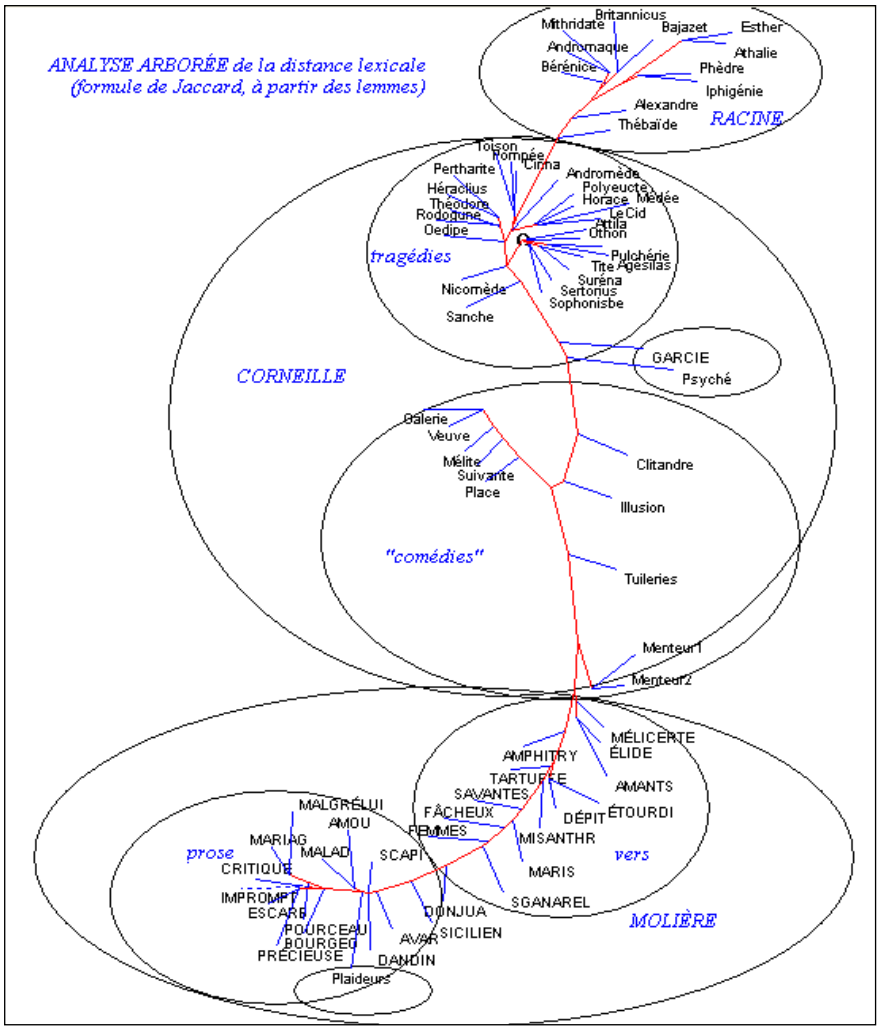


Figure 7. Distance du Menteur aux 74 autres pièces

Le plus surprenant est peut-être que l'aimantation du genre, si puissante qu'elle soit, n'ait pas dominé davantage la personnalité des trois écrivains et que le territoire de chacun soit ni nettement délimité. Les historiens de la littérature nous ont appris que leur entente a été médiocre et que chacun avait sa fierté, sa personnalité, ses ambitions, ses jalousies et aurait mal supporté qu'on lui fasse de l'ombre. Et chacun a son

originalité très reconnaissable sur le graphique. Ainsi bien loin de conforter la thèse de Pierre Louÿs, la statistique paraît plutôt l'infirmer[17].



[17] Poussé par un scrupule de dernière minute, nous avons soumis le corpus à un autre calcul de distance, connu sous le nom de corrélation de Bernouilli et proposé par Étienne Évrard dès 1966 (« Étude des dialectes bantous », in *Statistique et analyse linguistique, PUF, p.85-103).* Ce coefficient et sa variante simplifiée sont du type Jaccard (les relevés portent sur la présence/absence) et figurent en bon rang dans la liste établie par Baulieu (voir note 12). Les résultats sont tout à fait superposables à ceux du graphique 8.

Figure 8. Analyse de la distance Jaccard (appliquée aux lemmes)