Problème de la contamination dans le cadre de l'édition critique

Marc Le Pouliquen[1, 2], Jean-Pierre Barthélemy[1, 3]

[1]Département LUSSI TAMCIC, UMR CNRS 2872
ENST Bretagne, BP 832, 29 285 Brest Cedex

[2]Université de Bretagne Occidentale IUP GMP
6 av. Le Gorgeu - CS93837 29 238 Brest Cedex

[3]CAMS, UMR CNRS 8557, Ecole des Hautes Etudes en Sciences Sociales
54 bvd Raspail, 75 270 Paris cedex 06

Abstract

This paper presents two methods which make it possible to numerically study the contamination of a textual tradition. After a study of the variants observed between two versions of the text, we observe that some variants are genealogical, that is, they allow to establish the stemma codicum, while other variants are parallel or contaminated. The first method consists in building an index, which makes it possible to quantify the contamination of the corpus. The second uses phylogenetic methods in order to visualize the contaminations on the stemma by a graph.

Résumé

Ce papier présente deux méthodes permettant d'étudier numériquement la contamination d'une tradition textuelle. Après une étude des variantes observées entre deux versions du texte, nous observons que certaines variantes sont généalogiques, elles permettent d'établir le stemma codicum, tandis que d'autres sont parallèles voire contaminées et nuisent à la construction du stemma. La première méthode consiste à construire un indice permettant de quantifier la contamination du corpus. La deuxième utilise les méthodes de la phylogénétique afin de visualiser par un graphe les contaminations comme sur un stemma.

Mots-clés : contamination, hybridation, tradition textuelle, phylogénétique.

1. Introduction

Dans le cadre de l'édition d'anciens manuscrits, un des problèmes consiste à trier les différentes versions du texte appelées « témoins » afin d'essayer de reconstituer le manuscrit original avec fidélité. Dans l'approche « généalogique » ou « stemmatique », les éditeurs de texte emploient souvent des différences textuelles appelées « variantes », comme outils pour découvrir la parenté des versions du texte et dressent alors un arbre généalogique de cette filiation que l'on nomme « stemma codicum » (cf. Fig 1). L'ordinateur peut permettre de détecter toutes les variantes rapidement. Cependant, toutes les variantes ne sont pas « généalogiques », dans le sens où elles ne donnent pas d'information au sujet de la parenté des versions du texte. En effet, la reconstruction généalogique suppose que chaque copiste n'a utilisé qu'un modèle pour réaliser son exemplaire (filiation unique). On parle alors de transmission « verticale », et la tradition est dite « fermée ».

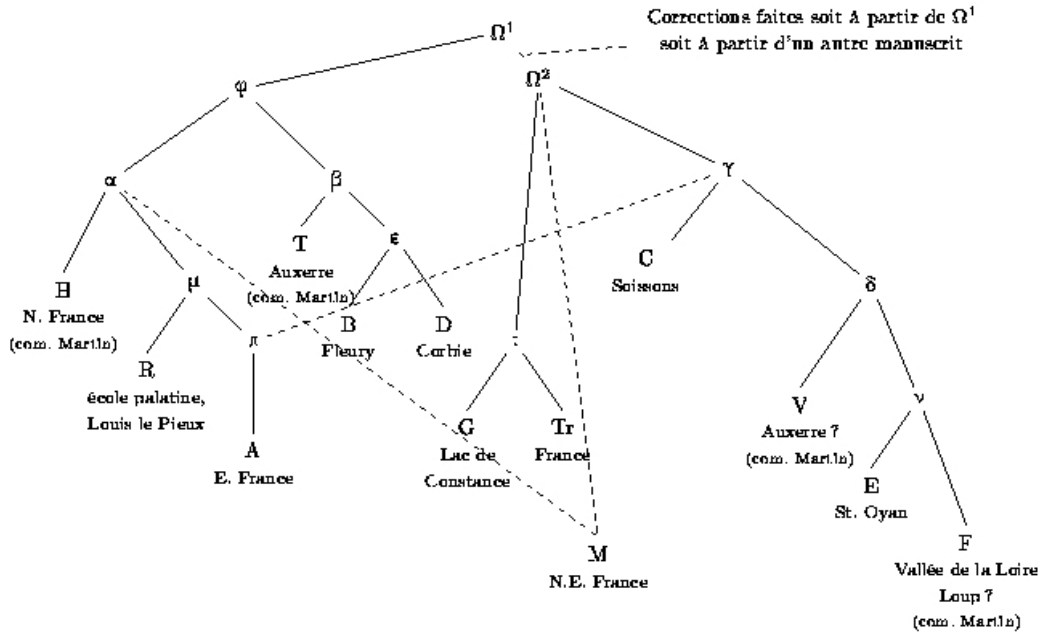

Figure 1 : Stemma de De Nuptiis Philologiae et Mercurii établi par *Danuta Shanzer*

Seulement, une fois que l'on admet la possibilité qu'un copiste ait employé plus d'un exemplaire pour obtenir son manuscrit, la construction de l'histoire du texte à partir des variantes devient plus compliquée voire impossible comme l'a constaté Paul Mass (1957) dans « Textkritick ». Ceci s'appelle la transmission « horizontale » ou « contamination », et une tradition de cette sorte est dite « ouverte ».

Dans ce papier, nous allons tenter, par différentes approches, de quantifier la contamination supposée d'une tradition textuelle puis de la modéliser à partir d'un stemma codicum.

Pour cela, nous examinerons dans un premier temps comment s'effectue la contamination au niveau des variantes. Dans un deuxième temps, nous proposerons deux méthodes informatiques permettant à l'éditeur de détecter la transmission horizontale. La première méthode s'attache à quantifier la contamination tandis que la deuxième permet de la visualiser sur un graphe. Nous finirons notre papier par un certain nombre de résultats expérimentaux obtenu sur un corpus réel avant de conclure.

2. La contamination au niveau philologique

2.1. Les différents types de variantes

Pour la transmission des textes avant l'imprimerie, les scribes ont eu un rôle majeur. Malgré tous leurs efforts, des erreurs de copie se sont infiltrées dans les différentes versions du texte et certains copistes ont modifié volontairement le texte pour plusieurs raisons. Commençons par examiner les différents types de variantes d'un texte que la critique textuelle a la tâche de repérer.

On trouve, parmi les erreurs de copie qui se reproduisent régulièrement, des erreurs au niveau des caractères ou des mots que l'on répertorie de la façon suivante :

- L'homophonie est la confusion de caractères ou de mots semblables.
- La métathèse consiste à inverser la position d'une lettre ou d'un mot.
- L'haplographie a lieu quand une répétition de caractères est omise par le copiste.
- La dittographie au contraire est la répétition involontaire de caractères ou de mots.
- La séparation d'un mot en deux et la combinaison de deux mots en un.

L'omission par homoeoteleuton où le copiste saute un paragraphe (ou un autre segment textuel) qui débute par le même mot qu'un autre.

Pour une définition plus complète de ces termes, on peut consulter Archer (1978).

On peut aussi classer les modifications volontaires de variantes au cours de la copie :

- Les corrections : Le copiste peut corriger une faute (orthographe, grammaire) qu'il observe dans le texte à copier.
- Les problèmes linguistiques : Le scribe peut moderniser un terme, remplacer un mot rare par un mot plus connu, ou changer légèrement la syntaxe qui évolue dans le temps.
- Les impératifs politiques ou religieux : Afin de valoriser une religion, un personnage ou un une école de pensée, un copiste peut modifier légèrement un texte.
- Les noms propres : Leur écriture évolue selon les régions le temps et la langue.
- Les rajouts explicatifs : Le copiste peut gloser un texte pour le préciser, glose qui sera par la suite intégrée au texte par un autre copiste.
- Les contaminations : Le scribe utilise une autre version du texte pour modifier la copie qu'il réalise.

Cette première étude des variantes est une classification en fonction des modifications réalisées sur les textes par les scribes. Elle permet de mieux comprendre l'évolution de la tradition textuelle. Pourtant, si l'éditeur souhaite déterminer le type de chaque variante (parfois fortement représentée), il lui faut faire preuve de beaucoup de patience et d'une connaissance experte du contexte entourant la tradition textuelle.

On constate aussi qu'un certain nombre de variantes permettent à l'éditeur d'inférer l'histoire de la tradition textuelle ; ceux sont les variantes dites « généalogiques ». En revanche, une partie des variantes que l'on nomme « variantes parallèles » contredisent l'histoire du texte et gênent sa construction.

2.2. Les variantes parallèles

Nous pouvons définir par « variantes parallèles », deux variantes identiques d'un même lieu variant telles que les deux textes qui les contiennent n'ont pas les mêmes « ancêtres ». Autrement dit, deux témoins de deux sous traditions textuelles différentes s'accordent sur la même variante que n'ont pas leurs « ancêtres ».

On peut distinguer trois types de variantes parallèles : les variantes accidentelles, les corrections et les contaminations.

- Parmi les variantes accidentelles, on trouve le cas de deux copistes qui peuvent avoir commis la même faute dans deux traditions différentes. On peut parler de coïncidence.
- Dans le cas des variantes parallèles corrigées, le copiste rectifie certaines fautes qui étaient apparues dans la tradition textuelle. C'est le cas dans une re-correction d'une faute grammaticale ou d'orthographe, dans une réadaptation historique ou géographique du texte, ou dans une réinterprétation sémantique des mots du texte.

  Par exemple, un copiste peut rectifier des mots évidemment incorrects dans le texte qu'il copie. Supposons que nous ayons six versions du texte avec 2 lieux variants sur la phrase suivante :

  « Cette phrase est créée pour l'exemple »

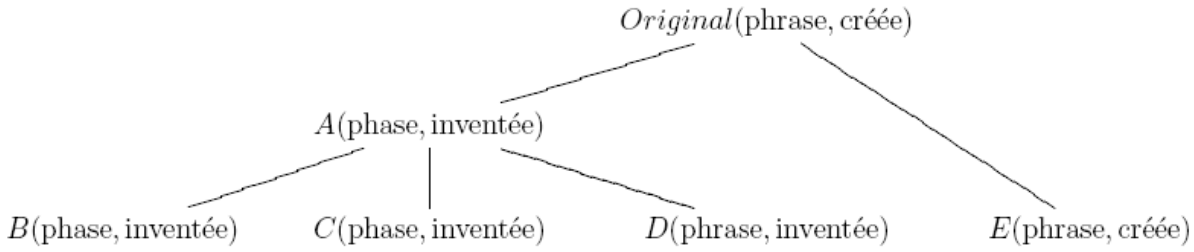

Figure 2 : Stemma codicum de 6 manuscrits O, A, B, C, D et E avec 2 lieux variants

A partir de l'original (cf. Fig. 2), on peut distinguer 2 sous-groupes, le sous-groupe « inventer » avec les témoins (A, B, C et D) et le sous-groupe « créer » avec les manuscrits E et O. L'histoire peut alors se résumer de la façon suivante : Le copiste de A a substitué *inventée* à *créée* et a oublié le r de *phrase*. Les copistes de B et C ont recopié correctement les leçons (autre mot pour variantes) sur A tandis que le copiste du texte D a identifié l'erreur et a reproduit la leçon originale. Ici, on peut conclure que l'on n'a pas de contamination, mais simplement que les éléments incorrects attirent l'attention et demandent à être corrigés, changés ou éliminés.

- Dans le cas des variantes parallèles contaminées, il n'est plus question de re-correction ou autre, mais d'un copiste qui utilise plusieurs textes pour construire sa propre version. Il introduit alors dans une copie des variantes remarquables qu'il a trouvées dans d'autres manuscrits.

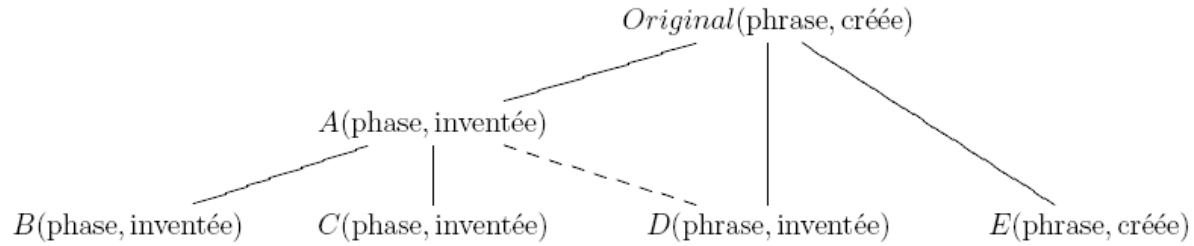

Figure 3 : Stemma codicum de 6 manuscrits O, A, B, C, D et E avec 2 lieux variants

On revient à l'exemple précédent. Une autre interprétation est envisageable pour l'éditeur avec le stemma de la figure 3. Le manuscrit D est envisagé comme étant une copie directe de l'original où le copiste a remplacé la variante *créée* par *inventée*. Cette variante provient du manuscrit A que le copiste connaissait quand il a réalisé sa

version. On constate ici la difficulté qu'il y a pour l'éditeur à reconstituer l'histoire du texte, c'est-à-dire de choisir entre la figure 2 et 3.

Dans le cadre de notre exemple, le faible nombre de lieux variants est un handicap pour le choix du stemma. En effet, c'est statistiquement sur le comportement d'un ensemble de lieux variants que l'on peut évaluer la validité d'une filiation par rapport à une autre.

Afin de détecter les variantes parallèles et en particulier la contamination, nous proposons deux méthodologies différentes. La première approche consiste à élaborer un indice qui caractérise le degré de contamination du corpus voire même d'un manuscrit. La deuxième approche utilise les méthodes phylogénétiques et les adapte à la stemmatique afin de visualiser le stemma et la contamination.

3. Méthode par construction d'un indice

3.1. Problématique

Nous allons construire un indice permettant de visualiser les contaminations, ou plutôt les variantes parallèles. Il est évident que dans l'exemple précèdent, une machine ne peut pas trancher entre contamination ou « corruption » accidentelle. En revanche, elle peut alerter l'éditeur sur une possible contamination.

Si l'on observe les trois manuscrits suivants avec trois variantes :

	Manuscrit A	Manuscrit B	Manuscrit C
Variante 1	phrase	phrase	phase
Variante 2	inventée	créée	créée
Variantes 3	exemple	modèle	exemple

Au vu des variantes 1 et 2, le stemma légitime est le stemma (b) de la figure 4. En effet, le copiste de A a modifié *créée* en *inventée* et le copiste de C a remplacé *phrase* par *phase*. Il est en revanche peu probable que l'on obtienne le stemma (a) car le copiste du manuscrit B et celui de C auraient tous les deux modifié *inventée* en *créée* !

Si l'on considère les variantes 2 et 3, on obtient logiquement le stemma (c) et avec les variantes 1 et 3 le stemma (a).

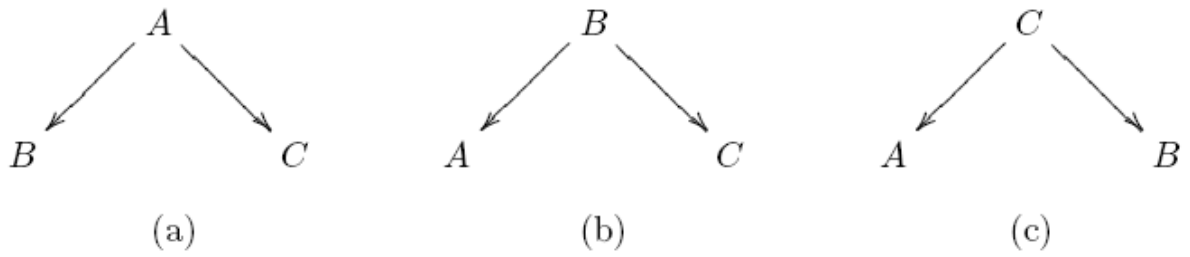

Figure 4 : Trois stemmas selon le choix des variantes.

Si l'on considère ensuite la troisième variante, aucun stemma ne convient, on a contamination. Selon l'orientation (détermination du manuscrit ancêtre), un des trois stemmas de la figure 5 est choisi.

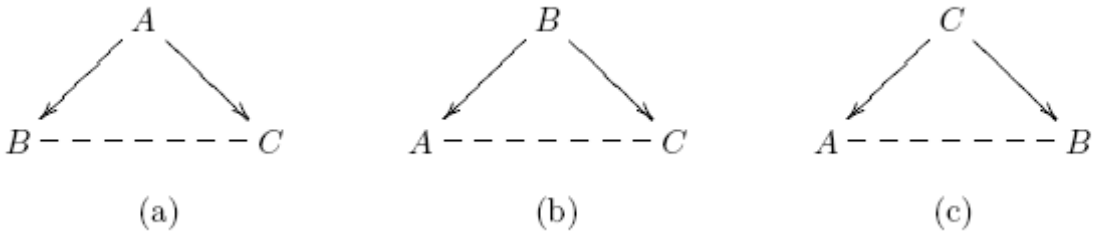

Figure 5 : Trois stemmas selon l'orientation.

Finalement pour déterminer la contamination entre deux manuscrits A et B, il suffit d'observer le problème précédent entre les variantes des manuscrits A, B et un manuscrit O « ancêtre » commun à A et B.

La méthode proposée consiste à utiliser le fichier de collation afin de déterminer pour chaque triplet de manuscrits si l'on a une « présomption » de contamination.

3.2. Méthode

Prenons A, B et C trois manuscrits de la tradition textuelle. On les compare deux à deux puis tous les trois afin de déterminer le nombre de variantes communes.

On pose alors :

$n_{AB} = Card(A \cap B)$

$n_{AC} = Card(A \cap C)$ où $\bigcap$ est l'ensemble des variantes communes aux 2 manuscrits.

$n_{BC} = Card(B \cap C)$

$n_{ABC} = Card(A \cap B \cap C)$ est l'ensemble des variantes communes aux 3 manuscrits.

Si $Ind_{ABC} = \min(n_{ab}, n_{bc}, n_{ac}) - n_{ABC} \geq 0$, cela signifie qu'il existe une variante commune aux manuscrits A et B qui n'appartient pas à C, une autre commune à A et C qui n'appartient pas à B et une dernière commune à B et C que l'on ne trouve pas dans A.

On peut alors construire un indice sur la tradition textuelle en moyennant l'indice précédent sur tous les triplets de manuscrits du corpus et en pourcentage par rapport au nombre de lieux variants :

$Ind_{Co} = \frac{100}{Nt * Nv} \sum_{A,B,C \in Co} Ind_{ABC}$ où Nt est le nombre de triplet du corpus et Co désigne le corpus et Nv est le nombre de lieu variants.

On peut aussi construire un indice par manuscrit en moyennant l'indice précédent sur tous les triplets de manuscrits du corpus contenant notre manuscrit.

$Ind_{A} = \frac{1}{N_A} \sum_{B,C \in Co} Ind_{ABC}$ où N_A est le nombre de triplet du corpus contenant A.

3.3. Exemple

Soit la tradition textuelle suivante composée de 9 manuscrits et de quatre lieux variants. Elle est représentée par le stemma de la figure 6.

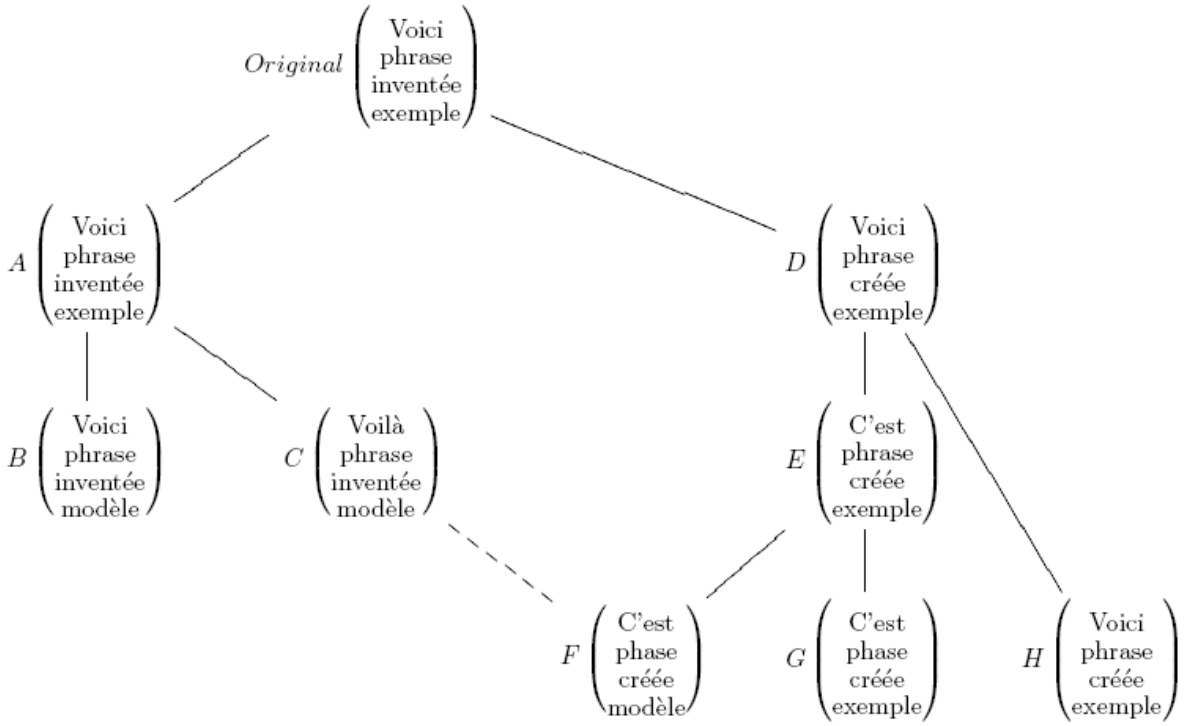

Figure 6 : Stemma codicum de 9 manuscrits avec 4 lieux variants

Le premier lieu variant (*Voici, Voilà, C'est*) partitionne le corpus en trois : OABD, C, EFGH.

Le deuxième lieu variant (*phrase, phase*) partitionne le corpus en deux : OABCDEH, FG.

Le troisième lieu variant (*inventée, créée*) partitionne le corpus en deux : OABC, DEFGH.

Le dernier lieu variant (exemple, modèle) partitionne le corpus en deux : OADEGH, BCF.

Voici une partie des calculs d'indice pour les 84 triplets ($C_9^3 = \frac{9x8x7}{6}$) du corpus. Seuls quatre triplets ne sont pas nuls : BFH, CDF, CEF et CFH.

triplets	BFG	BFH	BGH	CDE	CDF	CDG	CDH	CEF	CEG	CEH	CFG	CFH	…
Indices	0	1	0	0	1	0	0	1	0	0	0	1	0

La moyenne de contamination du corpus 1,78%.

Et la moyenne par manuscrit est celle du tableau suivant :

Manuscrits	O	A	B	C	D	E	F	G	H
Indices	0	0	3	3	2	2	6	0	2

On constate bien ici que le manuscrit F est le plus contaminé probablement avec le manuscrit B ou C.

4. Méthode par détection de l'hybridation en phylogénétique.

4.1. Introduction

Les éditeurs textuels et les biologistes évolutionnaires étudient le processus de la transmission (de gènes ou de textes) dans le temps. Il y a beaucoup d'analogies entre la façon dont un texte a été copié d'un manuscrit à l'autre, et la façon dont les espèces se sont transformées en de nouvelles espèces. Dans les trois dernières décennies, certains éditeurs ont utilisé les méthodes phylogénétiques dans le cadre de la stemmatique textuel. La contamination a bien entendu son pendant phylogénétique connu sous le nom d'hybridation et les variantes parallèles sont associées aux transferts de gènes horizontaux.

Un certain nombre de tentatives de modélisation de la contamination par méthodes phylogénétiques a déjà été effectué :

- Stephen Carlson (2004) a abordé le problème de la contamination en utilisant un enchaînement de méthodes dont celle des « hypertree » de Dickerman (1998). Le problème principal avec l'approche de Dickerman, est quelle est très coûteuse en calcul. Cela impose donc une limitation significative du nombre de manuscrits et du nombre de variantes.
- Spencer et al. (2004) ont utilisé la méthode « Reduced Median Network » de Bandelt et al. (1995). Cette approche est peu réaliste et difficilement exploitable car elle produit 8 517 manuscrits hypothétiques partant d'un corpus 82 manuscrits.

Nous allons regarder ce que peut donner la méthode de l'évolution réticulée de Legendre et Makarenkov (2000).

4.2. Méthode

On commence par construire une matrice de dissimilarité où la « distance » entre deux manuscrits est égale au nombre de variantes différentes. On utilise ensuite un algorithme de reconstruction phylogénétique (NJ de Saitou et Nei (1987) ou Addtree de Sattah et Tversky (1977)) qui nous permet d'obtenir un arbre de la filiation. Pour ajouter une nouvelle arête caractérisant la contamination, l'algorithme minimise la fonction des moindres carrés calculée comme étant la somme des différences quadratiques entre les dissimilarités d'origine et les distances de l'arbre obtenu. C'est-à-dire que l'ajout de la nouvelle arête permet une meilleure approximation de la dissimilarité que celle obtenue avec l'arbre.

L'algorithme s'arrête quand il a construit le nombre d'arêtes indiqué par avance ou à l'obtention du minimum d'un critère statistique.

Le logiciel utilisé afin de visualiser la contamination est TRex de Makarenkov (2001).

4.3. Exemple

Avec la tradition textuelle représentée par le stemma de la figure 5, nous avons obtenu la matrice de dissimilarité suivante :

	O	A	B	C	D	E	F	G	H
O	0	0	1	2	1	2	4	3	1
A	0	0	1	2	1	2	4	3	1
B	1	1	0	1	2	3	3	4	2
C	2	2	1	0	3	3	3	4	3
D	1	1	2	3	0	1	3	2	0
E	2	2	3	3	1	0	2	1	1
F	4	4	3	3	3	2	0	1	3
G	3	3	4	4	2	1	1	0	2
H	1	1	2	3	0	1	3	2	0

On peut par exemple constater que la distance du manuscrit B au manuscrit E est 3 car il y a deux variantes différentes, *Voici* à la place de *C'est*, *inventée* à la place de *créée* et *modèle* à la place de *exemple*.

En utilisant le logiciel TRex avec comme racine le manuscrit O, on obtient la figure 7.

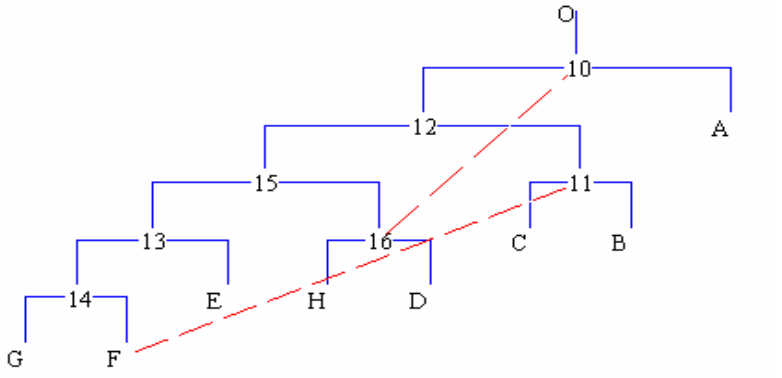

Figure 7 : Graphe réticulé obtenu avec TRex

On retrouve bien la contamination entre F et B ou F et C. Une autre contamination apparaît cependant entre O et H ou O et D qui n'est pas justifiée. Il faut donc affiner la méthode s'il l'on veut obtenir des résultats plus précis, comme par exemple en utilisant les indices précédents qui permettent de « confirmer » une contamination ou de la rejeter. Ceci dit, même confirmée par le programme, la contamination devra être vérifiée par l'éditeur, celui-ci étant le seul apte à valider le résultat.

5. Application à un corpus réel

Afin de réaliser les premières expérimentations, nous avons utilisé le corpus du poème qui ouvre le livre IX du *De Nuptiis Philologiae et Mercurii* Tertullianus de Martianus Capella. Ce corpus, collecté par Jean-Baptiste Guillaumin (2004), a l'avantage de disposer d'un stemma (cf. Fig. 1) élaboré par Danuta Shanzer (1986) qui propose une contamination.

Le corpus est constitué de 18 poèmes désignés par :

Co= (A,B,C,D,E,F,H,K,L,M,O,P,R,S,U,V,W,Z)

Après avoir réalisé le fichier de collation composé de 74 lieux variants et de 234 variantes, nous avons calculé les indices définis précédemment :

La moyenne de contamination du corpus 9,16%.

Et la moyenne par manuscrit est celle du tableau suivant :

A	B	C	D	E	F	H	K	L	M	O	P	R	S	U	V	W	Z
1 052	1 005	909	984	670	805	887	908	920	974	961	907	915	896	1 038	826	899	1 061

On constate bien ici que les poèmes A, B, U et Z sont les plus contaminés. Est-ce pour cela représentatif sachant qu'entre 14 poèmes sur 18, on a seulement 16% de différence entre les indices ? La seule chose vérifiable est pour le poème A qui semble en effet être contaminé d'après le stemma de la figure 1. Pour le poème B rien ne l'indique sur le stemma. Les poèmes U et Z ne sont pas représentés.

Le taux moyen de contamination du corpus est environ de 9% ce qui n'est pas en soit interprétable, car on manque d'ordre de grandeur. En effet, on ne dispose pas d'autres corpus avec lesquels réaliser la comparaison.

En utilisant TRex avec un calcul de dissimilarité, on obtient la figure 8.

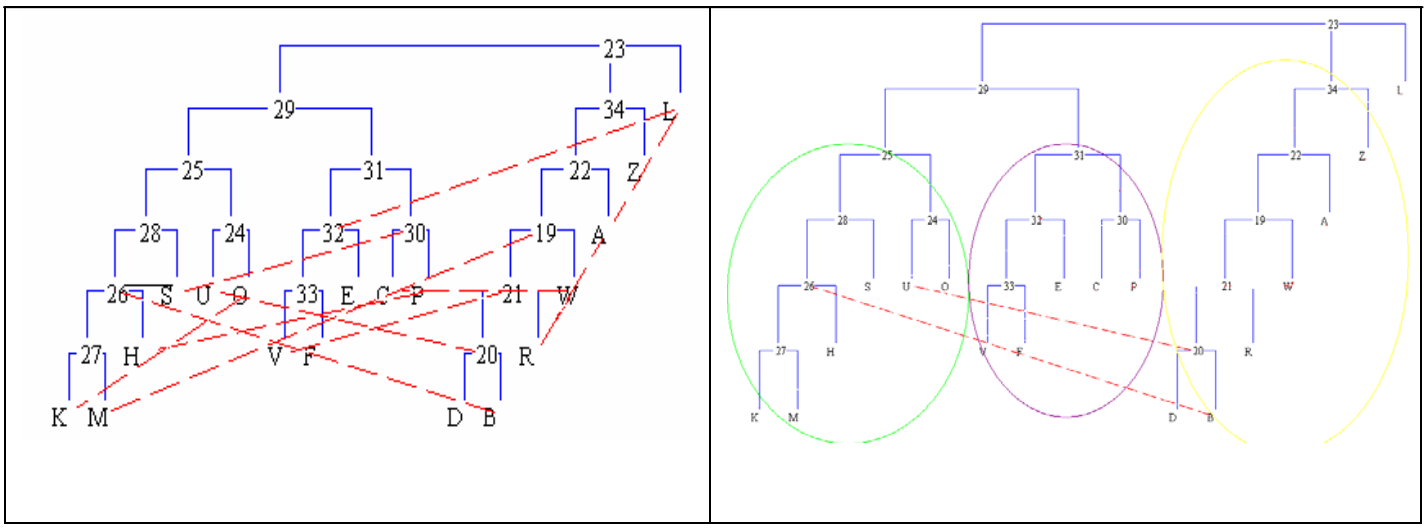

Figure 8 : Stemma et contamination par TRex

De nombreuses contaminations sont répertoriées sur le stemma obtenu (cf. Fig. 8) si l'on laisse la procédure aller jusqu'à son terme. Il vaut mieux l'interrompre avant afin de ne visualiser que les principales contaminations que l'on détermine avec l'indice des manuscrits associés. Seul l'éditeur est apte à déterminer le nombre de contaminations à détecter.

En ne conservant que les contaminations en accord avec les plus fortes valeurs de l'indice de chaque manuscrit (ici ABUZ), on obtient un graphe plus lisible. Deux contaminations sont alors plus présentes la B-26 et la U-28 qui correspondent sans doute à une contamination véritable de la sous tradition engendrée par le 25 (en vert et dont les seuls représentants dans la figure 1 sont le M et le H) avec la tradition engendrée par le 34 (en jaune).

La contamination entre A et C détectée par Danuta Shanzer n'est pas évidente sur notre poème, peut être peut-on l'associer à celle entre W et C ?

6. Conclusion

Le problème de détection de l'hybridation textuelle est loin d'être résolu par l'informatique uniquement. En effet les méthodes développées actuellement ne sont pas en mesure d'identifier clairement les contaminations, tout juste de les conforter.

Certains points restent encore à examiner :

- L'utilisation des treillis de Galois pour l'analyse des variantes (cf. Marc Le Pouliquen (2007)).
- L'utilisation des relations d'intermédiarité pour détecter les cycles dans les graphes. (cf. Marc Le Pouliquen et Barthélemy (2007)).
- L'utilisation de plusieurs corpus afin de comparer les résultats.

Finalement, le logiciel d'aide à l'édition critique doit permettre à l'éditeur de donner son avis d'expert tout au long du processus. C'est grâce à cette interaction que le programme pourra devenir utile à l'éditeur.

Références

Carlson S.C. (2004). *The Origin(s) of the 'Caesarean' Text*. SBL.

Dickerman A.W. (1998). Generalizing Phylogenetic Parsimony from the Tree to the Forest. *Systematic Biology* 47: 414-426.

Guillaumin J.-B. (1986). Corpus sur http://www.eleves.ens.fr/home/jguillau/index.html

Archer G. (1978). *Introduction à L'Ancien Testament*. Emmaüs, pages 52-56.

Le Pouliquen M. et Barthélemy J.-P. (2007). Construction d'arbres à partir de relations d'intermédiarité, Application au stemma codicum. *XIVe Rencontre de la Société francophone de classification, SFC2007*.

Le Pouliquen M. (2007). Using Lattices for Reconstructing Stemma. *Fifth International Conference on Concept Lattices and Their Applications, CLA*.

Makarenkov V., Legendre P. (2000). Improving The Additive Tree Representation of a Dissimilarity Matrix using Reticulations. In Kiers H.A.L., Rasson J.-P., Groenen P.J.F. and Schader M. (Eds), *Data Analysis Classification and Related Methods*, Springer, 35-40.

Makarenkov V. (2001). T-Rex: Reconstructing and Visualizing Phylogenetic Trees and Reticulation Networks, *Bioinformatics* 17, 664-668.

Mass P. (1957). Textkritik (Leipzig, 1927); 2nd ed. (idem, 1949); 3rd ed. (idem, 1957). Textual Criticism*, tr. Barbara Flower* (Oxford, 1958).

Quentin H. (1926). *Essais de critique textuelle*, Picard.

Shanzer D. (1986). *A Philosophical and Literary Commentary on Martianus Capella's De nuptiis philologiae et Mercurii.* Univ. of Calif. Press, book 1.

Saitou N. and M. Nei. (1987). The Neighbor-Joining Method: a new Method for Reconstructing Phylogenetic trees. *Mol Biol Evol* 4, 406-425.

Spencer M., Wachtel K., and Howe C.J. (2004). Representing Multiple Pathways of Textual Flow in the Greek Manuscripts of the Letter of James Using Reduced Median Networks. *Computers and the Humanities* 38: 1-14.

Sattah S. et Tversky A. (1977). Additive Similarity Trees. *Psychometrika* 42, 319-345.