MATHÉMATIQUES ET SCIENCES HUMAINES

J. P. BARTHELEMY
N. X. LUONG

Sur la topologie d'un arbre phylogénétique : aspects théoriques, algorithmes et applications à l'analyse de données textuelles

Mathématiques et sciences humaines, tome 100 (1987), p. 57-80
<http://www.numdam.org/item?id=MSH_1987__100__57_0>

Math. Sci. hum. (25^{e} année, n°100, *1987*, p.57-80)

SUR LA TOPOLOGIE D'UN ARBRE PHYLOGÉNÉTIQUE: ASPECTS THÉORIQUES, ALGORITHMES ET APPLICATIONS À L'ANALYSE DE DONNÉES TEXTUELLES.

J.P. BARTHELEMY *

N.X. LUONG**

INTRODUCTION.

Deux aspects essentiels de la discipline nommée usuellement "topologie" sont les notions de voisinage et d'équivalence de forme. Ce sont ces idées que nous voulons évoquer lorsque nous utilisons le mot "topologie" à propos d'arbres. Ce type d'étude a d'abord été développé dans le cadre de la biologie; c'est pourquoi nous mentionnons explicitement dans le titre de cet article les arbres phylogénétiques.

Ces arbres peuvent être obtenus de diverses manières (séquençage d'amino-acides, électrophorèse, hybridation d'ADN, etc.) et sont généralement examinés selon deux points de vue (Waterman et Smith, 1978):

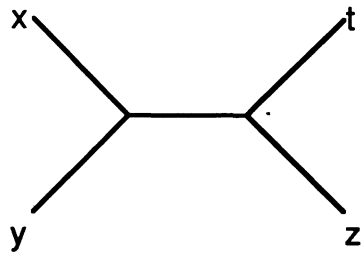
1°) Le point de vue métrique. L'ensemble X des unités d'évolution que l'on cherche à structurer est représenté aux sommets d'un arbre et on mesure la proximité entre deux éléments de X en calculant la longueur du chemin qui les relie (on obtient ainsi ce que Phipps, 1971, appelle la distance topologique et Farris, 1973, la distance cladistique tandis qu'une bonne partie du reste du monde parle de distance additive d'arbre). Cette voie a été fort explorée. Elle conduit, en particulier, à la caractérisation des distances additives d'arbre à l'aide de la fameuse condition des quatre points (Zaretskii, 1965; Buneman, 1971, 1974; Patrinos et Hakimi, 1972; Dobson, 1974) . Elle débouche sur de nombreux algorithmes d'approximation (ceux-ci sont décrits dans Barthélemy et Guénoche, 1987) .

2°) Le point de vue de la "topologie des bifurcations" (branching topology) liée à la structure de l'arbre. L'idée, ici, est que deux unités d'évolution sont "voisines" lorsqu'elles admettent un ancêtre commun proche de l'une et de l'autre. Cette notion est relative. Elle traduit

* E.N.S.Télécom. 46 rue Barrault. 75634 PARIS CEDEX

** U.R.L.9 - I.Na.L.F. et Département d'Informatique. Université de Nice. 98 bd Herriot BP 369 06007 NICE CEDEX

un dialogue entre "voisinage" et "opposition" et repose sur l'étude des configurations dans l'arbre, faisant intervenir les quadruplets d'éléments de X. Ainsi sur la figure ci-dessous x et y d'une part, z et t d'autre part sont voisins (les deux sommets non étiquetés jouent le rôle des ancêtres communs) et ces deux relations de voisinage sont complètement solidaires de l'opposition entre la paire {x,y} et la paire {z,t}. On notera que cette situation traduit le fait que x,y d'une part et z,t d'autre part peuvent être placés dans une même composante connexe, obtenue en gommant une arête de l'arbre.



Par ailleurs, à cause de la multiplicité des techniques permettant d'obtenir des arbres phylogénétiques, il est essentiel de savoir comparer diverses phylogénies construites sur l'ensemble X. Pour satisfaire à cette exigence, Estabrook et Meacham (1979) introduisent la notion d'*incompatibilité*. Il s'agit de l'impossibilité logique pour deux phylogénies de refléter une même configuration de bifurcations. Ici, la question est donc: deux arbres ont-ils la même forme? Là aussi les configurations sur quatre sommets vont jouer un rôle prépondérant. C'est à partir d'elles qu'Estabrook, McMorris et Meacham (1985) analysent cette question à l'aide de ces configurations et proposent quatre mesures de compatibilité. Cette étude a été complétée par Day (1985) qui propose deux autres indices et analyse tant d'un point de vue théorique que d'un point de vue empirique les fonctions de répartition de ces six indices.

En dehors du monde de la biologie, mentionnons que Dobson (1974) avait utilisé, surtout au niveau des notations, ces "configurations sur quatre sommets", lors de sa preuve de la condition des quatre points. Enfin ce type d'approche culmine avec le travail de Colonius et Schulze (1981) qui montrent que la considération "abstraite" de telle configurations permet de caractériser un arbre.

L'analyse des données a exporté les arbres hiérarchiques bien loin des terres de la classification des espèces animales et végétales. L'objet de ce travail est de montrer que l'on assiste au même phénomène pour l'approche topologique des arbres phylogénétiques.

Le premier paragraphe est consacré aux quelques éléments mathématiques nécessaires à la compréhension de la méthode, à la mise en place des algorithmes et à l'interprétation des résultats. Nous y étudions d'abord, sous le nom de *relation fondamentale*, les relations de voisinage succintement décrites ci-dessus. Puis nous définissons la notion de *groupement* qui

est en quelque sorte une "base de voisinage" pour une topologie des bifurcations. Nous introduisons ensuite la notion de *score* qui est une mesure du "niveau de voisinage" entre deux sommets de l'arbre . La connaissance des scores permet de déterminer la structure de base de l'arbre (ou, comme nous dirons, sa "forme"), c'est à dire un arbre libre qui lui est *équivalent.* Une généralisation de la notion de score permet d'établir une profonde relation entre les deux points de vue que nous avons mentionnés (le point de vue métrique et le point de vue de la topologie de bifurcation).

Le second paragraphe est consacré à deux algorithmes. Le premier porte sur la reconstitution d'un arbre phylogénétique à partir des scores et des groupements. Le second est une heuristique d'approximation d'une dissimilarité par une distance additive d'arbre . Le principe est de construire un arbre phylogénétique dont la "topologie" rende compte "au mieux" de la dissimilarité de départ. Chercher une heuristique, plutôt qu'un algorithme exact relativement à un critère d'approximation, est une option raisonnable, puisqu'on sait depuis Day (1986) que ce problème d'approximation , pour le critère des moindres carrés ou des moindres écarts est NP-difficile.

Ces algorithmes sont utilisés au troisième paragraphe, sur des données textuelles recueillies par Brunet (1987), à partir de mesures faisant intervenir la notion de connexion lexicale. Sont ainsi "arborisés" vingt textes complets de Victor Hugo et seize textes de Giraudoux.

Cette livraison de *Mathématiques et Sciences Humaines* est consacrée aux méthodes combinatoires en analyse des données. C'est dans cet esprit que nous avons considéré le premier paragraphe comme une préparation aux deux suivants. Son contenu rencontre d'ailleurs largement les contenus des travaux de Bandelt et Dress (1986), Barthélemy et Guénoche (1987) et Luong (1988) . Afin d'alléger cette partie "mathématique", nous avons systématiquement omis les démonstrations que l'on peut trouver dans cette dernière référence.

1 X-ARBRES: DÉFINITIONS ET PROPRIÉTÉS TOPOLOGIQUES.

1.1 X-arbres: définitions et notations.

On note par $A=(S,U)$ un arbre d'ensemble de sommets S et d'ensemble d'arêtes U . On désigne par (ab) le chemin de A entre les sommets a et b . L'arbre A est dit *valué* lorsqu'on considère une fonction π définie sur l'ensemble U , à valeurs réelles strictement positives. On le note par (A,π) . Lorsque toutes les arêtes sont de longueur 1 , on dit que π est la *valuation unitaire.*

Pour un arbre valué (A,π), on note par $\partial_\pi(a,b)$ la longueur, pondérée par π , du chemin qui joint les sommets a et b. On obtient ainsi l'espace métrique (S,∂_π). On dit que ∂_π est la

distance additive de A muni de π et que (S,∂_π) est un *arbre additif*. Si π est la valuation unitaire, on dit que ∂_π est la *distance unitaire* de A et on la note simplement par ∂.

Un X-*arbre* est un couple (A,f), où f est une application de X dans S telle que tout sommet dans S - f(X) est de degré au moins 3 .

Les sommets de f(X) sont appelés *sommets réels* et ceux de S-f(X) sont appelés *sommets latents* .

Du point de vue de la représentation de l'ensemble X "des objets" sur une structure d'arbre , toutes les feuilles d'un X-arbres sont image d'au moins d'un élément de X et il n'existe pas de sommet "isolé sur un chemin" (i.e. de degré 2) qui ne provienne pas de X . Lorsqu'aucune confusion n'est à craindre, le X-arbre (A,f) est simplement noté A .

On dit que (A,f) est un X-*arbre séparé* lorsque f est injective, que (A,f) est un X–*arbre libre* lorsque f(X) est l'ensemble des feuilles de A , que (A,f) est un X–*arbre valué* si A est valué.

Sur un X-arbre valué on pose pour x,y dans X : $d_\pi(x,y) = \partial_\pi(f(x),f(y))$;
d_π est appelé *écart additif* d'arbre induit par (A,π) sur X ; on le note par d lorsque π est la valuation unitaire . Si l'arbre est séparé, l'écart d_π est une *distance additive* ; on identifie alors X et f (X) et X devient un sous-ensemble de S .

On dit que les X-arbres (A,f) et (A',f') sont *égaux* lorsqu'il existe une bijection g de l'ensemble S des sommets de A dans l'ensemble S' des sommets de A' telle que:
(i) g(a)g(b) est une arête de A' si et seulement si ab est une arête de A et
(ii) gf = f'.

1.2 La relation fondamentale d'un X-arbre .

Soit $X^{(2)}$, l'ensemble des paires d'éléments de X . Une relation binaire R sur l'ensemble $X^{(2)}$ est un sous-ensemble de $X^{(2)}xX^{(2)}$. Par convention, la paire {x,y} de $X^{(2)}$ sera simplement notée xy . Lorsque $(xy,zt) \in R$ on écrit xy R zt et on dit, dans ce cas, que xy et zt sont *opposées* . On dit aussi qu'un sous–ensemble de X à quatre éléments, Y={x,y,z,t} est R-*dégénéré* si et seulement si aucune des expressions induites par R sur Y n'est vérifiée.

Considérons un X-arbre (A,f) .

DÉFINITION 1 . La *relation fondamentale* du X-arbre A est la relation binaire R_A sur $X^{(2)}$ définie par:

xy R_A zt si et seulement si

$d(x,y)+d(z,t) < \text{Min}\{d(x,z)+d(y,t) , d(x,t)+d(y,z)\}$.

Lorsque xy R_A zt , on dit que la paire zt est *opposée*, dans A , à la paire xy . A une permutation près entre x et y d'une part et entre z et t d'autre part, on a xy R_A zt si et seulement si x, y, z, t sont dans l'une des sept configurations représentées par la figure 1 .

Notons aussi que la relation R_A resterait la même si munissant A d'une valuation π quelconque, on remplaçait dans la définition 1 d par d_π.

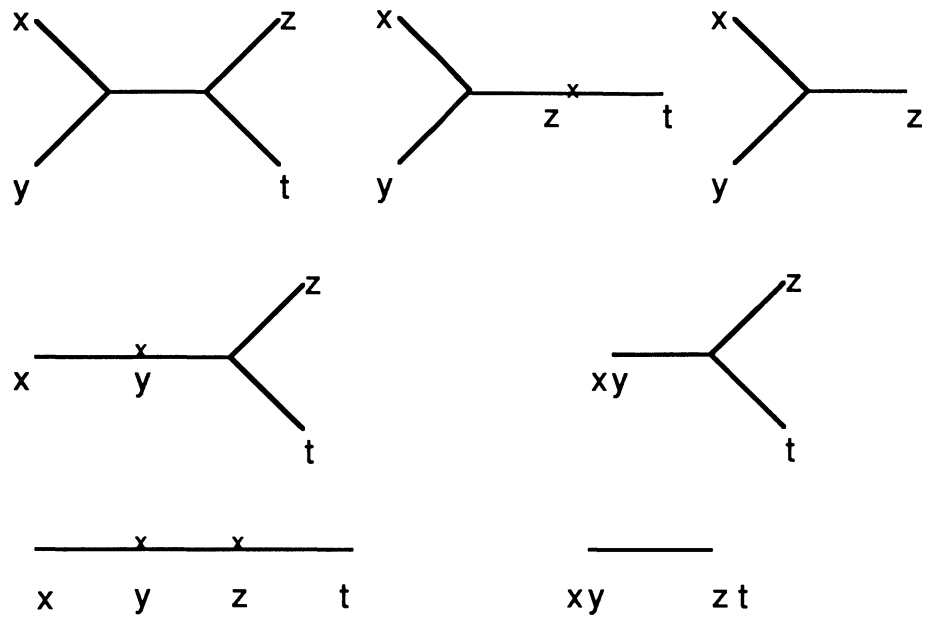


Figure 1

Soit α une arête du X-arbre A. En ôtant α de A, on obtient deux composantes connexes qui induisent une bipartition {B,B'} de X, on dit que {B,B'} est une *scission* de A et on la note $\sigma(\alpha)$. En examinant la figure 1 on se convainc aisément que:

LEMME 1. Pour un un X-arbre (A,f) on a l'équivalence suivante:
$xy\ R_A\ zt \Leftrightarrow$ il existe une scission $\sigma(\alpha) = \{B,B'\}$ de X, avec $f(x),f(y) \in B$ et $f(z),f(t) \in B'$.

Par ailleurs, on peut montrer que:

PROPOSITION 1 (Colonius et Schulze, 1981). La relation fondamentale R_A d'un X-arbre A possède les propriétés suivantes:

(i) pour tout $x,y,z,t \in X$, $xy\ R_A\ zt$ entraîne $zt\ R_A\ xy$. [*symétrie*]

(ii) pour tout $x,y,z,t \in X$, au plus une des trois expressions $xy\ R_A\ zt$, $xt\ R_A\ yz$, $xz\ R_A\ yt$ est vérifiée. [2-*antisymétrie*]

(iii) pour tout $x,y,z,t,u \in X$, $xy\ R_A\ zt$ entraîne soit $xy\ R_A\ zu$, soit $xu\ R_A\ zt$ (soit les deux). [*subtituabilité*]

(iv) pour tout $x,y,z,t,u \in X$, $xy\ R_A\ zt$ et $xy\ R_A\ tu$ entraînent $xy\ R_A\ zu$. [2-*transitivité*]

Les quatre conditions de la proposition 1 présentent quelque redondance puisque:

LEMME 2 (Bandelt et Dress, 1987). Toute relation binaire sur $X^{(2)}$ symétrique, 2–antisymétrique et substituable est aussi 2-transitive.

REMARQUES. Nous venons d'examiner, sur les X-arbres, une notion de voisinage. Elle est "relative": les sommets x et y sont "voisins" relativement aux sommets u et v si et seulement si ces deux paires sont opposées. L'introduction des groupements, au paragraphe 4 , permettra d'affiner cette notion.

Notons que cette relation fondamentale sur un arbre a été introduite par Dobson (1974) pour la démonstration de la condition des quatre points. L'algorithme de Sattah & Tversky (1977) a pour base cette relation. Colonius & Schulze (1981) en donnent une caractérisation . Estabrook, McMorris & Meacham (1985) , Day (1985) l'utilisent pour définir des indices de compatibilité entre les paires d'arbres dans le problème du consensus. McMorris (1985) montre les difficultés logiques rencontrées lorsque l'on base le problème du consensus sur cette relation.

1.3 Equivalence de X-arbres .

DÉFINITION 2. On dit que les X-arbres A et A' sont dits équivalents et on note $A \equiv A'$ lorsque $R_A = R_{A'}$.

En effectuant une récurrence sur le nombre n des éléments de X, on montre que:

PROPOSITION 2 . Si deux X-arbres libres et séparés A et A' sont équivalents alors ils sont égaux en tant que X-arbres.

Ainsi chaque classe d'équivalence, modulo $\equiv$, contient au plus un X-arbre libre et séparé. Les considérations qui suivent vont préciser l'allure de ces classes d'équivalence:
Chacune d'entre elles contient un X-arbre libre et séparé et un seul, et un procédé de construction permet d'engendrer tous les X-arbres d'une classe.

L'équivalence $\equiv$ correspond à la notion de compatibilité évoquée dans l'introduction. Une classe, modulo $\equiv$, représente , en quelque sorte, une "forme" d'arbre et les X–arbres libres et séparés deviennent des prototypes de formes.

Soit (A,f) un X-arbre d'ensemble de sommets S et d'ensemble d'arêtes U. Considérons un sommet s appartenant à f(X) et posons: $f^{-1}(s) = \{x_1,x_2,...,x_p\}$ avec $p \geq 1$. Dans ce cas on dit que s est *étiqueté* par $\{x_i \mid i=1,2,...,p\}$. Pour p>1, l'*expansion élémentaire* du X–arbre A par x_i est le X-arbre (A_{x_i},f_{x_i}) obtenu en ajoutant à A une feuille notée également x_i , en reliant x_i à s par une arête et en posant pour $y \in X$, $y \neq x_i$, $f_{x_i}(y) = f(y)$ et $f_{x_i}(x_i) = x_i$. Il est clair que $(A_{x_i},f_{x_i}) \equiv (A,f)$.

DÉFINITION 3 . Un X-arbre B est une *expansion* d'un X-arbre A lorsqu'on obtient B , à partir de A, par une suite d'expansions élémentaires.

En effectuant, à partir de A, une suite maximale d'expansions, on obtient un X-arbre libre et séparé; donc:

LEMME 3. Tout X-arbre est équivalent à un X-arbre libre et séparé.

Ainsi l'ensemble des classes d'équivalence de X–arbres, modulo $\equiv$, peut être assimilé à l'ensemble des X–arbres libres et séparés.
Du lemme , on déduit:

PROPOSITION 3. Deux X-arbres sont équivalents si et seulement si ils admettent une expansion commune.

1.4 Groupements et scores.

La notion de groupement dans un X-arbre A (Luong, 1983; Barthélemy et Luong, 1986) permet de considérer des amas d'objets voisins. Elle ne dépend que de la relation fondamentale et est caractéristique d'une "forme" d'arbre. Pour bien mettre ce fait en lumière, nous l'envisageons de manière abstraite en considérant une relation binaire (presque) quelconque sur X(2).

DÉFINITION 4. Soit R une relation binaire symétrique sur $X^{(2)}$ et soit G un sous–ensemble de X, $|G|\geq 2$.
- On dit que G est un *prégroupement* si et seulement si, pour tout $u,v\in G$ et pour tout $x,y,z\in X$:
(i) ux R yz entraîne vx R yz , et si {u,x,y,z} est R-dégénéré alors {v,x,y,z} est R–dégénéré.
- On dit qu'un préqroupement G est un *groupement* si et seulement si:
(ii) $|G| = n$ et la relation R est vide ou
$|G|<n$, pour tout $u,v\in G$ et pour tout tout $x,y\in X–G$, uv R xy .

REMARQUES. La condition (i) peut s'interpréter en disant que u et v ont un même comportement pour la relation R . Par ailleurs on vérifie que si R est en plus 2-antisymétrique la condition (i) est équivalente à
(iv) pour tout $u,v\in G$ et pour tout $x,y\in X$ on n'a ni xu R yv ni xv R yu .

EXEMPLE. La figure 2 représente un X-arbre et R est sa relation fondamentale. On y remarque les groupements {x,y,z} et {u,v,w} ; tandis que {t,h} est un prégroupement.

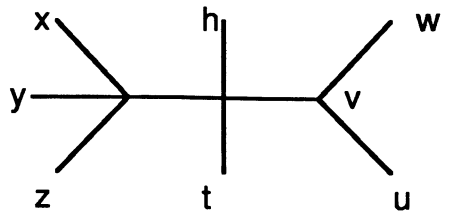


Figure 2

Les deux propriétés ci-dessous précisent la typologie des groupements de la relation R:
1°) Pour une relation binaire R sur $X^{(2)}$ symétrique, un groupement n'est jamais strictement inclus dans un prégroupement.
2°) Pour une relation binaire R sur $X^{(2)}$ symétrique, 2-antisymétrique et substituable, deux groupements sont toujours disjoints.

DÉFINITION 5. Soit R une relation binaire sur $X^{(2)}$, on appelle R-*score* de {x,y} et on note s*(x,y) , le nombre de paires de XxX qui sont opposées à {x,y} . La fonction s*, définie sur XxX, est appelé la *fonction score* de R.

On appelle R-*score large* de {x,y} et on note s(x,y) le nombre de paires de XxX opposées à {x,y} ou telles que {x,y,z,t} soit R-dégénéré.

EXEMPLE. Le tableau 1 donne les scores de la relation fondamentale de l'arbre de la figure 2.

	y	z	t	u	v	w	h
x	10	10	3	0	0	0	3
y		10	3	0	0	0	3
z			3	0	0	0	3
t				3	3	3	6
u					10	10	3
v						10	3
w							3

scores

	y	z	t	u	v	w	h
x	15	15	7	3	3	3	7
y		15	7	3	3	3	7
z			7	3	3	3	7
t				7	7	7	15
u					15	15	7
v						15	7
w							7

scores larges

Tableau 1

Scores et groupements sont liés par le résultat ci-dessous:

PROPOSITION 4. Soit R une relation binaire sur $X^{(2)}$, symétrique et 2–antisymétrique et soit G un sous-ensemble à k éléments de X , avec $k \geq 2$. G est un groupement si et seulement si les conditions (i) et (ii) ci-dessous sont vérifiées pour tout $x,y \in G$:

(i) $s^*(x,y)=\frac{1}{2}(n-k)(n-k-1)$.

(ii) $s(x,y)=\frac{1}{2}(n-2)(n-3)$.

1.5 Le théorème de Colonius et Schulze.

Considérons maintenant un X-arbre A. Et convenons de dire score (resp. groupement) dans A au lieu de score (resp. groupement) relatif à la relation fondamentale de A. De même, selon une convention déjà introduite, on dit que les paires xy et zt sont opposées dans A si elles le sont au sens de R_A. Ici les scores et les groupements sont donc liés non pas à un X–arbre particulier mais à une classe, modulo $\equiv$, de X-arbres (c'est à dire, selon notre interprétation, à une forme de X-arbres). Par ailleurs, en vertu de la proposition 1, toutes les propriétés que nous avons indiquées pour les scores et les groupements sont vraies pour les scores et les groupements dans un X-arbre. Le résultat ci-dessous montre qu'une relation sur $X^{(2)}$, symétrique, 2–anti-symétrique et substituable est exactement la même chose qu'une forme de X-arbre. Il achève de légitimer le point de vue relationnel que nous avons adopté.

THÉOREME 1 (Colonius et Schulze, 1981). Soit R une relation binaire sur $X^{(2)}$, symétrique, 2-antisymétrique et substituable. Il existe alors un X-arbre libre et séparé A tel que $R=R_A$.

1.6 Généralisation de la notion de score.

Nous avons remarqué, à propos de la définition 1 (1.1) que la relation fondamentale d'un X-arbre A pouvait être induite par n'importe quelle distance additive sur A. Cela constitue une passerelle qui mène du point de vue métrique, sur les X-arbres, au point de vue de la topologie des bifurcations. Dès lors se pose la question de l'existence d'une autre passerelle allant en sens inverse. Une réponse est fournie par le résultat ci-dessous, que nous énonçons en utilisant les notations introduites en 1.1. Ce résultat nous indique qu'à partir de la fonction score d'un X-arbre A (donc à partir de la relation fondamentale de A), on peut construire une distance additive d'arbres, dont A est le support.

PROPOSITION 5 (Colonius et Schulze, 1981; Furnas, 1985; Bandelt et Dress, 1987): Soit A un X-arbre libre et séparé et soit s^* la fonction score de A. Soit

$M>\mathrm{Max}\{s^*(x,y)\ ;\ x,y\in X\}$.

Posons $d(x,y)=M-s^*(x,y)$ pour $x\neq y$ et $d(x,x)=0$. Il existe alors une valuation π des arêtes de A telle que $d=d_\pi$.

Ce résultat a été mentionné, de manière assez peu explicite, par Colonius et Schulze (1981); retrouvé par Furnas (1985) , sa démonstration a été enfin publiée par Bandelt et Dress (1987), ces derniers auteurs font d'ailleurs explicitement référence à Colonius et Schulze. Nous allons

montrer ci-dessous comment il peut être réinterprété et généralisé. Pour ce faire nous allons introduire la notion de *décomposition forestière.*

Soit (A,f) un X-arbre, d'ensemble de sommets S et d'ensemble d'arêtes U. Considèrons un chemin $(ab) = s_1 s_2 ... s_k$ ($a = s_1$, $b = s_k$ et $s_i s_{i+1} \in U$ pour $1 \leq i < k$) de A. On note par d_i le degré du sommet s_i.

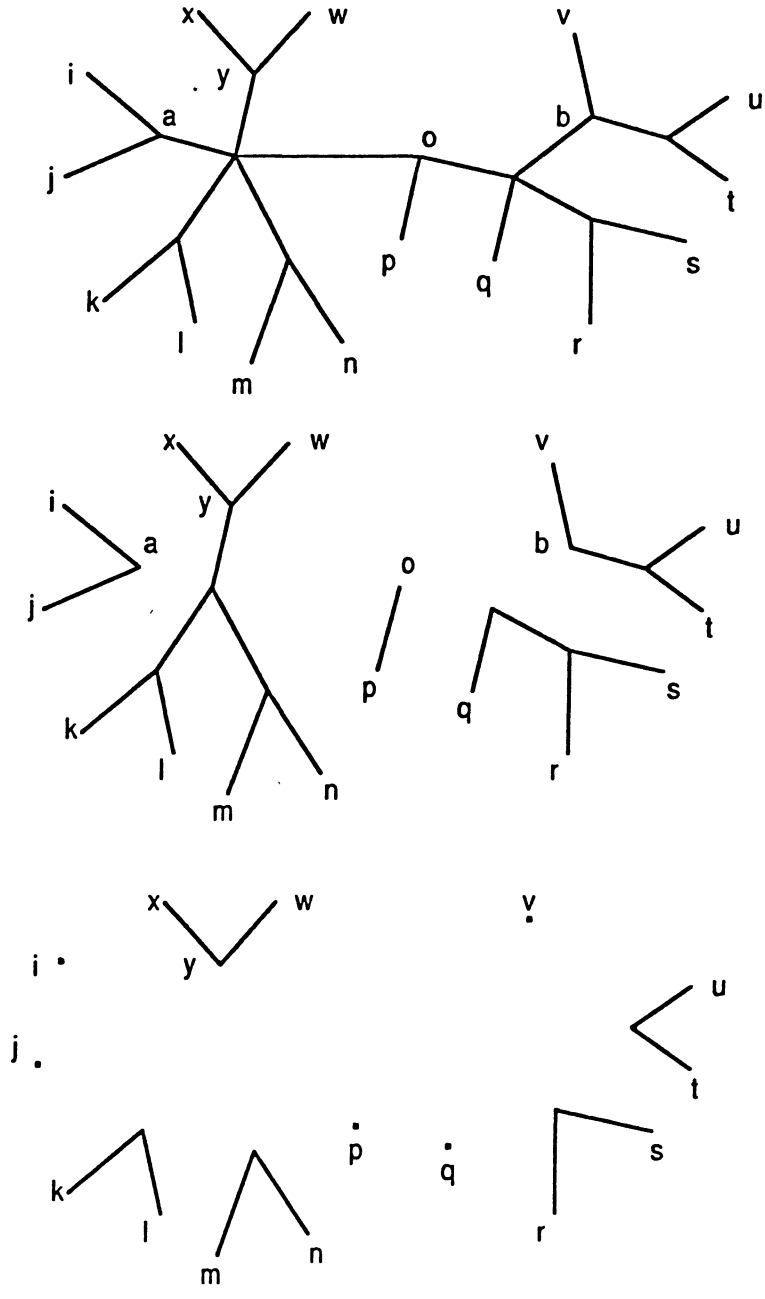


Figure 3 : *un X-arbre, les composantes connexes et la décomposition forestière pour un chemin (ab).*

En ôtant de A les (k-1) arêtes constituant (ab) , on obtient k composantes connexes $A_1, A_2, ..., A_k$, A_i étant la composante connexe qui contient le sommet s_i (cf figure 3). Lorsque a = b , en ôtant de A les arêtes incidentes en a, on obtient d_1 composantes connexes.

Soit s_i un sommet de (ab) . Pour $s_i \neq a$ et $s_i \neq b$, on définit les $d_i - 2$ *branches* en s_i, relativement à (ab) , comme les composantes connexes A_i^j de $A_i - \{s_i\}$ obtenues en ôtant les $d_i - 2$ arêtes incidentes à s_i .

Les $d_1 - 1$ ($d_k - 1$) branches en a (en b) sont les composantes connexes de A_1 (A_k) obtenues en ôtant les $d_1 - 1$ ($d_k - 1$) arêtes incidentes en a (en b) . Ces composantes connexes sont notées

A_1^j , $1 \leq j \leq d_1 - 2$ (A_k^j , $1 \leq j \leq d_k - 2$).

L'ensemble de tous les A_i^j forment une partition de $S - \{s_1, s_2, ..., s_k\}$. On pose: $X_i^j = A_i^j \cap X$. Les X_i^j forment une partition de l'ensemble X diminué des étiquettes de sommets réels situés sur le chemin (ab). Cette partition est appelée la *décomposition forestière* de X le long du chemin (ab). Pour $x,y \in X$, on note par $\boldsymbol{F}(x,y)$ la décomposition forestière de X le long de (f(x),f(y)). Les X_i^j sont appelés les *buissons* de $\boldsymbol{F}(x,y)$.

PROPOSITION 6 . Soient $x,y,z,t \in X$, la paire xy est opposée à la paire zt si et seulement si x et y appartiennent tous les deux à un même buisson de $\boldsymbol{F}(z,t)$. Si ϕ est la fonction de dénombrement des paires d'un ensemble: $\phi(Y) = \frac{1}{2}|Y|(|Y|-1)$,

alors $s^*(x,y) = \sum_{X_i^j \in \boldsymbol{F}(x,y)} \phi(X_i^j)$.

Ce résultat permet de réinterpréter la proposition 5. Il conduit aussi à sa généralisation.

DÉFINITION 6. Une fonction p définie sur l'ensemble 2^X des parties de X et à valeurs réelles est dite *strictement surmodulaire* si et seulement si :

$p(\emptyset) = 0$ et $B \cap C = \emptyset$ entraîne $p(B \cup C) > p(B) + p(C)$ pour tout B, C non vides de 2^X .

Soit (A,f) un X-arbre et soient $x,y \in X$. Soit p une fonction quelconque de 2^X à valeurs réelles.

Posons $p^*(x,y) = \sum_{X_i^j \in \boldsymbol{F}(xy)} p(X_i^j)$.

p^* généralise, à une fonction p quelconque, la fonction score d'un X-arbre définie à l'aide de la fonction de dénombrement des paires. Soit λ un nombre réel . On pose

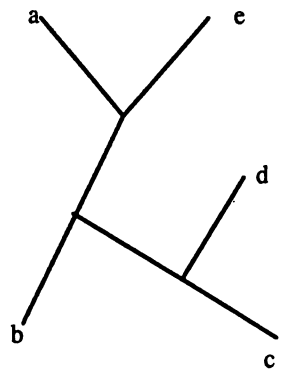
$d(p, \lambda)(x,y) = \lambda - p^*(x,y)$.

THÉOREME 2. Soit (A,f) un X-arbre et soit p une fonction strictement surmodulaire sur 2^X. Il existe un nombre réel $\lambda_0(p)$ tel que, pour tout $\lambda \geq \lambda_0(p)$, il existe un X-arbre $A(\lambda)$,

équivalent à A et une valuation π des arêtes de $A(\lambda)$ vérifiant: $d_\pi = d(p,\lambda)$. De plus le X-arbre $A(\lambda)$ est libre et séparé si et seulement si $\lambda > \lambda_0(p)$.

Ainsi ce résultat permet d'interpréter $d(p,\lambda)$ comme une distance additive relative, selon les valeurs de λ, à *deux* X-arbres équivalents à A. L'un est toujours libre et séparé, il correspond à $\lambda > \lambda_0(p)$. L'autre ne l'est jamais, il correspond à $\lambda = \lambda_0(p)$. Sa démonstration est assez technique, on peut la trouver, avec la détermination de $\lambda_0(p)$, dans Barthélemy et Guénoche (1987), dans le cas particulier où A est libre et séparé et, dans le cas général, dans Luong (1988). Pour faciliter sa compréhension, illustrons le par deux exemples, où l'on a choisi les scores généralisés définis par la fonction $p(E)=|E|^2$.

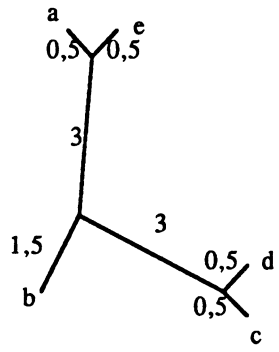
Dans la figure 4-(i) on a un X-arbre A et le tableau des scores généralisés correspondant.



	a	b	c	d	e
a		5	3	3	9
b			5	5	5
c				9	3
d					3

Scores généralisés

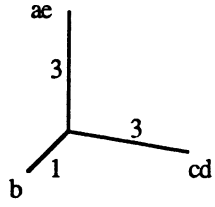
(i)



	a	b	c	d	e
a		5	7	7	1
b			5	5	5
c				1	7
d					7

Tableau des distances pour $\lambda = 10$

(ii)



	a	b	c	d	e
a		4	6	6	0
b			4	4	4
c				0	6
d					6

Tableau des distances pour $\lambda = 9$

(iii)

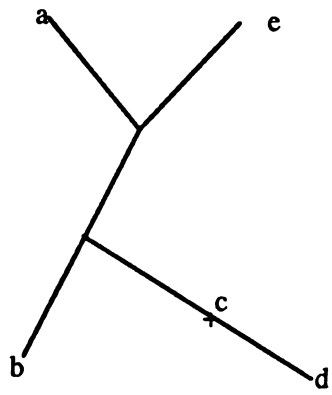
Figure 4

Ainsi pour toute paire (x,y) de X

$p^*(x,y) = \sum_{X_i^j \in F(xy)} |X_i^j|^{\lambda}$, les buissons X_i^j parcourant la décomposition forestière $F(xy)$.

Il vient ici $\lambda_0(p) = 9$ (nous demandons au lecteur de nous croire sur parole!). En prenant $\lambda = 10$ nous obtenons le X-arbre libre et séparé (ii). En prenant $\lambda=9$, on trouve le X-arbre non séparé (iii).

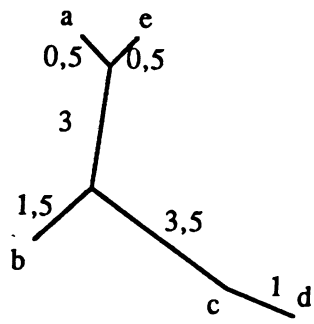
L'exemple suivant montre que les valeurs de la distance $d(p,\lambda)$ ne dépendent pas uniquement de la "forme" d'un X-arbre. En effet si l'on passe d'un X-arbre à un X-arbre équivalent, les valeurs de p restent les mêmes, mais la valeur de $\lambda_0(p)$ peut changer. La figure 5 représente un arbre (iv) équivalent à (i). Mais ici, on trouve $\lambda_0(p) = 10$ (nous supplions à nouveau le lecteur de nous croire!). Pour cette valeur nous obtenons l'arbre (v) qui n'est égal ni à l'arbre (ii), ni à l'arbre (iv).



	a	b	c	d	e
a		5	3	2	9
b			5	4	5
c				9	3
d					2

Scores généralisés

(iv)



	a	b	c	d	e
a	5	7	8	1	
b		5	6	5	
c			1	7	
d				8	

Tableau des distances pour $\lambda = 10$

(v)

Figure 5

2. ALGORITHMES .

Nous allons maintenant utiliser les notions de groupement et de score pour:
1°) Reconstituer un X-arbre à partir de sa distance additive.
2°) Approcher une dissimilarité quelconque par la distance additive d'un X-arbre.
Les considérations du paragraphe 1 ne permettent, en principe, que de reconstituer un X–arbre équivalent à A. L'utilisation de la notion de groupement pointé permettra de retomber exactement sur A. Le second algorithme (exposé succintement dans Barthélemy et Luong, 1986) se présentera comme une adaptation du premier dans le cas où la donnée de départ n'est plus une distance additive d'arbre.

2.1 Reconstruction d'un X-arbre.

La donnée d'une distance additive d sur un X-arbre A est suffisante pour déterminer la relation R_A et les scores dans A. Avec les scores, on peut déterminer tous les groupements de A .

On sait que les éléments d'un groupement ont un même comportement pour cette relation, il est alors possible de construire A de manière itérative. On calcule les scores pour en dégager les groupements. On prendra comme *représentant* de chaque groupement l'un de ses éléments. Soit Y l'ensemble formé de ces représentants et les éléments de X n'entrant dans aucun groupement. La connaissance des distances entre les éléments de Y ramène la construction de l'arbre A à celui d'un arbre B, B étant un arbre qui peut être construit à partir de A en effaçant un certain nombre de feuilles et d'arcs adjacents à ces feuilles.

Techniquement, on peut aussi prendre comme représentant d'un groupement G, $|G| = k$, avec $G=\{x,y,z,...\}$, un élément "moyen" g obtenu de la manière suivante:

$$\partial(g,t) = \frac{1}{k} \sum_{x \in G} d(x,t) \quad \text{, pour tout } t \in X .$$

Il est clair que g a le même comportement pour la relation fondamentale de A que n'importe quel élément x de G.

A l'itération m , supposons que l'arbre ait p groupements, on les notera alors par $G^{(m)}_i$, $i \in I$, $I = \{1,2,...,p\}$.

S'il existe dans un groupement G un élément qui est un sommet intérieur de l'arbre on dit que G est un groupement *pointé* . Cet élément est appelé *sommet* du groupement G.

Soit Δ une matrice d'une distance additive sur l'ensemble X , à n éléments. On utilisera les procédures suivantes:
- Procédure Calcul-Scores(Δ ; σ^*, σ) . On calcule les scores σ^* et et les scores larges σ. Notons que le score large de la paire (x,y) est le nombre de fois où elle vérifie

$$d(x,y)+d(u,v) \leq \text{Min}\{ d(x,u)+d(y,v) , d(x,v)+d(y,v)\} .$$

-Procédure Dégager-Groupement (σ^*, σ ; $G^{(m)}_i : i \in I$). Utiliser les scores larges σ pour dégager le prégroupements et les scores σ^* pour en sélectionner les groupements.

-Procédure Calcul-Long-Group ($G^{(m)}_i : i \in I$) Calculer la longueur des arcs dans les groupements $G^{(m)}_i$. Soient u et v deux feuilles liées directement à un sommet intérieur p de l'arbre et soit un sommet h quelconque. On a

$$d(u,p) = \frac{1}{2}(d(u,v)+d(u,h)-d(v,h)) .$$

-Procédure Repré-Group ($G^{(m)}_i : i \in I$). Pour chaque groupement $G^{(m)}_i$, déterminer un représentant g_i . Si $G^{(m)}_i$ est pointé, on choisit pour g_i son sommet, sinon on choisit pour g_i la médiane des éléments de $G^{(m)}_i$.

-Procédure Réduction-Mat (Δ; $G^{(m)}_i : i \in I$). Réduction de la matrice Δ . Pour chaque groupement $G^{(m)}_i$ on efface dans Δ les lignes (et les colonnes) correspondantes et on ajoute à Δ une ligne (et une colonne) composée des distances $d(g_i,t)$, pour tout t appartenant à $X-G^{(m)}_i$.

- Procédure Etoile(Y, T_m) : on construit une étoile (éventuellement composée de deux branches) reliant les éléments de Y , sous-ensemble de X . On a ainsi un arbre T_m .
- Procédure Développer(T_m): On examine les feuilles $a^{(m)}_i$ de l'arbre T_m .

$a^{(m)}_i$ étant un représentant d'un groupement $G^{(m-1)}_j = \{f_1,f_2,...,f_k\}$:

- à $a^{(m)}_i$ correspond un sommet virtuel de l'arbre: on le renomme et on trace à partir de ce sommet des arcs reliants à $f_1,f_2,...,f_k$ (considérés commes des feuilles de l'arbre T_{m-1}).
- à $a^{(m)}_i$ correspond à un sommet réel de l'X-arbre: on trace à partir de $a^{(m)}_i$ des arcs le reliants aux autres éléments du groupement (considérés commes des feuilles de l'arbre T_{m-1}).

ALGORITHME 1.

```
m := 1 ;
TANT QUE |X| ≥4 FAIRE
  DEBUT Calcul-Scores( Δ ; σ*, σ);
        SI σ* = 0 ALORS ALLER À arbre ;
        Dégager-Groupement (σ*, σ ; G(m)i : i ∈ I) ;
        Calcul-Long-Group(G(m)i : i ∈ I) ;
        Repré-Group(G(m)i : i ∈ I) ;
        Réduction-Mat(Δ; G(m)i : i ∈ I) ;
        m := m+1 ;
  FIN ;
arbre :   Etoile(X, Tm) ;
  TANT QUE m < n FAIRE
    DEBUT
        Développer( Tm) ;
        m:=m+1 ;
    FIN .  (* fin de l'algorithme *)
```

2.2 Représentation d'une dissimilarité par un X-arbre.

Les données constituent maintenant la matrice de dissimilarités $\Delta = (d_{ij})$. On peut s'inspirer de l'algorithme 1 pour avoir un algorithme de représentation sur un arbre. On construit des "groupements" (ce mot est ici utilisé par abus de langage) en s'inspirant du cas "exact" et en l'enrichissant des nuances suivantes:

- On montre aisément que si a et b sont dans un groupement G , |G|=k, et si x dans X–G alors $s^*(a,b) - s^*(a,x) \geq n-k-1$. Ainsi à chaque étape de l'algorithme, on peut utiliser la majoration précédente pour avoir une certaine tolérance dans la formation des groupements. Dans la pratique nous prenons souvent pour ce seuil $\frac{1}{2}(n-k-1)$.

- Lorsqu'aucun sous-ensemble de X n'apparaît comme groupement, on considérera comme groupement tout sous-ensemble dont les paires ont même score et ce score étant maximum.

L'algorithme va utiliser les mêmes procédures que pour le cas exact, avec les modifications suivantes:

-Procédure Dégager-Groupement 2(σ^*,σ ; $G^{(m)}_i$: $i \in I$) A partir des matrices σ^* et σ des scores on construit, à l'étape m, les groupements $G^{(m)}_i$. Chaque groupement G , |G|=k , est formé avec une certaine tolérance de seuil $\lambda_x(n-k-1)$. On choisira λ entre 0 et 1 . Si aucun groupement en ce sens n'apparait, on considérera les paires dont les scores sont deux à deux maximum. Compte tenu des propriètés des groupements établies en 1.4, il semble raisonnable d'appliquer le principe suivant : on remplace par $G \cup G'$ deux groupements G et G' tels que: $G \cap G' \neq \varnothing$.

- Procédure Calcul-Long-Group2($G^{(m)}_i$: $i \in I$) : si l'on a regroupé deux éléments u et v , la longueur de l'arc up , p étant le sommet intérieur lié directement à u,v est estimé par

$$d(u,p) = \frac{1}{2}\left(d_{uv} + \frac{1}{n-2} \sum_{h \neq u,v} d_{uh} - \partial_{vh}\right)$$

- Procédure Indice-d'agrégation($G^{(m)}_i$: $i \in I$) : Soit $s^{*'}$ le score calculé d'une paire (x,y) d'un sous-ensemble G de X , G est soit un groupement (formé avec un certain seuil de tolérance), soit une paire de score maximal. Soit s^* le score théorique, c'est à dire le score que devrait avoir (x,y) s'ils appartenaient tous deux à un groupement. On considère comme indice de qualité arborée, le rapport $\frac{s^{*'}}{s^*}$, appelé *indice d'agrégation* . Cet indice permet de mesurer la manière dont les objets vont s'agréger et d'apprécier la qualité de leur représentation sur l'arbre. Si l'indice est égal à 1 alors x et y respectent parfaitement la relation fondamentale dans tout quadruplet d'éléments de X contenant {x,y} .

ALGORITHME 2.

m := 1 ;
TANT QUE |X| ≥4 FAIRE
DEBUT Calcul-Scores(Δ ; σ*, σ);
SI S* = 0 ALORS *arbre* ;
Dégager-Groupement2 (σ*,σ ; $G^{(m)}_i$: i ∈ I) ;
Calcul-Long-Group2($G^{(m)}_i$: i ∈ I) ;
Indice-d'agrégation($G^{(m)}_i$: i ∈ I)
Repré-Group($G^{(m)}_i$: i ∈ I) ;
Réduction-Mat(Δ; $G^{(m)}_i$: i ∈ I) ;
m := m+1 ;
FIN ;
arbre : Etoile(X, T_m) ;
TANT QUE m < n FAIRE
DEBUT Développer(T_m) ;
m:=m+1 ;
FIN. (* fin de l'algorithme *)

2.3 Quelques remarques.

L'algorithme 2 s'inscrit dans une lignée d'algorithmes, dont ADDTREE (Sattah et Tversky, 1977) a été le point de départ. Dans ceux-ci la priorité est mise sur l'obtention de la structure d'arbre la mieux "appropriée" plutôt que sur l'estimation directe d'une distance additive . Il a, dans cette mouvance, l'avantage d'être:
- exact sur les distances additives;
- plus rapide en pratique grâce à l'utilisation des groupements.
De plus, il permet d'obtenir des arbres non binaires et fournit un indice de qualité de l'arbre dont l'utilisateur peut utilement tenir compte.

Enfin, une fois l'arbre construit, on peut toujours réévaluer ses arêtes de manière à ce que la distance additive induite approche au mieux (mais relativement à cet arbre là), au sens des moindres carrés, la dissimilarité de départ. Il ne s'agit plus que d'un simple problème d'optimisation quadratique sous contrainte de positivité. On obtient ainsi une heuristique pour le problème NP–difficile (c.f. Day (1986)) de l'approximation directe.

Cet algorithme est comparé à d'autres dans Guénoche (1987), sur des données de petite taille (n=7). Dans Luong (1988) de telles comparaisons sont également effectuées sur un échantillon plus vaste et des données de tailles plus diversifiées (n de 18 à 32).

3 APPLICATION À DES ANALYSES DE DONNÉES TEXTUELLES: LA CONNEXION LEXICALE.

3-1 Les indices de la connexion lexicale.

Soient deux textes T_1 et T_2 et soient L_1 et L_2 les lexiques (i.e. les ensembles de formes lexicales -mots et ponctuation-) correspondants. Le problème ici est de savoir dans quelle mesure deux textes partagent le même contenu lexical. Plus nombreux sont les mots communs aux deux textes, plus forte est leur connexion lexicale . Au contraire si l'on considère les vocables du premier texte que l'on ne rencontre pas dans le second, on mesure l'indépendance du premier par rapport au second. On peut d'ailleurs considérer le problème aussi bien au niveau des occurrences qu'à celui des vocables.

Introduits par Muller [1967], *l'indice de connexion lexicale* entre L_1 et L_2 est la quantité

(i) $|L_1 \cap L_2| / |L_1 \cup L_2|$

et *l'indice d'indépendance de* T_1 *par rapport à* T_2 la quantité

(ii) $|L_1 - L_2| / |L_1|$.

Le premier est une distance, la deuxième est un indice non symétrique; ils ne tiennent compte que de la présence ou de l'absence des vocables dans le vocabulaire. On remarque que premier est le fameux indice de Jaccart [1908].

Pour faire intervenir la fréquence des vocables (i.e. les occurrences), ce qui est important dès que les textes sont assez longs, Muller [1967] et Brunet [1971] introduisent une autre mesure appelée *connexion lexicale* basée sur l'hypothèse que la répartition des occurrences des vocables dans les deux textes suit la loi binômiale. Cet indice est construit par une procédure informatique élaborée, que l'on peut trouver dans Brunet [1971] , écrite en PL1. Elle consiste à:

(1) Etablir un tableau de fréquences à deux dimensions (f_{ij}): f_{ij} , pour i,j fixés, est le nombre de vocables ayant la fréquence i dans le premier texte et la fréquence j dans le second.

(2) Calculer les effectifs théoriques à partir de la loi binômiale. En pratique, on se limitera à des effectifs théoriques supérieures à un certain seuil σ (σ=5 ici).

(3) Calculer les χ^2 entre les (f_{ij}) et les effectifs théoriques. Déterminer le nombre de degrés de liberté ν.

(4) Sommer les χ^2 (lorsque les effectifs théoriques sont supérieurs à σ). En déduire les écarts-réduits par la transformation $(2\chi^2)^{1/2} - (2\nu - 1)^{1/2}$.

Nous l'illustrons ici par un exemple extrait de Brunet [1987].

Le tableau 2 représente la répartion de 4729 vocables qui appartiennent à *Lettres à la fiancée* et *Hernani* de Hugo . Les lignes représentent les effectifs observés dans les *Lettres* et les colonnes ceux d'*Hernani* . Notons que l'indice d'indépendance lexicale (ii) ne fait intervenir que le total de la première ligne (988), le total de la première colonne (2253) et le total général (4729) et l'indice de connexion lexicale (i) utilise seulement ce total général et le total de tout ce

qui n'appartient ni à la première colonne, ni à la première ligne, soit 1478 vocables. Si l'on admet "le schéma de l'urne":

-les hapax (i.e. formes de fréquence 1 , respectivement ici 1009 et 640 occurrences) devraient se distribuer proportionnellement à l'étendue de chacun des deux textes (notées respectivement e_1 et e_2 ; on a ici e_1=93890 et e_2=22833 occurrences; notons aussi $p=e_1/(e_1+e_2)$ et $q=e_2/(e_1+e_2)$),

- la répartition de la classe 2 (391,165,161 suivant la sous-diagonale montante liant les fréquences 2) doivent correspondre à la distribution théorique

$$(p+q)^2 = p^2 + 2p\,q + q^2$$

et celles des classes de fréquence n selon le développement de $(p+q)^n$.

(les effectifs de la classe 20 englobent les classes supérieures)

fréq.	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	0	640	161	76	41	17	15	11	7	4	3	3	3	1	0	0	3	1	0	2	10
1	1009	145	44	16	10	6	8	6	4	5	2	0	0	0	0	0	0	1	0	0	4
2	391	68	31	19	7	9	2	0	0	3	1	1	2	1	0	0	0	0	0	0	4
3	199	56	18	17	8	5	3	4	2	1	3	0	0	0	0	0	0	0	0	0	4
4	121	34	17	9	6	3	3	2	0	1	0	1	0	0	0	0	0	0	0	0	2
5	105	34	12	11	6	5	3	2	2	2	1	0	0	1	1	0	0	0	0	0	0
6	88	23	12	7	3	4	2	0	2	1	0	2	2	0	1	0	0	1	1	0	1
7	54	21	4	8	7	6	2	1	1	1	0	2	1	0	0	0	0	1	0	0	0
8	41	13	6	6	8	3	0	1	1	1	0	0	0	0	1	0	1	0	0	0	0
9	33	11	8	5	4	1	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0
10	30	4	10	3	2	1	0	1	2	2	0	0	0	1	0	0	1	0	0	1	1
11	17	9	4	2	4	1	1	2	0	1	0	0	1	0	0	0	0	0	0	0	0
12	20	7	4	3	4	0	1	1	0	1	0	0	0	1	0	0	1	0	0	0	1
13	14	4	7	5	2	3	3	0	2	0	0	0	0	0	0	0	0	0	0	0	1
14	17	8	4	0	1	2	1	0	0	0	0	1	0	0	0	0	0	0	0	0	2
15	10	3	3	6	2	2	0	0	3	0	0	0	0	1	0	0	0	0	0	0	0
16	15	4	4	2	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
17	5	4	3	0	1	1	1	0	1	0	0	1	0	1	1	0	0	0	0	0	0
18	8	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2
19	6	2	1	2	2	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	2
20	60	44	26	29	28	17	13	18	11	7	14	8	8	7	10	8	6	8	1	8	137

Tableau 2: *Répartition de 4729 vocables qui appartiennent à Lettres à la fiancée et Hernani.*[1]

Le tableau 3 rend compte d'un autre couple qui croise les *Contemplations* et les *Travailleurs de la mer.* (on n'a représenté ici que jusqu'à la fréquence 9 dans les deux textes). Les effectifs théoriques sont rapprochés des observations et les écarts transformé en χ^2 (lorsque la fréquence théorique est supérieure au seuil de 5) . La somme des χ^2 partiels donne une distance $\Delta(T_1,T_2)$ entre les deux textes, avec un degré de liberté $\Lambda(T_1,T_2)$ qui est égal à 38 ici . Pour atténuer l'effet de taille, la distance entre deux textes est convertie en écart réduit par la transformation

$$\Gamma(T_1,T_2) = (2\,\Delta\,(T_1,T_2))^{1/2} - (2\Lambda(T_1,T_2) -1)^{1/2} .$$

Ainsi, avec la terminologie de Muller [1967], les *χ^2 de la connexion lexicale* constituent la distance Δ et Γ est *l'écart réduit de la connexion lexicale* . Le tableau 4 représente cet écart entre 22 oeuvres de HUGO. Nous traitons ce type de données dans les applications présentées ici.

---

[1]Les tableaux 2,3 et 4 donnent une image des sorties d'ordinateur, les auteurs s'excusent du manque de lisibilité de celles-ci compte tenu du taux de réduction nécessaire.

Croisement CONTEMPLATIONS et TRAVAILLEURS DE LA MER

tableau réel

fréq.	0	1	2	3	4	5	6	7	8	9
0	0	2651	1000	468	253	149	101	68	55	43
1	1051	560	205	107	95	52	64	26	16	
2	295	167	103	61	50	38	31	20		
3	110	65	52	26	27	27	25			
4	48	53	28	21	25	12				
5	39	33	22	16	18					
6	26	14	9	19						
7	17	16	17							
8	19	6								
9	5									

tableau théorique

fréq.	0	1	2	3	4	5	6	7	8	9
0	0	2237	604	212	82	38	15	9	4	2
1	1465	791	416	215	124	60	39	21	12	
2	259	272	211	162	99	77	48	30		
3	59	92	106	86	84	63	46			
4	15	35	42	55	52	46				
5	5	11	22	27	30					
6	1	5	9	13						
7	0	2	4							
8	0	1								
9	0									

tableau des écarts

fréq.	0	1	2	3	4	5	6	7	8	9
0	0	414	396	256	171	111	86	79	51	41
1	-414	-231	-211	-68	-29	-8	25	5	4	
2	36	-105	-108	-61	-49	-39	-17	-10		
3	61	-27	-54	-60	-57	-36	-21			
4	33	18	-14	-34	-27	-34				
5	34	22	0	-11	-12					
6	25	9	0	6						
7	17	16	13							
8	19	7								
9	5									

tableau des x2

fréq.	0	1	2	3	4	5	6	7	8	9
0	0.0	76.5	259.6	309.6	355.3	327.1	476.2	756.7	0.0	0.0
1	116.9	234.8	107.1	21.6	6.7	1.2	15.6	1.2	1.7	
2	5.0	40.8	55.5	40.5	24.2	19.8	6.2	3.5		
3	61.6	8.0	27.6	42.2	38.8	20.7	9.9			
4	71.7	9.6	4.9	21.1	13.8	24.7				
5	0.0	43.1	0.0	4.5	4.7					
6	0.0	0.0	0.0	2.7						
7	0.0	0.0	0.0							
8	0.0	0.0								
9	0.0									

données relatives au couple (totaux des diagonales)

freq. 1	2	3	4	5	6	7	8	9	
3702	1654	960	630	466	316	291	226	185	total:8617

nbre consécutifs dans le couple : 8617
nbre consécutifs dans CONTEMPLATIONS : 3609 (sur 5692)
nbre consécutifs dans TRAVAILLEURS : 6797 (sur 8824)

x2 (regroupement des effectifs faibles): 2402
x2 total : 6315
degrés de liberté : 58

Tableau 3: *connexion lexicale, détail des calculs*

	LET	MER	AUT	DAM	BOR	TUD	RUY	RAY	RHI	COI	CON	LS1	MI1	MI2	MI3	FIN	RUE	CO2	MER	CO3	LS2	LS3
LET		79.1	126.2	131.5	65.4	56.7	106.7	132.5	132.6	60.1	151.2	192.0	118.6	122.2	131.1	154.3	168.4	84.0	157.5	100.1	163.1	156.9
MER	79.1		45.3	40.4	26.4	30.6	26.3	48.4	43.1	65.4	52.2	43.3	46.0	48.5	49.1	54.8	55.2	70.3	54.1	66.1	43.6	47.1
AUT	126.2	45.3		56.7	57.5	65.4	54.9	8.7	54.8	99.9	22.9	28.6	64.3	59.8	56.8	31.6	35.7	117.8	65.3	139.0	24.9	28.4
DAM	131.5	40.4	56.7		37.1	43.7	36.6	48.4	81.2	114.5	90.9	80.1	71.4	74.5	76.0	76.6	71.5	156.8	105.9	123.8	90.7	63.3
BOR	65.4	26.4	57.5	37.1		22.8	30.2	62.0	36.9	54.0	51.5	46.1	36.6	37.4	39.9	55.5	58.0	50.1	45.0	51.4	42.9	51.5
TUD	56.7	30.6	65.4	43.7	22.8		39.2	68.4	44.3	42.6	56.1	57.1	43.0	43.6	43.3	66.3	68.6	44.5	54.6	43.9	49.4	62.5
RUY	106.7	26.3	54.9	36.6	30.2	39.2		50.8	41.3	77.5	51.8	49.2	36.4	35.0	35.2	62.5	56.1	78.9	52.6	79.5	48.7	48.2
RAY	132.5	48.4	8.7	48.4	62.0	68.4	50.8		43.0	112.6	15.9	26.9	53.8	50.5	52.6	31.9	28.5	117.2	66.2	106.0	22.4	26.5
RHI	132.6	43.1	54.8	81.2	36.9	44.3	41.3	43.0		122.0	81.6	76.5	69.8	80.3	77.8	68.3	64.6	146.8	101.2	110.2	83.8	49.4
COI	60.1	65.4	99.9	114.5	54.0	42.6	77.5	112.6	112.0		136.1	164.1	93.2	98.3	102.6	147.3	144.3	42.9	138.3	46.1	151.2	133.1
CON	151.2	52.2	22.9	90.9	51.5	56.1	51.8	15.9	81.6	136.1		36.4	86.1	86.9	89.1	15.9	34.3	144.0	101.9	134.7	21.0	22.6
LS1	192.0	43.3	28.6	80.1	46.1	57.1	49.2	26.9	76.5	164.1	36.4		85.8	84.0	77.2	21.5	30.0	189.9	90.3	152.4	8.7	4.3
MI1	118.6	46.0	64.3	71.4	36.6	43.0	36.4	53.8	69.8	93.2	86.1	85.8		30.0	25.6	74.0	74.6	111.5	73.4	90.8	67.8	63.2
MI2	122.2	48.5	59.8	74.5	37.4	43.6	35.0	50.5	80.3	98.3	86.9	84.0	30.0		28.2	74.3	62.0	116.0	70.5	95.5	50.1	57.4
MI3	131.1	49.1	56.8	76.0	39.9	43.3	35.2	52.6	77.8	102.6	89.1	77.2	25.6	28.2		65.2	56.6	123.8	61.0	96.8	61.6	54.7
FIN	154.3	54.8	31.6	76.6	55.5	66.3	62.5	31.9	68.3	147.3	15.9	21.5	74.0	74.3	65.2		41.2	137.5	77.0	136.9	14.8	15.2
RUE	168.4	55.2	35.7	71.5	58.0	68.6	56.1	28.5	64.6	144.3	34.3	30.0	74.6	62.0	56.8	41.2		144.4	72.6	137.3	30.3	20.0
CO2	84.0	70.3	117.8	156.8	50.1	44.5	78.9	117.2	146.8	42.9	144.0	189.9	111.5	116.0	123.8	137.5	144.4		169.7	16.8	144.1	133.0
MER	157.5	54.1	65.3	105.9	49.0	54.6	52.8	66.2	101.2	136.3	101.9	90.3	73.4	70.5	61.0	77.0	72.6	169.7		132.4	93.2	59.7
CO3	100.1	66.1	139.0	123.8	51.4	43.9	79.5	106.0	110.2	46.1	134.7	152.4	90.8	95.5	90.8	136.9	137.3	16.8	132.4		136.1	114.2
LS2	163.1	43.6	24.9	90.7	42.9	49.4	48.7	22.4	83.8	151.2	21.0	8.7	87.8	90.1	81.6	14.8	30.3	144.1	93.2	138.1		4.9
LS3	156.9	47.1	28.4	63.3	51.5	62.5	48.2	26.5	49.4	133.1	22.6	4.3	63.2	57.4	54.7	15.2	20.0	133.0	59.7	114.2	4.9	

Tableau 4: *Ecarts réduits de la connexion lexicale entre 22 textes de Victor Hugo*

3-2 Victor HUGO.

Nous utilisons les données de Brunet [1987] provenant d'un vaste dépouillement de l'oeuvre de Hugo : vingt textes complets (donnant plus de deux millions d'occurrences!), dont trois romans (y compris l'intégralité des *Misérables*), quatre pièces de théâtre, huit textes poétiques, et quatre recueils de correspondances. S'y ajoute un récit de voyage, le *Rhin* . Comme le poids écrasant des Misérables risquait de détruire l'équilibre, on a réparti en trois sous-textes leur masse imposante . La *Légende des siècles* compte à elle seule pour trois textes, eu égard à l'échelonnement des dates de publication.

La figure 6 est la représentation de ces données par un arbre. La corrélation avec la mesure de départ est excellente (r=0,95): cette dernière était "presque" une distance additive d'arbres. A chaque sommet intérieur est associé son indice d'agrégation. A part celui de l'étoile centrale qui est faible (0,65), tous les autres sont proches de 1 , ce qui montre la qualité de la structure de l'arbre. Celui-ci est d'une très bonne lisibilité: on sera sensible à la netteté des groupes constitués par les branches de l'arbre, notamment à celui qui réunit les textes poétiques et à celui qui fédère les textes épistolaires. Dans les deux cas les sous–groupes s'imposent avec la même clarté: c'est la chronologie qui explique la relation privilégiée entre les *Lettres* et le *Correspondance* 1 , entre *Correspondance* 2 et *Correspondance* 3 , entre les *Feuilles* et *Rayons* . L'unité des trois sous-ensembles de la *Légende des siècles* est nettement affirmée comme le rapport étroit qui lie la *Fin de Satan* aux *Contemplations* que Hugo a explicitement reconnu. L'arbre indique l'autonomie des éléments romanesques qui ont un lien lâche avec la structure. C'est qu'une grande distance est établie entre *Notre-Dame* et les *Travailleurs de la*

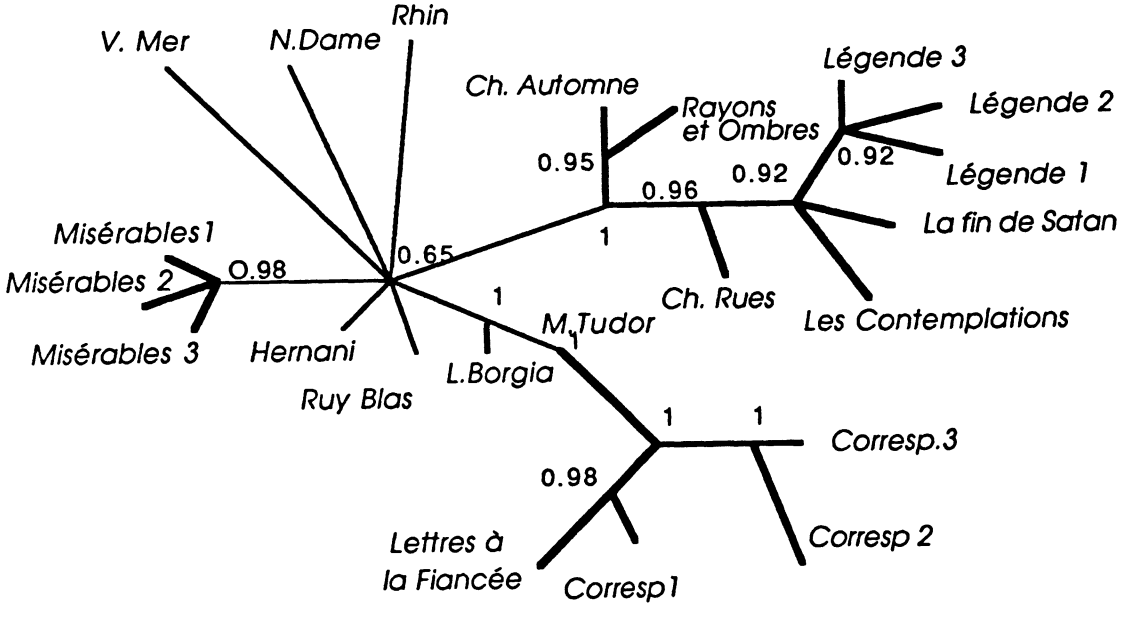


Figure 6 : *Représentation arborée de l'écart réduit de la connexion lexicale de 22 textes de Hugo.*

mer, qui divergent par le thème, l'écriture et la date de composition. Quant au *Rhin* , il est unique en son genre. Le théâtre n'est pas non plus solidement rattaché à la structure, surtout les pièces en vers dont le statut propre implique une division intérieure. Les pièces en proses, elles, se trouvent sur le chemin entre la correspondance et les autres textes. Ce sont les indications de distance qui regroupent les quatre pièces de théâtre.

3-3 Jean GIRAUDOUX.

L'analyse arborée de la connexion lexicale de 16 textes de Giraudoux donne aussi de résultats remarquables. Les données sont dans Brunet [1971]. Elles proviennent du dépouillement de quatres romans et douze pièces de théâtre de Giraudoux.

Ici encore la corrélation est très élevée (0,95). Lisons l'arbre de la figure 7 de bas en haut. Les deux premiers romans de l'ordre chronologique *Simon le Pathétique* et *Suzanne et le Pacifique* sont dans un groupement; il en est deux même pour les deux romans suivants, *Siegfried et le Limousin* et *Bella* . Ces deux groupements se réunissent avec *Siegfried* et *Intermezzo* , adaptations de romans pour le théâtre, pour former l'une des deux branches principales de l'arbre. Au dessus de cette branche, les pièces courtes *Cantiques des Cantiques* et *Apollon de Bellac* sont proches par leur distance, ce dernier se regroupe avec *La Folle de Chaillot* , toutes deux pièces modernes et comiques. En haut de l'arbre, les pièces nobles se rejoignent à un embranchement de cinq branches dont l'un est constitué par les pièces tragiques et sobres: *Electre* , *Sodome et Gomorrhe* et *Pour Lucrèce* , toutes les trois s'inspirant de l'antiquité judéo–gréco–latine.

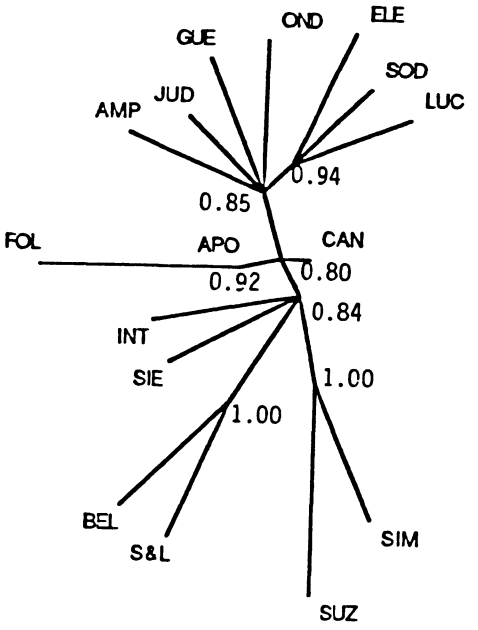


Figure 7 : *Représentation arborée de l'écart réduit de la connexion lexicale de 16 textes de Giraudoux.*

REMERCIEMENTS. Nous remercions chaleureusement Etienne Brunet grâce auquel nous avons pu accéder au vaste ensemble des données qu'il a constitué, qui nous a permis de reproduire ici certains de ses tableaux et qui nous a apporté une aide précieuse lors de l'interprétation des résultats. Nos remerciements vont aussi aux deux référés "anonymes" pour leur aide aimable et fructueuse.

BIBLIOGRAPHIE

BANDELT, H.J. & DRESS A. "Reconstructing the shape of a tree from observed dissimilarity data" , *Advances in Applied Mathematics*, (1987), (à paraîttre).

BARTHELEMY, J.P. & GUENOCHE, A. *Arbres et représentations des proximités* , Collection Méthodes et Programmes, Masson, (1987), Paris.

BARTHELEMY, J.P. & LUONG, N.X. "Représentations arborée des mesures de dissimilarités" , *Statistique et Analyse des Données* , 11, 1, (1986), 20-41.

BRUNET, E. "La connexion lexicale", *C.U.M.F.I.D.*, 4, Université de Nice, (1971), 173-207.

BRUNET, E. "Une mesure de la distance intertextuelle: la connexion lexicale", Actes du Coll. "*Le Nombre et le Texte*", (1987) , Université de Liège.

BUNEMAN, P. "The recovery of trees from measures of dissimilarity" . in *Mathematics in Archeological and Historical Sciences*. F.R. Hodson, D.G. Kendall, P. Tautu eds. , Edinburgh University Press, (1971) , 387-395.

BUNEMAN, P. "A note on metric properties of trees" , *Journal Comb. Theory* (B), 17, (1973), 48-50.

COLONIUS, H. & SCHULZE, H.H. "Tree structure for proximity data" , *British Journ. of Math. and Stat. Psychology*, 34, (1981), 167-180 .

DAY, W.H.E. *Analysis of quartet dissimilarity measures between undirected phylogenetic trees*, FCAR report, CRM-1315,(1985) , Newfoundland.

DAY, W.H.E. *Computational complexity of infering phylogenies from dissimilarity matrices* . Technical Report 8610, Memorial University of Newfounland, St John's, Newfoundland, (1986) , Canada.

DOBSON, A.J. "Unrooted trees for numerical taxonomy" , *Journ. of Applied Proba.*, 11, (1974) , 32-42 .

ESTABROOK, G.F., F.R. & MEACHAM, C. "How to determine the compatibility of undirected character state trees", *Math. Biosc.*, 46, (1979), 251-256.

ESTABROOK, G.F., McMORRIS, F.R. & MEACHAM, C. "Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units", *Systematic Zoology*, 22, 1, (1985), 193-200.

FARRIS, J.S. "On comparing the shape of taxonomic trees", *Systematic Zool.*, 22, 1, (1973), 50-54.

FURNAS, G.W. "Neighbour counts, tree metrics and pairwise cluster structure", Communication au *4th Europeen Meeting of the Psychometric Soc. and Classification Soc.*, (1985), Cambridge.

GUENOCHE, A "Etude comparative de cinq algorithmes d'approximation des dissimilarités par les arbres à distances additives", *Math. Sci. hum.*, 98, (1987), 21-40.

JACCART P. "Nouvelles recherches sur la distribution florale", *Bull. Soc. Vand. Sci. Nat.*, 44, (1908) , 223-270.

LUONG, N.X. "Voisinage lâche, score et famille scorante", *Cahiers du S.U.R.F.*, 2, (1983), Université de Besançon.

LUONG, N.X. *Méthodes d'analyse arborée. Algorithmes. Applications* . Thèse d'Etat , (1988), Université de Paris V.

McMORRIS, F.R. "Axioms for consensus functions on undirected phylogenetic trees", *Mathematical Biosciences* 74, (1985), 17-21.

MULLER, C. *Etude de Statistique Lexicale* . Larousse, Paris, (1967) , 379 p .

PATRINOS, A.N. & HAKIMI, S.L. "The distance matrix of a graph and its tree realization" , *Quart. Appl. Math.*, (1972), 255-269 .

PHIPPS, J.P. "Dendrogram topology", *Systematic Zool.*, 20, 3, (1971), 306-308.

SATTAH, S. & TVERSKY, A. "Additive similarity tree", *Psychometrika* , 42, 3 , (1977), 319-345 .

SIMOES-PEREIRA, J.M.S. "A note on tree realizability of a distance matrix", *Journ. Comb. Theory* (B) , 6, (1967), 303-310 .

WATERMAN, M.S. & SMITH, T.F. "On the similarity of dendrograms", *J.Theor. Biology*, 73, (1978), 369-381.

ZARETSKII, K. "Constructing a tree on the basis of a set of distances between hanging vertices" , *Upekki Math. Nauk.*, 20, (1965), 90-92, (en Russe).