Entre linguistique et littérature, un tunnel sous les mots

Etienne Brunet

Entre linguistique et littérature, un tunnel sous les mots[1]

Etienne Brunet
BCL (CNRS), Université de Nice, brunet@unice.fr

Beaucoup de gens n'aiment pas s'engager dans un tunnel. À plus forte raison si le tunnel est un tuyau où ne passent que des chiffres. Ceux qui abhorrent la statistique peuvent quitter la salle. Ils ont mon temps de parole pour prendre un café. La linguistique et la littérature sont aux deux bouts du tunnel. Aux linguistes le premier tiers de mon exposé : occasion pour les littéraires de rejoindre le café. Le second tiers est voué à la littérature : c'est le moment pour les linguistes de prendre l'air. Le dernier tiers est consacré, s'il reste du monde dans la salle, aux deux disciplines ensemble et aux rapports souterrains ou sous-marins qu'elles entretiennent et que les chiffres aident à découvrir.

La statistique est une servante, une auxiliaire ancillaire, qui ne prétend à rien d'autre qu'à servir les autres disciplines, au moins celles qui reposent sur l'observation. Elle a été bien accueillie par les sciences dures et beaucoup des sciences humaines. La médecine, la géographie humaine, l'économie, les sciences politiques, la sociologie, la psychologie en font un usage régulier, d'autant que la technologie offre peu d'autres outils pour accéder à la rigueur quand il s'agit d'étudier les faits humains. Mais un dernier carré résiste où l'on trouve l'art, la religion, la philosophie et la littérature. Là encore domine le sentiment que le jugement suffit appuyé sur le raisonnement, la culture et la sensibilité. Ce n'est pas que ces disciplines ignorent le concret et l'observation des faits, mais elles se contentent de l'intelligence pour les interpréter. La linguistique, elle qui a servi de modèle scientifique à quelques-unes des sciences humaines, n'a pas réservé à la statistique l'accueil qu'on pouvait attendre. Pendant plusieurs décennies, dans le sillage de Chomsky, les linguistes ont promené de colloque en colloque leur exemplier, abondamment garni de phrases forgées pour la circonstance et soumis à la fameuse conscience linguistique. Pourquoi s'attarder, pensaient-ils, à chercher des exemples naturels dans les textes, quand les illustrations artificielles remplissent le même office à moindres frais. Et les lexicographes n'étaient pas loin de penser la même chose quand il leur fallait agrémenter les définitions.

Je n'oublie pas pourtant – c'était avant Chomsky – une chiquenaude initiale qui avait poussé en avant sinon la statistique, du moins l'observation chiffrée des faits langagiers. C'était dans les années 60. Bernard Quemada et les chercheurs bisontins avaient mis en route le Centre d'étude du vocabulaire français, après avoir participé à une entreprise antérieure initiée dès 1953 par Wagner et Guiraud et vouée à l'établissement d'un Index du Vocabulaire du théâtre classique (les promoteurs ne pouvaient pas prévoir que ces mêmes données allaient servir, soixante ans plus tard, à une thèse audacieuse qui voit dans Molière la plume de Corneille et que nous examinerons plus loin, quand les littéraires seront revenus dans la salle). Au même moment le Recteur Imbs démarrait le chantier lexicographique qui allait devenir le TLF et où aucun exemple ne devait se trouver qui ne fût daté et signé. Pourtant dès 1959 mon maître Pierre Guiraud notait non sans quelque dépit : « La linguistique est la science statistique type ; les statisticiens le savent bien ; la plupart des linguistes l'ignorent encore. [2]», propos cités en exergue par Charles Muller dans son *Initiation à la statistique linguistique*. Si

---

[1] Article publié dans les Actes du colloque *Linguistique et Littérature, Cluny, 40 ans après,* Ablali D. et Kastberg M. (éd.), Besançon, 2010, p. 127-151.

[2] Pierre Guiraud, *Problèmes et méthodes de la statistique linguistique*, Reidel publishing Company, Dordrecht, 1949, p.15.

Guiraud vivait[3] encore (il serait quasi centenaire comme Muller ou Fontenelle), il pourrait formuler à peu près le même constat.

Certes la linguistique s'est ouverte à l'informatique et le traitement automatique du langage est en pleine expansion, comme aussi tout ce qui touche à la recherche documentaire. L'explosion d'Internet a contribué encore au développement de l'industrie linguistique. Mais la statistique reste dans l'ombre. Sans doute les moteurs de recherche l'utilisent-ils pleinement, mais sans le dire et sans montrer leurs recettes et leurs algorithmes. Un brevet ou un secret dans ce domaine peut rapporter, comme à Google, la fortune. On voit bien que la statistique bien conduite parvient à donner l'information utile en une seconde, mêlée il est vrai à quelques scories. Mais la communauté linguistique s'est peu investie dans cette technologie, qui est restée l'affaire des techniciens et des informaticiens. Sans doute les disciples lointains de Guiraud, de Muller, de Wagner ou de Quemada[4] se regroupent-ils périodiquement - et la dernière fois c'était précisément à Besançon – dans les JADT (Journées d'Analyse des Données Textuelles). La statistique y apparaît comme truchement privilégié pour accéder au texte, mais le texte c'est aussi l'affaire des historiens, des sociologues, des politologues et les linguistes n'ont pas la majorité dans cette société savante, non plus que les littéraires.

Le langage offre pourtant, comme le disait Guiraud, des facilités exceptionnelles pour les méthodes statistiques. Les données sont bon marché ou même gratuitement disponibles. En quelques minutes on peut faire son marché sur Internet. Ici et là des bases gigantesques sont interrogeables, qui délivrent non seulement l'information secondaire, par exemple bibliographique, mais aussi la source primaire, le texte original. L'abondance s'ajoute à la disponibilité, or la statistique est particulièrement à l'aise dans les grands nombres, quand le hasard est facile à circonscrire. Qu'on songe, à l'opposé, à l'inconfort du sociologue qui dépouille une enquête onéreuse, limitée et parfois lacunaire ou infidèle. Abondantes et souvent gratuites, les données en matière de langage sont sûres, complètes, contrôlables, faciles à corriger, à copier, à transmettre. En revanche, si un protocole en médecine, un entretien en psychologie, une enquête d'opinion ont été conduits incorrectement, il n'est guère possible de recommencer et le travail est perdu. Enfin les données linguistiques ont une vertu rare dans les sciences dures : c'est leur lisibilité immédiate, en dehors même du traitement statistique. Que les programmes viennent à s'égarer, en proposant des résultats absurdes, le chercheur a les moyens de récuser leur témoignage aberrant, parce la simple lecture du texte lui donne une idée point trop imprécise du résultat escompté. Dans bien des disciplines au contraire l'homme est à la merci des instruments et rien ne le protège contre la défaillance éventuelle des mesures et du traitement.

- I – Statistique et linguistique

Le texte n'est un privilège ni de la linguistique ni de la littérature. Mais certaines sortes de texte ne relèvent que de la première. Ainsi on voit mal l'intérêt que la critique littéraire pourrait porter à un dictionnaire, en tant qu'objet d'étude. Les contraintes du genre y laissent trop peu d'espace aux variations stylistiques. Mais un lexicologue y trouve son miel. Et la statistique fournit un outil approprié quand les données textuelles sont structurées ce qui est le cas des bases et des dictionnaires.

1 - À titre d'exemple on se propose d'aborder la question de l'emprunt, problème éminemment linguistique, puisqu'il met en jeu plusieurs langues. Certes l'emprunt pourrait être recherché dans les textes, dans la publicité et dans la rue. Mais il est plus rapide de

[3] Ouvrage publié d'abord chez Larousse en 1968, puis chez Hachette en 1973, dans une édition revue et augmentée en deux volumes, et enfin chez Champion où il reste disponible à la vente.

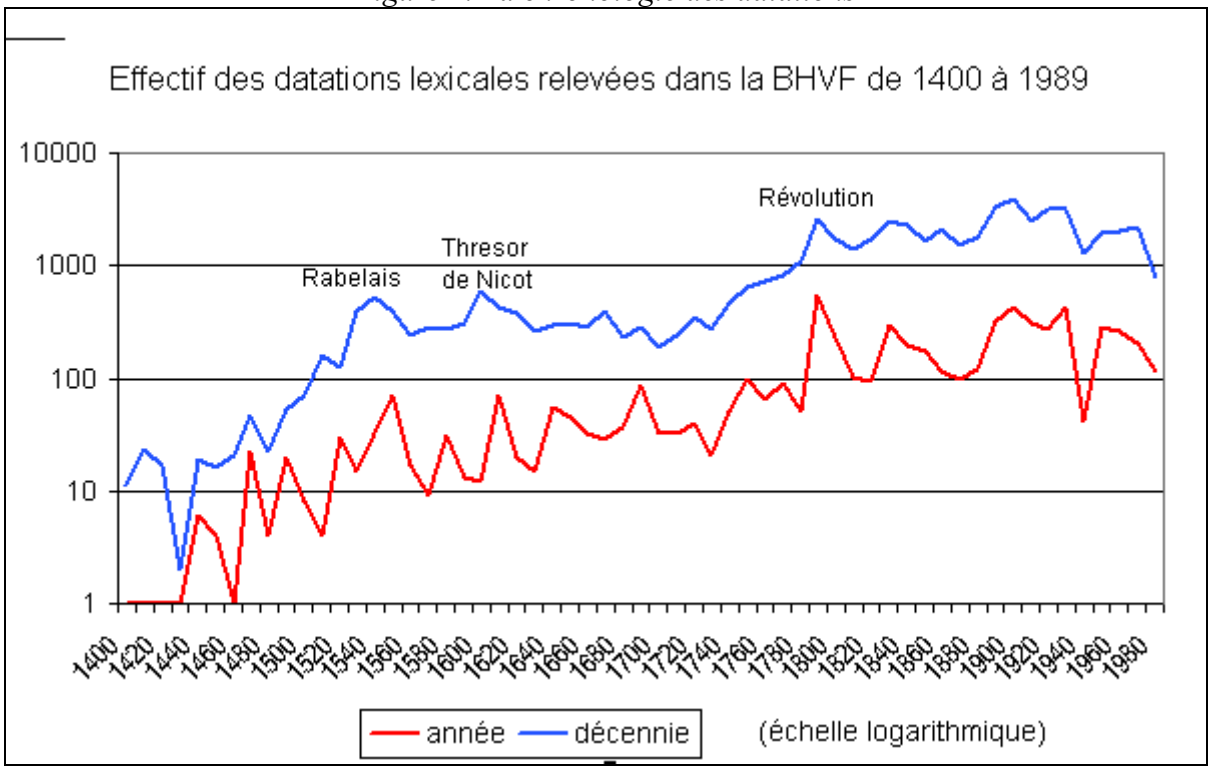
[4] À Besançon, d'autres noms viennent à l'esprit : Gentilhomme, Peytard, Massonie.

l'étudier dans les relevés déjà réalisés, par exemple dans la Base historique du vocabulaire français (BHVF), realisée à Nancy et accessible sur Internet. À vrai dire cette base issue d'un recensement gros de 50 volumes entrepris par Quemada à partir de 240 ouvrages lexicologiques, dont 42 dictionnaires, est moins un catalogue d'emprunts que de datations. Mais si l'on excepte le maigre fonds gaulois qui ne résulte pas d'un emprunt mais d'un héritage et qui ne comprend guère que 160 unités, tout dans la langue française vient de l'extérieur, principalement du latin, et de la multiplication interne des éléments importés.

Malheureusement la trace individuelle des mots se perd au-delà du Moyen Âge, parce que manque le témoignage des textes. Les datations relevées se comptent par unités au Moyen Âge pour une même décennie, par dizaines au XVe siècle, par centaines à partir de Rabelais et par milliers après la Révolution. Au total c'est plus de 50 000 mots recensés. À chacun une carte d'identité est délivrée au moment où il entre dans les textes du patrimoine et, pour la plupart et pour quelque temps, dans le dictionnaire. On a la trace contrôlable de leur passage, la preuve de leur existence à un moment de l'histoire. Mais leur survie n'est pas garantie et leur descendance n'est pas évoquée. Au reste ce contrôle douanier, fait *a posteriori*, est sujet aux lacunes et aux révisions. Bien des mots ont passé la frontière sans se faire remarquer, et ceux qu'on a épinglés n'en sont peut-être pas à leur première tentative. Pour certains d'entre eux la base signale leurs datations successives, la première attestation étant toujours considérée comme provisoire, en attendant un éventuel document antérieur.

Mais la statistique ne s'embarrasse pas de ces incertitudes. Elle est faite pour voir à travers le brouillard et l'image qu'elle donne dans la figure 1 est d'une grande clarté. La pente a été adoucie par la représentation logarithmique, mais les accidents n'en sont pas gommés pour autant. Des accès de fièvre laissent la trace des événements littéraires ou historiques, au passage de Rabelais, du *Thrésor* de Jean Nicot, ou de la Révolution. À vrai dire la statistique a-t-elle son mot à dire ici ? Nullement, le relevé, effectué année par année, décennie par décennie, n'a nul besoin de pondération, ni de pourcentages. Les effectifs absolus parlent d'eux-mêmes.

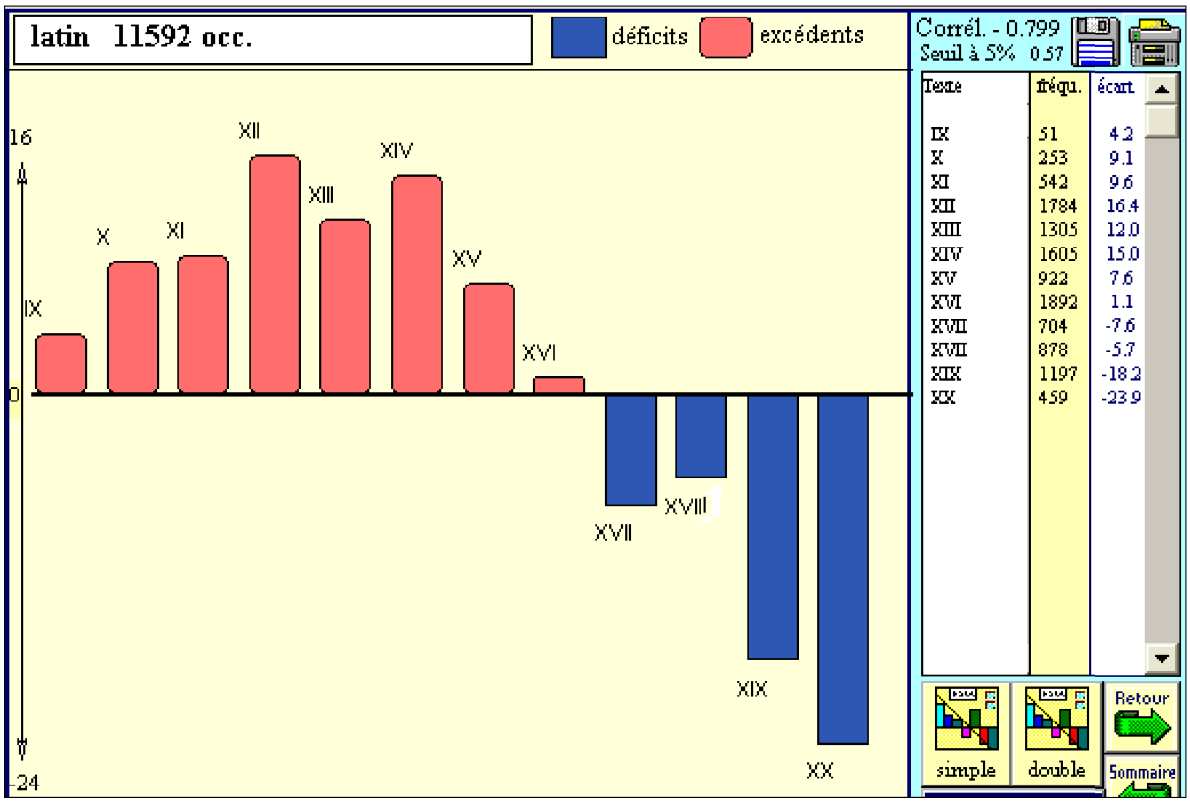
Figure 1. La chronologie des datations



2 - Il n'en va pas de même lorsque des comparaisons plus complexes doivent être faites et que les faits se présentent dans un tableau dit de contingence, avec des données inscrites au croisement des lignes et des colonnes. En donnant cette fois au mot *emprunt* son sens propre, consultons le *Petit Robert*, dans sa version cédérom. Parmi les rubriques offertes à l'interrogation figure l'origine du mot. Le recensement est laborieux parce qu'il faut renouveler la consultation pour chacune des 200 langues étrangères qui interviennent dans l'enrichissement, massif ou anecdotique, du français. L'apport du latin est bien sûr le plus lourd et apparemment constant, puisque bon an mal an chaque siècle apporte son lot d'au moins mille mots, sauf les premiers et le dernier.

siècle	IX	X	XI	XII	XIII	XIV	XV	XVI	XVII	XVIII	XIX	XX
effectif	51	253	542	1784	1305	1605	922	1892	704	878	1197	459

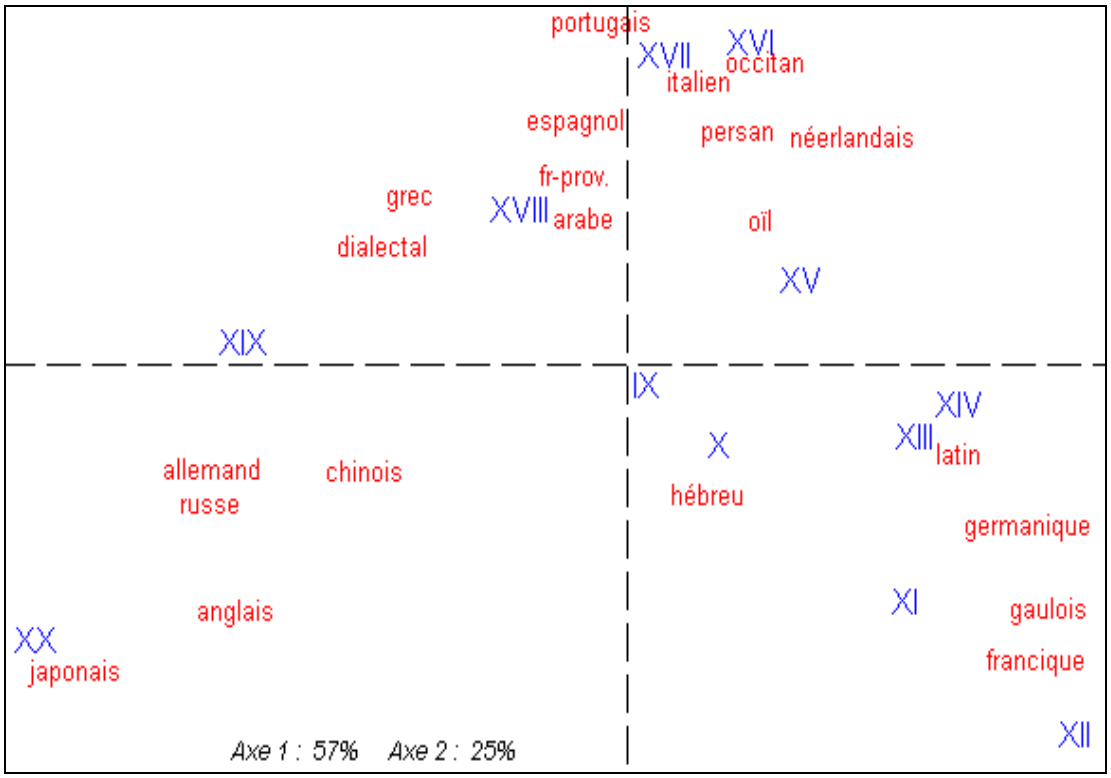
Mais la statistique propose une autre lecture car elle tient compte des autres langues, dont la plupart progressent quand le latin stagne. Les parts du marché se rétrécissent pour le latin quand on se réfère au total des emprunts. En réalité le calcul n'est pas un simple pourcentage mais un écart réduit, plus fiable.



La courbe pour le grec – qu'on associe souvent au latin - s'oppose radicalement à celle du latin. C'est au XXe siècle que le grec est le plus sollicité, surtout pour la médecine. En cela le grec se pose en rival de l'anglais, même si l'essor de l'anglais est plus rapide. Pour chacune des langues on peut dresser la courbe de l'évolution du marché. Et de la même façon on peut établir le profil de chacun des siècles et montrer en parallèle ses sources d'approvisionnement. L'analyse factorielle, réalisée dans la figure 3, rend compte en même temps de toutes les lignes (c'est à dire les langues) et de toutes colonnes (les époques). L'interprétation en est aisée : c'est le temps qui gouverne la distribution, de la droite à la gauche : dans le quadrant inférieur droit sont concentrés tous les siècles antérieurs à la Renaissance, du IXe au XIVe. La droite est accaparée par les temps modernes, XXe et XIXe, tandis que les siècles intermédiaires, du XVIe au XVIIIe, occupent la zone médiane, tout en s'écartant vers le haut du graphique. Ce profil linéaire, en forme de croissant, est caractéristique des données sérielles ou chronologiques, qu'un même courant anime.

Or sur cette chaîne du temps, dont les maillons sont régulièrement ordonnés d'un bord à l'autre, prennent place trois groupes distincts. Le gaulois, le francique et les parlers germaniques accompagnent le latin dans la zone consacrée aux origines. À l'opposé l'allemand, le russe, le japonais et le chinois font cercle autour de l'anglais et du XXe siècle. Enfin les langues latines, italien en tête, ont la faveur de la Renaissance et des siècles classiques[5]. On prêtera une attention particulière aux éléments intermédiaires : par exemple le néerlandais et les langues d'oïl tendent à se rapprocher des origines (le XVe est à mi-chemin entre les deux premières divisions). Et de l'autre côté le grec est partagé entre l'époque classique et les temps modernes.

Graphique 3. Analyse factorielle des emprunts relevés. Temps et espace.



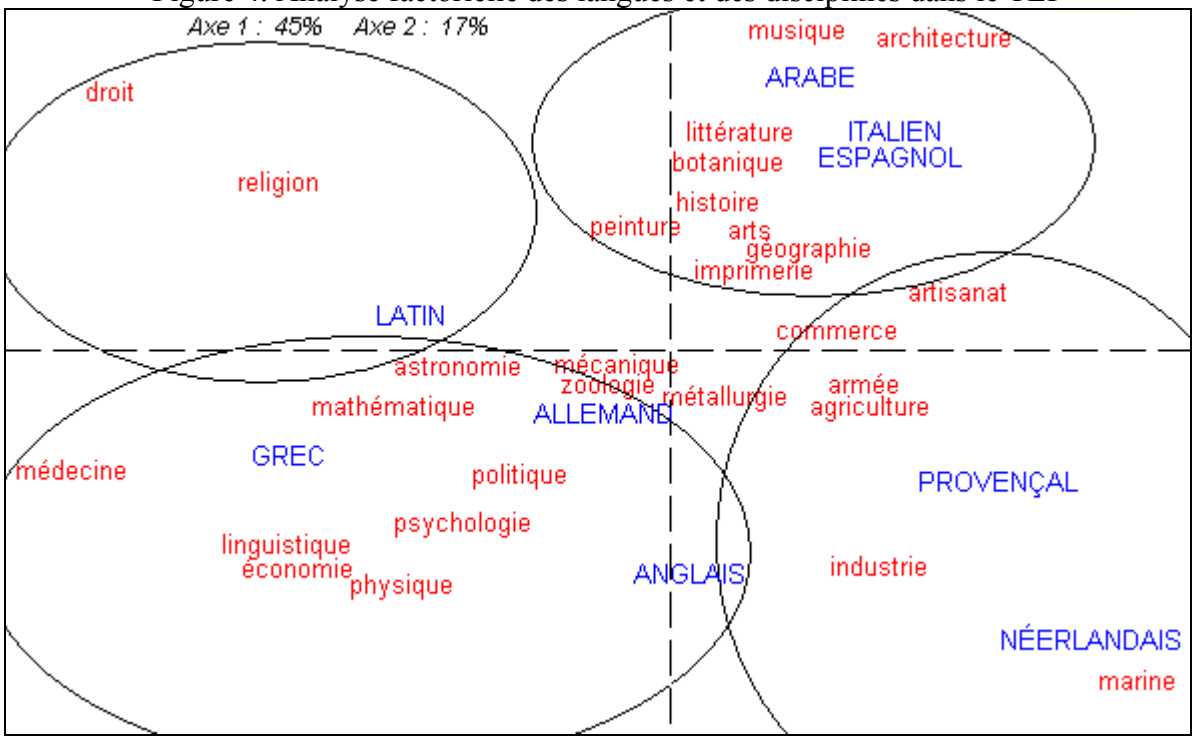
Ce n'est pas le lieu de développer comment les méthodes multidimensionnelles font la synthèse d'un tableau. Il suffit de savoir décrypter les résultats. Ici la connaissance du sujet est nécessaire pour donner un nom aux oppositions qui apparaissent dans le graphique entre la droite et la gauche, le haut et le bas et qu'on appelle facteurs, sans autre précision. La proximité de certains points, lignes ou colonnes, n'est généralement pas le fruit du hasard. Elle est gouvernée par le principe : qui s'assemble se ressemble. Mais c'est au chercheur de trouver la raison de ce rapprochement.

3 - Nous donnerons un second exemple de cette méthode, appliquée pareillement à l'emprunt. Mais l'analyse portera sur un autre tableau, venu d'une autre source. Nous abandonnons ici le Petit Robert pour le TFL dans sa version électronique, c'est à dire le TLFI. Ce dictionnaire est riche d'information concernant le domaine, lorsque cette mention est nécessaire pour un terme technique. En croisant le domaine et la langue d'origine, comme précédemment les siècles et les langues, on dispose d'un tableau qui dit à quel pays telle ou telle discipline s'est adressée pour ses emprunts. La réponse est dans la figure 4.

[5] Ces pavillons latins, surtout l'espagnol, couvrent souvent la marchandise arabe.

Les diverses sciences, groupées dans le quadrant inférieur gauche, font surtout appel au grec et à l'anglais pour satisfaire leurs besoins de terminologie spécialisée. Le latin jouit quasiment d'un monopole pour tout ce qui touche au droit et à la religion (quadrant supérieur gauche). De l'autre côté de l'axe vertical, on aborde le royaume des arts, tous réunis autour de l'italien, de l'espagnol et de l'arabe. De la littérature à la musique, de la peinture à l'architecture, toutes les voies où s'engage la création artistique mènent vers le sud. Enfin la boucle se ferme avec les activités traditionnelles (commerce, agriculture, artisanat, industrie, armée, marine) auxquelles le pays s'est consacré depuis toujours et dont la terminologie puise dans le fonds ancien de la langue. Le provençal et le néerlandais ont été alors les fournisseurs, à côté du latin et de l'héritage gaulois et germanique[6].

Figure 4. Analyse factorielle des langues et des disciplines dans le TLF



4 - Reste une question sensible : les immigrés lexicaux ont-ils un faciès ? Peut-on à première vue les reconnaître à leur timbre ou à leur coloration ? Ceux qu'irrite le franglais ont vite fait de dénoncer les graphies *oo, oa* ou *ing*, et les intrus qui ne savent pas cacher leur *k* ou leur *w*. On peut leur objecter que ces caractéristiques s'effacent au bout de quelques générations et que l'orthographe tend à les fondre dans le moule français. Beaucoup des mots qui ont franchi la Manche au XIXe ou avant ne sont plus reconnaissables, comme la *redingote* qui recouvre un *riding-coat*. Certains, que les Normands ont d'abord exportés, ont franchi deux fois la frontière, et reviennent au pays après un long séjour en Angleterre. Il n'en reste pas moins que la structure phonétique des langues empruntées laisse des traces que l'analyse statistique peut déceler, même si la conscience linguistique n'y est guère sensible. Ainsi les mots d'origine anglaise ou germanique préfèrent les sourdes *t* et *p*, quand la latinité est plus accueillante aux sonores *d* et *b*. Les voyelles accentuées (surtout *é* et *è*) sont propres au fonds ancien de la langue et les emprunts récents n'y ont guère recours.

Pour mener cette enquête, on a interrogé de nouveau le TLF mais cette fois, au lieu de se contenter de noter les effectifs, on a enregistré la liste de tous les mots qui répondaient au critère retenu : ici la nationalité des emprunts. Cependant les données obtenues pouvaient difficilement être assimilées à un texte suivi. C'était des listes, chaque mot étant un hapax.

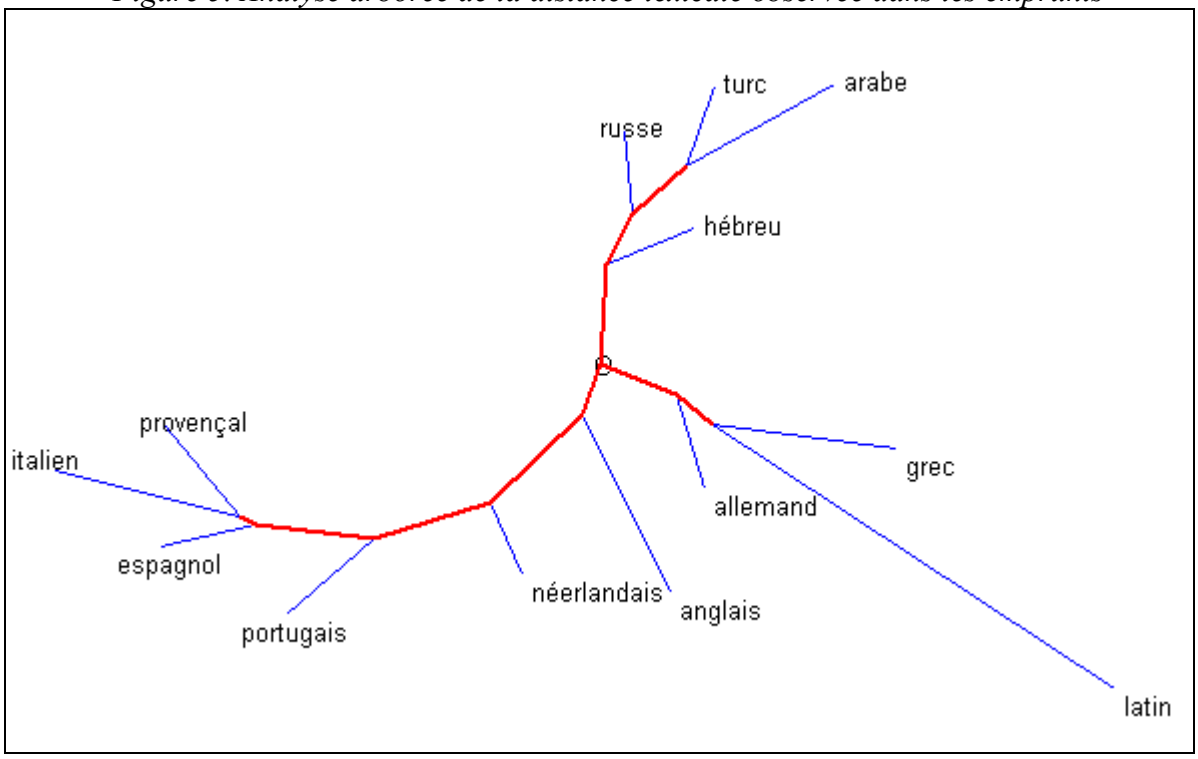
[6] Les effectifs pour cet héritage étaient trop étroits pour rentrer dignement dans les calculs.

Dès lors les calculs de distance, fondées sur la répétition des mots, n'étaient plus applicables. Le découpage en n-grammes de quatre lettres a permis de sortir de l'impasse et d'obtenir des fréquences pour les séquences retenues. En déplaçant la fenêtre progressivement d'une lettre à la suivante, le mot *découpage* aurait ainsi généré six segments élémentaires : *déco, écou, coup, oupa, upag* et *page*. Ces segments, pouvant être communs à d'autres mots, donnent lieu à une indexation et à différents calculs statistiques, comme celui des spécificités. L'anglais est friand des combinaisons *ting, pper, aker, unch* ; le grec des séquences *ique, oïde, poly, logi* et de toutes celles qui contiennent *ch*, *ph* ou *th*, par exemple *phor, graph, chro* ; l'allemand est senti confusément dans les mots qui contiennent *eiss*, *huss* ou *hall* et l'arabe se devine derrière une initiale en *ar* ou dans les combinaisons *abou, hamm, chou*. Quant aux langues latines elles partagent quelques séquences communes : *esqu, ille, elle, nade, asse, aill, ette.*

Le calcul de distance est réalisé sur la totalité des n-grammes recensés dans le corpus. Le principe qui établit cette distance est fondé sur le coefficient de Jaccard, c'est-à-dire un rapport entre les n-grammes communs et les n-grammes exclusifs. Pour deux langues dont on mesure la distance, les mots communs tendent au rapprochement, les mots privatifs à l'éloignement. Quand toutes les combinaisons des langues deux à deux ont été épuisées, on obtient un tableau triangulaire analogue à celui que fournissent les cartes géographiques, pour les distances kilométriques de ville à ville.

L'analyse arborée (figure 5), exploitant au mieux un tel tableau, en constitue un graphe, qui équivaut à une sorte de carte où la parenté explique les regroupements et les oppositions. Les langues romanes (italien, espagnol, portugais et provençal) se retrouvent entre elles à la gauche du graphique. Le grec et le latin font cause commune à droite, tandis que les langues germaniques se portent au centre. Restent les inclassables (hébreu, arabe, turc et russe), qui sont reléguées dans un *no man's land* en haut de la figure. Le *melting pot* n'est donc pas allé au bout du processus d'intégration et quelques traits physiques indélébiles dénoncent encore les origines.

Figure 5. Analyse arborée de la distance lexicale observée dans les emprunts



- II - Statistique et Littérature

Quand intervient la littérature, la statistique change de visage. Elle n'a plus affaire à des bases structurées, dont les rubriques se laissent facilement enfermer dans des tableaux, mais à du texte intégral, encadré faiblement par le réseau mou des ponctuations[7]. Il est alors indispensable de procéder à l'indexation du texte, afin d'obtenir une liste alphabétique où tous les mots seront engrangés avec leur fréquence et leurs références. On aura alors les lignes du tableau, ce qui reste insuffisant. Car le statisticien, comme le cinéaste, a des lunettes rectangulaires, il lui faut des colonnes. Pour les constituer il faut aussi procéder à la partition des données, à la séparation des textes à l'intérieur du corpus. Au croisement d'une ligne on aura donc la fréquence du mot i dans le texte j. Tous les traitements ultérieurs découlent de ce tableau initial, sans lequel la statistique, qui est toujours comparative et contrastive, n'a rien à dire. Ce tableau brut peut être livré tel quel à l'analyse factorielle, et c'est une bonne approche pour voir la disposition des textes et le compromis qui s'établit entre les forces chronologiques, génériques ou thématiques auxquelles les textes sont soumis. Mais beaucoup d'autres tableaux sont dérivés du premier, quand au gré de la recherche on exclut, isole ou cumule certains mots ou certains textes.

C'est un de ces tableaux dérivés que nous nous proposons d'examiner. Mais il est carré au lieu d'être rectangulaire. Les textes se trouvent en ligne mais aussi en colonne et à l'intersection on a l'effectif des mots que le texte i partage avec le texte j. En réalité la mesure est plus complexe, car l'effectif brut doit être pondéré par la fréquence des mots et la taille des textes. C'est une mesure de distance qui devrait permettre de classer les textes en présence et d'aider à résoudre des problèmes de datation ou d'attribution. Car les littéraires n'attendent pas des jugements de valeur sur la qualité poétique ou littéraire d'une oeuvre mais ils aimeraient disposer d'une technique sûre pour désigner la date ou l'auteur d'un texte, comme les empreintes digitales ou l'ADN peuvent contribuer à démasquer le criminel. Or il y a quelques années la presse s'est emparée d'une affaire de ce genre qui n'était pas nouvelle mais qu'on abordait pour la première fois avec l'approche lexicométrique. Il s'agit de la thèse selon laquelle Corneille aurait écrit bon nombre de pièces que la tradition attribue à Molière. Bien entendu c'est aux historiens de la littérature d'en décider et il semble que la cause, magistralement argumentée par Georges Forestier, soit entendue de ce côté-là. Du côté de la technique utilisée, un champion bisontin, Jean Marie Viprey, s'est dressé pour défendre Molière et je pourrais lui laisser la parole sur ce sujet puisqu'il est parmi nous. Mais lui et moi sommes d'accord pour dénoncer l'imprudence – voire l'impudence - d'une technique auxiliaire quand elle parle en maîtresse. 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[8] ». Or les

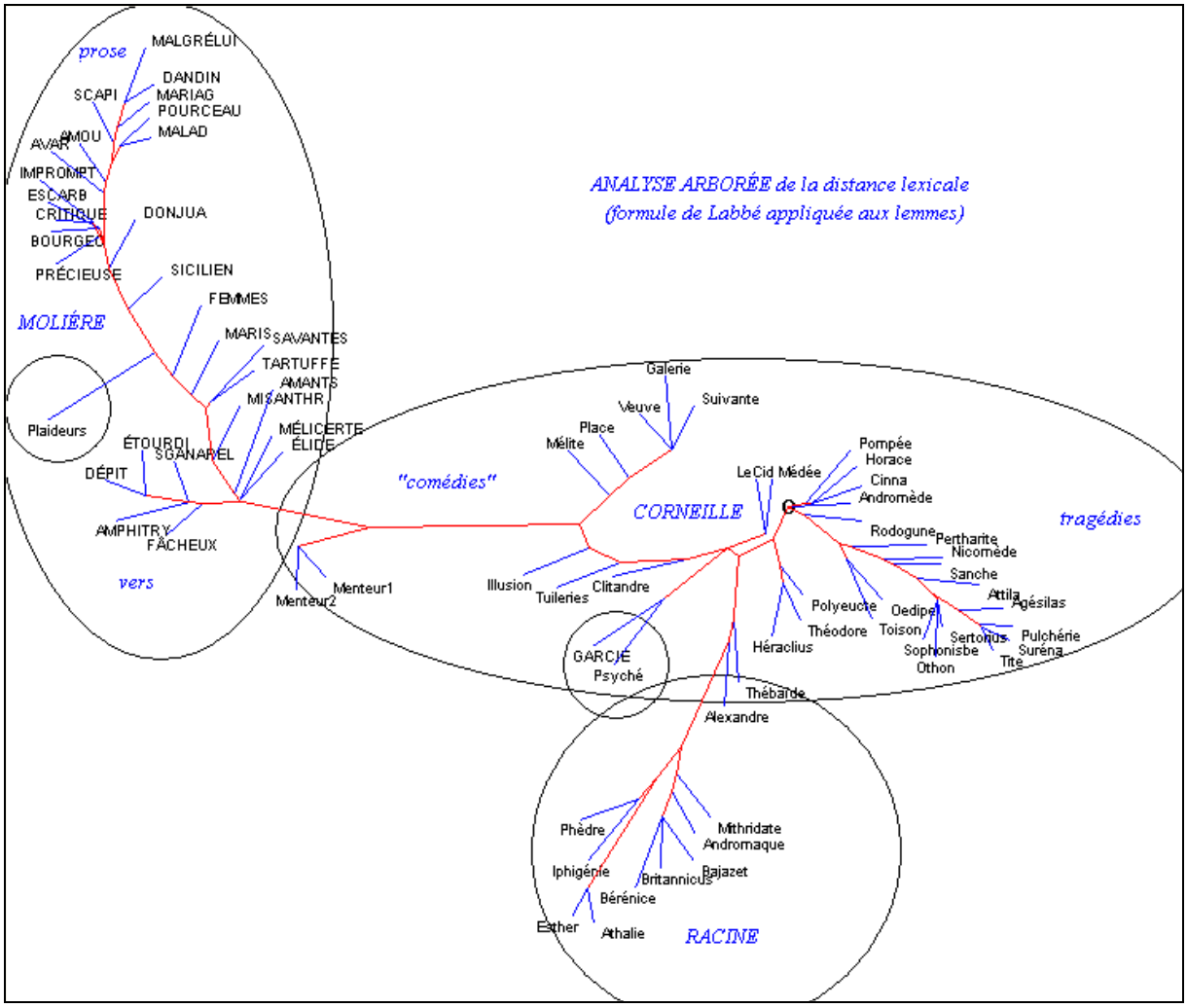
[7] L'histoire littéraire peut cependant rencontrer parfois des bases structurées, quand elle prend pour objet non les œuvres elles-mêmes, mais leur répartition dans l'espace des bibliothèques, des citations ou des livres empruntés ou dans le temps de la réception, des rééditions, des ventes, des traductions. La statistique emprunte alors les méthodes de la sociologie, appliquées au monde du livre.

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

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. L'essentiel du débat – dans la presse comme dans la discussion technique engagée entre spécialistes – a porté sur cette prétention de prouver, jugée imprudente et abusive. Tout est affaire d'interprétation et la nôtre, avec les mêmes données et les mêmes résultats, est assez différente de celle qui a provoqué le scandale.

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 utilisée tendrait plutôt à distinguer les deux écrivains. Considérons en effet la carte des proximités, telle que la dessine l'analyse arborée (figure 6).

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



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.

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[9]. Comme Molière 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[10], 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 2775 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.

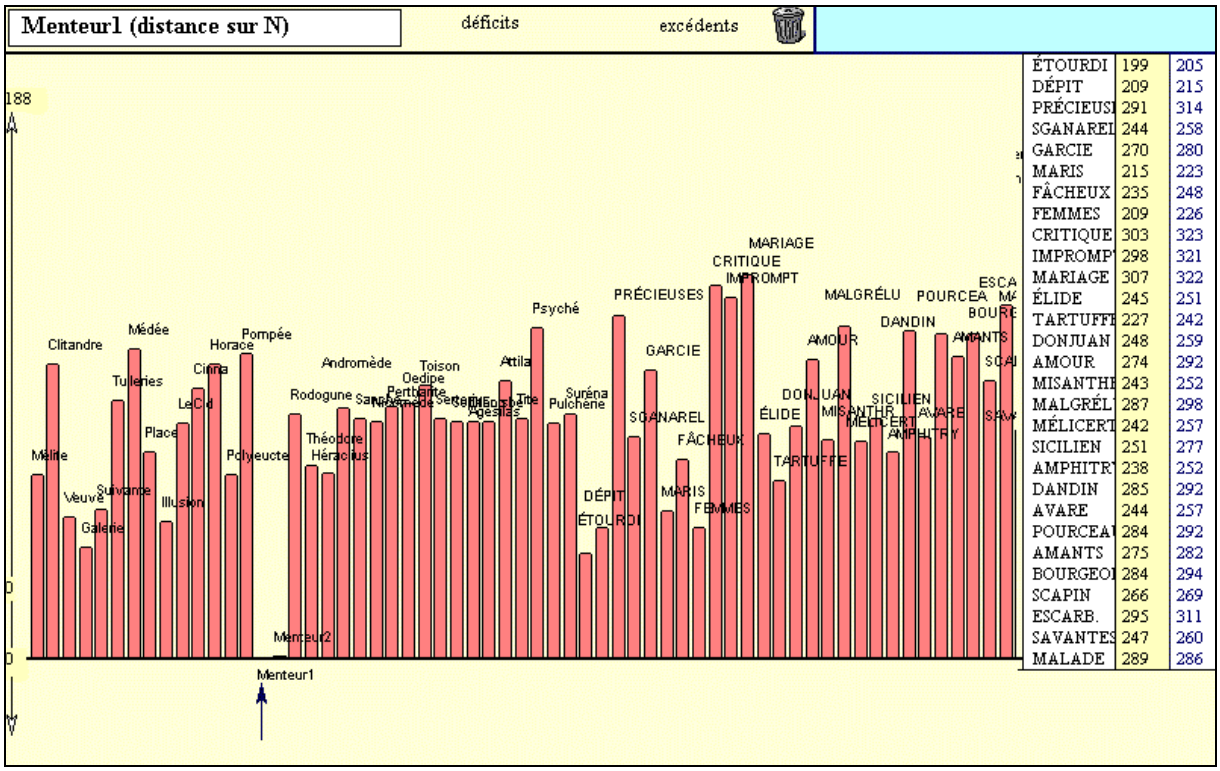
Au besoin, avant ou après cette synthèse, rien n'interdit de concentrer son attention sur une ligne ou une colonne du tableau, par exemple celle qui correspond au *Menteur*, comme dans la figure 7. On constate en effet que cette comédie a des accointances fortes non seulement avec les autres comédies 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-

[9] 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.

[10] 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 XVIIe siècle).

d'œuvre[11]. 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 2775 coefficients entrelacés qu'il faut dénouer et il ne suffit pas de tirer sur un fil.

F*igure 7. Distance du* Menteur *aux 74 autres pièces*



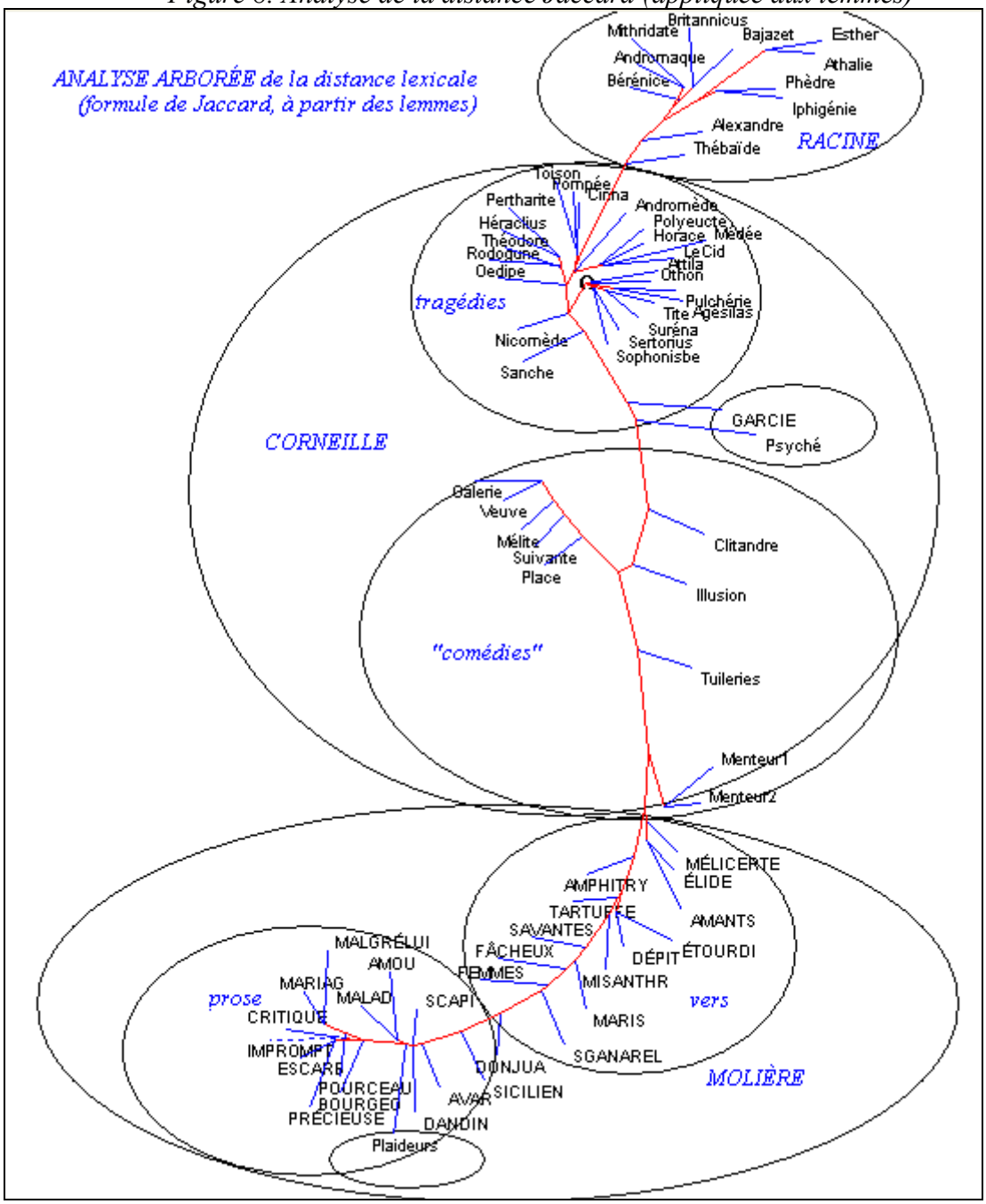
Beaucoup d'autres analyses viennent renforcer l'interprétation qui s'impose dans la figure 6[12]. Celle qui suit (figure 8) reprend le même corpus en lui appliquant un calcul de distance différent, connu sous le nom de Jaccard. Il faut bien se persuader que le programme d'analyse arborée place automatiquement 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. 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

[11] 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.

[12] 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.

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



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.

- III - Littérature et linguistique mêlées

La statistique est comme Arlequin : serviteur de deux maître qui s'ignorent. Elle peut les servir tour à tour, comme une femme de ménage qui distribue son temps d'une maison à l'autre. Ainsi peut-elle aider la littérature dans les questions de datation, d'attribution ou de caractérisation, comme elle peut guider la linguistique dans la description, l'évolution et l'extension des faits langagiers. Dans ces deux interventions les domaines sont clairement distingués : si les observations et les comptages portent sur les dictionnaires, la visée ne peut être que linguistique ; s'il s'agit au contraire de données primaires, comme l'œuvre d'un écrivain, il y a chance qu'on s'oriente vers la littérature.

En réalité, même dans les cas les plus exemplaires comme ceux que nous avons choisis, littérature et linguistique ne sont pas très éloignées. S'occuper de datations et d'emprunts c'est à n'en pas douter une entreprise de linguiste. Mais qui a déposé les matériaux sur la table du linguiste sinon la tradition littéraire, à qui on doit l'essentiel des textes sauvegardés et des témoignages. Et dès la figure 1 le lexicologue ne pouvait éviter de citer Rabelais. Inversement l'étude comparée de Corneille et de Molière débouchait sur l'opposition de la comédie et de la tragédie, sur la distinction des genres. Il y a un abus de langage à toujours parler des « genres littéraires », comme si la linguistique n'avait rien à voir avec le genre. Il lui revient pourtant d'expliciter ce qui sépare l'écrit de l'oral, l'utilitaire du littéraire, le document et la fiction.

Or la lexicométrie peut aussi jouer, comme la stylistique, un rôle fédérateur et réunir dans une même approche l'objet littéraire et l'objet linguistique. Depuis que la linguistique de corpus a fait du texte son champ exploratoire en lui accordant valeur de témoignage, sinon de preuve, tout texte réellement produit relève de sa compétence, qu'il s'agisse de brochures techniques, d'entretiens oraux ou de productions littéraires. Les textes littéraires n'ont pas droit à quelque privilège, comme s'ils étaient la quintessence du langage ; la linguistique ne doit pas non plus les exclure, comme s'ils en étaient la perversion, le ver esthétique ayant corrompu le fruit. Or s'appliquant à un texte littéraire, les observations et les leçons de la lexicométrie ont une portée ambiguë, et l'on hésite à décider si leur intérêt est linguistique ou littéraire. Certes traditionnellement, depuis l'exemple de Muller, on distingue les faits de structure lexicale qui mettent en effectifs des propriétés formelles et qui relèvent de la linguistique, et l'étude du contenu lexical qui considère les mots individuellement avec leur signification propre et leur environnement, ce qui se rapproche de l'interprétation thématique. Mais un mot remarquable rencontré dans son contexte peut susciter conjointement l'intérêt du linguiste et du littéraire, le premier s'attachant généralement à la forme ou à la syntaxe, le second à la signification et aux connotations sémantiques, quoique les points de vue puissent être échangés.

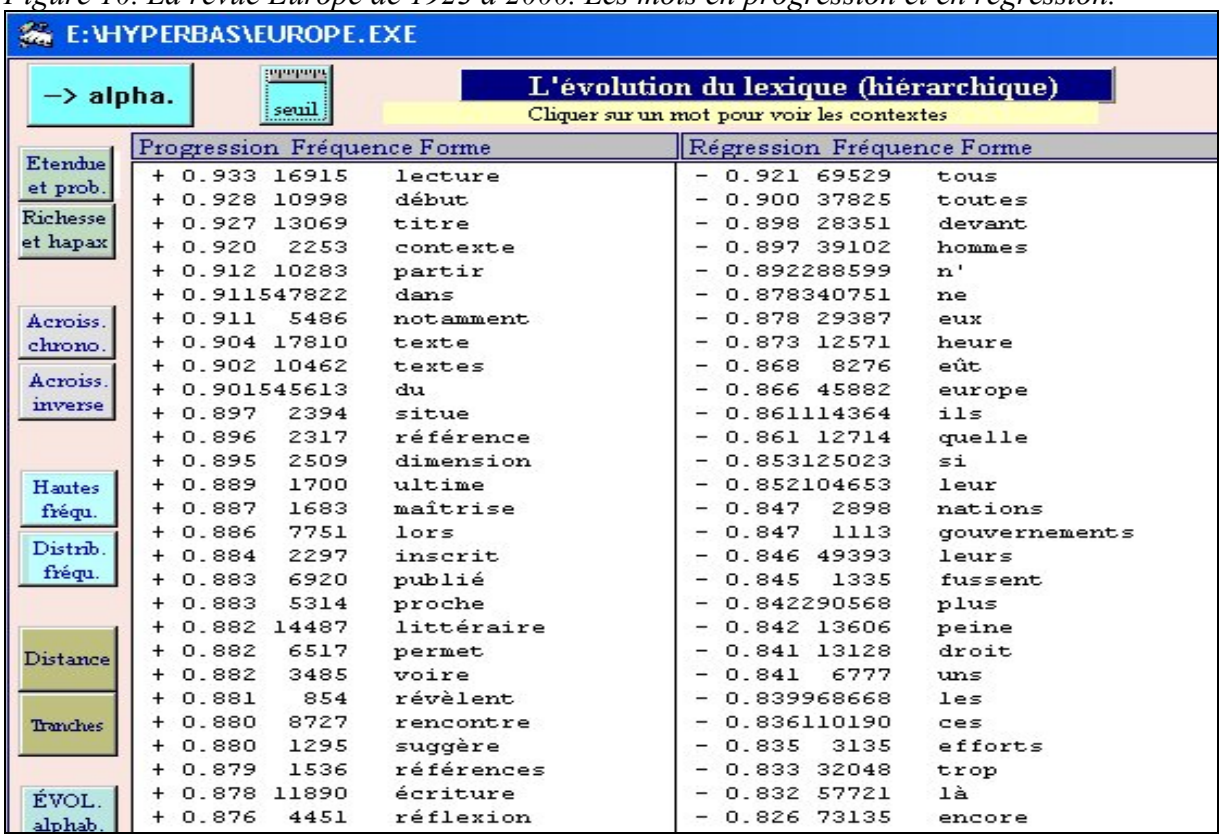
Si les perspectives ne sont pas échangées ou inversées, du moins sont-elles juxtaposées. Et la statistique a l'habitude de produire ses résultats sans faire la distinction. Ainsi dans le tableau 9 elle livre une liste de spécificités proustiennes en mêlant les mots outils et les mots sémantiques. La recherche thématique s'attachera plutôt aux acteurs, aux activités, et aux sentiments de la comédie mondaine (*princesse, baron, duc, amies, prince, marquise, tante, dîners, relations, plaisirs, jalousie*), tandis que le grammairien relèvera les ingrédients qui entrent dans la longue phrase proustienne : principalement les conjonctions et les relatifs et les négations (*que, qu', si, quand, qui, ou, laquelle, dont ne, n', pas*). En réalité l'exégète dans le même mouvement recueillera toutes ces informations et bien d'autres, moins évidentes, qui suivent ce court extrait.

Tableau 9. Les spécificités de Proust

83.30	1624	703	princesse
72.25	124605	9601	avait
66.23	431159	24669	que
65.21	27426	3104	avais
64.32	1094	448	baron
62.03	1639	542	duc
60.15	13585	1855	eût
59.43	300597	17631	qu'
56.85	235834	14223	elle
53.67	148707	9597	était
45.51	5755	878	jusqu'
43.72	25043	2272	chez
43.47	103601	6604	si
43.13	997	298	eusse
42.36	295523	15613	qui
42.30	1279768	57701	de
40.94	2518	488	personnes
39.32	6414	845	ayant
37.22	73938	4739	être
36.75	788	227	amies
36.69	95805	5805	même
36.26	72689	4626	où
35.76	6432	792	laquelle
35.55	1712	346	eussent
35.32	2329	414	prince
34.79	557	178	marquise
34.70	136680	7637	comme
34.05	22559	1833	car
34.03	157506	8557	mais
33.58	267396	13464	ne
33.33	31993	2353	été
33.08	2423	403	amie
32.84	3644	517	dîner
32.22	2936	445	tante
32.20	8573	895	plaisir
32.17	14637	1299	celle
31.59	201936	10374	pour
31.52	12214	1129	pu
31.39	408692	19341	d'
30.87	974	221	jalousie
29.97	29466	2101	avoir
29.71	889	203	plaisirs
29.04	23016	1715	elles
28.45	305816	14610	pas
28.37	15416	1258	seulement
28.22	174180	8859	plus
27.84	51596	3166	quand
27.77	3456	443	autrefois
27.58	1683	280	relations
27.41	205106	10154	n'
27.12	3253	419	réalité
27.09	129511	6777	par

Encore la *Recherche du temps perdu* constitue-t-elle un corpus clairement littéraire. Mais il arrive que le corpus ait un statut ambigu, comme celui de la revue *Europe* que Henri Béhar a publié récemment sur DVD[13]. Il s'agit d'un corpus énorme, de 58 millions de mots (7500 auteurs pour 28000 articles). Et de 1823 à 2000 c'est presque un siècle qui s'offre à la vue.

Figure 10. La revue Europe de 1923 à 2000. Les mots en progression et en régression.



[13] *La revue* Europe *en texte intégral*, Europe , 4 rue Marie-Rose 75014 Paris, 2005.

C'est certes un panorama littéraire puisque le mot littéraire est dans le sous-titre de la revue et que les numéros accueillent des comptes-rendus et des critiques de l'actualité littéraire. Il contient même, du moins au début, des œuvres courtes, poèmes ou nouvelles. Mais la revue constitue aussi un témoignage sur l'évolution de la pensée et de la langue au XX^{e} siècle. La figure 10 dresse la liste des mots qui progressent (à gauche) ou qui régressent (à droite) au cours de la période étudiée. La revue tend visiblement vers la critique littéraire, alors qu'à ses débuts elle ne cachait pas ses idées politiques, sociales et philosophiques.

Les deux perspectives se croisent d'autant plus facilement que la lexicométrie ne se contente plus de traiter les graphies, ce matériau composite où les marques syntaxiques se mêlent aux éléments sémantiques. Le traitement s'étend maintenant au lemme, en rejoignant l'ambition première des pionniers de la discipline. Le progrès n'est pas dans la sûreté des analyses, mais dans l'automatisation du codage. Qu'il s'agisse, pour le français, de Cordial ou de TreeTagger, que le processus de lemmatisation repose sur des règles linguistiques ou sur des algorithmes stochastiques, on a accès présentement à des corpus très larges, dûment désambiguïsés et étiquetés. La station d'épuration trie et sépare les homographes, dépouille la graphie de sa gangue flexionnelle et propose des produits dérivés et raffinés, dont le lemme et le code grammatical. Le lemme mène à l'analyse thématique, le code à l'analyse syntaxique, de quoi satisfaire à la fois les exigences littéraires et linguistiques. Mieux même un codage sémantique est parfois disponible, quoique encore insuffisant, qui devrait permettre une saisie directe des thèmes. Et pareillement les combinaisons des codes, systématiquement relevées, donnent accès aux structures syntaxiques.

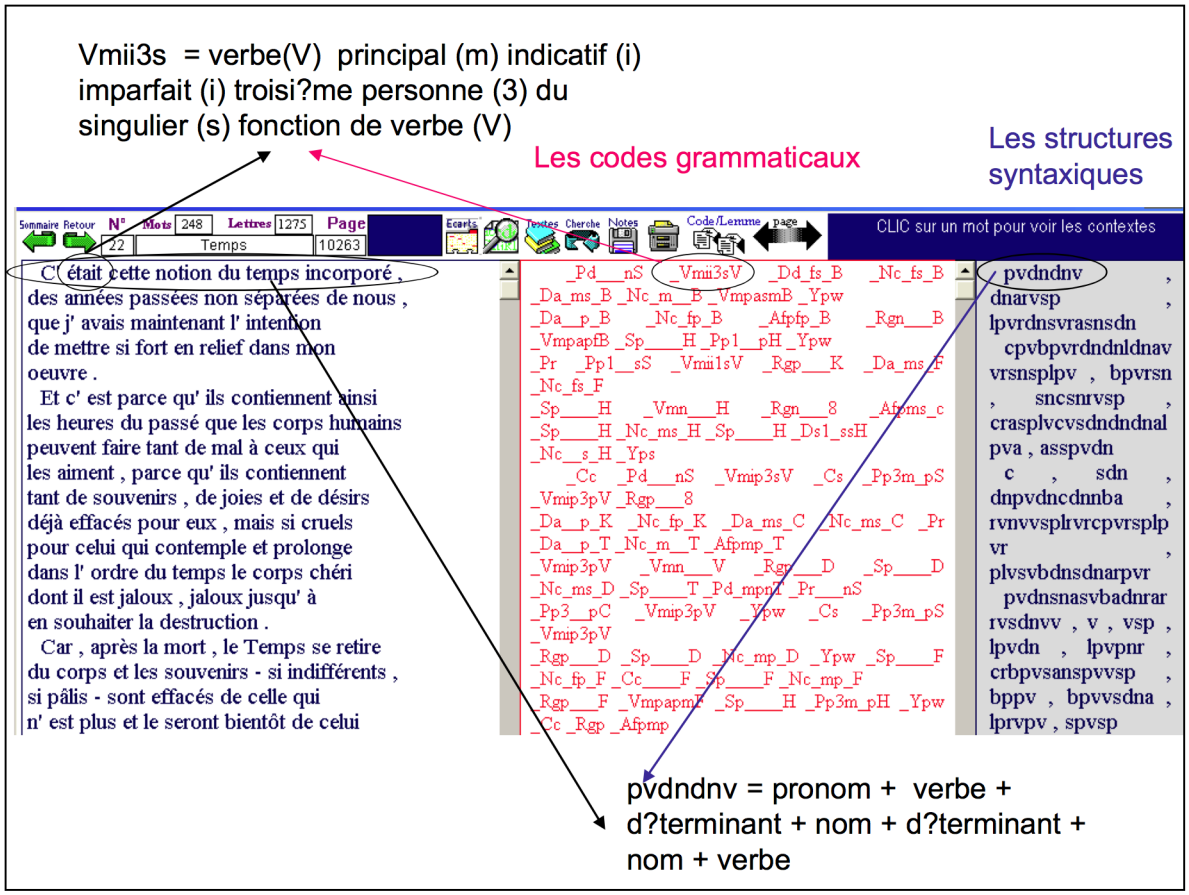


Figure 11. Un texte codé et lemmatisé (traité par Cordial*)*

Voici ainsi transformée une des dernières pages du *Temps retrouvé*, représentée en trois colonnes : graphies, codes grammaticaux, séquences syntaxiques (figure 11). Une quatrième colonne est utilisée qui n'a pas trouvé place sur le graphique et qui est facile à reconstituer : Elle correspond aux lemmes, aux entrées du dictionnaire avec ajout d'un code simplifié pour distinguer les homographes. La première ligne s'y inscrit comme suit : *ce_5 être_1 ce_7 notion_2 du_7 temps_1 incorporer_1.*

La recherche documentaire s'applique à chacun de ces quatre niveaux qui sont rigoureusement alignés et qui communiquent entre eux. Il est ainsi possible d'extraire les contextes qui contiennent tel ou tel code grammatical ou telle séquence syntaxique. La statistique, bien entendu, se donne libre cours sur les quatre plans et c'est l'occasion de voir s'ils s'accordent et si les littéraires qui portent leur regard plutôt sur les graphies et les lemmes auront les mêmes résultats que les linguistes dont l'intérêt risque d'aller aux codes et aux séquences syntaxiques. , Nous choisirons un corpus témoin qui est associé habituellement à notre logiciel Hyperbase et qui est particulièrement adapté à l'expérimentation de la distance intertextuelle, puisqu'on y trouve réunis une vingtaine d'œuvres romanesques, à raison de deux textes par auteur (choisis au début et à la fin de la carrière de l'écrivain afin d'augmenter la distance entre les textes de la même plume). Il s'agit de tester si la machine peut reconnaître la parenté des textes écrits par le même auteur, le choix allant de Marivaux à Proust. C'est le même programme qui nous a servi précédemment à comparer Molière à Corneille et à Racine.

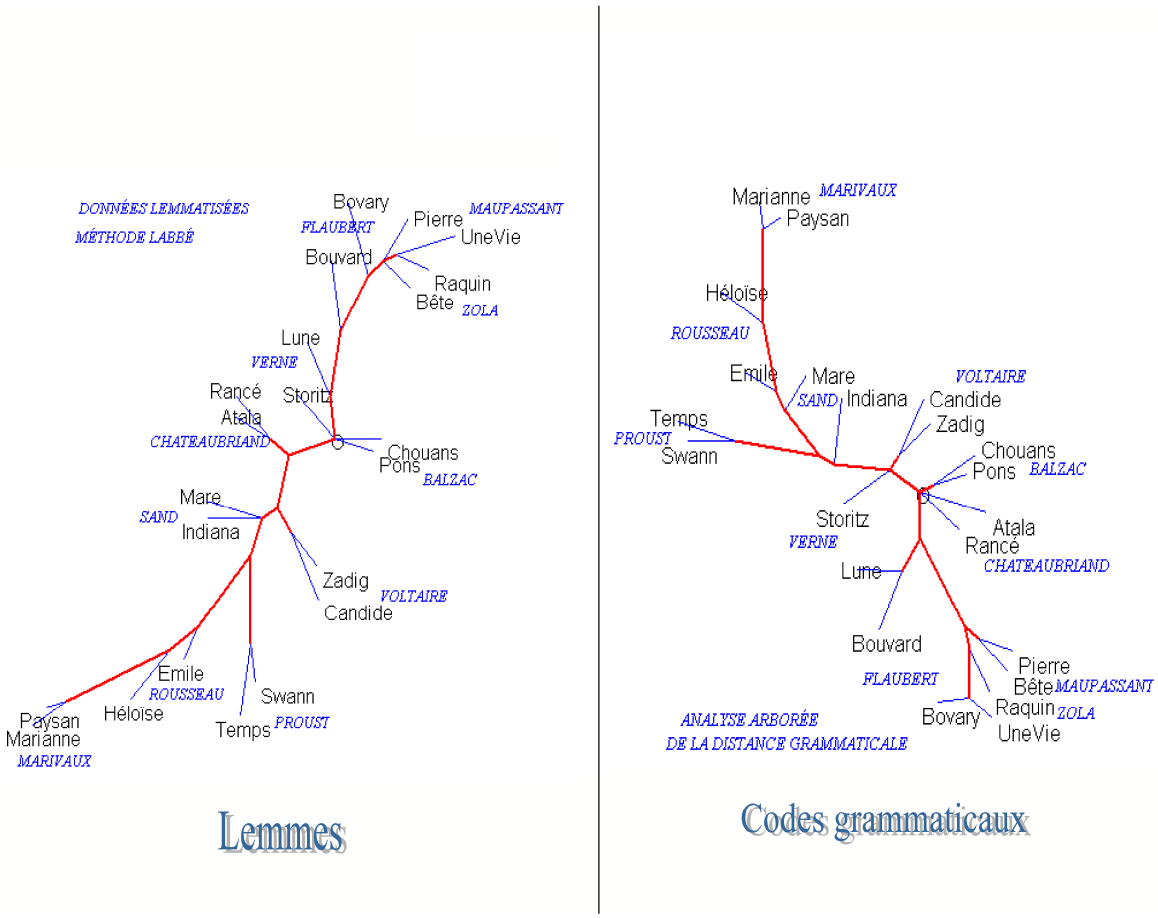


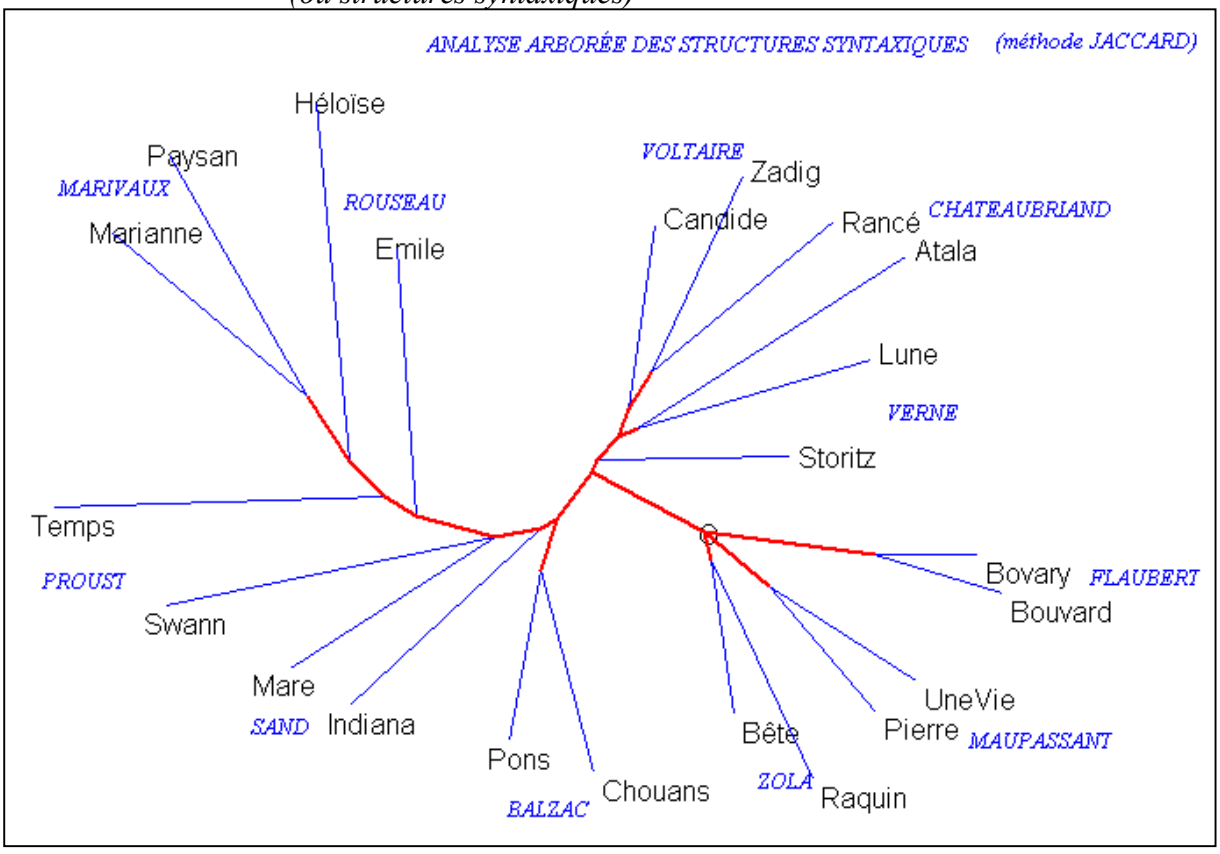
Figure 12. Convergence des analyses faites sur les lemmes et les codes grammaticaux

Or qu'il s'agisse d'un objet plutôt littéraire (les lemmes) ou d'un objet plutôt linguistique (les codes), le classement et la représentation des textes sont les mêmes. Sans savoir le moins

du monde que *Zadig* et *Candide* appartiennent à la même plume, le calcul réunit ces deux textes comme il apparie *Les Chouans* et le *Cousin Pons* et tous les autres couples. Il n'y a guère d'hésitation que pour Jules Verne, ce qui se comprend car il y a 40 ans *entre De la terre à la lune* et *Le secret de Wilhelm Storitz* et Verne qui devenu vieux se porte candidat à l'Académie française a changé sa manière et n'écrit plus pour les enfants. Ailleurs la distance entre deux textes que trente ans séparent dans la carrière d'un écrivain n'est pas assez grande ni dans les thèmes ni dans le style pour tromper la perspicacité de la machine. Mais le plus étonnant n'est pas dans la justesse du calcul. Il est dans la convergence des deux représentations[14]. Et pourtant aucun lien nécessaire n'est exigé a priori entre le sens et la forme, entre la thématique et la grammaire[15].

Un pas de plus et, le contenu sémantique s'éloignant encore, on ne livre plus à la machine que des séquences de codes. Difficile de savoir de quoi on parle quand on lit de tels rébus, comme le segment *pvdndnv* qui signifie *pronom +verbe +déterminant +nom +déterminant +verbe* et qui correspond à la première ligne de la figure 11. Le calcul reste imperturbable et dans ces segments vides de sens, où ne subsistent que des bouts ne syntaxe, la plume de Voltaire se distingue de celle de Rousseau et Proust se tient très loin de Flaubert et de Zola. Car si l'orientation générale suit la chronologie, cela n'empêche certains auteurs d'être en avance sur leur temps (c'est le cas de Voltaire) ou bien en retard ou à côté (c'est le cas de Sand et surtout de Proust qui loin de continuer la veine réaliste et naturaliste se rapproche du roman psychologique en faveur au XVIIIe siècle).

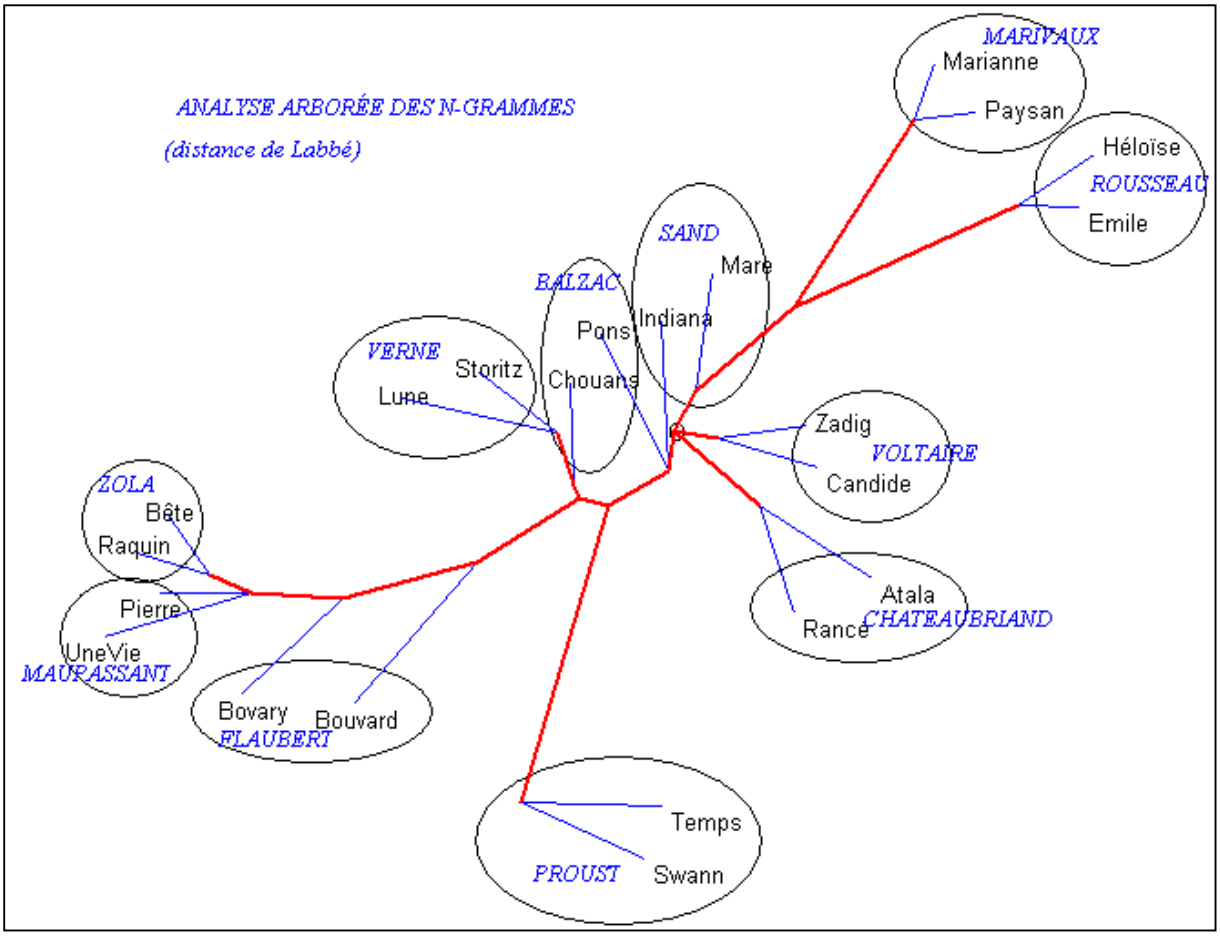
Figure 13. Analyse factorielle prenant en compte les séquences de codes (ou structures syntaxiques)



[14] On n'attachera aucune importance à l'orientation. Les deux graphiques sont superposables, si une rotation des 90 degrés s'applique au second.

Certains esprits peuvent s'étonner de cette convergence inattendue. Ils crieront sans doute au scandale si on pousse l'analyse jusqu'à l'extrême limite, soit en ignorant la notion de mot et en découpant la chaîne graphique en n-grammes[16] qui n'ont ni queue ni tête, ni sens, ni codes, soit en ignorant l'alphabet pour ne plus garder qu'un codage binaire qui ne veut connaître que deux caractères : voyelle et consonne. La première expérience est une perversion des données qui met la machine dans la situation de Champollion devant les hiéroglyphes. Certes le sens du message ne sera pas décrypté aussi bien que celui de la pierre de Rosette. Mais les textes seront classés exactement comme ils le sont dans les graphiques 12 et 13. La dernière épreuve est la plus radicale. Certes on a gardé les blancs et les contours des mots. Mais on a perdu l'alphabet. Ne restent que deux symboles C et V pour signifier l'emplacement des consonnes et des voyelles. L'appauvrissement est considérable puisque tous les mots de deux lettres n'ont le choix qu'entre CV, VC ou VV. Il n'y a pas beaucoup plus de variétés de 3 lettres et au total les 19156 hapax du texte original ne sont plus que 607.

Figure 14. Analyse arborée d'un corpus réduit à un codage consonne/voyelle.



[15] Des calculs supplémentaires permettent de préciser les thèmes ou les parties du discours préférés ou sous-représentés dans chacun des textes. On constate que Flaubert, Balzac et Chateaubriand se rangent dans le camp du substantif, quand Rousseau, Marivaux, Sand et Proust choisissent le verbe.

[16] Nous avons choisi des n-grammes de 4 lettres, avec fenêtre coulissante. Ainsi le mot écrivain génère 5 n-grammes : écri criv riva ivai vain.. Les mots de moins de 4 lettres sont ignorés. Les paramètres n'ont rien d'obligatoire : les fenêtres peuvent être successives et le blanc peut être considéré comme un caractère ordinaire. Les résultats n'en sont guère affectés.

La confusion est générale : la même configuration peut désigner des dizaines ou des centaines de mots différents. Nul ne reconnaîtra la dernière ligne de la *Recherche du temps perdu*[17] dans une chaîne plus obscure encore que celle du génome (lequel au moins dispose de quatre lettres) : vcccv cvccvvccvc cvcc cv cvvcc cvcc cvcvc cv ccvcvc cvcc cv cvcc. Les deux millions de mots du corpus ont pris cette forme sibylline. Et pourtant le miracle – ou le scandale – se produit : imperturbable le calcul discerne un ordre dans cet amas informe et retrouve la trace des textes et de leur auteur (figure 14).

Devant de tels prodiges, le littéraire et le linguiste sont réunis dans la même perplexité. Ils croyaient avoir loti le texte, chacun ayant sa part bien définie. La statistique passait de l'un à l'autre sans déplacer la frontière et dévoiler le secret du voisinage, comme Arlequin dans la pièce de Goldoni. Mais en servant séparément ses deux maîtres, littérature et linguistique, elle finit par les servir ensemble, à la même table. Comme dans la pièce de Goldoni.

[17] « …entre lesquelles tant de jours sont venus se placer – dans le temps ».