

Qui a écrit quoi ? L'attribution d'auteur et la distance intertextuelle

(Juillet 2002)

(texte soumis à la revue Corpus)

Dominique LABBE
Institut d'Etudes Politiques de Grenoble
dominique.labbe@iep.upmf-grenoble.fr

Avertissement (mai 2003)

Au début de l'année 2002, E. Brunet et moi-même, avons organisé en commun, l'expérience suivante : il m'envoyait une série de textes "anonymés", à charge pour mes programmes de reconnaître les textes appartenant à un même auteur.

Le texte qu'on lira ci-dessous a été adressé à E. Brunet en juillet 2002. Au vu de ce compte-rendu, E. Brunet a mené de son côté sa propre expérience sur ces mêmes fichiers. Il a rédigé son interprétation qu'il m'a remise en août de la même année. C'est à ce moment que j'ai pu découvrir les ouvrages et les auteurs sur lesquels portait l'expérience...

Il ne s'agit donc pas, à proprement dit, d'une "expérience en double aveugle" puisque j'étais le seul aveugle et que le tiers arbitre manquait. Etant donné l'hostilité d'E. Brunet à mes formules, procédures et méthodes, il ne pouvait être soupçonné de bienveillance à mon égard, c'est pourquoi j'ai accepté cette asymétrie et l'absence d'arbitre.

Le dossier complet devait paraître en début 2003 dans la revue du laboratoire de Nice. Les délais s'accumulant, les attaques malveillantes, dont je fais l'objet actuellement, m'obligent à dévoiler ce texte. Afin de permettre au lecteur de juger du degré de réussite de mes tests, je place en annexe, la présentation du corpus, telle qu'elle figure dans la note que m'a fait parvenir d'E. Brunet en août 2002.

*E. Brunet m'a refusé l'ajout d'une post-face dans laquelle j'aurais aimé souligner certaines "difficultés" entraînés par ses choix, notamment le fait qu'il a utilisé les "éditions de référence" comportant souvent des "avant-propos" ajoutés lors des éditions successives — textes parfois très longs, comme dans *Châteaubriand* — qui ajoutent de la distance, pour un même roman, entre le premier extrait (où figure cette préface) et le deuxième extrait ; ou encore le caractère très particulier des six derniers textes.*

Si E. Brunet pense que cette expérience "ne prouve rien", comme il le répète à l'envi, il pourra publier son propre texte à l'appui de ses dires (en dehors de corrections de formes, cette publication sera certainement fidèle au texte qu'il m'a remis en août 2002).

Naturellement, je tiens à la disposition du lecteur sceptique, les textes originaux, les fichiers lemmatisés et le programme de calcul de la distance.

Etienne Brunet a bien voulu se prêter à l'expérience suivante : constituer un corpus de 50 textes anonymés afin de mettre à l'épreuve l'application de la distance intertextuelle à la question de l'attribution d'auteur. Ce corpus a été préalablement normalisé et lemmatisé¹. Le calcul de la distance est appliqué sur les vocables et non sur les formes graphiques brutes. Il a été complété par deux expériences de classification.

LA DISTANCE INTERTEXTUELLE

Rappelons que le calcul de la **distance** entre plusieurs textes réunis en corpus vise à répondre à la question : quels sont ceux les plus proches et les plus éloignés ? Le calcul consiste à comparer le vocabulaire des textes, pris deux à deux, en neutralisant les différences de taille². Une faible distance indiquera que la plus grande partie de la surface du couple considéré est commune. A l'inverse, plus la distance s'accroît, plus les textes sont décalés. Les valeurs de l'indice varient entre 0 et 1 :

— une valeur de 0 signifie que les deux textes utilisent le même vocabulaire avec les mêmes fréquences d'emploi. Il ne s'agit pas forcément d'un décalque exact : les mots peuvent être placés dans un ordre différent ; les temps des verbes modifiés ou le genre et le nombre des adjectifs... Autrement dit, deux textes dont la distance est faible, s'ils ne disent pas la même chose, partagent certainement le même univers intellectuel et sont écrits dans un style semblable.

— un indice de 1 signifie que les deux textes n'ont aucun mot en commun. Cette situation est aussi théorique que la précédente car, s'ils utilisent la même langue, les locuteurs sont condamnés à utiliser les outils de cette langue (déterminants, pronoms, verbes auxiliaires...) Mais plus on se rapprochera de 1, plus les textes appartiendront à des genres et à des univers intellectuels différents, plus ils développeront des thèmes éloignés ;

— un indice de 0,5 signifie que les textes ont en commun la moitié de leur surface.

L'interprétation des résultats doit tenir compte de trois dimensions principales :

En premier lieu, la dimension temporelle : la langue change au cours du temps et ces changements sont sensibles à l'échelle même d'une vie un peu longue. Certains auteurs sont un peu comme des paysans qui labourent toujours les mêmes champs et ne sortent pratiquement pas de leur campagne. En cas de thème(s) unique(s), la classification de leur œuvre risque fort d'être chronologique.

En second lieu, la dimension des genres : la langue offre plusieurs registres possibles, avec des vocabulaires différents. Il s'agit d'abord de l'opposition entre l'oral et l'écrit (pour un même auteur, s'exprimant sur un même thème, les textes oraux et écrits sont habituellement séparés par une distance minimale de 0.3). Au sein de ces deux catégories, de nombreux genres sont encore possibles (soutenu, familier ; tragédie, comédie ; scientifique, fiction,

¹ Dominique LABBE, *Normes de saisie et de dépouillement des textes politiques*, Grenoble, Cahier du CERAT, 1990.

² Labbé C., Labbé D., "Inter-Textual Distance and Authorship Attribution. Corneille and Moliere", *Journal of Quantitative Linguistics*, 8-3, December 2001, pp 213-231.

romanesque...) En toute rigueur, la comparaison des distances ne devrait se faire que dans un même registre et un même genre.

Enfin, la dimension thématique : le thème traité entraîne avec lui tout un vocabulaire. Il y a un lexique de l'amour, de la guerre, de la maladie, des affaires, de la politique, du crime, des voyages en train, en avion, en bateau...) Naturellement, pour un même auteur, un changement de thème créé de la distance et, pour des auteurs différents, le fait de traiter le même thème, à une même époque, engendre une proximité...

L'application de ce calcul, à plus d'un millier de textes de toute nature, nous a permis d'étalonner l'échelle suivante :

— une distance inférieure ou égale à 0,20 indique que les textes appartiennent aux mêmes registre et genre, qu'ils ont un thème unique et qu'ils ont été écrits par un seul auteur. En effet, jusqu'ici, nous n'avons jamais rencontré de textes d'auteurs différents séparés par une distance aussi faible et il paraît certain que, si le cas devait se présenter, l'on pourra conclure avec certitude que le second aura plagié le premier ou qu'il l'aura utilisé comme "nègre"...

— une distance comprise entre 0,20 et 0,25 indique que les textes appartiennent au même registre mais, s'ils sont d'un même auteur, ils développent des thèmes un peu plus éloignés ou ont pu être écrits à des époques différentes. S'ils appartiennent aux mêmes registre et genre, mais sont d'auteurs différents, les thèmes sont encore très proches et l'on peut soupçonner le plagiat ou de sérieuses "réminiscences". On ne peut cependant écarter totalement l'hypothèse d'une collision : des auteurs différents, mais s'exprimant de manière contemporaine sur un même sujet avec les mêmes sources, etc (nous avons rencontré ce cas dans les corpus de presse pour des articles contemporains et sur un même sujet) ;

— de 0,25 à 0,35 : pour un même auteur, les textes appartiennent probablement à des genres ou à des registres différents. En tous cas, les thèmes développés sont assez éloignés. Pour des textes de même registre et de même genre, mais d'auteurs différents, les thèmes sont encore assez proches ;

— au-dessus de 0,35 : les auteurs ou les registres sont différents ; pour un même auteur, dans un même registre : les thèmes ou les époques de rédaction sont éloignés.

Pour nous résumer : la distance entre deux textes est fonction des auteurs, des époques, des genres et des thèmes.

L'échelle ci-dessus vaut pour des textes de taille supérieure à 1.000 mots et inférieure à 100.000. De plus, il est préférable de ne pas comparer entre eux des textes ayant des différences de taille supérieures à 1/12. En dehors de cet intervalle, la distance peut être significativement corrélée à la taille des textes. Le corpus établi par E. Brunet présente le cas le plus favorable : des textes de dimensions très proches³.

Appliqué à ce corpus, le calcul aboutit donc à une matrice de 50 lignes par 50 colonnes dont la reproduction intégrale est évidemment impossible. L'exploitation de cette matrice peut se faire en deux temps. En premier lieu, on extrait quelques valeurs remarquables puis on lui applique des procédés de classification automatique.

³ La normalisation, la lemmatisation et l'exclusion des signes de ponctuation modifient légèrement la taille des textes. Celle-ci varie entre 8108 mots (texte 18) et 9150 (texte 30).

VALEURS REMARQUABLES

Puisque le jeu consiste à repérer les textes qui sont certainement ou très probablement d'un même auteur, on peut utiliser l'échelle normalisée qui vient d'être présentée et rechercher dans la matrice les distances les plus faibles (tableau I ci-dessous).

Tableau I. Les distances remarquables (inférieures à la moyenne diminuée de deux écarts-types : auteurs probablement identiques)

Couple	distance
02 24	0,195
02 23	0,200
01 23	0,207
23 24	0,208
01 24	0,217
06 28	0,219
05 27	0,223
01 02	0,227
46 47	0,238
48 50	0,242
45 47	0,244
47 49	0,246
45 46	0,247
46 49	0,247
04 26	0,247
47 48	0,247
49 50	0,250
47 50	0,250
45 49	0,251
46 50	0,251
20 42	0,252
46 48	0,252
45 48	0,252
45 50	0,257
48 49	0,258
16 38	0,259
22 44	0,267
11 33	0,268
03 25	0,270
15 37	0,270
14 36	0,271
18 40	0,272

Les distances inférieures ou égales à 0.2 permettent d'affirmer avec certitude que l'auteur est le même ainsi que l'époque à laquelle les textes ont été écrits et, enfin, que leur genre et leur(s) thème(s) sont semblables. Dans le corpus « Brunet », seuls deux couples de textes entrent dans cette catégorie de la « certitude raisonnable » : 02 et 23 ainsi que 02 et 24. La distance intertextuelle étant transitive⁴, on peut affirmer avec certitude que ces trois textes ont le même auteur. De plus le texte 01 étant séparé des numéros 23, 24 et 02 par une distance à peine supérieure à 0.2, la transitivité permet de lui appliquer la même conclusion. Ces quatre

⁴ Sur les propriétés de la distance intertextuelle, on se reportera à l'article de présentation...

textes sont donc d'un même auteur utilisant le même genre et développant des thèmes semblables ou proches...

Pour les distances inférieures ou égales à 0.25, on sort de la « certitude » pour entrer dans le « très probable ». Soit l'auteur est semblable mais, alors, au moins l'un des paramètres (genre, thème ou époque) aura légèrement changé.. Sinon, l'un des deux auteurs s'est fortement "inspiré" de l'autre... Dans le corpus « Brunet », une dizaine de couples de textes se trouvent dans cette situation : 23 et 24 (mais ils sont déjà élucidés) ; 06 et 28 ; 05 et 27 ; 04 et 26 et un groupe de six textes dont les numéros sont compris entre 45 et 50. Pour ce dernier groupe, la propriété de transitivité évoquée plus haut permet de subodorer un même auteur, des genres et des thèmes proches, voire semblables.

Si nous nous plaçons dans une perspective « policière », l'enquête doit s'interrompre ici. Il peut paraître décevant de n'avoir pu « marier » que 16 textes alors que, très probablement, E. Brunet a glissé beaucoup plus de « couples » dans ce corpus ! Mais, pour reprendre la métaphore policière, il est important de disposer d'outils qui désignent les coupables avec certitude, même si l'on doit, pour cela, laisser échapper certains « suspects ». Telle est la raison pour laquelle nous avons étalonné notre échelle de manière restrictive en fixant des seuils sévères. A ce prix, nous disposons d'une méthode répondant au problème principal de l'attribution d'auteur qui est de pouvoir conclure avec un haut degré de certitude.

Toutefois, nous nous trouvons ici devant un corpus. Nous pouvons soupçonner qu'il a été construit selon une certaine logique. Nous proposons de retrouver cette logique en procédant en deux temps.

Tout d'abord, examinons les valeurs centrales. La distance moyenne est de 0,376. L'écart-type autour de cette moyenne de 0,053 (ce qui donne un coefficient de dispersion relative de 14%). Ces valeurs indiquent que :

— en moyenne, les textes sont nettement différents les uns des autres — ce qui permet d'écarter absolument l'hypothèse d'un auteur unique, voire celle d'un petit nombre d'auteurs contemporains. A l'inverse, on peut être certain que ces textes ont été écrits à des époques éloignées, qu'ils développent des thèmes différents, ou encore qu'ils appartiennent à plusieurs genres littéraires... Ces variables pouvant naturellement additionner leurs effets ou se neutraliser en partie. En effet, les moyennes obtenues sur d'autres corpus comportant un grand nombre d'auteurs différents — les discours des Premiers ministres sur un demi-siècle, les articles de la presse économique et sociale... — font apparaître des moyennes plus basses.

— l'écart-type indique une variation relativement importante autour de cette moyenne (ou encore une dispersion assez forte). Cette valeur permet d'élargir un peu nos investigations. En effet, pour ce corpus, on peut considérer comme « remarquables » — c'est-à-dire que l'on peut déclarer "anormales" avec moins de 5% de chances de se tromper — les distances qui sortent de la plage de variation "normale" autour de la moyenne (\pm deux écarts type), c'est-à-dire inférieures 0,271 ou supérieures à 0,483 (Tableaux 1 et 2)..

Ceci permet de récupérer, toujours avec une probabilité élevée, quelques couples supplémentaires : 20 et 42 ; 16 et 38 ; 22 et 44 ; 11 et 33 ; 03 et 25 ; 15 et 37 ; 14 et 36 ; 18 et 40... Sous réserve de confirmation par E. Brunet, la méthode permet donc de « marier » en "couples certains" — c'est-à-dire avec moins de 5% de chances de se tromper — 32 des 50 textes, ce qui n'est pas un si mauvais résultat...

A l'opposé de la plage de variation « normale », et sous réserve que le corpus ne mélange pas de l'oral et de l'écrit, on peut aussi rejeter absolument certains mariages (tableau II ci-dessous)..

Tableau II. Les couples de textes les plus éloignés. Distances supérieures à la moyenne augmentée de deux écarts type (auteurs très probablement différents)

Couple	distance
14 27	0,495
05 17	0,495
01 15	0,496
09 34	0,496
14 23	0,497
27 36	0,497
09 24	0,500
10 34	0,500
02 36	0,501
10 23	0,502
05 37	0,503
05 10	0,503
34 41	0,503
19 27	0,504
27 35	0,505
01 36	0,506
05 09	0,507
05 13	0,508
09 23	0,509
15 27	0,510
01 19	0,510
24 41	0,510
01 41	0,510
15 23	0,511
02 41	0,512
02 19	0,512
19 24	0,514
05 41	0,515
19 34	0,516
23 41	0,525
05 35	0,525
05 36	0,533
05 14	0,534
05 19	0,534
19 23	0,535
05 15	0,535

Remarques :

— le seuil de 0.5 fixé dans notre échelle pré-étalonnée se trouve empiriquement vérifié. Sous réserve d'une validation par E. Brunet, une distance supérieure à ce seuil permet de conclure avec certitude que les auteurs sont différents. En effet, il peut être aussi utile d'écarter une hypothèse que de la valider. Pour reprendre la métaphore policière, l'enquêteur va pouvoir écarter avec certitude certains « suspects »...

— si le corpus avait été constitué de 25 auteurs différents, nous aurions pu à ce stade, « marier » la quasi-totalité des textes, en combinant les associations et les exclusions. Mais les

cas {01-02-23-24} et {45-46-47-48-49-50} signalent que certains auteurs ont très probablement plus de deux textes dans le corpus, ce qui complique la recherche...

— il reste donc une « zone grise » où se trouvent encore un nombre important de « suspects » potentiels. D'autres instruments plus sophistiqués sont nécessaires pour apporter un peu de lumière dans cette zone grise.

Avant d'examiner ces instruments, on signalera que la matrice des distances apporte beaucoup d'autres informations intéressantes. Par exemple, imaginons que nous ayons traité un échantillon prélevé aléatoirement dans une base de données contenant "toute" la littérature française. Chacun des échantillons forme un point du nuage dont les coordonnées sont définies par les distances le séparant de tous les autres. On peut classer chacun des échantillons en fonction de sa distance par rapport au centre de gravité du nuage, du plus central au plus périphérique (tableau III ci-dessous).

Tableau III. Position de chaque texte par rapport à tous les autres

Les 20 textes les plus centraux			Les 20 textes les plus décalés		
N°	Titre	Distance au centre G	N°	Titre	Distance au centre G
1	50	0,328	31	24	0,384
2	46	0,329	32	01	0,386
3	47	0,329	33	28	0,388
4	45	0,332	34	39	0,389
5	49	0,332	35	31	0,391
6	48	0,332	36	34	0,391
7	22	0,352	37	06	0,391
8	21	0,354	38	35	0,393
9	04	0,355	39	29	0,393
10	25	0,356	40	23	0,397
11	12	0,358	41	27	0,398
12	03	0,359	42	08	0,399
13	11	0,360	43	15	0,403
14	33	0,361	44	14	0,404
15	43	0,364	45	10	0,408
16	38	0,364	46	36	0,412
17	30	0,365	47	09	0,415
18	44	0,365	48	05	0,426
19	16	0,366	49	19	0,427
20	26	0,366	50	41	0,433

Par exemple, il est intéressant de constater que les six derniers textes, qui ont été identifiés comme ayant très probablement un même auteur, sont également les plus centraux et qu'ils sont situés quasiment à la même distance du centre du nuage. En quelque sorte, ces 6 textes comportent le plus grand nombre de mots communs à tous les autres, ou à une partie importante d'entre eux et, de plus, ces mots « communs » se trouvent en proportion à peu près semblable dans chacun de ces textes... Ils sont donc très singuliers et, probablement pas tout à fait de même nature que les autres. Pour en savoir plus concernant ces individus à la fois « étranges » et si ressemblants, il faudrait examiner leurs vocabulaires, ce qui sort de cette étude.

L'examen direct de la matrice des distances apporte donc déjà un grand nombre d'informations mais laisse en suspens un certain nombre de cas. Pour les résoudre, on peut avoir recours à la classification automatique.

CLASSIFICATIONS

Deux classifications ont été opérées sur la matrice des distances : la procédure classique (classification automatique) et l'analyse arborée.

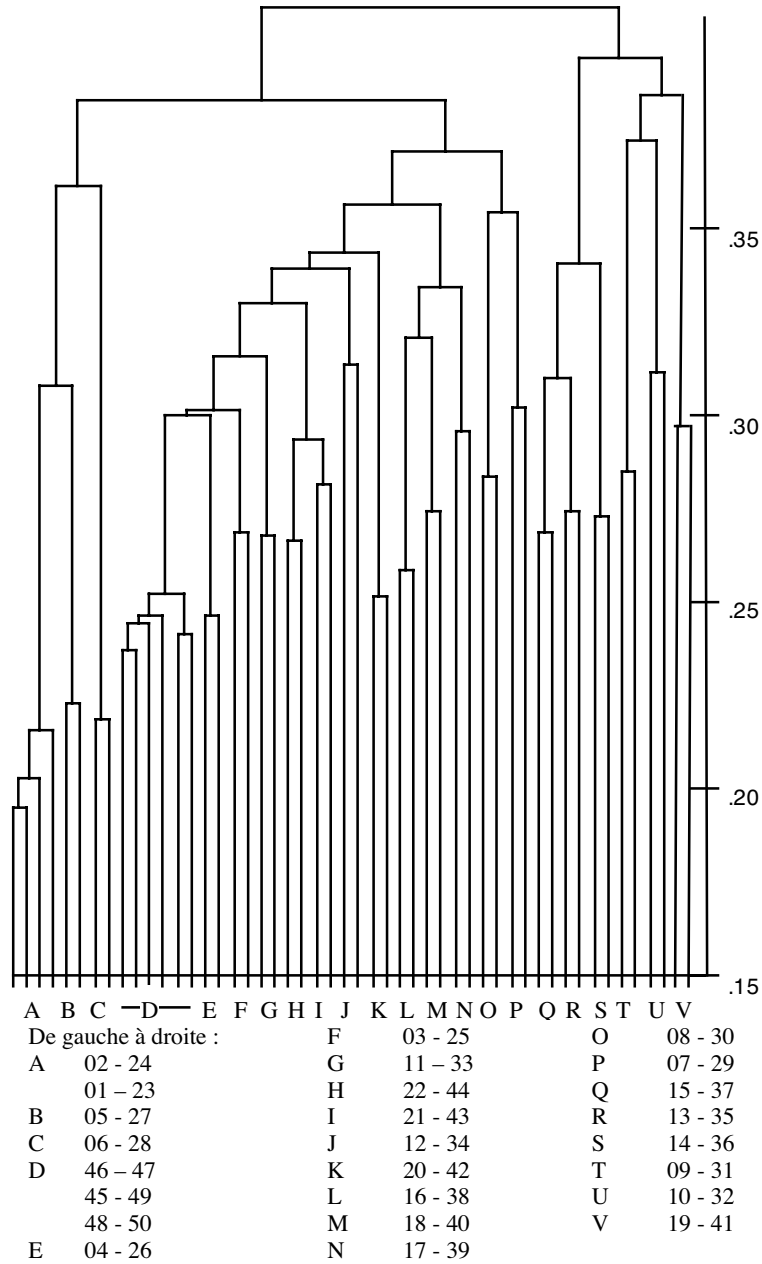
La **classification automatique ascendante** est opérée sur la matrice des distances. L'algorithme procède à la construction d'une classe en regroupant les deux textes séparés par la distance la plus faible (ici 02 et 24), puis il recalcule les distances des autres textes par rapport à ce nouvel ensemble, etc. Et ceci jusqu'à la constitution d'un ensemble unique. Ces regroupements successifs — par la technique dite de la « moyenne simple avec saut minimal » — sont représentés par un « dendogramme » (tableau IV). L'ordre d'agrégation se lit de gauche — les textes les plus proches — à droite (les textes les plus singuliers ou les plus éloignés des autres) et les distances correspondantes aux différents niveaux d'agrégation se lisent en ordonnées.

La distance est indiquée sur l'axe vertical : plus la jonction entre deux traits est élevée plus les textes ou groupes de textes sont éloignés. En coupant le graphe, horizontalement et au plus près de l'un des seuils mentionnés ci-dessus, on pourra isoler les groupes de textes très proches, relativement proches, etc. Ces groupes étant isolés, on pourra étudier en quoi leurs vocabulaires diffèrent grâce à l'étude de leurs spécificités. On notera que, plus l'on s'élève dans le graphe, plus les classes constituées sont hétérogènes et plus l'interprétation des différences deviendra complexe. D'où plusieurs lectures possibles selon la hauteur à laquelle on se place sur la graphe.

Au niveau "micro" (au plus près de l'axe horizontal), l'algorithme a repéré plusieurs choses :

- un bloc constitué des quatre documents que l'on a attribués au même auteur (A) ;
 - un autre bloc constitué des six derniers textes (45 à 50) que l'on peut aussi très probablement considérer comme du même auteur (nous avons souligné plus haut l'étrangeté de ces textes par rapport au reste du corpus)...
 - les autres textes sont groupés en 20 "paires" et l'algorithme ne laisse aucun "orphelin".
- On peut donc affirmer que le corpus comporte probablement 22 ou 23 auteurs différents et/ou extraits de 22 ou 23 textes différents. Comme nous l'avons indiqué ci-dessous, certaines paires peuvent être attribuées avec un degré raisonnable de certitude au même auteur. Pour d'autres, le groupement est réalisé « par défaut » : c'est l'hypothèse la plus probable mais, plus le nœud est situé haut, plus la conclusion devra être tirée avec prudence. C'est le cas notamment pour les couples J et U dont les "jambes" se rejoignent au-dessus de 0.3 et qui sont donc fortement hétérogènes.

Tableau IV. Classification ascendante sur les 50 textes



Au niveau « meso » — les nœuds immédiatement supérieurs aux paires — les conclusions sont beaucoup moins solides tant ces agrégations s'opèrent haut dans le graphe. On peut soupçonner un même auteur — ou des auteurs différents écrivant à la même époque et sur des thèmes assez proches — dans {H-I} puis, de façon moins certaine, dans les ensembles {D-E-F}, {A-B-C}, {L-M-N}, {Q-R-S}. De même, en utilisant le raisonnement par défaut discuté ci-dessus, on peut également « marier » : O avec P ainsi que les paires T et U ;

Enfin, au niveau "macro", on peut distinguer quelques grands groupes. Globalement, trois vastes ensembles s'opposent. Le premier va de A à C ; le second de D à P et le troisième de Q à U. Seuls se placent nettement à l'écart les deux textes formant le couple V (19-41) qui est manifestement "à part" dans ce corpus. A ce niveau macro, plusieurs conclusions sont possibles : si l'on subodore un même auteur, ces textes n'ont pas été écrits à la même époque,

le genre ou les thèmes traités sont divers. L'existence de plusieurs auteurs demeure l'hypothèse la plus probable. Sont-ils regroupés à cause de leur écriture ? du genre ? des époques ? D'autres outils que nous ne pouvons décrire ici — comme les spécificités du vocabulaire, les syntagmes répétés, les structures de phrases... — devront être utilisés pour répondre à ces questions et expliquer les proximités relatives, les principales oppositions.

Du point de vue méthodologique, on remarquera que la classification automatique traditionnelle produit des "effets de chaîne". Certaines proximités entre textes ne sont pas discernables car les sommets qui les relient sont effacés par des agrégations effectuées à un niveau inférieur (autrement dit quand un texte est "marié" à un autre, on recalcule la distance de ce "couple" à tous les autres, de telle sorte que les liens existant entre l'un ou l'autre des membres de ce couple avec un "tiers" sont effacés). L'arbre ne doit donc pas être utilisé aveuglément. L'appartenance de chacun des textes à une classe donnée doit éventuellement être contrôlée sur la matrice des distances. Il n'est pas mauvais non plus de recalculer la distance moyenne de chacun des textes à l'ensemble de ses "voisins" supposés. Cette réserve admise, l'algorithme correspond exactement à ce que nous cherchons : une technique exploratoire permettant de repérer dans un corpus des "familles" de textes plus ou moins homogènes. Pour aller plus loin, il faut recourir à des outils plus sophistiqués comme l'analyse arborée.

L'ANALYSE ARBOREE

L'algorithme mis au point par Xuan Luong combine l'analyse topologique et la classification arborée⁵. Il s'agit d'obtenir, dans un plan, la meilleure représentation possible des distances de chacun des textes à tous les autres. Chaque texte est représenté par une feuille terminale de l'arbre. La distance qui le sépare d'un autre est matérialisée par la longueur du chemin à parcourir sur l'arbre pour unir ces deux textes. Les textes qui sont rattachés à un même nœud forment des groupes plus ou moins homogènes en fonction de la longueur des branches (tableaux V et VI).

⁵ Jean-Pierre Barthélémy, Alain Guénoche, *Les arbres et les représentations des proximités*, Paris, Masson, 1988. Jean-Pierre Barthélémy, Xuan Luong, "Représenter les données textuelles par des arbres", in Sylvie Melley (ed), *4e journées internationales d'analyse statistique des données textuelles*, Université de Nice, 1998, p. 49-71. Xuan Luong, "L'analyse arborée des données textuelles : mode d'emploi", *CUMFID*, 1994, 16, p 25-42.

Tableau V. Analyse arborée sur les distances originales du corpus "Brunet".

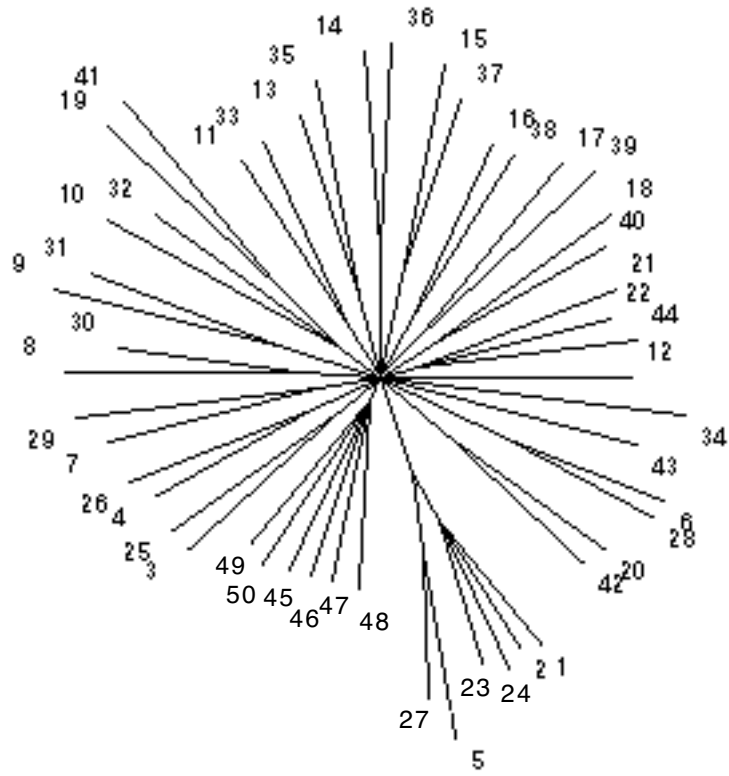
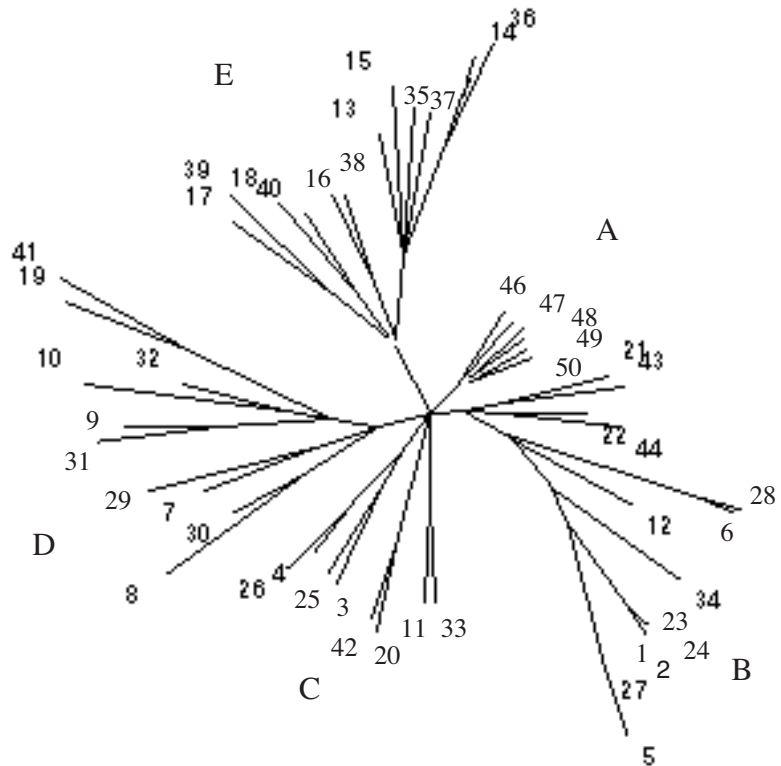


Tableau VI. Analyse arborée sur les distances diminuées de 0.19.



Le premier graphe a été établi avec les distances "brutes" et donne une étoile presque parfaite mais il permet déjà de confirmer certaines conclusions déjà acquises. Par exemple, au bas du diagramme et au plus près du centre, on trouve le groupe formé par les 6 derniers textes {45 à 50} ou, dans le quart sud-est, le "quatuor" {01-02-23-24} qui a été reconnu comme étant certainement du même auteur. Pour le reste, on retrouve la structure par couples qui a déjà été identifiée, les branches se rejoignant plus ou moins près du centre en fonction de la distance relative séparant les deux textes attribués au même auteur (ainsi le couple formé des textes 19 et 41 est à la fois le plus hétérogène et le plus décalé par rapport au centre comme nous l'avions déjà constaté sur le dendrogramme).

Le second graphe (tableau 6) opère une sorte de "grossissement". L'algorithme de X. Luong utilise le "théorème" selon lequel la topologie de l'arbre est inchangée lorsque l'on retranche de toutes les distances une quantité légèrement inférieure à la plus petite d'entre elles, ici celle du groupe {02-24} soit 0,19 (de même que, dans le dendrogramme ci-dessus, l'origine est placée à cette distance minimale). Pour les arbres, cette opération ne change rien à la disposition des textes dans le plan, mais elle réduit la longueur des branches reliant les "feuilles" terminales — les feuilles les plus proches apparaissent maintenant comme un point unique (01-02-23-24 et 05-27) — et elle grossit les "troncs", c'est-à-dire les sections unissant les principaux nœuds. Naturellement, le "grossissement" obtenu ne doit pas faire oublier la relative "équidistance" des couples mise en valeur dans le graphe précédent.

Le graphe fait apparaître une série de groupes remarquables. Outre celui formé par les 6 derniers textes (A) qui figure au plus près du nœud central, on identifie quatre ensembles assez nettement séparés et qui confirment la plupart des conclusions tirées de la classification automatique.

— B : dans le quart sud-est, au groupe {01-02-23-24} viennent se joindre {06-28, 12-34, 05-27} et {21-43, 22-44} ;

— le groupe C constitue la famille la plus hétérogène (les chemins reliant ces feuilles sont les plus longs et les branches se rejoignent quasiment au centre du graphe) : {03-25, 04-26} {20-42}, {11-33} ;

— D regroupe : {07-29-08-30} avec {09-31-10-32} et {19-41}, ce dernier couple restant le plus éloigné du nœud central ;

— E rassemble {14-36-13-15-35-37} avec {17-39, 18-40, 16-38} et constitue une famille relativement homogène mais aussi nettement décalée par rapport au nœud central ;

La moisson est donc sensiblement plus riche qu'avec la classification automatique classique mais elle se heurte aux mêmes limites en ce qui concerne l'interprétation de ces grandes familles de textes. L'analyse devra donc se poursuivre avec d'autres instruments lexicométriques que nous renonçons à évoquer pour ne pas nous écarter de la question posée.

CONCLUSIONS

Le lecteur trouvera dans la partie rédigée par E. Brunet les éléments qui lui permettront de juger la fiabilité de notre méthode en substituant aux numéros, les auteurs et les titres des extraits correspondant⁶.

En ce qui concerne l'attribution d'auteur, nous rappellerons que nous cherchons à conclure à coup sûr, quitte à conclure moins souvent. Il ne s'agit pas de reconnaître tous les auteurs mais de ne pas se tromper quand on en reconnaît un... Naturellement, des expériences comme celle qui vient d'être présentée permettront d'améliorer les résultats et d'étudier en détail les propriétés de la distance.

La principale caractéristique de ce corpus demeure la relative "équidistance" entre tous ces textes. Cependant, l'arbre fait mieux ressortir que le dendrogramme un certain nombre d'"affinités" pour lesquelles il est évidemment impossible de trancher entre les quatre facteurs principaux qui agissent sur la distance : auteur, époque, genre et thème... En fonction des éléments révélés par E. Brunet, il sera possible de réfléchir aux moyens de neutraliser l'un ou l'autre de ces facteurs — notamment auteurs et époques — ce qui permettra de faire apparaître des "familles littéraires", des "filiations", etc.

La taille relativement limitée des extraits n'autorisera sans doute pas de conclusions définitives sur les textes et les auteurs en question. Cependant, si notre analyse ne contient pas d'aberrations, on aura prouvé que, au-delà de la question de l'attribution d'auteur qui nous intéressait ici, la classification automatique combinée avec la distance intertextuelle pourront être des outils intéressants pour la critique littéraire.

⁶ Le texte que l'on vient de lire est le compte-rendu que nous avons adressé à Etienne Brunet à la fin de l'expérience. Certes, nous aurions facilité la tâche du lecteur en fusionnant nos deux articles et levant l'anonymat des textes, mais, à la réflexion, il nous a semblé préférable de permettre au lecteur de juger "sur pièces". Le tableau en annexe a été communiqué par E. Brunet après la remise du compte-rendu qu'on vient de lire.

Annexe. La composition du corpus "Brunet"

Numéro	Auteur	Extraits de :
1	1 Marivaux	<i>La Vie de Marianne (L.1)</i>
2	2Marivaux	<i>Le Paysan parvenu (L.1)</i>
3	1Voltaire	<i>Zadig</i>
4	2Voltaire	<i>Candide</i>
5	1Rousseau	<i>La Nouvelle Héloïse (L.1)</i>
6	2Rousseau	<i>Emile (L.5)</i>
7	1Chateaubriand	<i>Atala</i>
8	2Chateaubriand	<i>La Vie de Rancé</i>
9	1Balzac	<i>Les Chouans</i>
10	2Balzac	<i>Le Cousin Pons</i>
11	1Sand	<i>Indiana</i>
12	2Sand	<i>La Mare au diable</i>
13	1Flaubert	<i>Madame Bovary</i>
14	2Flaubert	<i>Bouvard et Pécuchet</i>
15	1Maupassant	<i>Une Vie</i>
16	2Maupassant	<i>Pierre et Jean</i>
17	1Zola	<i>Thérèse Raquin</i>
18	2Zola	<i>La Bête humaine</i>
19	1Verne	<i>De la terre à la lune</i>
20	2Verne	<i>Secrets de Wilhelm Storitz</i>
21	1Proust	<i>Du côté de chez Swann</i>
22	2Proust	<i>Le Temps retrouvé</i>
23	3 Marivaux	<i>La Vie de Marianne (L.1)</i>
24	4 Marivaux	<i>Le Paysan parvenu (L.1)</i>
25	3 Voltaire	<i>Zadig</i>
26	4 Voltaire	<i>Candide</i>
27	3 Rousseau	<i>La Nouvelle Héloïse (L.1)</i>
28	4 Rousseau	<i>Emile (L.5)</i>
29	3 Chateaubriand	<i>Atala</i>
30	4 Chateaubriand	<i>La Vie de Rancé</i>
31	3Balzac	<i>Les Chouans</i>
32	4Balzac	<i>Le Cousin Pons</i>
33	3Sand	<i>Indiana</i>
34	4Sand	<i>La Mare au diable</i>
35	3 Flaubert	<i>Madame Bovary</i>
36	4 Flaubert	<i>Bouvard et Pécuchet</i>
37	3 Maupassant	<i>Une Vie</i>
38	4 Maupassant	<i>Pierre et Jean</i>
39	3 Zola	<i>Thérèse Raquin</i>
40	4Zola	<i>La Bête humaine</i>
41	3Verne	<i>De la terre à la lune</i>
42	4Verne	<i>Secrets de Wilhelm Storitz</i>
43	3Proust	<i>Du côté de chez Swann</i>
44	4Proust	<i>Le Temps retrouvé</i>
45	Collages de Brunet	pages 1 de tous les textes
46		pages 10 de tous les textes
47		pages 20 de tous les textes
48		pages 30 de tous les textes
49		pages 40 de tous les textes
50		pages 50 de tous les textes