Peut-on mesurer la distance entre deux textes?

Étienne Brunet

Peut-on mesurer la distance entre deux textes?

Étienne BRUNET
UMR 6039, CNRS, Université de Nice

Résumé : Le présent exposé tente d'explorer et de comparer les méthodes qu'on a proposées jusqu'ici pour mesurer la distance entre deux textes. Les formules sont diverses, et s'appliquent tantôt à la fréquence, tantôt à la présence/absence. Et l'objet mesuré varie grandement (graphies, n_grammes, lemmes, classes de fréquence, codes grammaticaux, structures syntaxiques ou sémantiques). L'expérience montre pourtant que la convergence est au rendez-vous.

La distance entre deux textes, c'est comme la distance entre deux êtres ou entre deux cultures. Il ne semble pas qu'on puisse appliquer là la mesure. C'est pourtant à la mesure qu'on soumet les ossements, les tombes, les ruines et toutes les traces que peut laisser l'homme derrière lui, des excréments à l'ADN, et qui permettent d'établir des distances dans le temps et l'espace. N'y a-t-il pas dans les textes des éléments contrôlables et matériels qui puissent donner prise à l'observation et à la quantification? C'est le rôle que beaucoup ont voulu confier à l'ordinateur: détecter les sources et les emprunts à la manière du détecteur de mensonges, dater les textes à la manière du carbone 14, délivrer des certificats d'authenticité ou de paternité, à la manière des empreintes digitales ou des séquences d'ADN. Hélas, les problèmes d'attribution ou de datation sont les plus épineux qui soient quand on ne dispose pas de témoignages externes ou historiques et qu'on doit se fonder sur la seule analyse interne du texte. Cette difficulté ne tient pas seulement au caractère approximatif des mesures, que cache la précision illusoire des décimales, mais surtout à la multiplicité des points de vue, des angles et des perspectives, l'objet à cataloguer étant aussi rebelle à la géométrie et à la régularité qu'un rhizome de gingembre ou de topinambour. On peut tout au plus isoler un caractère et le mesurer, en appliquant à tel ou tel caractère du texte ce que Proust dit du temps, que nous n'hésitons pas à

mesurer, même si la mesure est approximative:"…des gens sans perspicacité spéciale, voyant deux hommes qu'ils ne connaissent pas, tous deux à moustaches noires ou tout rasés, disent que ce sont deux hommes, l'un d'une vingtaine, l'autre d'une quarantaine d'années. Sans doute on se trompe souvent dans cette évaluation, mais qu'on ait cru pouvoir la faire signifie qu'on concevait l'âge comme mesurable." Ainsi le discours est-il senti confusément comme mesurable, même si la mesure n'est pas attachée à un élément unique, aisément identifiable, non plus que l'âge des gens. L'âge se devine en un clin d'œil aux rides, à la fatigue du regard ou de la peau, à la démarche, à la silhouette, à de multiples détails dont la synthèse est immédiate. Mais dans la fiche signalétique d'un individu, l'âge n'est qu'un paramètre parmi des milliers d'autres, dont la mesure est difficile et dont on ne sait s'ils sont liés ou indépendants. Mis à part le cas des vrais jumeaux, comment évaluer la distance entre deux individus? Faut-il se fonder sur des antécédents communs, des goûts partagés, des similitudes physiques ou morales? Le problème est si complexe qu'on renonce souvent à démêler les causes qui engendrent les amitiés et les amours.

Quand il s'agit des textes les difficultés sont du même ordre. Certes le texte a une réalité matérielle qui se prête à l'analyse. Les éléments comptables peuvent être soumis aux instruments de mesure, des plus menus (les atomes des lettres) aux plus volumineux (les grosses molécules des structures syntaxiques, des schémas narratifs ou topoi, des constellations lexicales ou thématiques). Mais il y a tant de tests, tant de mesures, d'indices, de dosages et de stylogrammes que le jugement reste en suspens. On sort perplexe du laboratoire, comme on sort perplexe, cardiogrammes sous le bras, après un bilan de santé. Le logiciel Cordial - qui est sans doute le meilleur qu'on puisse trouver sur le marché français - propose ainsi plus de 200 mesures ou pourcentages, prolongeant à l'extrême l'analyse. Mais à l'heure de la synthèse, la baudruche se dégonfle et l'appréciation stylistique qui en résulte n'échappe ni à l'impropriété, ni à la trivialité, ni à l'incohérence. Car il reste à établir une hiérarchie dans la batterie des tests, en retenant ceux qui semblent les plus discriminants.

Pour notre part nous craignons non seulement la parcellisation et l'incohérence des résultats mais le danger de se tromper dans l'interprétation, même quand ces résultats semblent clairs et convergents. Même lorsqu'une distance paraît établie solidement entre deux textes, on ne sait pas toujours à quoi la rattacher. À l'auteur? À l'époque? Au sujet traité? Au genre littéraire? Une expérimentation réalisée avec Charles Muller il y a dix ans laisse entendre que l'influence du genre est souvent prépondérante. Un corpus avait été constitué en croisant trois écrivains et trois genres. Le facteur chronologique avait été neutralisé, puisque les trois écrivains choisis : Hugo, Lamartine et Musset appartiennent à la même époque et sont généralement classés dans la même école romantique. De même le facteur thématique ne trouvait guère à s'exercer, puisque l'expérience portait sur les 60 mots les plus fréquents, qui sont largement indépendants du sujet traité (les pronoms personnels, jugés trop sensibles à cette influence, avaient été écartés). Il s'agissait donc d'exploiter un tableau de 60 lignes et de 9 colonnes. À l'intersection on avait relevé, par exemple, la fréquence du mot "mais" (ligne i) dans le roman (ou le théâtre ou la poésie) de Hugo (ou de Lamartine ou de Musset) (colonne j). Soumis à l'analyse factorielle, le tableau révélait qu'on était en présence de trois auteurs: un romancier, un dramaturge et un poète. Le genre avait effacé les vraies signatures.

- I – Les méthodes

Nous nous proposons aujourd'hui de renouveler l'expérience sur une base beaucoup plus large. L'examen portera cette fois sur tout le vocabulaire et non pas seulement les hautes fréquences. Et divers corpus, monographiques ou non, seront explorés où les variables en cause: auteurs, époques, genres, registres entreront en jeu. Voir liste ci-dessous (figure 1). Et surtout on s'efforcera de varier les méthodes de mesure, répertoriées parmi beaucoup d'autres dans la figure 2.

Figure 1 - Les corpus exploités…

Figure 2 - Les formules proposées:

JACCARD (A,B) = $|\{\mathbf{A}\} \cap \{B\}| \ / \ |\{A\} \cup \{B\}|$
= AB / A+B , en notant par AB l'intersection des vocabulaires A et B
En particulier D = ((A - AB) / A) + ((B - AB) / B)
(proportion du vocabulaire exclusif dans A + proportion du vocabulaire exclusif dans B)

SALTON : Espace Vectoriel
Voir la thèse de Bénédicte Pincemin, sous presse.

MINKOSWSKI (A,B) = $[\Sigma^{n}_{i} p_j \times (x_{aj} - x_{bj})^{q}]^{1/q}$
cas particulier : la distance euclidienne pondérée pour q = 2
$d(A,B) = [\Sigma^{n}_{i} p_j \times ((x_{aj} - x_{bj})^{2}]^{1/2}$ (la sommation Σ concerne les i lignes dans les j colonnes, la pondération pj s'impose quand les colonnes n'ont pas le même poids)

MAHALANOBIS…KOLMOGOROF…HAMMING…

MULLER Calcul de la connexion lexicale par la méthode binomiale, en mesurant pour chaque classe de fréquence l'effectif, observé et théorique, dans les deux textes A et B comparés[1].

[1] Voir le chapitre consacré à la connexion lexicale in *Principes et méthodes de statistique lexicale,* Hachette Université, 1977, p. 145-154, ouvrage réédité dans la collection *Unichamp* des éditions Champion.

LABBÉ[2] (A,B) = Σ^{n}_{i} (réel$_j$ - théo$_{jk}$) / (réel$_j$ + théo$_{jk}$)
La sommation Σ^{n}_{i} concerne les i lignes dans les j colonnes. L'indice k désigne la colonne mise en relation avec la colonne j, la fréquence théorique du mot i dans le texte j étant proportionnelle à la fréquence de ce mot dans le texte k.

ANALYSE FACTORIELLE, étendue au tableau lexical entier (mis à part les fréquences basses). Méthode appliquée par L. Lebart et A. Salem[3].

1 - La méthode qui vient le plus naturellement à l'esprit est celle de Jaccard. Sans citer ce devancier, Muller l'avait proposée il y a trente ans tout en soulignant son défaut majeur qui est d'être très sensible aux écarts d'étendue entre les deux textes comparés et que nous avons heureusement pu corriger. Le principe est simple. Il se borne à établir, pour deux textes à comparer, le rapport entre les mots qui sont communs aux deux textes et ceux qui n'appartiennent qu'à l'un des deux. Chacun des deux quotients (dont la somme constitue la mesure de la distance) est le rapport, pour un texte donné, du vocabulaire exclusif au vocabulaire total. Il évolue nécessairement entre 0 et 1. La somme a donc pour limites 0 et 2 (et la moyenne 0 et 1). Pour chaque paire considérée, la distance obtenue tient compte de l'étendue de l'un et l'autre vocabulaires, selon la formule:

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). Dans cette formulation améliorée la somme se situe autour de 1 et reste insensible aux différences d'étendue des deux textes mis en parallèle. Observons en effet que les deux quotients évoluent en sens inverse et d'un même pas, quand s'accroît l'inégalité d'étendue des textes. En une telle situation le plus petit texte aura du mal à affirmer son indépendance face au plus gros, et son quotient d'exclusivité se rapprochera de zéro. Mais pour la même raison, le texte le plus long aura un gros contingent de termes exclusifs qui échapperont par la force des choses au plus

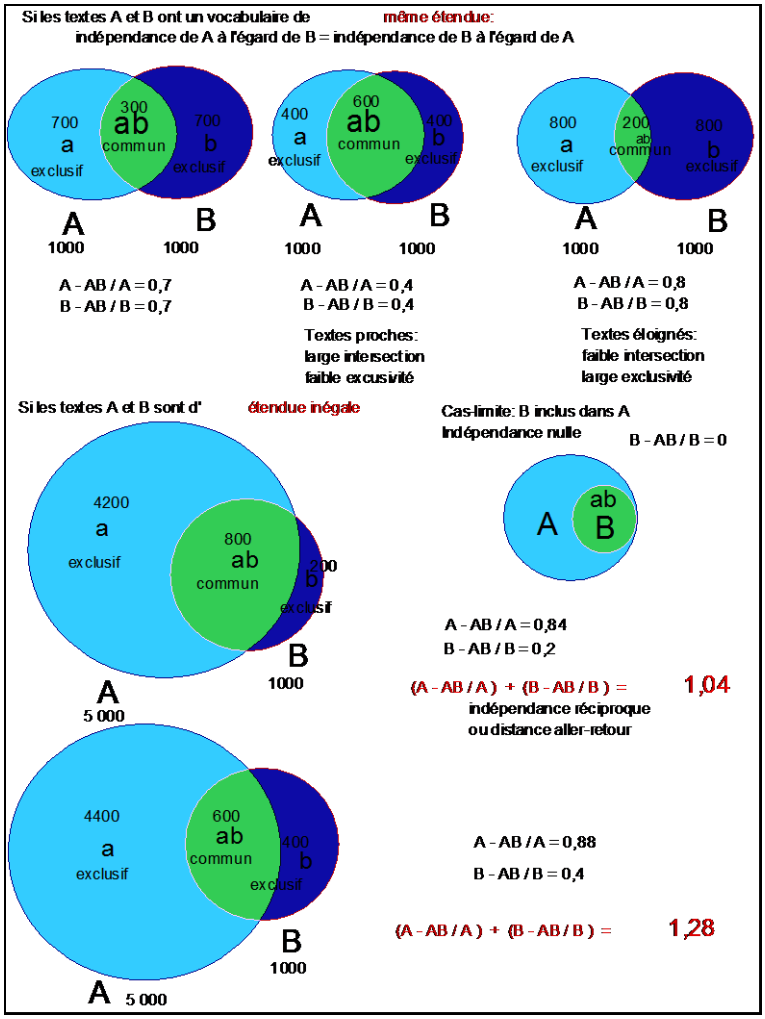
---

[2] D. Labbé, D. Monière, *La connexion intertextuelle…*, Actes du Colloque *JADT 2000*, Lausanne, p. 85-94. Dans cet article la formule qu'il nous prête - et qu'il conteste - nous est parfaitement étrangère et ne figure aucunement dans la publication qu'il cite.

[3] L. Lebart, A. Salem, *Statistique textuelle*, Dunod, 1994, p.137-147.

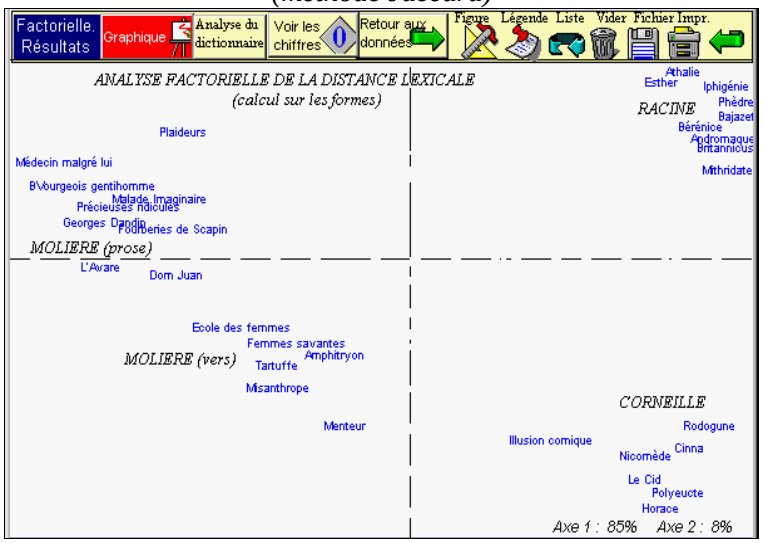
petit, et son quotient d'exclusivité tendra vers 1. Au total on observera une neutralisation mutuelle de ces deux mouvements opposés.

Figure 3 La distance de Jaccard (améliorée) (test sur présence/absence)



C'est au théâtre classique, issu de FRANTEXT, que sera emprunté l'exemple (figure 4) qui illustre cette première méthode. Pour un corpus donné la distance est calculée pour tous les appariements de textes deux à deux et la mesure prend place dans un tableau triangulaire assez analogue au tableau à double entrée qui, dans les atlas géographiques, rend compte des distances kilométriques de ville à ville. L'analyse factorielle, en se fondant sur ce tableau des distances, redéploie les villes (ou les textes) sur la carte. Cette carte montre que les trois auteurs accaparent chacun un coin du graphique. Les deux auteurs de tragédies s'opposent d'abord à la comédie de Molière (c'est le premier facteur, gauche vs droite), avant de se distinguer l'un de l'autre sur le second facteur (haut vs bas). Ce second facteur agit aussi sur les comédies: celles qui sont en prose (y compris les *Plaideurs*) se détachent des pièces en vers (y compris le *Menteur*). Remarquons au passage que le premier axe reflète plutôt l'influence du genre que celle de l'auteur, puisque les *Plaideurs* et le *Menteur*, infidèles à leur créateur, rejoignent à gauche le clan des pièces comiques (en s'éloignant cependant le moins possible de leur origine).

Figure 4. Analyse factorielle de la distance lexicale (méthode Jaccard)

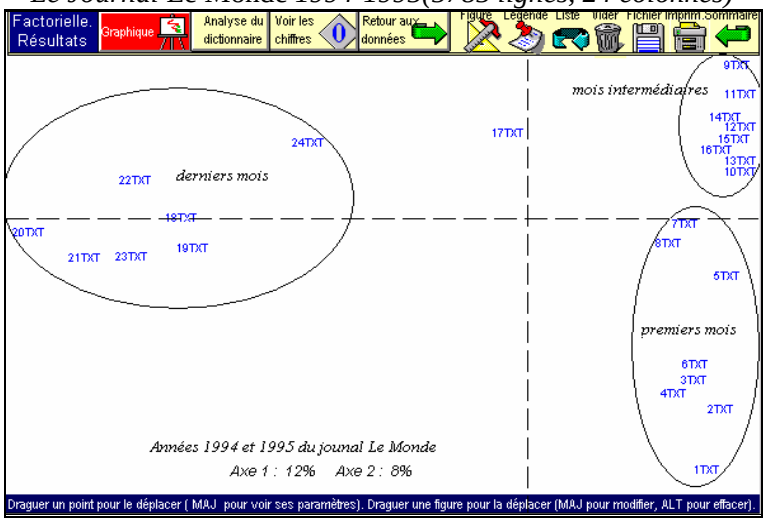


2 - D'aucuns ont observé que ce calcul faisait la part belle aux raretés du vocabulaire et particulièrement aux hapax, au détriment des fréquences plus courantes. Les classes de fréquence élevée perdent ainsi tout poids dans le calcul, puisqu'elles se trouvent nécessairement dans la partie commune et inévitable du vocabulaire (*ab*). On estime que la distance ainsi mesurée est surtout sensible aux variations thématiques, les paramètres stylistiques s'attachant plutôt aux mots de fréquence supérieure. Mais ce calcul peut être jugé trop sensible aux artefacts que peuvent produire l'inconstance de l'orthographe, les fautes de frappe, l'abondance des noms propres, bref tous les phénomènes, parfois mineurs et négligeables, qui engendrent la multiplication des formes. Certains considèrent que c'est donner trop d'importance à l'excentricité et qu'une véritable appréciation de la distance entre deux textes doit considérer, pour un même mot, le dosage des fréquences dans les deux textes comparés. Il est évident que si le partage des fréquences est inégal (par exemple 1 occurrence dans le texte A et 19 dans le texte *B*), il contribue moins à rapprocher les deux textes que si la répartition était équilibrée, soit 10 occurrences dans chacun (en considérant que les deux textes sont de même étendue). Dans les deux cas le calcul précédent rangeait le mot à l'intersection des deux textes, ne tenant compte que de la présence/absence, en ignorant les disparités des fréquences.

Pour tenir compte de la fréquence, il y a une solution radicale qui consiste à considérer ce que A. Salem appelle le *TLE* (Tableau Lexical Entier). Dès qu'un corpus lexicométrique est constitué, on dispose sous une forme ou sous une autre du dictionnaire des fréquences dont chaque ligne est consacrée à la distribution fréquentielle d'un mot ou d'un lemme dans le corpus. Il sert généralement à extraire les spécificités d'un texte ou la représentation graphique d'un mot. Mais on peut le soumettre en entier au programme d'analyse factorielle, lequel ne répugne pas à traiter de grandes masses, comptant plusieurs milliers de lignes (même au delà de 100 000). Il y a pourtant une limite qui tient non au programme, mais à la nécessité méthodologique d'écarter les fréquences trop faibles, où le calcul de l'écart réduit perd sa légitimité, à moins qu'on propose à l'analyse les

fréquences brutes, auquel cas on risque d'être abusé par l'effet de taille, c'est-à-dire la domination des textes les plus étendus et des mots les plus fréquents. L'analyse qui suit (figure 5) rend compte de deux années du journal Le Monde, soit un corpus de 6 millions de mots. La séquence chronologique y est fort visible qui groupe les huit premiers mois dans le quadrant inférieur droit, les huit mois suivants dans le quadrant supérieur droit et les sept derniers à l'extrême gauche, le 17e mois faisant la transition. Seuls les mots de fréquence supérieure à 75 ont été pris en considération (soit 5785 lignes dans le TLE). Mais le résultat reste stable si l'on baisse la barrière jusqu'à la fréquence 10 avec un effectif cinq fois supérieur.

Figure 5 - Analyse du Tableau Lexical Entier
Le Journal Le Monde *1994-1995(5785 lignes, 24 colonnes)*



3 - La troisième méthode est encore empruntée à Muller. Elle a rarement été appliquée, sinon dans nos études sur Giraudoux et Hugo. Car elle exige des calculs très lourds, sans offrir toutes les garanties. La procédure suit pourtant une logique rigoureuse qui tient compte des fréquences en leur appliquant la loi binomiale. On doit d'abord réunir les deux textes à comparer dans un corpus et y établir la gamme de distribution des fréquences (c'est-à-dire l'effectif des mots employés 1 fois, 2 fois, n fois). Le calcul binomial permet alors de prévoir

comment théoriquement devrait se répartir cet effectif dans les deux textes. Pour les hapax l'alternative est simple et leur effectif dans chaque texte est proportionnel à l'étendue de chacun. Avec la fréquence 2 le choix est plus complexe: les deux occurrences peuvent être distribuées dans les deux textes ou se concentrer dans le texte A ou dans le texte B. La fréquence 3 présente quatre possibilités (3 occurrences en A 0 en B, 2 en A 1 en B, 1 en A 2 en B et 0 en A 3 en B). Nous renvoyons à Muller pour la suite des opérations. Une fois établis les effectifs réel et théorique pour chaque case du tableau, un écart est calculé, puis un Chi2. La somme des Chi2 donne la mesure de la distance entre les deux textes, compte tenu des degrés de liberté, qui peuvent varier d'une paire de textes à l'autre. Or c'est là que le bât blesse. Seule les basses fréquences en effet ont des effectifs suffisants pour permettre le calcul du Chi2. Quand le seuil n'est pas atteint, il y a lieu de regrouper les effectifs trop faibles. Quand toutes les paires ont été examinées, les Chi2 obtenus ne sont pas tout à fait indépendants de l'étendue des textes, par l'effet de la loi des grands nombres. Pour un écart proportionnellement semblable, la valeur du Chi2 (comme celle de l'écart réduit) augmente avec la taille des observations. Pour neutraliser cette distorsion, force est de recourir à quelque pondération. Celle que nous proposons en cette occasion est la racine carrée du vocabulaire.

La figure 6 rend compte de la distance lexicale, ainsi calculée, dans l'œuvre de Giraudoux. Le tableau des distances là aussi a été confié à l'analyse factorielle, où l'on reconnaît aisément l'influence du genre dans le premier facteur (romans à gauche, théâtre à droite), et celle de la chronologie dans le second facteur. Certes les deux premiers facteurs d'une analyse factorielle n'épuisent pas nécessairement la totalité de la variance. Mais dans le cas présent ils accaparent 78% de l'inertie et il n'est guère utile de considérer les autres facteurs. Mais pour exploiter ce même tableau des distances, on peut recourir à une autre méthode qui rend compte au mieux de leur complexité, ce qu'elle exprime dans un graphe (figure 7). Cette analyse dite arborée a été proposée par Luong et Barthélémy. Tous les textes dans le graphe sont liés les uns aux autres, par

des chemins plus ou moins directs, qui reproduisent les distances initiales. Seule est à prendre en compte la longueur des segments, et non pas les angles et les directions, qui suivent une courbure arbitraire, liée aux contraintes de l'espace disponible. L'interprétation en est limpide et l'analyse précédente s'en trouve confirmée.

Figure 6 - La loi binomiale (méthode Muller).Giraudoux.

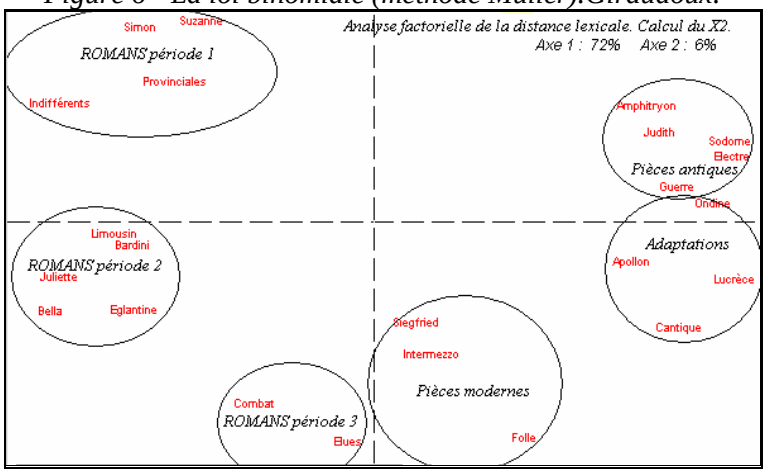
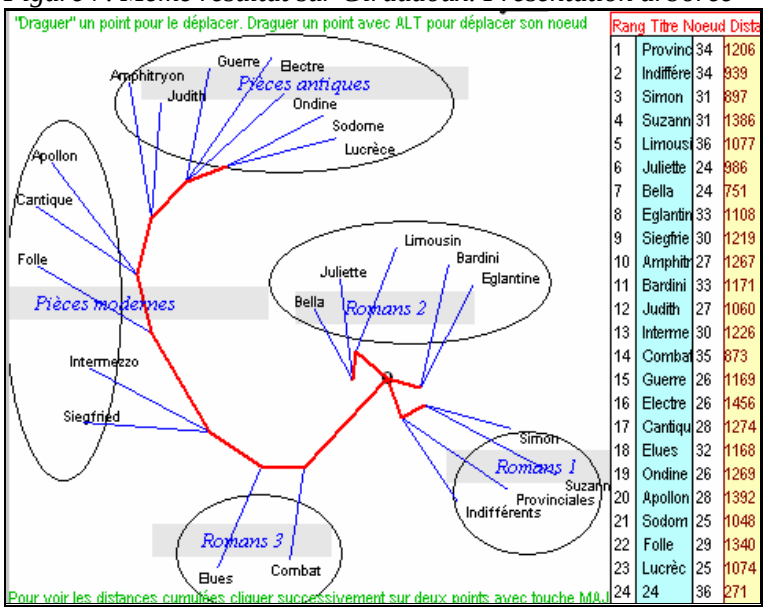


Figure 7. Même résultat sur Giraudoux. Présentation arborée

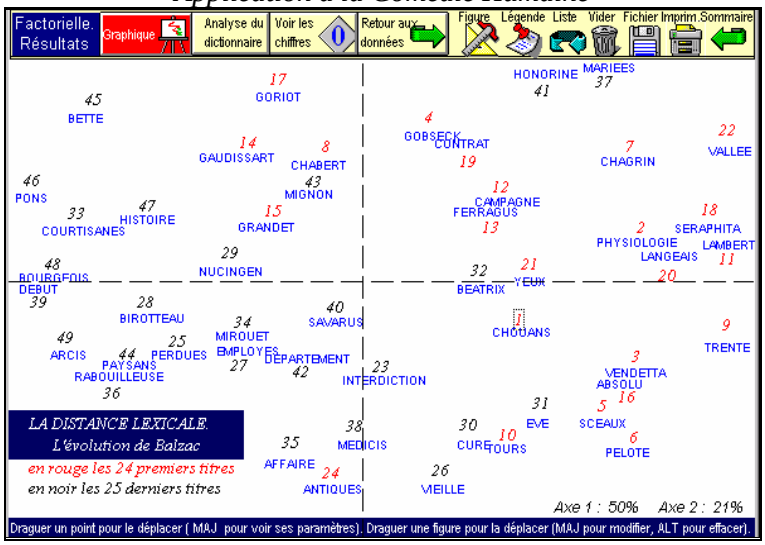


4 - Les deux méthodes qui précèdent tiennent compte de la fréquence des mots et lui appliquent les lois traditionnelles le la statistique (normale, hypergéométrique ou binomiale). Dominique Labbé, bien connu pour ses travaux sur les hommes politiques, principalement Mitterrand et de Gaulle, a pensé qu'on pouvait s'affranchir de ces schémas classiques. Il a proposé récemment (Actes du Colloque *JADT 2000*, Lausanne, p. 85-94) un algorithme efficace qui pour chaque mot apprécie la distribution réelle des fréquences dans les deux textes en la comparant non plus à la répartition théorique mais à l'écart maximal possible dans cette distribution

$$D_{(A,B)} = \sum d_i \quad / \sum dmax_i$$

pour i variant du premier au dernier mot du vocabulaire des textes A et B. Sans reprendre le détail des calculs et des explications, pour lesquels nous renvoyons à l'article cité, nous nous bornerons à montrer le résultat obtenu sur le corpus de Balzac.

Figure 8 - La méthode Labbé.
Application à la Comédie Humaine

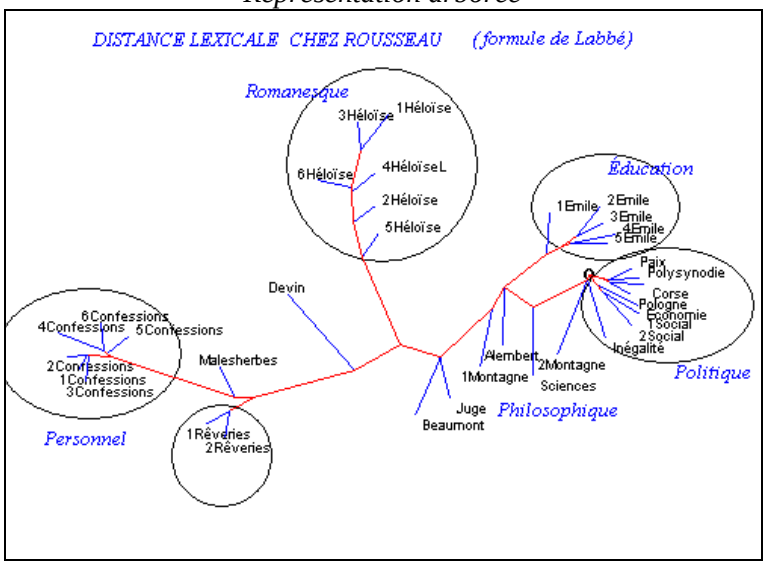


Les 49 titres considérés sont numérotés dans l'ordre chronologique. Or au lieu de se distribuer équitablement dans l'espace du graphique 8, ils apparaissent soumis à une

gravitation temporelle qui concentre les premiers romans dans la moitié droite et les derniers dans la moitié gauche. La composition des constellations observées permet d'interpréter cette évolution générale et les exceptions qu'elle accepte. Les deux pôles qui gouvernent la gravitation semblent être l'amour et l'argent, même si ces deux thèmes sont souvent mêlés dans le nœud balzacien. Mais le dosage varie: l'amour domine à droite (avec quelques transfuges de la dernière période), l'argent s'impose à gauche (avec quelques romans anticipatifs de la première période).

L'évolution de la thématique est moins sensible chez Rousseau, d'abord parce qu'elle est cachée par la prépondérance du genre littéraire, mais aussi parce que la carrière de Rousseau est faite de ruptures, de retours et d'un va-et-vient permanent entre les considérations morales, sociales, politiques, religieuses et l'épanchement des sentiments personnels.

Figure 9 - La méthode Labbé. Application à Rousseau. Représentation arborée



La typologie de ses œuvres n'en apparaît pas moins clairement dans la figure 9, établie sur les coefficients de la méthode Labbé. À un bout de la chaîne on trouve les écrits politiques, des premiers (Discours sur les Sciences et les Arts,

et sur l'Inégalité) aux derniers (Considérations sur le gouvernement de Pologne, Projet de constitution pour la Corse) en passant par le Contrat Social. De ce côté se trouvent aussi les développements sur les arts (Lettre à d'Alembert), l'éducation (l'Emile) et la religion (Lettre à Ch. De Beaumont). À l'autre bout se tiennent les écrits personnels: les Rêveries du promeneur solitaire, les Lettres à Malesherbes et les Confessions, et, sur une branche latérale, la Nouvelle Héloïse.

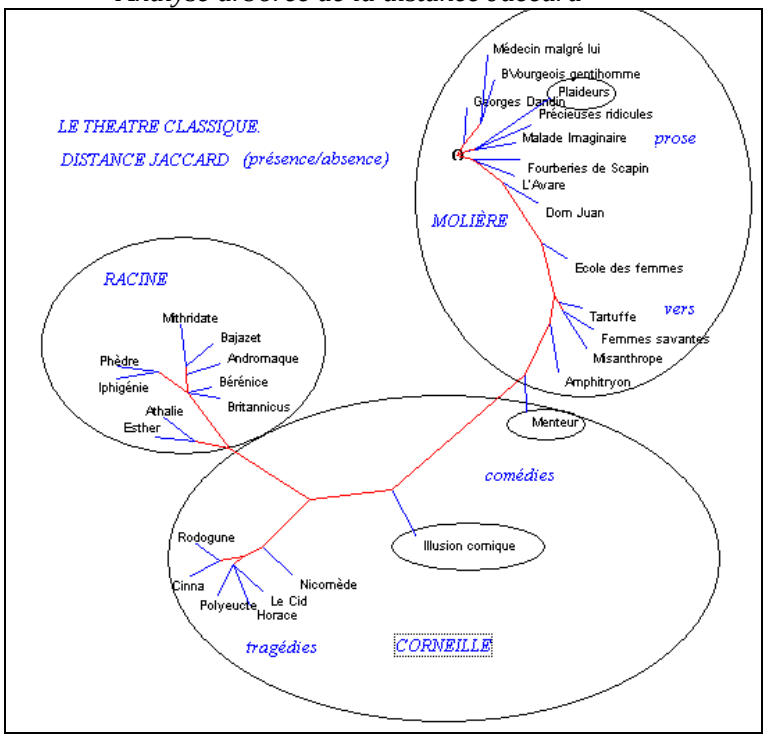
- II –Evaluation critique

1 - Reste à faire un choix parmi ces méthodes. Tout serait remis en question, s'il advenait que les résultats fussent incohérents ou contradictoires. Or l'expérience montre que les différences sont peu sensibles, même lorsque l'objet considéré n'est pas exactement le même. S'appuyer sur les effectifs des classes de fréquence (méthode du Chi2 de Muller), c'est ignorer complètement la signification des mots pour n'envisager que la structure lexicale. À l'inverse s'en tenir au relevé des mots communs et exclusifs (méthode Jaccard), c'est négliger les mots fréquents et particulièrement les mots-outils (nécessairement communs) et ainsi ignorer les faits de syntaxe et beaucoup des faits de style. Les imperfections du corpus n'auront pas les mêmes conséquences selon les traitements. Les fautes d'orthographe accidentelles auront plus d'importance dans la méthode Jaccard, les fautes systématiques de codage ou l'incohérence de la norme (par exemple le mélange des imparfaits en *oit* et en *ait* dans certaines éditions de Rousseau) compteront davantage dans les calculs qui envisagent la fréquence. Pourtant la convergence est au rendez-vous.

Reprenons en effet l'exemple du théâtre classique, et comparons la figure 10 obtenue par la méthode Jaccard, à la figure 11 fondée sur la méthode Labbé. Les mêmes commentaires que nous avait inspirés l'analyse factorielle de la figure 1 s'imposent pareillement au vu des deux graphes, qui , tout en séparant Corneille et Racine sur deux branches adjacentes, les éloigne l'un et l'autre de Molière, le *Menteu*r et plus encore les *Plaideurs* faisant exception. Même la distinction entre les comédies en vers de Molière et celles en prose est soulignée

parallèlement dans les deux figures. Le recouvrement des deux méthodes s'observe dans la grande majorité des corpus que nous avons pu étudier[4].

Figure 10. Le théâtre classique.
Analyse arborée de la distance Jaccard

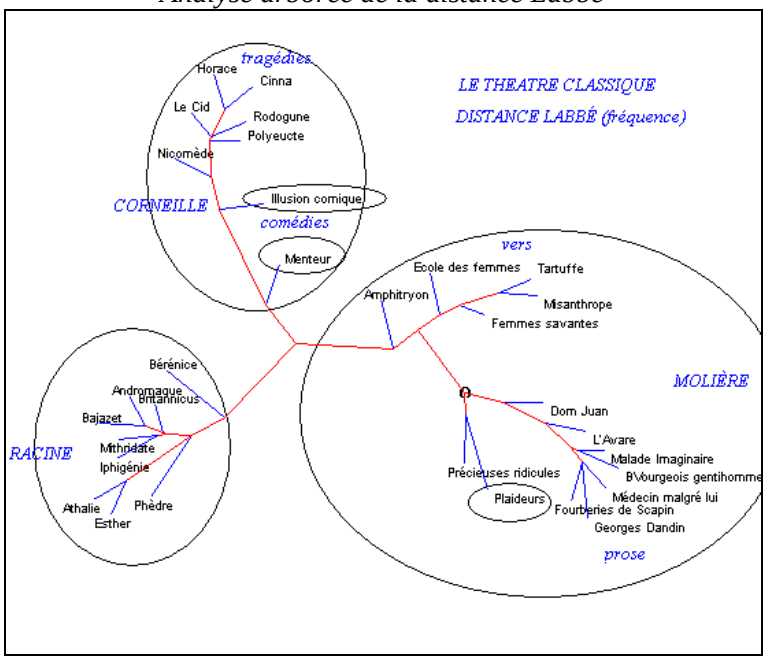


La méthode du Chi2 (Muller) participe au concert, quand on a le courage de la mettre en œuvre. On a un peu moins de confiance dans la méthode de l'analyse factorielle appliquée au tableau lexical entier (méthode Salem), pour des raisons qu'on a

[4] Cette étude sur le théâtre classique a été conduite antérieurement à celle de Dominique Labbé, à partir des données fournies par l'INaLF. En puisant les siennes à la même source, Labbé a constitué un corpus plus complet, qu'il a soumis en outre à la lemmatisation. Si les résultats ne changent guère, Labbé en donne une interprétation qui diffère de la nôtre et que nous discuterons dans un article ultérieur.

expliquées. L'expérience montre que les résultats cessent d'être stables, si l'on fait entrer les basses fréquences dans les données (dans le cas où celles-ci sont pondérées).

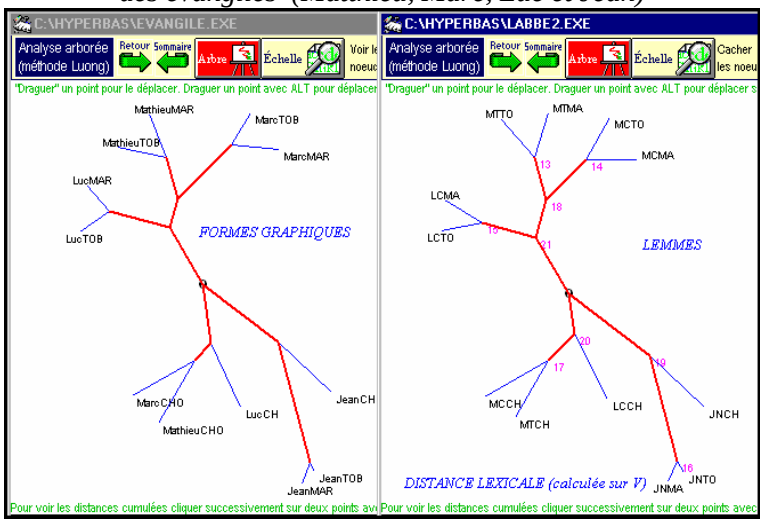
Figure 11. Le théâtre classique.
Analyse arborée de la distance Labbé



2 - Cette convergence des approches est illustrée par le corpus des Évangiles (figure 12). Cette fois le genre littéraire n'exerce plus son influence écrasante. Le sujet est commun aux quatre évangélistes qui racontent la même histoire et enseignent la même leçon aux mêmes populations, à la même époque et dans la même langue. La plupart des variables sont ainsi neutralisées, pour mieux isoler les faits qui relèvent de l'auteur et du traducteur. Car on a mis en parallèle trois traductions des quatre évangiles (la traduction œcuménique, celle de l'abbaye de Maredsous et celle d'E. Chouraqui), ce qui porte à douze le nombre de textes comparés. La question qui se pose est de mesurer le poids respectif de l'auteur et du traducteur. Or la réponse est la même dans les deux analyses de la figure 12. Cette fois on ne compare

pas deux méthodes à propos du même objet mais deux objets différents soumis au même traitement. La même méthode de Jaccard est en effet appliquée aux formes graphiques dans la figure de gauche, puis aux vocables dans celle de droite. Dans les deux cas le premier clivage sépare les auteurs, l'évangéliste Jean s'écartant des trois synoptiques, quelle que soit la traduction, tandis que le second clivage est dû à l'originalité de la traduction de Chouraqui qui entraîne tout ce qu'il touche sur une branche détournée.

Figure 12 . Trois traductions (TOB, Maredsous, Chouraqui) des évangiles (Matthieu, Marc, Luc et Jean)

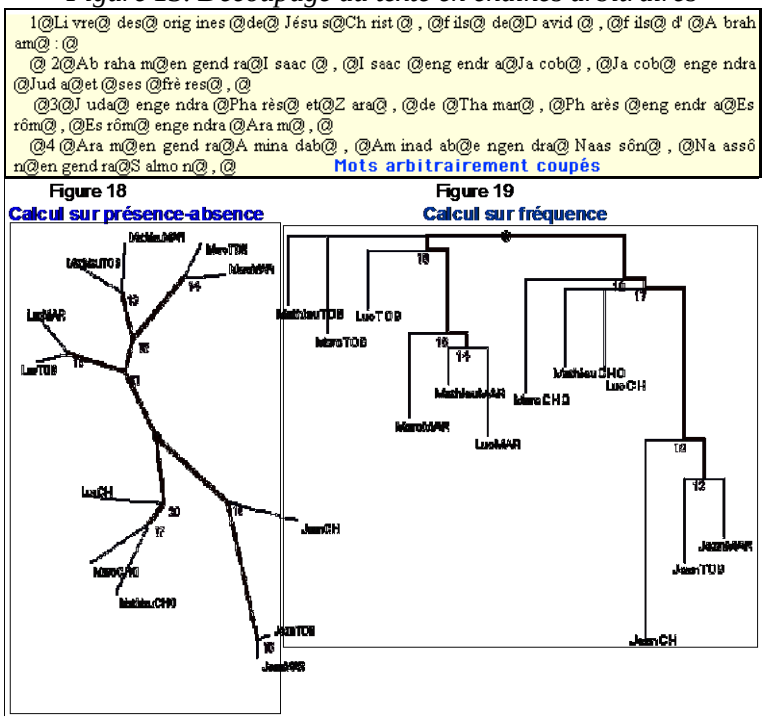


Le même texte évangélique traité successivement avec et sans lemmatisation donne la même image des distances intertextuelles. Qu'on traite 9622 formes graphiques ou 5014 vocables, les résultats restent stables. Comme il y a superposition pure et simple, l'effort de la lemmatisation ne semble pas s'imposer à qui veut seulement établir une typologie des textes fondée sur le vocabulaire. Il en va autrement si l'ambition est autre et si l'on veut établir les distances en mesurant les catégories grammaticales ou quelque objet qui nécessite un codage préalable.

3 - On a voulu pousser plus loin l'expérimentation et mettre à l'épreuve la robustesse de l'analyse et son indifférence à la

nature codée ou non des données. Cette stabilité se maintient-elle lorsqu'on dénature le texte en faisant éclater la structure du mot. Remplaçons tous les blancs par un caractère arbitraire, par exemple le signe @, et découpons la chaîne en tronçons de quatre lettres, qu'on va considérer comme des "mots", même s'ils n'ont ni queue ni tête, ni forme ni sens. En cela nous reprenons la démarche que A. Lelu applique aux N-grammes. Au lieu d'ajouter une information (ce que fait l'étiquetage), nous retranchons un élément essentiel : la segmentation en unités lexicales. Pis encore: le continuum graphique est rompu et perverti par de fausses coupures, comme si on voulait crypter le texte. On peut en juger à partir des premières lignes de l'Évangile qu'on a ainsi transcrites au haut de la figure 13.

Figure 13. Découpage du texte en chaînes arbitraires

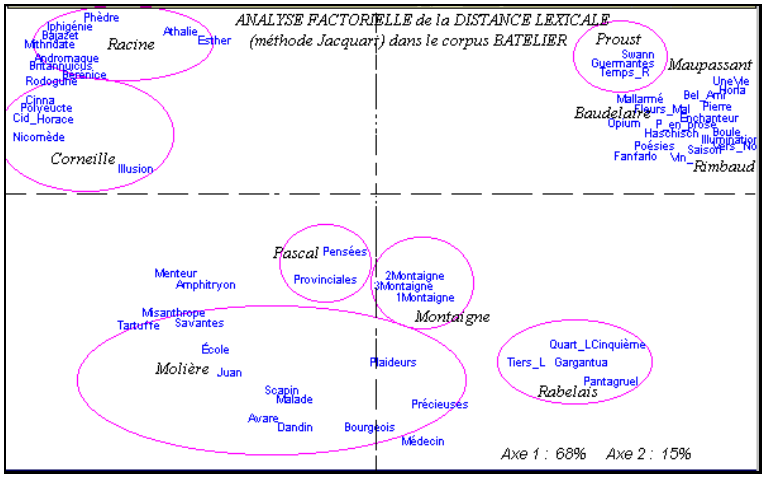


Et pourtant, dans cette eau boueuse où aucun mot n'est reconnaissable, la décantation des évangélistes et des traducteurs se fait limpide, qu'on envisage les 400 000 occurrences selon le procédé Labbé (partie droite du graphique 13) ou les 16000 "mots" découpés, selon le procédé Jaccard (partie gauche). La conclusion de cette expérience est encourageante: à l'heure d'Internet, où circulent tant de textes de qualité médiocre, les méthodes multidimensionnelles sont assez puissantes et robustes pour souffrir sans dommage les impuretés et les erreurs.

5 - Que se passe t-il lorsqu'on épure les données au lieu de s'amuser à les corrompre? On a vu que les corpus lemmatisés ne se distinguent guère des corpus bruts où le brouillard de l'homographie gêne peu l'appréciation des distances lexicales. La lemmatisation permet cependant d'autres approches qui font abstraction de l'individualité des mots pour ne retenir que leur fonction dans la phrase. La neutralisation de la valeur individuelle des mots était aussi la caractéristique de la méthode Muller qui comme, nous l'avons vu, ne considérait que l'effectif des classes de fréquence (voir figure 6). En procédant ainsi on s'affranchit du thème, en isolant des phénomènes plus directement linguistiques et stylistiques. On peut craindre en effet qu'à trop s'attacher au lexique, on ne différencie les textes qu'à partir du sujet, en distinguant par exemple les romans d'amour, d'aventure, de mer ou d'argent, que le même écrivain peut fort bien choisir tour à tour sans se départir de sa manière propre. Certes quelques-unes des approches précédentes s'appuyaient aussi sur les mots grammaticaux, dont le contenu sémantique a tendance à s'effacer au profit de la syntaxe. Mais il est sans doute de meilleure méthode d'étudier directement les structures grammaticales. C'est ce que réalise la figure 15, à partir d'un corpus réunissant une soixantaine de textes de notre littérature, de Rabelais à Proust. Comme il s'agit des parties du discours, sur lesquelles le temps n'a pas d'influence directe, la typologie des textes échappe à la dérive du temps, dont le vocabulaire, au contraire, ne peut guère s'affranchir, même lorsqu'on normalise l'orthographe, et dont la figure 14 porte témoignage: chaque siècle s'y dessine un territoire propre: le

XVIe dans le quadrant inférieur droit, le XVIIe dans la moitié gauche, et les temps modernes dans le quadrant supérieur droit.

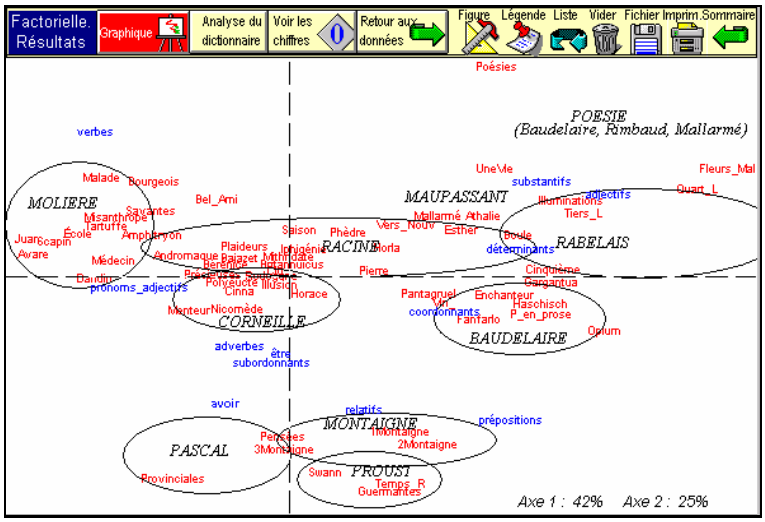
Figure 14. La distance lexicale dans le corpus BATELIER (méthode Jaccard)



À rebours de cette évolution trop prévisible, la répartition des catégories grammaticales dans la figure 15 met ensemble des textes et des auteurs très éloignés dans le temps; Proust avec Pascal et Montaigne, Rabelais avec Baudelaire et Rimbaud. Et un même écrivain peut s'y trouver distendu et écartelé, ce qui arrive à Racine, d'abord favorable au verbe, comme Corneille et Molière, et passant progressivement dans le clan du substantif. C'est en effet un champ clos où s'affrontent deux camps: d'un côté le verbe et ses acolytes: pronoms et adverbes, de l'autre le substantif et sa suite: adjectifs, déterminants et prépositions. Le théâtre est dans le premier camp, la poésie et le roman dans le second. Mais certains prosateurs refusent ce choix et manifestent d'autres préférences. Quand il s'agit d'analyser et de raisonner, non d'évoquer, de raconter ou de dialoguer, le discours a besoin d'articulations logiques et en particulier des subordonnants, des coordonnants et des relatifs. C'est ce goût commun qui réunit au

bas du graphique la phrase de Pascal, de Montaigne et de Proust.

Figure 15 - La distance grammaticale. Le corpus BATELIER



****

1 - La distance dans le discours est ce qu'elle est en peinture: une perspective, un point de vue. Plus encore que le monde physique, l'univers du discours est soumis à la relativité. Faute d'un point d'appui unique, les mesures varient selon l'objet isolé, et la méthode choisie. Pourtant les paramètres qu'on croit isoler sont souvent liés entre eux, par l'effet d'une redondance ou surdétermination qui explique la convergence des résultats, comme si l'on photographiait une boule en variant les angles et les points de vue.

2 - À supposer qu'on puisse avec sûreté répartir les textes dans l'espace, comme on distribue les villes sur une carte, il resterait à décrire et à expliquer les oppositions et les rapprochements. Beaucoup des analyses factorielles que nous avons produites ne montrent que les variables, c'est-à-dire les textes, sans montrer

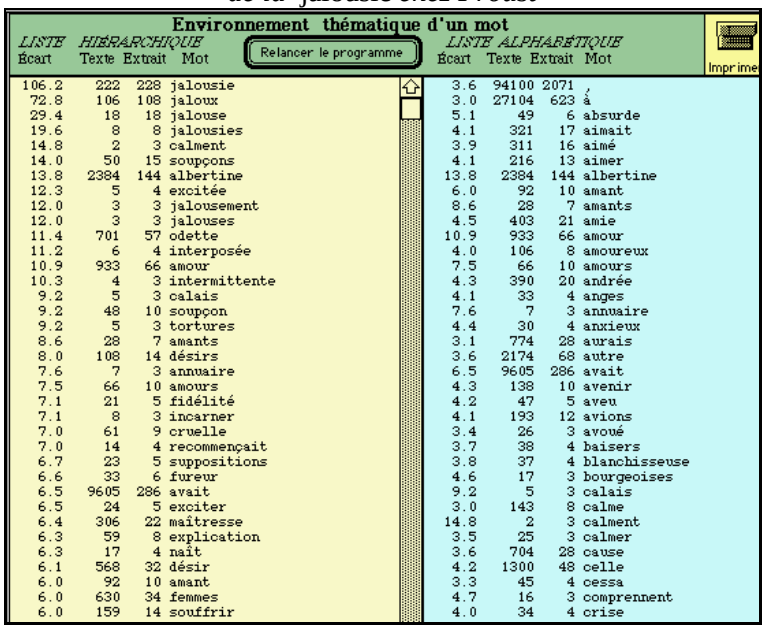
les individus, c'est-à-dire les mots ou les lignes du tableau. Or le pouvoir explicatif de l'analyse de correspondance est précisément dans la correspondance qu'on peut établir entre les uns et les autres, entre les colonnes et les lignes. Et c'est pourquoi la figure 15, qui établit ce rapport, est plus riche d'informations que les précédentes. Car lorsqu'il s'agit de distances brutes, on a souvent affaire à des tableaux carrés et symétriques où lignes et colonnes désignent les mêmes objets, ce qui rend sans effet la correspondance. L'analyse du TLE (méthode Salem) n'a pas cet inconvénient (c'est un tableau rectangulaire de contingence), mais le nombre de lignes est si élevé que leur représentation graphique serait trop encombrante.

3 - Pour compléter et préciser les calculs de distance, la lexicométrie propose heureusement d'autres voies, et notamment le calcul des spécificités, qui, si l'on réfléchit, est encore un calcul de distance, cette fois entre les mots et les textes. Dans l'univers gravitationnel du discours, certains mots se trouvent pris dans le champ d'attraction d'un texte ou d'un autre et contribuent à lui donner sa coloration particulière. L'étude de ce spectre thématique ou stylistique est devenue la voie royale de la discipline.

4 - Enfin il peut être légitime de s'interroger non seulement sur la distance entre les textes, mais, à travers les textes, sur la distance entre les mots. Pour un mot donné, par exemple le mot *jalousie* chez Proust (figure 16), la démarche consiste à explorer tous les contextes qui enveloppent ce mot et à comparer cet ensemble de phrases particulières à celles de la totalité du texte. Le résultat se lit dans une liste de corrélats triés selon la plus ou moins grande accointance qui les rapproche du thème proposé. Ces alliances sont souvent des relations familiales de dérivation ou de synonymie (parfois d'antonymie), parfois des rapports de bon voisinage syntaxique, parfois des rencontres provoquées par la rime ou le calembour. On peut même imaginer une étude systématique de toutes les relations fréquentielles que les mots tissent entre eux dans le fil d'un texte, soit qu'ils se repoussent , soit qu'ils s'attirent, et dessiner en fin de compte une carte du

texte peuplée de constellations lexicales en orbite. Et c'est ce que fait le logiciel *Alceste*, avec un certain succès.

Figure 16 La distance entre les mots. L'environnement lexical de la jalousie *chez Proust*



Environnement thématique d'un mot

Relancer le programme — Imprime

LISTE HIÉRARCHIQUE

Écart	Texte	Extrait	Mot
106.2	222	228	jalousie
72.8	106	108	jaloux
29.4	18	18	jalouse
19.6	8	8	jalousies
14.8	2	3	calment
14.0	50	15	soupçons
13.8	2384	144	albertine
12.3	5	4	excitée
12.0	3	3	jalousement
12.0	3	3	jalouses
11.4	701	57	odette
11.2	6	4	interposée
10.9	933	66	amour
10.3	4	3	intermittente
9.2	5	3	calais
9.2	48	10	soupçon
9.2	5	3	tortures
8.6	28	7	amants
8.0	108	14	désirs
7.6	7	3	annuaire
7.5	66	10	amours
7.1	21	5	fidélité
7.1	8	3	incarner
7.0	61	9	cruelle
7.0	14	4	recommençait
6.7	23	5	suppositions
6.6	33	6	fureur
6.5	9605	286	avait
6.5	24	5	exciter
6.4	306	22	maîtresse
6.3	59	8	explication
6.3	17	4	naît
6.1	568	32	désir
6.0	92	10	amant
6.0	630	34	femmes
6.0	159	14	souffrir

LISTE ALPHABÉTIQUE

Écart	Texte	Extrait	Mot
3.6	94100	2071	,
3.0	27104	623	à
5.1	49	6	absurde
4.1	321	17	aimait
3.9	311	16	aimé
4.1	216	13	aimer
13.8	2384	144	albertine
6.0	92	10	amant
8.6	28	7	amants
4.5	403	21	amie
10.9	933	66	amour
4.0	106	8	amoureux
7.5	66	10	amours
4.3	390	20	andrée
4.1	33	4	anges
7.6	7	3	annuaire
4.4	30	4	anxieux
3.1	774	28	aurais
3.6	2174	68	autre
6.5	9605	286	avait
4.3	138	10	avenir
4.2	47	5	aveu
4.1	193	12	avions
3.4	26	3	avoué
3.7	38	4	baisers
3.8	37	4	blanchisseuse
4.6	17	3	bourgeoises
9.2	5	3	calais
3.0	143	8	calme
14.8	2	3	calment
3.5	25	3	calmer
3.6	704	28	cause
4.2	1300	48	celle
3.3	45	4	cessa
4.7	16	3	comprennent
4.0	34	4	crise

Références bibliographiques

Barthélémy J.P. (1987). « Sur la topologie d'un arbre phylogénétique : aspects théoriques, algorithmes et applications à l'analyse des données textuelles », *Mathématiques et Sciences humaines* 100: 57-80.

Barthélémy J.-P. & Luong X. (1998). « Représenter les données textuelles par les arbres... », in S. Mellet (éd.) *JADT 1998, 4èmes Journées Internationales d'Analyse statistique de Données Textuelles.* Nice : Université de Nice, pp. 49-71.

Brunet É. (1988). « Une mesure de la distance intertextuelle : la connexion lexicale », *Revue Informatique et Statistique dans les Sciences humaines, (Le nombre et le texte.* Hommage à Étienne Évrard), 24, 1-4 : 81-116.

Brunet É. (2002). «Un texte sacré peut-il changer ? Variations sur l'Évangile », in J. Cook (éd.), *Bible and Computer*. Leiden / Boston : Brill, pp. 79-98.

Brunet É. & Muller C. (1988). «La statistique résout-elle les problèmes d'attribution ? », *Strumenti critici III*, 3 : 367-387.

Labbé D. & Monière D. (2000). « La connexion intertextuelle. Application au discours gouvernemental québécois », in M. Rajman & J.-C. Chappelier (éds.), *JADT 2000 Actes des 5èmes journées internationales d'Analyse statistique des Données Textuelles*, Lausanne : EPFL, vol. 1 : 85-94.

Lebart L. & Salem A. (1994). *Statistique textuelle.* Paris : Dunod.

Luong X. (1994). « L'analyse arborée des données textuelles : mode d'emploi », *Travaux du cercle linguistique de Nice* 16 : 27-42.

Muller C. (1997). *Principes et méthodes de statistique lexicale*. Paris : Hachette. Ouvrage réédité dans la collection *Unichamp* des éditions Champion.