Peut-on se fier aux arbres ?

Cyril Labbé, Dominique Labbé

Cyril LABBE
Université Grenoble I (cyril.labbe@imag.fr)

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

Peut-on se fier aux arbres ?

In

Serge HEIDEN et Bénédicte PINCEMIN (eds). *Actes des 9e journées internationales d'analyse statistique des données textuelles (Lyon, 12-14 mars 2008)*. Lyon : Presses Universitaires de Lyon, 2008, tome II, p. 635-645.

Version soumise au comité scientifique du congrès et acceptée sans modification.

Peut-on se fier aux arbres ?

Cyril LABBE[1], Dominique LABBE[2]

1. Université Grenoble I (cyril.labbe@imag.fr)

2. Institut d'Etudes Politiques de Grenoble (dominique.labbe@iep.grenoble.fr)

Abstract

Intertextual distance provides a simple and interesting solution to measure proximities and oppositions in large text corpora. Its properties make it a good tool for text classification, and especially for tree-analysis which is presented and discussed in this paper. In order to measure the quality of this classification, two indices are proposed. The method presented provides an accurate tool for literary studies and authorship attribution - as is demonstrated by its application to a blind test.

Résumé

La distance intertextuelle fournit une solution simple et intéressante pour mesurer les proximités et les oppositions dans un grand corpus de textes. Ses propriétés en font un bon outil pour la classification des textes, spécialement pour l'analyse arborée qui est présentée et discutée. Deux indices sont proposés pour mesurer la qualité de ces classifications. La méthode fournit un outil efficace pour les études littéraires et l'attribution à des auteurs connus de textes d'origine douteuse ou inconnue, ainsi qu'il est démontré grâce à une expérience en aveugle.

Mots clefs : Distance intertextuelle – classification arborée – attribution d'auteur – qualité des graphes.

1. Introduction

Grâce à l'ordinateur, les méthodes de classification ont connu un essor considérable. Parmi celles-ci, la classification arborée est classique en génétique (Felsenstein 2004a et 2004b ainsi que le site : http://evolution.genetics.washington.edu) ou en linguistique historique (Embleton 1986 et pour une revue récente : Holm 2007).

Cet outil a été appliqué à l'analyse des entretiens sociologiques (notamment : Bergeron & Labbé 2000, Labbé & Labbé 2001b), au discours politique (notamment Labbé & Monière 2000, Labbé & Monière 2003) à l'attribution à un auteur connu de textes inconnus ou d'origine douteuse (notamment : Labbé & Labbé 2001, Merriam 2002, Merriam 2003a, Merriam 2003b, Monière & Labbé 2006, Lafon & Peeters 2006).

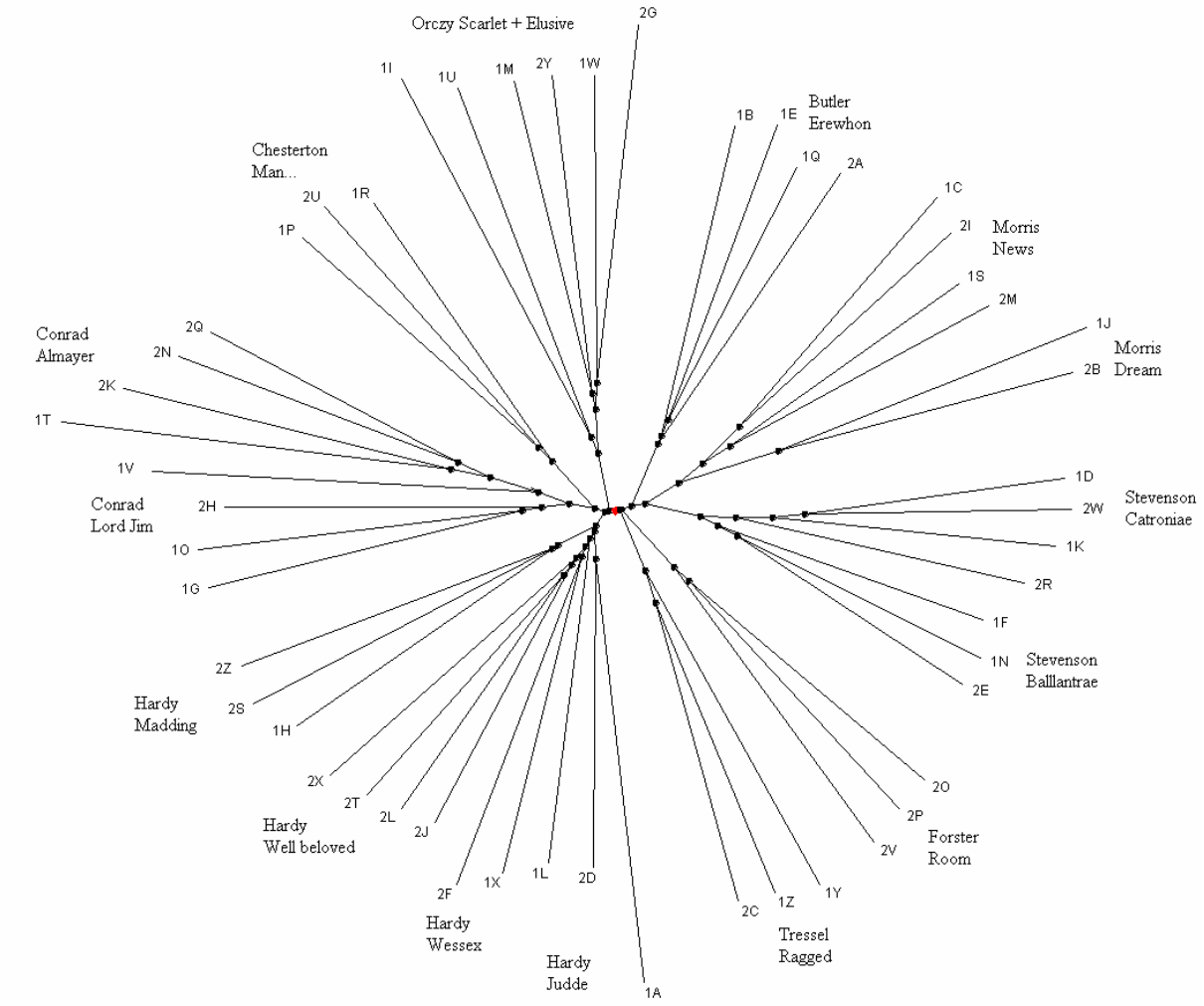
Quelle confiance accorder à ces classifications arborées ? Certaines mesures permettent de répondre à ces questions en évaluant la fiabilité des résultats de cette classification. On utilise comme exemple les résultats d'une expérience en aveugle réalisée en 2004 avec deux chercheurs anglais.

2. Les expériences Oxquarry

A la demande de Gerard Ledger et Thomas Merriam, une série d'expériences en aveugle ont été réalisées (Labbé 2007). Lors de la première expérience, G. Ledger a soumis 52 textes anonymés en demandant lesquels de ces textes étaient écrits par les mêmes auteurs et, par conséquent, lesquels étaient d'auteurs différents. Ce corpus – nommé par G. Ledger : "Oxquarry1" - est décrit en annexe 1. Ces textes avaient été choisis parce qu'il semblait difficile de distinguer les auteurs de certains d'entre eux. Deux indications étaient fournies : il y avait plusieurs auteurs et chacun de ces auteurs avait au moins deux textes.

Ces textes ont été traités selon la norme "OCP" (Hockey & Martin 1988). Leurs distances mutuelles ont ensuite été calculées. Enfin la classification arborée a été réalisée sur ces distances (tableau 1).

Tableau 1. Classification arborée du corpus Oxquarry1



Ce résultat est obtenu de manière entièrement automatique : aucune opération manuelle sur les textes ; l'opérateur intervient à aucun stade ; les calculs portent sur la totalité des textes sans qu'aucun mot en soit retiré. L'expérience s'est déroulée sous le contrôle de G. Ledger et T. Merriam qui ont pu en vérifier l'entière reproductibilité.

Pour la présente communication, les noms des auteurs et les titres des oeuvres ont été ajoutés au document original qui ne comportait que les numéros des textes soumis au test. Avant que l'information contenue dans l'annexe 1 soit fournie par G. Ledger et T. Merriam, un certain nombre de questions se posaient. Ce graphique traduit-il fidèlement l'organisation du corpus ? Peut-on se fier à un graphique de ce genre ? Pour répondre à ces questions, il est nécessaire d'exposer la manière dont le graphique a été tracé. Comme toute opération de classification, cet arbre repose d'abord sur le calcul d'une distance entre les objets classifiés.

3. Distances entre textes et classification arborée

La distance entre deux textes est mesurée par le nombre de mots ("tokens") différents qu'ils contiennent (formules dans Labbé & Labbé 2001). Cette mesure est une *distance* – et non pas une simple mesure de dissimilarité - car elle présente trois propriétés caractéristiques :

- positivité : $d_{(a,b)} \geq 0$ et $d_{(a,a)} = 0$ (la distance d'un texte à lui-même est nulle ; si $d_{(a,b)} = 0$, alors A et B contiennent les mêmes mots avec les mêmes fréquences) ;

- symétrie : $d_{(a,b)} = d_{(b,a)}$ (le résultat est le même que la mesure soit effectuée en considérant d'abord A ou B) ;

- inégalité triangulaire : $d_{(a,b)} \leq d_{(a,c)} + d_{(c,b)}$ (l'égalité n'est possible que si le texte C est un sous-ensemble de A et de B).

Ce calcul appliqué aux 52 textes du corpus Oxquarry1 – pris deux à deux – génère un tableau de 2 704 cellules – 52 colonnes et 52 lignes - dont la taille interdit une reproduction intégrale. Du fait de la propriété d'identité, la diagonale de ce tableau est nulle (soit 2 652 cases non nulles) et du fait de la propriété de symétrie, il y a 1 326 distances différentes (2 652/2). Comme indiqué dans Labbé 2007, les distances les plus courtes permettent de regrouper correctement la quasi-totalité des textes. Cependant, devant des populations aussi vastes, le recours à des classifications est une nécessité (pour une présentation de la question : Sneath & Sokal 1973 et Benzecri 1980).

La méthode usuelle consiste à représenter l'ensemble des textes par des points dont les coordonnées dans l'espace sont déterminées par leur position relative par rapport à tous les autres. Ici les 52 textes forment un "nuage" de points comprenant 1 326 distances différentes. Par la méthode de l´"analyse factorielle des correspondances" (Lebart et Salem, 1994), on détermine d'abord le plan qui passe au plus près de tous ces points (et par le barycentre du nuage) puis l'on projette orthogonalement chacun des points sur ce plan, ce qui donne une représentation plane du nuage. Cette méthode a un inconvénient évident : un point sera d'autant plus fidèlement représenté qu'il sera proche du plan d'ajustement ; en revanche, les points les plus éloignés risquent d'être "mal" représentés.

La classification arborée ne présente pas de tels inconvénients.

L'analyse arborée repose sur le théorème suivant : si tous les individus étudiés sont séparés par des distances (présentant toutes les trois propriétés énoncées ci-dessus), il existe un "arbre" qui représente exactement les positions respectives de ces individus les uns par rapport aux autres (Pour la démonstration : Luong 1988). Cependant, la construction d'un tel arbre "parfait" exigerait que toutes les combinaisons possibles soient examinées alors que leur nombre augmente exponentiellement en raison de l'effectif de la série (la conclusion revient sur ce point). Divers algorithmes ont été imaginés pour construire cet arbre sans avoir à examiner toutes ces combinaisons. Nous utilisons l'algorithme mis au point par X. Luong (code source dans Luong 1988, les principes et les formules sont également présentées dans Luong 1994). Notre logiciel a été réalisé avec son aide et avec celle de M. Ruhlman (Ruhlman 2003).

Pour présenter l'analyse arborée, examinons d'abord 4 textes du corpus Oxquarry1, tous extraits du roman de Morris (*News*). Le problème de la représentation dans un plan ne se pose qu'à partir de 4 individus (3 donnent un plan et 2 une droite). On verra plus bas que ces 4 textes sont parmi ceux qui posent le plus de problèmes à l'algorithme de construction de l'arbre. Dans le tableau 2, la distance est exprimée en pour 10.000 mots. Par exemple, les

textes 1C et 1S ont 3 531 mots différents pour 10 000 (ou encore ils partagent 6 469 mots en commun), etc.

Tableau 2. Distances intertextuelles entre les 4 extraits de Morris (pour 10 000 mots)

	A (1C)	B (2I)	C (1S)	D (2M)
A (1C)	-	2 881	3 531	3 513
B (2I)	2 881	-	3 030	2 972
C (1S)	3 531	3 030	-	2 809
D (2M)	3 513	2 972	2 809	-

Pour déterminer la position de ces 4 points, l'algorithme utilise la formule suivante (dite "condition des 4 points" :

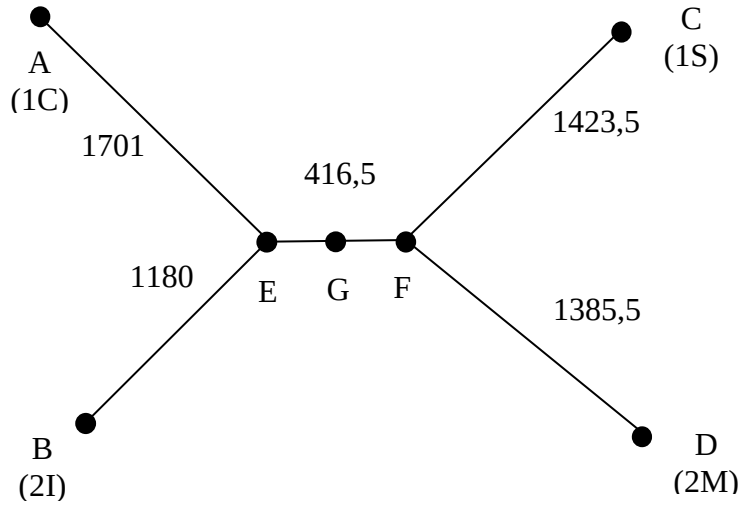
$$(1)\ d(a,b) + d(c,d) \leq \text{MIN}[d(a,c) + d(b,d),\ d(a,d) + d(b,c)].$$

L'algorithme affecte, aux 6 couples possibles, un "score" établi en fonction du nombre de fois que deux couples de textes - considérés par rapport à tous les autres possibles - se trouvent remplir cette condition des quatre points. Dans le tableau 2 ci-dessus, on a effectivement :

$$(AB + CD) < (AC + BD)\ ;\ (AB + CD) < (AD + BC), \text{ etc.}$$

Les couples AB et CD obtiennent un score de 1 et les 4 autres (AC, AD, BC et BD) un score nul. Les arêtes joignant A à B et C à D seront tracées d'abord, ce qui détermine la structure de l'arbre (tableau 3).

Tableau 3 Classification arborée des 4 textes du tableau 2



Pour tracer l'arbre, le calcul de la longueur des arêtes se fait de la manière suivante (pour les formules générales, voir Luong 1988 et Ruhlman 2003) :

AE = AB/2 + ((AC + AD) – (BC + BD))/4 = 1701

BE = AB - AE = 1180

Les textes A et B sont maintenant représentés par le point E. Les arêtes reliant le point F aux points B et D sont calculées de la même manière :

CF = CD/2 + ((CA + CB) – (DA + DB))/4 = 1423,5

DF = CD - CF = 1385,5

Enfin, les points E et F sont joints et la racine est placée au milieu du segment central :

EF = ((AC + AD + BC + BD) – (2AB + 2CD))/4 = 416,5

La topologie de l'arbre n'aurait pas été modifiée si l'on avait commencé par placer le couple CD au lieu du couple AB. Autrement dit, au cours d'une même itération, l'ordre, dans lequel les feuilles sont placées et les groupes constitués, est indifférent.

Dans cet arbre, A, B, C, D sont les **feuilles** ; E et F les **nœuds** figurant respectivement les groupements de A avec B et de C avec D ; G est la **racine** de l'arbre. Les segments de droite, ou **arêtes,** sont des **branches** quand elles relient des feuilles à des nœuds et des **troncs** quand elles relient des nœuds entre eux. La distance entre deux points quelconques est figurée par le **chemin** unissant ces points et la longueur de ce chemin est proportionnelle à la distance originelle correspondante (**arbre valué**). Par exemple, le chemin A-C est égal à : 1701 + 416,5 + 1423,5 = 3541. Le fait que cette distance arborée soit très légèrement différente de la distance intertextuelle (3531) suggère que la représentation n'est pas parfaite.

En étudiant la **contribution** des mots **à la distance** (Labbé & Labbé 2003) et le vocabulaire propre des quatre textes de Morris, il apparaît que certaines distorsions proviennent de deux choses. En premier lieu, quelques variantes graphiques font que le même mot est compté comme deux mots différents, ce qui expliquerait les légères distorsions constatées dans le tableau 3. En second lieu, les difficultés à classer les textes de Morris dans le tableau 1 proviendraient de ce qu'ils sont en orthographe américaine alors que les autres sont à la norme anglaise (*neighbour*, *labour*, *honour*, *splendour*, *centre*, *recognise*, etc.). Autrement dit, les incertitudes ne proviendraient pas du procédé de mesure mais de l'hétérogénéité du matériau, ou encore de différences dans le "calibrage" des textes. C'est pourquoi, sur les textes français, nous opérons une standardisation ("normalisation") des graphies avant tout calcul statistique.

Enfin, les longueurs des arêtes AB et CD sont égales aux distances correspondantes dans le tableau 2. Cela révèle une caractéristique importante de l'algorithme de Luong : les arbitrages, nécessités par les distorsions évoquées ci-dessus, sont repoussés à la dernière étape (le tracé du tronc central EF) et portent donc sur une portion réduite de l'arbre (dans le tableau 3, ce tronc central est trois à quatre fois plus court que les branches terminales). Cette méthode peut donc engendrer certains problèmes.

4. Contrôle de la qualité des opérations

Pour contrôler la qualité d'un arbre, X. Luong propose de calculer un indice d'agrégation (ci-dessous : *Agreg*) en utilisant la formule (1) (condition des quatre points). A une étape donnée, soit *N* le nombre de textes (ou groupes de textes représentés par un nœud) restant à classer et $d_{(a,b)}$, la distance séparant deux textes et/ou noeuds A, B non encore agrégés à cette étape. Le score de l'arête potentielle A-B ($Sobs_{(a,b)}$) sera le nombre de fois que $d_{(a,b)}$ satisfera (1). Le score maximum théorique ($Stheo_{(a,b)}$) est égal à ((N-2).(N-3))/2, c'est-à-dire le nombre de couples possibles pour une population composée de N-2 individus séparés par des distances. Pour ce couple (*A*,*B*), l'indice de Luong est :

$$Agreg = \frac{Sobs_{(a,b)}}{Stheo_{(a,b)}}$$

Cet indice varie entre 0 et 1. Toute valeur inférieure à 1 signale que certaines parties de l'arbre ne satisfont pas complètement à la condition (1). Pour le graphique 1, l'indice moyen est égal à 0,997. Dans 152 cas sur plus de 12 000, la condition (1) ne s'est pas trouvée remplie. Cette situation se produit dès que l'équation (1) aboutit une inégalité inverse ne serait-ce que d'un mot. Or, dans un corpus comme Oxquarry1, une telle situation est pratiquement inévitable. En

effet, 95% des distances sont comprises entre 3 320 et 4 820 soit une forte concentration autour de la moyenne (4 067). La plupart des violations de (1) s'expliquent par de très faibles différences entre les distances concernées et l'examen des textes correspondant permet de localiser la source de ces légères discordances – comme indiqué ci-dessus à propos des 4 extraits de Morris. A titre d'exemple, voir les 11 nœuds formant le cluster Hardy sur la figure 1 : certains d'entre eux sont séparés par des chemins extrêmement courts (quelques mots) qui sont le résultat de plusieurs centaines de calculs comme ceux présentés sous le tableau 3.

On en tire qu'il est nécessaire de définir un seuil en dessous duquel on pourra considérer que la relation (1) est acquise sans l'inégalité stricte.

Ce seuil aurait une seconde utilité. En effet, la formule (1) interdit de regrouper plus de deux textes à la fois. Dès que le nombre d'individus à classer dépasse quelques dizaines, cela donne un grand nombre de noeuds et des arbres difficiles à lire. Pour avoir des arbres plus clairs, l'introduction d'un seuil "de tolérance" permet de rattacher plus de deux textes à un même nœud. Dans le code source du programme publié dans Luong 1988, il apparaît que l'auteur a introduit une tolérance de 10% pour autoriser le tracé d'une arête malgré la violation de la condition des 4 points, ce qui aboutit à un nombre beaucoup plus réduit de nœuds dans les arbres tracés à l'aide du logiciel de Luong. Nous nous sommes interdit cette facilité au moins dans la phase expérimentale actuelle.

Dans l'expérience Oxquarry1, un seuil de tolérance de 5% aurait suffi pour obtenir un taux d'agrégation de 100%. Cela aurait aussi permis, par exemple, de rattacher les textes de Hardy à trois nœuds correspondant aux trois œuvres dont sont tirés les 12 extraits de cet auteur et celles de Orczy à deux nœuds correspondant aux deux œuvres présentes dans le corpus Oxquarry.

Cette discussion suggère de modifier l'indice de Luong afin de répondre à deux questions :

- l'arbre obtenu est-il le plus efficace et le plus simple possible ? Nous revenons en conclusion sur cette première question.

- avec quelle fidélité l'arbre représente-t-il les distances originales ? La réponse est donnée par le rapport entre la distance initiale et la longueur du chemin reliant les deux feuilles correspondantes sur l'arbre. Le tableau 4 donne les résultats de ce calcul pour l'arbre obtenu sur les 4 textes de Morris (tableau 2 et figure 3 ci-dessus).

Tableau 4. Calcul des indices de confiance des chemins de l'arbre du tableau 2

Noeud	Chemins induits	Distances initiales	Distances arborées	Qualités des chemins	Qualité du noeud
E	A - B	2 881	2 881	1	1
F	C - D	2 809	2 809	1	1
G	A - C	3 531	3 541	0,9972	0,9969
	A - D	3 513	3 503	0,9972	
	B - C	3 030	3 020	0,9967	
	B - D	2 972	2 982	0,9966	
Total		18 736	18 736	0,9979	

Les chemins reliant les arêtes opposées (A-B et C-D) restituent intégralement l'information contenue dans la matrice originale ; le tronc central de l'arbre en restitue 99,7 % et l'arbre total 99,8 %. Pour l'arbre présenté au début de cette communication, les valeurs sont les suivantes :

- qualité moyenne de l'arbre (98,17%). L'information initiale contenue dans la matrice des distances est donc restituée avec une incertitude inférieure à 2%.

- pour les nœuds, l'indice le plus faible (95,3%) est atteint par le noeud reliant les quatre textes extraits de News par Morris qui ont servi d'exemple ci-dessus. Tous les autres nœuds ont un indice supérieur. Par exemple, le dernier nœud placé avant la racine (reliant l'œuvre de Orczy au tronc central) induit 662 chemins. Il restitue ces 662 distances avec un indice de 98,2% qui est la moyenne des 662 indices tous supérieurs à 95%.

- Pour les 1 326 chemins unissant chaque feuille terminale à toutes les autres (tableau 5), 69 seulement ont un indice de qualité inférieur à 95% (mais tous supérieurs à 90%). Autrement dit, en acceptant le "seuil de tolérance" utilisé par Luong dans la construction de ses arbres (10%), on peut affirmer que l'arbre, présenté au début de cette communication, est fiable.

Tableau 5. Indices de confiance des chemins entre feuilles
(classement par indices décroissants)

Indices	Effectifs absolus	%
$X \geq 0{,}9999$	65	4,90
$0{,}9999 > X \geq 0{,}990$	388	29,26
$0{,}990 > X \geq 0{,}980$	368	27,75
$0{,}980 > X \geq 0{,}970$	236	17,80
$0{,}970 > X \geq 0{,}960$	133	10,03
$0{,}960 > X \geq 0{,}950$	67	5,05
$0{,}950 > X \geq 0{,}900$	69	5,21
	1326	100,00

5. Conclusions

Certaines objections sont souvent opposées à ces expériences. Par exemple, un résultat, comme celui de la figure 1, peut être un simple coup de chance. Ou encore, à une époque donnée, tous les auteurs utiliseraient à peu près le même vocabulaire, par conséquent, les indices, comme la distance intertextuelle, ne permettraient pas de rendre compte des "vraies" différences entre les auteurs, etc.

Ces objections reviennent à considérer comme équiprobables toutes les combinaisons possibles dans le corpus "Oxquarry". Or, la figure 1 identifie 13 couples, 8 trios, etc., soit 62 combinaisons représentées par autant de nœuds sur l'arbre. Combien a-t-on de chances d'obtenir ces combinaisons au hasard (en admettant que toutes les combinaisons possibles sont équiprobables) ? Il y a 52! ($8{,}066\ e^{67}$) manières différentes de combiner 52 objets. La probabilité de tirer successivement les 62 objets recherchés - en 62 tirages successifs dans une urne contenant $8{,}066e^{67}$ objets différents - est :

$$(62/8{,}066\ e^{67}) \cdot (61/(8{,}066\ e^{67} - 1)) \ldots (1/(8{,}066\ e^{67} - 61)) = 1{,}91^{e\text{-}4030}$$

De plus, le test organisé avec G. Ledger et T. Merriam comportait deux expériences organisées selon le même principe. Toutes deux ont été couronnées de succès. En admettent que toutes les combinaisons sont équiprobables, la probabilité pour que l'enchaînement de ces succès soit le fait du hasard est donc le carré du résultat ci-dessus…

Ce calcul est évidemment absurde car ce que démontrent des expériences comme "Oxquarry", c'est justement que certaines combinaisons sont plus "probables" que d'autres et cela essentiellement pour trois raisons. Premièrement, les auteurs – même contemporains et traitant de sujets proches dans un même genre – n'utilisent pas exactement les mêmes mots avec les mêmes fréquences : il est donc possible, grâce à une mesure judicieusement calibrée de rendre compte de ces différences. Deuxièmement, la distance intertextuelle – dans les limites de validité définies par Labbé & Labbé 2001 (pour le français) et Labbé 2007 (pour

l'anglais) – peut rendre compte de ces différences parce qu'elle possède les propriétés d'une *distance*. Enfin, la classification arborée offre une représentation – en deux dimensions - très satisfaisante d'un nuage de points séparés par plusieurs milliers de *distances* différentes (à condition qu'il s'agisse effectivement de *distances*). Dès lors, la combinaison de la distance intertextuelle et de la classification arborée offre un outil efficace pour l'attribution, à un auteur connu, de textes d'origine plus ou moins douteuse ou inconnue.

Il reste deux sources possibles d'erreur.

D'une part, certains auteurs peuvent se masquer et tenter de "brouiller les pistes". Il existe plusieurs exemples dans l'histoire littéraire qui permettent de "tester" cette hypothèse, notamment le cas Gary-Ajar. Cet exemple suggère que ces tentatives - même menées avec beaucoup de talent - sont vaines dès lors que les textes dépassent quelques dizaines de pages (Bona 1987 ; Lafon & Peters 2006).

D'autre part, des imperfections dans les traitements et calculs peuvent être source d'incertitude. De ce point de vue, un résultat fiable dépend de deux conditions.

Premièrement, l'orthographe des textes doit avoir été soigneusement révisée et les graphies normalisées. En français – langue fortement flexionnelle – il apparaît également nécessaire de travailler sur les vocables et non sur les formes graphiques. Enfin, les textes contenant une proportion significative de mots étrangers ou de "jargon" doivent être exclus des analyses.

Deuxièmement, l'algorithme de classification doit être efficace et introduire le minimum de distorsions dans les données initiales. Nous travaillons actuellement sur quelques améliorations importantes de l'algorithme de Luong.

- Des méthodes du type "branch and bound" (Minoux 1989) peuvent accroître considérablement l'efficacité actuelle de l'algorithme de classification.

- La construction des arbres se fera en deux étapes. Dans un premier temps, tous les points et les nœuds sont placés et une longueur provisoire est affectée à chaque arête. Puis ces arêtes sont recalculées afin de répartir les ajustements nécessaires sur la totalité des chemins concernés et non plus seulement sur les parties centrales de l'arbre comme actuellement.

- A chaque étape de la classification, on s'assure que les solutions choisies sont effectivement les meilleures, c'est-à-dire celles qui aboutissent aux plus petites distances possibles au sein de chaque regroupement et qui maximisent les distances avec les textes maintenus à l'extérieur de ce regroupement. Ceci est réalisé grâce à l'analyse de la variance totale de la matrice des distances, analyse qui permet également le calcul d'un indice d'agrégation plus significatif que ceux présentés dans cette communication.

Remerciements :

Gérard Ledger et Tom Merriam ont organisé les expériences Oxquarry et nous ont aidé à rédiger le compte rendu ; X. Luong nous a introduit à la topologie et a réalisé nos premiers arbres ; M. Ruhlman a écrit avec nous le logiciel d'analyse arborée utilisé pour cette expérience ; E. Arnold, G. Bensimon, J.-G. Bergeron, M. Brugidou, P. Hubert, F. Lapierre, J. et N. Leselbaum, D. Monière, G. Paéquin, B. Peeters… ont participé aux premières expériences.

Références

Barthélémy J.-P. et Guénoche A. (1988). *Les arbres et les représentations de proximité*. Paris, Dunod.

Benzecri J.-P. (1980). *L'analyse des données. 1. La taxinomie*. Paris, Dunod.

Bergeron J.-G. & Labbé D. (2000). "L'évaluation de la négociation raisonnée par les acteurs : une analyse lexicométrique ". In Bernier C. & Al (éds). *Formation, relations professionnelles à l'heure de la société-monde*. Paris-Québec : Paris-Québec, L'Harmattan - Presses de l'Université Laval, p. 239-252.

Bona Dominique (1987). *Romain Gary*. Paris, Mercure de France.

Embleton S. (1986). *Statistics in Historical Linguistics*. Bochum, Brokmeyer.

Felsenstein J. (2004a). *Inferring Phylogenies*. Sunderland, Sinauer Ass.

Felsenstein J. (2004b). *Package of Programs for Inferring Phylogenies (PHYLIP).* Seattle, University of Washington.

Hockey S. & Martin J. (1988). *OCP Users' Manual*. Oxford, Oxford University Computing Service.

Holm H. J. (2007). "The New Arboretum of Indo-European "Trees". Can New Algorithms Reveal the Phylogeny and Even Prehistory of Indo-European ?". *Journal of Quantitative Linguistics*. 14-2, p. 167-214.

Labbé C. & Labbé D. (2001a). "Inter-Textual Distance and Authorship Attribution Corneille and Molière". *Journal of Quantitative Linguistics*. 8-3, p. 213-231.

Labbé C. & Labbé D. (2001b). *Discrimination et classement au sein d'un groupe d'entretiens. Le cas du confort électrique*. Grenoble, Journées d'études du CIDSP, 9 mars 2001.

Labbé C. & Labbé D. (2003). "La distance intertextuelle". *Corpus*. 2, p. 95-118.

Labbé C. & Labbé D. (2006). "A Tool for Literary Studies. Intertextual Distance and Tree Classification. *Literary and Linguistic Computing*. 21-3, p. 311-326.

Labbé D. (2007). "Experiments on Authorship Attribution by Intertextual Distance in English". *Journal of Quantitative Linguistics*, April 2007, 14-1. p. 33-80.

Labbé D. & Monière D. (2000). "La connexion intertextuelle. Application au discours gouvernemental québécois". In Rajman M. et Chappelier J.-C. (eds). *Actes des 5e journées internationales d'analyse des données textuelles*. Lausanne, Ecole polytechnique fédérale, vol. 1, p. 85-94.

Labbé D. & Monière D. (2003). *Le vocabulaire gouvernemental. Canada, Québec, France (1945-2000)*. Paris, Champion.

Lafon M. & Peeters B. (2006). *Nous est un autre*. Paris, Flammarion.

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

Ledger G. R. (1995). "An Exploration of Differences in the Pauline Epistles", *Literary and Linguistic Computing*. 10-2, p. 85-97.

Luong X. (1988). *Méthodes d'analyse arborée. Algorithmes, applications.* Thèse pour le doctorat ès sciences. Paris, Université de Paris V.

Luong X. (1994). "L'analyse arborée des données textuelles : mode d'emploi". *Travaux du cercle linguistique de Nice*. 16, p. 25-42.

Merriam T. (2002). "Intertextual Distances between Shakespeare Plays, with Special Reference to *Henry V* (verse)". *Journal of Quantitative Linguistics*. 9-3, p. 260-273.

Merriam T. (2003a). "An Application of Authorship Attribution by Intertextual Distance in English". *Corpus*. 2, 2003, p. 167-182.

Merriam T. (2003b). "Intertextual Distance, Three Authors". *Literary and Linguistic Computing*. 18-4, p. 379-388.

Minoux M. (1989). *Programmation mathématique : Théorie et Algorithmes*. Paris, Dunod.

Monière D. & Labbé D. (2006). "L'influence des plumes de l'ombre sur les discours des politiciens". In Condé C. & Viprey J.-M. *Actes des 8e Journées internationales d'Analyse des données textuelles*. Besançon, Vol. II, p. 687-696.

Sneath P. & Sokal R. (1973). *Numerical Taxonomy*. San Francisco, Freeman.

Ruhlman M. (2003). *Analyse arborée. Représentation par la méthode des groupements*. Grenoble, Polytech' – CERAT.

Annexe 1. Corpus Oxquarry1*

Set 1				Set 2			
Except	Author	Titles	Chap.	Except	Author	Titles	Chap.
A	Hardy	Jude	I	A	Butler	Erewhon revisit.	XIV
B	Butler	Erewhon revisit.	II	B	Morris	Dream of JB	
C	Morris	News	XIII	C	Tressel	Ragged TP	
D	Stevenson	Catrionae	V	D	Hardy	Jude	
E	Butler	Erewhon revisit.	XVIII	E	Stevenson	Ballantrae	IV
F	Stevenson	Ballantrae	II	F	Hardy	Wessex Tales	
G	Conrad	Lord Jim	XIV	G	Orczy	Elusive P	VII
H	Hardy	Madding	III	H	Conrad	Lord Jim	XXI
I	Orczy	Scarlet P	I	I	Morris	News	VIII
J	Morris	Dream of JB	VII	J	Hardy	Well beloved	I
K	Stevenson	Catrionae	X	K	Conrad	Almayer	VI
L	Hardy	Jude	VII	L	Hardy	Well beloved	XII
M	Orczy	Scarlet P	XIV	M	Morris	News	XIX
N	Stevenson	Ballantrae	V	N	Conrad	Almayer	XI
O	Conrad	Lord Jim	VII	O	Forster	Room with view	I
P	Chesterton	Man who was	I	P	Forster	Room with view	IV
Q	Butler	Erewhon revisit.	VII	Q	Conrad	Almayer	IX
R	Chesterton	Man who was	VII	R	Stevenson	Catrionae	XVI
S	Morris	News	I	S	Hardy	Madding	X
T	Conrad	Almayer	II	T	Hardy	Well beloved	2 VI
U	Orczy	Elusive P	I	U	Chesterton	Man who was	III
V	Conrad	Lord Jim	II	V	Forster	Room with view	VIII
W	Orczy	Elusive P	XIV	W	Stevenson	Catrionae	I
X	Hardy	Wessex Tales		X	Hardy	Well beloved	VIII
Y	Tressel	Ragged TP		Y	Orczy	Scarlet P	VII
Z	Tressel	Ragged TP		Z	Hardy	Madding	XVIII

* G. Ledger et T. Merriam nous ont fourni ce tableau à la fin de l'expérience.