

# Stemma codicum et relation d'intermédiation, utilisation de la méthode de Don Quentin

Marc Le Pouliquen <sup>1</sup>, Marc Csernel <sup>2</sup>

<sup>1</sup> EP Telecom Bretagne, Labsticc UMR 3192, BP 832, 29285 Brest Cedex - France

<sup>2</sup> AXIS Inria-Rocquencourt, BP-105- 78180 Le Chesnay Cedex - France

## Résumé

Dans cet article, nous allons modéliser la relation ternaire d'intermédiation dans le cadre de l'édition critique de texte. L'éditeur doit essayer de reconstituer au mieux, à partir des manuscrits préservés, le manuscrit original tel que l'auteur l'a écrit. Une des méthodes utilisées consiste à élaborer un arbre de filiation des manuscrits restants, appelé le « *stemma codicum* ». Nous proposons, comme le suggère Don Quentin, de construire cet arbre à partir des relations d'intermédiation entre les manuscrits : Un manuscrit B est entre les manuscrits A et C si le manuscrit C a été copié à partir du manuscrit B et que lui-même a été copié sur le manuscrit A. Avec un peu de dénombrement, nous nous apercevons que le nombre de relations d'intermédiation peut vite devenir trop important pour pouvoir envisager leur construction et leur comparaison manuellement. Cela fait sans doute partie des obstacles qui ont gênés les éditeurs dans la formalisation du *stemma* et son utilisation. Profitant des possibilités de calculs actuels des ordinateurs, la méthode de Don Quentin peut être modifiée et adaptée pour tenter de reconstruire le *stemma*. Nous observons enfin que ces relations permettent d'obtenir une sériation des manuscrits dont l'analyse oriente l'éditeur vers une version proche du manuscrit original.

## Abstract

The goal of this paper is to model the ternary betweenness relation within the framework of critical edition of manuscripts. The editor tries to reconstruct, as well as possible, the original manuscript using a corpus of various preserved manuscripts. This corpus is made up from manuscripts which have been copied one from the other. To achieve such a goal, it is interesting to draw up a family of filiations trees called "stemma codicum". As suggested by Don Quentin, we propose to build this tree using the betweenness relation within the manuscripts. Manuscript B is between manuscripts A and C, if manuscript C was copied from manuscript B which itself was copied from A. We notice that the number of betweenness relations grows rather quickly with the number of manuscripts to be compared. It is usually too large to allow hand made comparison and construction. Thanks to the capabilities of current computers calculations, the method of Don Quentin can be modified and adapted to build the *stemma* by computers. We finally observe that these relations provide a seriation of the manuscripts set which can direct an editor towards a text which is rather close from the original one.

**Keywords :** *stemma codicum*, filiation of manuscripts, betweenness, tree, seriation

## 1. Introduction

L'édition critique de manuscrits anciens reconstitue, au mieux, à partir des différents manuscrits conservés, l'œuvre telle que l'auteur l'a voulue. Pour retrouver le texte original, l'éditeur confronte les différentes versions du texte qui nous ont été transmises par les copistes. En effet, l'auteur a pu écrire différentes versions de son texte, les copistes ont fait des erreurs (oubli de mot, saut de ligne, amélioration...), et les évolutions du temps et de l'espace (trous

dans le papier, paragraphe illisible, évolution de la langue...) multiplient exponentiellement les dissemblances entre les versions conservées.

La réalisation de l'édition critique classique se fait en plusieurs étapes dont l'une consiste à établir un arbre de filiation des manuscrits pour savoir lequel a été copié sur l'autre et de détecter les chaînons manquants : c'est l'établissement du *stemma codicum*. Le *stemma codicum* (cf. Fig. 1) se présente donc sous la forme d'un arbre généalogique des manuscrits représentant la filiation entre les différentes versions du même texte.

Cet article s'intéresse à la modélisation mathématique de l'intermédierité entre manuscrits permettant de construire et de préciser le *stemma codicum*. Les méthodes phylogénétiques ou les treillis précédemment utilisés (cf. Le Pouliquen, 2007 ; Le Pouliquen et al., 2006) ne permettent pas de détecter les manuscrits ancêtres communs qui apparaîtraient comme des nœuds intérieurs dans les graphes ou les arbres servant de modèles au *stemma codicum*. Nous allons utiliser la méthode de Don Quentin (Quentin, 1926) qui propose de reconstituer des petites chaînes de trois manuscrits dont l'un est l'intermédiaire des deux autres puis d'assembler ces petites chaînes afin d'inférer le *stemma* complet.

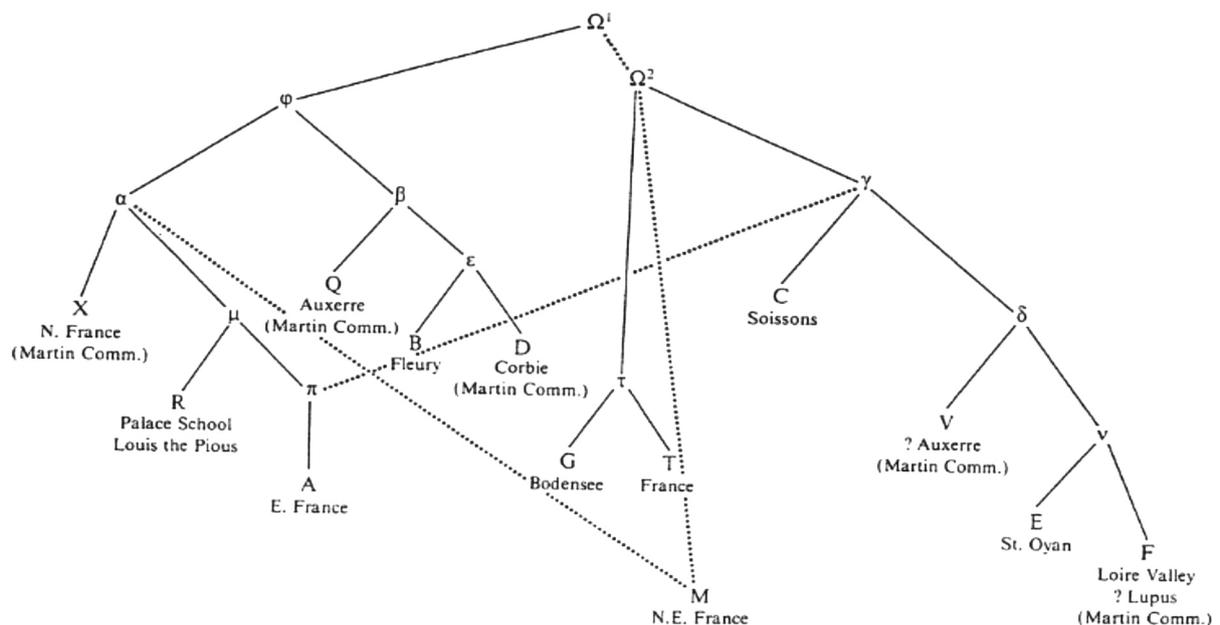


Figure 1 : Exemple de *Stemma codicum*<sup>1</sup>. Chaque lettre correspond à un manuscrit, les lettres grecques indiquent les manuscrits perdus ou supposés, les pointillés représentent les contaminations<sup>2</sup>

En premier, nous allons tenter dans cet article de modéliser cette relation dans le cadre de l'intermédierité de textes. Sur le *stemma* de Fig. 1, nous constatons que le manuscrit *V* est intermédiaire entre les manuscrits  $\delta$  et *F*, c.-à.-d. que le manuscrit *F* a été copié à partir du manuscrit  $\delta$  qui, lui-même, a été copié sur le manuscrit *V*. C'est cette notion d'intermédierité par la copie que nous souhaitons modéliser.

Dans la deuxième partie, nous réalisons les premières expérimentations sur deux corpus réels, la *Kāśhikāvṛitti* ou *Glose de Bénarès*, un corpus de manuscrits sanskrits et *De Nuptiis*, un corpus

<sup>1</sup> *Stemma* établi par Danuta Shanzer (1986).

<sup>2</sup> Contamination : copie d'un manuscrit sur plusieurs modèles. (syn.: corruption, hybridation).

latin déjà étudié par des philologues. Constatant que la modélisation ne permet pas de retrouver d'intermédiaires, nous définissons des relations de presque-intermédiarité. Ces relations, classées selon un indice de pertinence, nous donnent alors des informations philologiques sur les corpus.

Afin de déterminer le manuscrit original <sup>3</sup>, nous nous intéressons, dans un troisième temps, aux possibilités d'obtenir une sériation des manuscrits à partir des relations de presque-intermédiarité obtenues. Les relations obtenues *n'étant pas toutes compatibles*, l'examen de toutes les possibilités de sériation pour déterminer la moins mauvaise étant trop long (en factoriel du nombre de manuscrits), nous proposons une solution approchée dont le résultat est soumis à l'expertise de l'*éditeur*.

Dans la dernière partie, nous présentons les conclusions propres aux deux traditions textuelles que nous avons utilisées. Nous regardons ensuite l'apport de la méthodologie pour l'éditeur ainsi que les perspectives envisagées.

## 2. Différentes modélisations de l'intermédiarité au niveau des manuscrits

### 2.1. La méthode de Don Quentin

Parmi les méthodes classiques de construction du *stemma codicum*, celle de Don Quentin propose de déterminer des chaînes de trois manuscrits dont l'un est intermédiaire entre les deux autres au sens de la copie, puis d'assembler ces chaînes afin de reconstruire le *stemma* complet.

La méthode de Don Quentin s'attache à reconstituer l'enchaînement des manuscrits au moyen d'une comparaison trois par trois des manuscrits, qui permet de déterminer les intermédiaires. Pour savoir si un triplet  $(A, B, C)$  de manuscrits est tel que  $B$  est intermédiaire entre  $A$  et  $C$ , Don Quentin impose deux conditions :

Pour que  $B$  soit intermédiaire entre  $A$  et  $C$ , il suffit que :

- (i)  $A$  et  $C$  s'accordent tour à tour avec  $B$  sur les variantes ;
- (ii)  $A$  et  $C$  ne s'accordent jamais contre  $B$ .

Exemple: Soient les 3 phrases suivantes correspondants aux trois mêmes phrases de différents manuscrits copiés les uns sur les autres.

$A$  = « Voici une phrase courte inventée pour l'exemple »

$B$  = « Voici une phrase inventée pour cet exemple »

$C$  = « Voici une phrase créée pour cet exemple »

Nous avons ici 3 lieux variants : *courte*, *inventée/créée* et *l'/cet* correspondant à 6 variantes et résumés dans Tab. 1.

Pour les lieux variants 1 et 3, nous voyons que  $B$  s'accorde avec  $C$  ; pour le lieu variant 2,  $B$  s'accorde avec  $A$  donc la condition (i) est remplie. La condition (ii) est elle aussi validée car  $A$  et  $C$  ne s'accordent jamais contre  $B$ .  $B$  est donc bien intermédiaire entre  $A$  et  $C$ . Ces deux conditions correspondent bien à la notion d'intermédiaire au sens de la copie qui, ici, peut être vu comme : Le copiste de  $B$  a omis *courte* et a modifié *l'* en *cet* et le copiste de  $C$  a remplacé *inventée* par *créée*.

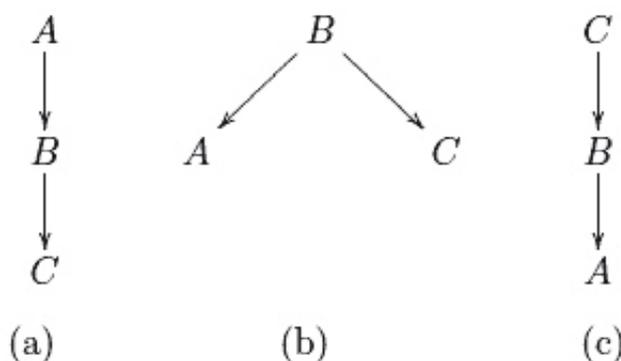
<sup>3</sup> Le manuscrit original est la version écrite de la main même de l'auteur, il est en général perdu (syn. : le manuscrit autographe ou l'archétype).

N° du lieu variant	Variantes	Phrases associées	Variantes	Phrases associées
1	courte	$A$	$\{\emptyset\}$	$B, C$
2	inventée	$A, B$	créée	$C$
3	$L'$	$A$	cet	$B, C$

Tableau 1 : Table de collation des trois phrases

Il est en revanche peu probable que  $C$  soit l'intermédiaire, car le copiste de  $C$  a supprimé *inventée* qui serait réintroduit par le copiste suivant. Dans ce cas,  $A$  et  $B$  s'accordent sur la variante *inventée* contre  $C$  qui contient lui, la variante *créée*, donc le (ii) n'est pas valide et  $C$  n'est pas considéré comme intermédiaire entre  $A$  et  $B$ .

Nous pouvons cependant remarquer que la méthode de Don Quentin ne permet pas de détecter le sens de la copie. Trois filiations différentes peuvent être construites conformément à *B est intermédiaire entre A et C* (cf. Fig. 2).

Figure 2 : Trois filiations différentes si «  $B$  est intermédiaire entre  $A$  et  $C$  »

Nous allons maintenant utiliser différentes définitions mathématiques de l'intermédialité pour tenter de modéliser l'intermédialité au sens de la copie entre manuscrits.

**Notation 1** À partir de maintenant, pour exprimer le fait que  $B$  est intermédiaire entre  $A$  et  $C$ , on se contentera d'écrire  $(A, B, C)$ .

## 2.2. Intermédialité en géométrie

Des nombreuses caractérisations géométriques de l'intermédialité, celle qui nous intéresse est celle introduite par Menger (cf. Menger, 1928) sous le nom de relation métrique d'intermédialité de la façon suivante :

**Définition 1** Une relation ternaire  $\{a, b, c\}$  sur un ensemble  $E$  est dite **relation métrique d'intermédialité**  $(a, b, c)$  s'il existe une métrique  $d$  sur  $E$  telle que :

$$(a, b, c) \Leftrightarrow d(a, b) + d(b, c) = d(a, c) \quad (1)$$

Il faut désormais définir une métrique qui permet de comparer les manuscrits de telle sorte que l'intermédialité au sens de la copie corresponde à celle de la métrique. Pour cela, reprenons l'exemple précédent.

Soit la distance  $d$  défini par le nombre de mots insérés, supprimés ou substitués entre 2 textes alignés (une sorte de distance d'édition ; cf. Levenshtein, 1966) au niveau des mots. Si nous réalisons les calculs sur l'exemple précédent, nous obtenons alors Tab. 2.

A	Voici une phrase	courte	inventée	pour	L'	exemple
B	Voici une phrase		inventée	pour	cet	exemple
C	Voici une phrase		créée	pour	cet	Exemple
$d(A,B)=2$		$d(B,C)=1$		$d(A,C)=3$		

Tableau 2 : Alignement des trois phrases pour compter le nombre de variantes

Nous constatons en effet que  $(A, B, C)$  puisque  $3 = d(A, C) = d(A, B) + d(B, C) = 2 + 1$ . Nous n'avons pas, en revanche,  $(A, C, B)$  puisque  $3 = d(A, C) = d(A, B) + d(B, C) = 2 + 1$ .

Pour les applications, nous utiliserons l'indice

$$I_M = \frac{d(A, B) + d(B, C) - d(A, C)}{d(A, C)} \quad (2)$$

qui est nul si  $(A, B, C)$ . Il permet aussi de déterminer si un manuscrit est « plus ou moins intermédiaire » et de comparer les indices entre eux.

### 2.3. Intermédierité définie par une relation d'ordre

Nous pouvons aussi présenter une relation ternaire avec une relation d'ordre.

**Définition 2** Une relation ternaire  $(a, b, c)$  est appelée **intermédierité d'ordre** si et seulement si  $a \leq b \leq c$  ou  $c \leq b \leq a$ .

Si l'on désire utiliser cette définition dans le cadre des manuscrits, il convient de définir une relation d'ordre entre les manuscrits au sens de :

$$M1 \leq M2 \Leftrightarrow M1 \text{ a été copié sur } M2 \text{ ou est égal à } M2.$$

Malheureusement, étant donné deux manuscrits, nous n'avons aucune façon de savoir si l'un a été copié sur l'autre ou vice versa à l'examen du texte. Nous pouvons nous douter que deux manuscrits trop dissemblables ne sont pas des copies l'un de l'autre. Mais les copistes étant plus ou moins précis, les manuscrits copiés sont plus ou moins proches et leurs proximités ne traduisent pas forcément une filiation directe. Il n'y a donc pas de modélisation envisageable de la relation d'ordre et de l'intermédierité qu'elle induirait.

### 2.4. Intermédierité en théorie des ensembles

Nous allons nous intéresser à la relation d'intermédierité introduite par Restle (1959). Cette relation définit la notion d'intermédierité au niveau des ensembles.

**Définition 3** Soient trois ensembles  $A, B$  et  $C$ . On considère que  $B$  est intermédiaire entre  $A$  et  $C$  si et seulement si :

$$A \cap \bar{B} \cap C = \emptyset$$

$$\bar{A} \cap B \cap \bar{C} = \emptyset$$

Finalement, la modélisation peut se schématiser par Fig. 3.

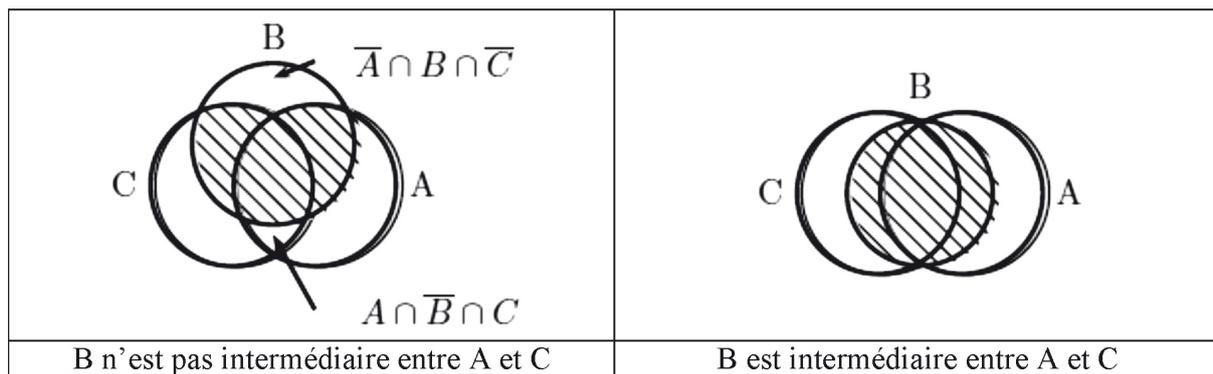


Figure 3 : Intermédiation au sens de Restle

Reprenons notre exemple précédent. Pour adapter la Définition 3 aux manuscrits, nous associons à chaque manuscrit, l'ensemble de ses variantes. L'ensemble A est constitué des variantes  $\{courte, inventée, l'\}$ , B de  $\{inventée, cet\}$  et C de  $\{créée, cet\}$ . Nous pouvons alors le modéliser par Fig. 4.

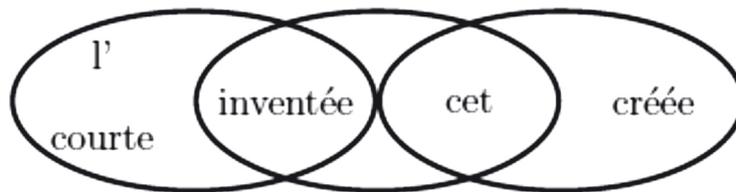


Figure 4 : Visualisation de l'ensemble intermédiaire de l'exemple

Nous constatons que les intersections sont vides ce qui correspond à nouveau au fait que B est entre A et C.

Pour la suite, nous utiliserons l'indice

$$I_E = \frac{\text{Card}(A \cap \bar{B} \cap C) + \text{Card}(\bar{A} \cap B \cap \bar{C})}{\text{Card}(A)} \quad (3)$$

qui est nul si  $(A, B, C)$ . Comme avec  $I_M$ , nous pouvons comparer les indices entre eux grâce à une pondération par le cardinal de A et nous pouvons savoir si un manuscrit est « plus ou moins intermédiaire ».

### 2.5. Intermédiation définie directement à partir des conditions de Don Quentin

Les indices créés à partir des définitions précédentes semblent cohérents avec la notion d'intermédiation recherchée. Cependant, ils ne correspondent pas exactement aux deux conditions (i) et (ii) (cf. § 2.1.) établies par Don Quentin. Nous voulons aussi construire un indice directement à partir des conditions de Don Quentin. L'indice doit être nul en cas d'intermédiation. Plus l'indice est important, plus le manuscrit est éloigné d'un manuscrit intermédiaire hypothétique. On peut ainsi déterminer si un manuscrit est « plus ou moins

<sup>4</sup>  $\text{Card}()$  correspond au *cardinal* d'un ensemble qui est le nombre d'éléments de cet ensemble.

intermédiaire». Enfin, on veut pouvoir comparer les indices obtenus selon les triplets sans être influencé par la taille des manuscrits.

Supposons que  $B$  soit intermédiaire entre  $A$  et  $C$ .

- Soit  $n_1$  le nombre de variantes de  $B$  qui n'appartiennent ni à  $A$ , ni à  $C$  et  $n$  le nombre de lieux variants de  $B$ . Posons  $I_{Q1} = \frac{n_1}{n}$ . Alors  $I_{Q1} \in [0,1]$  et si  $I_{Q1} = 0$  la condition (i) de Don Quentin est satisfaite.
- Soit  $n_2$  le nombre de variantes communes à  $A$  et  $C$  qui n'appartiennent pas à  $B$ . Posons  $I_{Q2} = \frac{n_2}{n}$ . Alors  $I_{Q2} \in [0,1]$  et si  $I_{Q2} = 0$  la condition (ii) de Don Quentin est satisfaite.
- En prenant,  $I_Q = 0.8 * I_{Q2} + 0.2 * I_{Q1}$  l'indice est nul si les deux conditions de Don Quentin sont vérifiées. Nous pondérons davantage (80%) la condition (ii) car elle semble plus indispensable à l'intermédialité que la condition (i).

En reprenant l'exemple précédent, nous obtenons quand  $B$  est intermédiaire entre  $A$  et  $C$  :

$$\left. \begin{array}{l} n = 3, n_1 = 0 \text{ donc } I_{Q1} = 0 \\ n = 3, n_2 = 0 \text{ donc } I_{Q2} = 0 \end{array} \right\} \Rightarrow I_Q = 0. \text{ Donc } B \text{ est intermédiaire entre } A \text{ et } C$$

Si on examine le cas où  $C$  est intermédiaire entre  $A$  et  $B$  :

$$\left. \begin{array}{l} n = 3, n_1 = 1 \text{ donc } I_{Q1} = \frac{1}{3} \\ n = 3, n_2 = 1 \text{ donc } I_{Q2} = \frac{1}{3} \end{array} \right\} \Rightarrow I_Q = \frac{1}{3}. \text{ Donc } C \text{ n'est pas intermédiaire entre } A \text{ et } B$$

### 3. Première expérimentation

#### 3.1. Présentation des corpus

Notre travail s'effectue sur deux corpus différents : Un corpus sanskrit d'une cinquantaine de manuscrits jamais édité actuellement, la *Kāśhikāvṛitti* et un corpus latin, *De Nuptiis*, déjà étudié par des philologues, et qui nous permettra de tester et de vérifier la validité des méthodes proposées.

Le corpus sanskrit est constitué de manuscrits d'un célèbre traité de grammaire nommé *Kāśhikāvṛitti* ou *Glose de Bénarès*. La *Glose de Bénarès* est le plus répandu et le plus pédagogique des commentaires de *Pāṇini*. Elle date probablement du VII<sup>e</sup> siècle de notre ère et les manuscrits sont écrits en utilisant différentes écritures, parmi lesquelles la devanagāri, le *tēlugu* ou l'écriture tamoule. Les manuscrits sont nommés de la façon suivante : *bh8*, *bh* est une abréviation du nom de la bibliothèque et 8 désigne le 8<sup>ième</sup> manuscrit.

Pour nos premières expérimentations, la tradition textuelle de la *Kāśhikāvṛitti* n'est pas toujours adaptée :

- c'est une tradition importante en nombre et quant à la longueur des manuscrits ;
- on ne dispose pas d'information philologique ou d'étude de l'histoire du texte ;
- les problématiques du sanskrit compliquent la tâche.

C'est pourquoi, nous avons décidé d'employer une autre tradition textuelle plus simple. Nous avons utilisé le corpus engendré par le poème qui ouvre le livre IX du *De Nuptiis Philologiae et Mercurii* de Martianus Capella (V<sup>ième</sup> siècle ap. J.-C.). Ce corpus, collecté par Jean-Baptiste

Guillaumin <sup>5</sup>, a l'avantage de disposer d'un *stemma*, certes incomplet, élaboré par D. Shanzer (Fig. 1) et est constitué de 18 poèmes désignés par les lettres suivantes :

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

### 3.2. Résultat des premiers tests sur le corpus De Nuptiis

Les résultats obtenus pour l'intermédiarité caractérisée par l'indice  $I_Q$  ne donnent aucune relation d'intermédiarité. En effet, nous disposons d'une table de collation qui contient 18 manuscrits et 234 variantes. Cela nous donne donc 2448 triplets ( $18 \cdot 17 \cdot 16 / 2$ ) ou relations d'intermédiarité potentielles sans tenir compte des symétriques. Nous avons calculé les indices dont un extrait est visible sur Tab. 3.

Triplets	SFV	MFV	KFV	UFV	LFV	PFV	RFV	HFV	SEV	MEV	...
$I_Q$	0,086	0,087	0,087	0,087	0,087	0,087	0,088	0,088	0,090	0,090	...

Tableau 3: Tableau des 10 indices d'intermédiarité les plus faibles

Comme nous n'obtenons aucune relation d'intermédiarité stricte, nous avons examiné les relations obtenues pour les indices les plus faibles. Les manuscrits *E* et *F* sont présents dans les 70 premiers triplets ayant les indices les plus faibles. Nous pouvons alors nous poser des questions sur le rôle d'intermédiaire des manuscrits *E* et *F* entre la sous-tradition textuelle *CEFV* et le reste du corpus. Cette hypothèse semble en contradiction avec le *stemma codicum* de D. Shantzer (cf. Fig. 1) qui utilisait un corpus incomplet. Il reste donc à l'appréciation de l'éditeur d'examiner cette hypothèse.

### 3.3. Résultat des premiers tests sur le corpus Kāshikāvṛitti

Pour ce corpus plus important (49 manuscrits et 2933 variantes), nous obtenons 55274 triplets ( $49 \cdot 48 \cdot 47 / 2$ ). Dans le cas de l'indice  $I_M$ , nous n'obtenons pas non plus de relations d'intermédiarité comme nous pouvons le voir sur Tab. 4.

Triplets	tri1-st1-tri21	tri2-st1-tri45	tri21-st1-tri32	tri1-st1-tri32	tri1-st1-tri45	...
$I_M$	0,23	0,24	0,24	0,25	0,26	...

Tableau 4 : Tableau des 5 premiers triplets d'indice intermédiarité les plus faibles

En revanche, l'étude des indices les plus faibles peut vraisemblablement apporter quelques informations philologiques que l'éditeur doit confirmer. Le cas du manuscrit *st1* semble être intéressant. Écrit en *nandinagārii*, il apparaît comme un manuscrit central dans la sous-tradition textuelle des manuscrits en *malayālam*. Nous savons que le manuscrit *st1* a longtemps été catalogué, par erreur, comme un manuscrit en écriture *grantha*, écriture ayant évolué vers le *malayālam*. Il y a peut-être une proximité à étudier ce qui expliquerait le résultat ?

### 3.4. Conclusion sur les premiers tests

Nous avons utilisé les différents indices d'intermédiarité sur les deux corpus étudiés et nous constatons que nos modélisations ne permettent pas de retrouver les intermédiaires (s'ils existent). Cela n'est pas étonnant puisque c'est le reproche qui était fait à Don Quentin et

<sup>5</sup> Le corpus a été collecté par J.-B. Guillaumin en 1986.

à sa méthode par de nombreux philologues de ne pas trouver assez d'intermédiaires pour reconstruire le *stemma*.

Cependant, grâce aux indices les plus faibles, nous détectons plus globalement, si un manuscrit est intermédiaire entre deux branches de la tradition textuelle ou s'il a un rôle de manuscrit central dans une sous-tradition textuelle. Les hypothèses faites sur les indices les plus faibles doivent être soumise à l'éditeur pour une validation.

Comme souvent dans le passage des définitions théoriques aux corpus réels, l'expérimentation montre qu'il faut assouplir les définitions. Il semble alors intéressant de définir la notion de presque-intermédiarité :

**Définition 4** *A partir d'un indice d'intermédiarité  $I$  ( $I = I_M, I_E$  ou  $I_Q$ ), nous définissons la relation de presque-intermédiarité  $[A, B, C]$  sur un corpus  $C$  par :*

Pour que  $B$  soit presque-intermédiaire entre  $A$  et  $C$ , il suffit que :

- (i)  $I(A, B, C) \leq I(B, A, C)$  et  $I(A, B, C) \leq I(A, C, B)$
- (ii) Il existe un indice seuil  $I_S$  tel que  $I(A, B, C) \leq I_S$

Il reste à définir un indice seuil au-delà duquel l'éditeur considère que l'on ne trouve plus de relations d'intermédiarité significatives. Pour les applications, nous utiliserons un indice seuil permettant de conserver 20% des relations potentielles.

## 4. Application à la sériation

Une des problématiques de l'édition critique est la détermination du manuscrit original. Nous espérons que la sériation des manuscrits basée sur les relations de presque-intermédiarité éclaire la tradition textuelle étudiée et permet de détecter, le cas échéant, le manuscrit original.

### 4.1. Introduction

La sériation archéologique permet d'ordonner, généralement par rapport au temps, un ensemble de vestiges dont on dresse préalablement un inventaire sous la forme d'une table binaire (présence/absence) de données ou d'objets représentatifs des vestiges au cours du temps. On suppose naturellement qu'un vestige est d'autant plus ancien qu'il est constitué des données ou des objets les plus vieux. Il s'agit alors, pour l'expert, de retrouver par association des données ou des objets caractéristiques, l'ordre chronologique des vestiges.

Deux caractéristiques majeures sont à noter dans notre application :

- L'ensemble, sur lequel nous allons travailler, n'est pas ensemble de vestiges, mais un ensemble de manuscrits pour lequel on dispose d'information interne aux manuscrits (table de collation), mais aussi d'information externe (information codicologique et paléographique).
- La méthodologie est aussi originale puisqu'elle utilise l'intermédiarité (ou la presque-intermédiarité) pour obtenir la sériation.

Classiquement, la sériation de données archéologiques est obtenue en utilisant une table de contingence entre les vestiges et leurs critères descriptifs et des méthodes comme l'analyse factorielle de Benzécri (1973) reprise en archéologie par DjinDjian (DjinDjian and Leredde, 1980) ou la méthode de classification par bloc utilisée par Bertin (1969) et formalisée par Lerman (1972).

#### 4.2. Méthode du Tri Bulle

L'algorithme du tri bulle consiste à regarder les différentes valeurs adjacentes d'une liste, et à les permuter si la première est supérieure à la seconde. On commence par les deux premières valeurs de la liste et ensuite, on compare et permute éventuellement les valeurs 2 et 3, 3 et 4 jusqu'à la fin de la liste. Cette première série de comparaisons effectuées, le dernier élément de la liste est le plus grand. Il faut donc réitérer l'opération pour les (n-1) premières valeurs. L'algorithme se termine quand il n'y a plus de permutations possibles.

Nous disposons d'une liste de manuscrits ainsi que d'une relation ternaire (de presque-intermédierité) entre nos manuscrits. Nous avons adapté le tri bulle de telle sorte que, pour chaque triplet de manuscrits consécutifs, le manuscrit « le plus intermédiaire » se trouve bien au milieu. Les trois manuscrits consécutifs sont alors permutés s'il y a nécessité et parmi les 6 permutations possibles, nous choisissons celle dont l'indice est le plus faible et qui n'annule pas la permutation précédente. L'algorithme s'arrête lorsque, dans tous les triplets consécutifs de manuscrits, le manuscrit central est presque-intermédiaire entre les deux autres.

Si nous testons l'algorithme sur des nombres, la convergence est rapide et les nombres s'ordonnent naturellement. Sur les corpus de manuscrits, les relations de presque-intermédierité peuvent être contradictoires : On peut avoir  $(A, B, C)$ ,  $(A, D, C)$  et  $(B, C, D)$  alors que  $(A, B, C)$  et  $(A, D, C)$  entraînent forcément  $(B, D, C)$  ou  $(D, B, C)$ . En effet, pour établir une sériation, l'ensemble des relations ternaires doit respecter des conditions établies par Defays (1979). Ces relations contradictoires (qui ne respectent pas les conditions de Defays) se visualisent sur le tri bulle par des cycles infinis de permutations que l'on stoppe en choisissant celle qui minimise un score global de la sériation (somme des indices de tous les triplets de manuscrits consécutifs de la sériation proposée). Nous considérons que les relations contradictoires qui ne sont pas retenues sont moins importantes que celles que nous conservons.

#### 4.3. Application au corpus De Nuptiis

En lançant l'algorithme sur le corpus *De Nuptiis*, nous obtenons les résultats visibles sur Tab. 5. Ne disposant pas de datation précise de l'ensemble du corpus, nous avons utilisé, pour analyser les résultats, une classification arborée (Fig. 5) réalisée à partir de la distance  $d$  utilisée au § 2.2 et de l'algorithme Neighbor-Joining (Saitou and Nei, 1987). Deux informations philologiques semblent pertinentes. Nous observons que le manuscrit L est bien intermédiaire entre deux sous-traditions textuelles dans la sériation comme dans l'arbre obtenu. De la même façon, nous retrouvons le triplet  $(F, E, V)$  dans nos deux représentations ; cela semble envisageable au vu du stemma codicum de D. Schanzer (Fig. 1).

#### 4.4. Application au corpus Kāshikāvṛitti

Le corpus étant plus important, la visualisation des résultats est difficile à interpréter. Pour les tests, nous avons donc réduit le corpus aux seuls manuscrits datés (au nombre de 15). Nous obtenons, selon l'ordre des manuscrits dans la liste initiale, des sériations différentes. Sur les deux sériations proposées sur Tab. 6, certains groupes de manuscrits restent stables, les autres manuscrits se positionnant par rapport à ces groupes.

Si nous regardons le résultat de la sériation par rapport à la datation, globalement, le résultat ne permet pas d'obtenir d'information correcte sur la chronologie des manuscrits.

On constate que pour certain triplet comme (hp1, bh8, G3) dans la première sériation, le manuscrit intermédiaire au niveau de la date n'est pas le même qu'avec les indices construits ce

qui n'est pas le cas dans la deuxième sériation. D'autres triplets comme (io4, io5, G2) dans la deuxième sériation ne sont pas correctement orientés au niveau des dates, bien qu'ils le soient au niveau des indices. Dans ce cas, la modélisation de l'intermédiarité ne correspond pas à la sériation temporelle.

B
A
P
D
R
W
L
C
H
K
F
E
V
Z
S
O
U
M

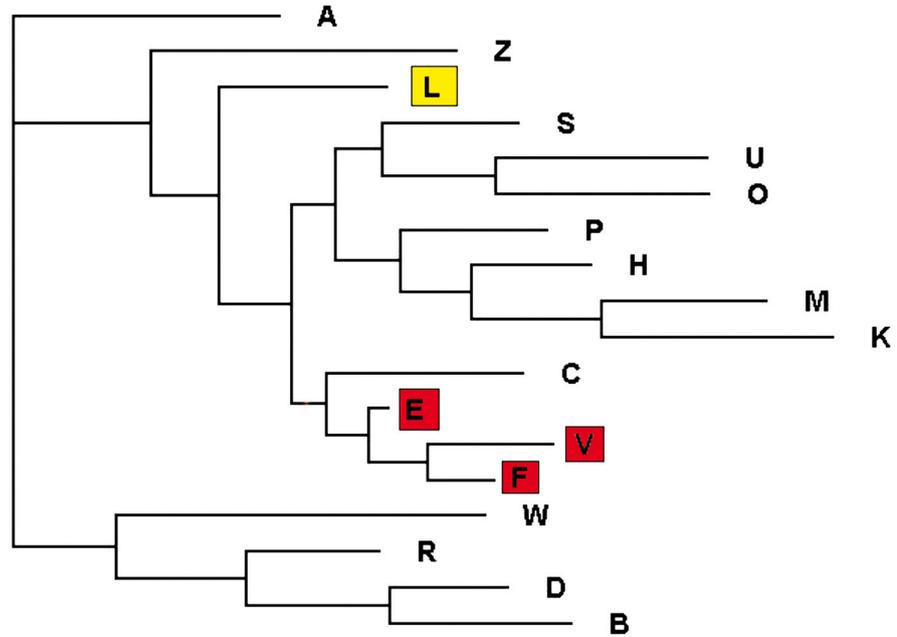


Tableau 5: Sériations sur le corpus De Nuptiis

Figure 5: Classification arborée obtenue sur le corpus De Nuptiis

jm6 1762	G1 1408
bh6 1486	ba2 1437
G1 1408	bh6 1486
hp1 1756	bh7 1707
bh8 1868	bh8 1868
G3 1809	G3 1809
bh7 1707	hp1 1756
io5 1632	jm6 1762
io2 17--	io7 18--
io4_17	io4 17--
io1 1863	io5 1632
G2 1870	G2 1870
io7 18--	io1 1863
ba2 1437	io2 17--
km2 old	km2 old

Tableau 6: Sériations obtenues sur le corpus réduit

## 5. Conclusion

Les méthodes d'intermédiarité constituent un outil intéressant dans la classification de tous les documents. Il est toujours avantageux de savoir qu'un objet s'intercale entre deux autres. Dans le cadre de documents textuels, le classement peut s'effectuer selon plusieurs axes : les thèmes, les auteurs, la grammaire, le style, les dates...

Dans tous ces cas, il faut définir l'intermédiarité pour positionner notre document parmi les autres. Dans l'édition critique et, plus précisément, dans la reconstitution de l'histoire d'un texte, nous avons tenté de modéliser la notion de « copie intermédiaire ». Nous constatons que cette modélisation stricte ne permet pas de retrouver les intermédiaires dans le cas d'un corpus réel. Nous pouvons alors rechercher des manuscrits presque-intermédiaires. Cette modélisation peut permettre de détecter si un manuscrit est intermédiaire entre deux branches de la tradition textuelle ou s'il a un rôle de manuscrit central dans une sous-tradition textuelle.

Pour la sériation, les résultats obtenus ne tiennent pas vraiment leur promesse. On est loin d'obtenir une sériation permettant de détecter le manuscrit original. En utilisant un classement global, certaines relations de presque-intermédiarité obtenues sont vraisemblablement trop relaxées. Il faudrait que l'éditeur puisse les ajuster de façon plus ponctuelle selon les manuscrits du triplet proposé pour ne pas les dénaturer. D'autre part, la détection des relations incompatibles par la méthode proposée peut permettre à l'éditeur d'affiner son choix et de proposer alors une sériation plus intéressante.

## Références

- Benzécri, J.-P. (1973). *Analyse des Données*. Tome 3. Paris : Dunod.
- Bertin J. (1969). Traitement graphique de l'information, *Annales*, 24e année, Paris, Armand Colin, vol. janvier février.
- Defays D. (1979). Tree representation of ternary relations. *Journal of Mathematical Psychology*, 19 : 208-218.
- DjinDjian F. and Leredde H. (1980) Traitement automatique des données en archéologie. *Les Dossiers de l'Archéologie*, Dijon, vol. 42 : 52-69.
- Le Pouliquen M. (2007). Using Lattices for Reconstructing Stemma. In *Actes de la Fifth International Conference on Concept Lattices and Their Applications, CLA*, CEUR Workshop Proceedings, CEUR WS.org, <http://www.lirmm.fr/cla07/>.
- Le Pouliquen M., Barthélemy J.P. and Bertrand P. (2006). Filiation de manuscrits sanskrits et arbres phylogénétiques. In *RNTI-C-2 Classification : points de vue croisés, 2008, Actes de la XIIIe rencontre de la Société Francophone de Classification*, <http://lita.sciences.univ-metz.fr/~sfc06/>.
- Lerman I.C. (1972). Analyse du phénomène de la sériation à partir d'un tableau d'incidence. *Mathématiques et Sciences Humaines*, vol. 38 : 39-57.
- Levenshtein V.I. (1966). Binary Codes capable of correcting deletions, insertions, and reversals. *Soviet Physics -Doklady*, 10, 8 : 707-710.
- Menger K. (1928). Untersuchungen über allgemeine Metrick. *Mathematische Annalen*, 100 : 75-163.
- Quentin H. (1926). Essais de critique textuelle. Paris : Picard.
- Restle F. (1959). A metric and an ordering on sets. *Psychometrika*, 24 : 207-220.
- Saitou N. and Nei M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology Evolution*, 4 : 406-425.
- Shanzer D. (1986). A Philosophical and Literary Commentary on Martianus Capella's *de Nuptiis Philologiae et Mercurii*. University of California Press.