

# Filiation de manuscrits sanskrits et arbres phylogénétiques

M. Le Pouliquen\* J.P. Barthélemy\*,\*\* P. Bertrand \*\*\*

\*Département LUSSI TAMCIC, UMR CNRS 2872

ENST Bretagne, BP 832, 29285 Brest Cedex

marc.lepouliquen@enst-bretagne.fr

\*\*CAMS, UMR CNRS 8557, Ecole des Hautes Etudes en Sciences Sociales

54 bvd Raspail, 75270, Paris cedex 06

\*\*\*Département LUSSI TAMCIC, UMR CNRS 2872

ENST Bretagne, 2 rue de la Châtaigneraie, CS 17607,35576 Cesson Sévigné Cedex

**Résumé.** La fabrication d'un stemma codicum est l'une des approches les plus rigoureuses de la critique textuelle. Elle exige la reconstruction de l'histoire du texte en classifiant le corpus pour décider si un groupe de manuscrits est engendré par un intermédiaire perdu. Pour classifier notre corpus, nous employons des méthodes de l'analyse textuelle informatisée et de la reconstruction phylogénétique afin d'établir un arbre de la filiation. Les techniques employées sont dédiées à un corpus de manuscrits sanskrits avec toutes les spécificités de cette langue.

## 1 Introduction

Dans le cadre de l'édition d'anciens manuscrits, un des problèmes consiste à trier les différentes versions du texte afin d'essayer de reconstituer le manuscrit original avec fidélité. L'analyse des différents manuscrits pour réaliser l'édition critique est un travail colossal et se fait en plusieurs étapes dont l'une consiste à établir un arbre de filiation de ces manuscrits pour savoir lequel a été copié sur l'autre et de détecter les chaînons manquants : c'est l'établissement du *stemma codicum*.

Le projet consiste à utiliser les méthodes de la phylogénétique et de l'analyse textuelle informatisée afin de proposer un arbre qui permette d'établir un premier classement automatique de manuscrits sanskrits. Nous présentons dans cet article, un rapide aperçu de la philologie.



FIG. 1 – *Fragment d'un manuscrit*

Puis, on s'intéresse aux manuscrits, aux particularités du sanskrit et aux différentes opérations préparatoires menées sur le corpus. On s'attache à la description des méthodes d'alignements afin de déterminer les dissimilarités voire les distances que l'on peut obtenir entre nos textes. On s'intéresse ensuite aux algorithmes d'inférence d'arbres utilisés en phylogénie afin de les adapter à notre problématique. Nous décrivons finalement les expérimentations réalisées sur plusieurs corpus avant de conclure.

## 2 Méthodes philologiques d'établissement du *stemma codicum*

Un texte qui a été copié des centaines de fois constitue ce que l'on appelle une tradition textuelle et tous les exemplaires qui nous sont parvenus sont appelés les *témoins* du texte.

Les témoins sont généralement différents. 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 témoins existants.

L'éditeur critique doit à partir de ce corpus reconstituer au mieux le manuscrit original encore appelé *archétype*.

Pour construire le texte critique à partir des témoins, on part du constat suivant, à savoir que toutes les copies qui contiennent, aux mêmes endroits, les mêmes fautes, ont été faites les unes sur les autres et donc dérivent toutes d'une copie où ces fautes existaient. Pour classer les témoins, on recourt donc à la méthode de la comparaison des fautes appelées *variantes* qui classées selon leur influence sur l'acte de copie permettent de dresser un arbre généalogique des manuscrits. Cette méthode dite « Lachmannienne » [Salemans (1999)] présente l'avantage de préparer le travail de l'édition critique car pour reconstituer le texte le plus proche de l'original, on évalue laquelle des variantes convient le mieux.

Une autre méthode historique est celle de Quentin (1926). Elle s'attache à reconstituer l'enchaînement des manuscrits au moyen de comparaison trois par trois. C'est donc la recherche des intermédiaires qui permettra de reconstituer le *stemma*. Un manuscrit A est intermédiaire entre B et C (cf. fig 2) si B et C s'accordent (au niveau des variantes) tour à tour avec A et qu'ils ne s'accordent jamais ensemble contre A.

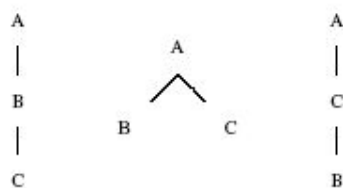


FIG. 2 – Trois *stemmas* différents où le manuscrit A est intermédiaire entre le mns. B et le C

Plusieurs *stemmas* différents peuvent être produits à partir d'une étude philologique selon le choix du manuscrit ancêtre commun. Bédier (1928) l'a découvert lors de l'étude des textes

du « Lai de l'Ombre ». Il a essayé de démontrer que des méthodes généalogiques, menant à beaucoup de résultats différents, sont sans valeur. Il a alors préconisé de s'en tenir aux variantes d'un témoin unique, celui qu'il a jugé le meilleur et de renoncer aux stemmas trop nombreux et donc trop approximatifs.

Après ce rapide panorama et pour mieux cerner la difficulté du problème, imaginons une oeuvre dont on possède 150 exemplaires non identiques et des milliers de variantes. Combien d'années de travail seraient nécessaires à un homme pour la réalisation d'une édition critique ?

L'importance des méthodes informatiques pour aider l'éditeur critique s'impose alors.

### 3 Les Manuscrits et leurs traitements

Le corpus de mon étude est constitué d'un certain nombre de manuscrits sanskrits relatif à la Glose de Bénarès. C'est le plus ancien commentaire complet sur le traité grammatical de Panini (5<sup>e</sup> siècle av Jésus Christ). Il existe environ 150 manuscrits de la Glose de Bénarès recensés en Inde et en Occident dans une dizaine d'écritures différentes. Les manuscrits comportent environ 800 pages, 245000 lignes et 1,5 millions de caractères. Pour le projet, 3 chapitres ont été sélectionnés d'environ 20 pages chacun. Pour réaliser cette classification, certaines caractéristiques du sanskrit sont importantes :

#### 3.1 Caractéristiques du sanskrit

L'alphabet sanskrit Devanagari ou plutôt le *syllabaire* est très complexe, à cause de la très grande quantité de sons à exprimer. Il s'agit cependant de lettres (syllabes et voyelles, plus signes de ponctuations) et non pas d'un système d'idéogrammes comme en Mandarin. Le sanskrit possède un alphabet de 46 lettres ce qui oblige lors de sa translittération à faire correspondre une lettre sanskrite à une séquence de lettres latines et complique par la même la comparaison des textes et la reconnaissance des mots. Contrairement au français où les lettres s'ajoutent simplement les unes aux autres, la syllabe sanskrite est le fruit d'une ligature, comme notre Æ. Le sanskrit est une langue qui peut s'écrire sans espace ni ponctuation entre les mots.

क	का	कि	की	कु	कू	कृ	कृ	क्ल
ka	kā	ki	kī	ku	kū	kr̥	kr̄	kl̥

FIG. 3 – Les ligatures de la consonne K avec voyelles et diphtongue et leur transcription

L'absence de blanc peut rendre les textes ambigus voire parfois difficilement compréhensibles.

Ajoutons que les initiales et les finales de chaque mot influent les unes sur les autres, apportant de nombreuses modifications ; d'où la première difficulté en lecture de reconnaître chaque terme et de restituer sa forme originale. Les sandhi (jonctions) sont une difficulté supplémentaire dans la reconnaissance des mots. Ils correspondent aux liaisons que l'on fait par oral dans les « z' autres » .

Pour en finir avec les spécificités du sanskrit, les spécialistes dont Filliozat (1941) parlent de l'orthographe « vicieuse mais traditionnelle des scribes » et d'autres spécialistes nous ont

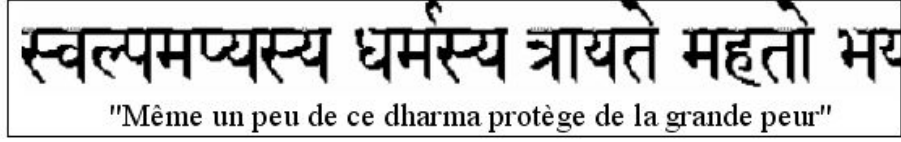


FIG. 4 – Exemple de sanskrit

montré que les mots difficiles à comprendre sont ceux où les manuscrits proposent une autre solution : preuve que loin de recopier sans comprendre, le scribe a remplacé le mot qu'il ne comprenait pas...

Pour pouvoir lancer des analyses textuelles sur ces manuscrits, nous leur avons fait subir un certain nombre d'opérations que nous allons détaillées.

### 3.2 La translittération

La translittération est l'opération consistant à transcrire lettre à lettre les mots d'une langue à une autre. On utilise alors un alphabet romain augmenté de signes diacritiques, qui permet de transcrire la Devanagari sous une forme facilement lisible et imprimable pour des européens (cf. fig 3)

Afin de procéder à la transcription, nous nous sommes contentés de suivre les conventions de transcription de Veltuis (1991) pour les caractères devanagari en utilisant le logiciel TeX/LaTeX. Cette procédure présente l'avantage de rendre les données utilisables sur n'importe quelle plate-forme sans conversion.

### 3.3 La lemmatisation et le Padapatha

La lemmatisation permet de comparer des textes au niveau des lemmes et non au niveau des formes. Ainsi si l'on veut compter le nombre d'occurrences de la forme « avoir » , on compte le nombre de fois où l'on retrouve le mot « avoir » dans le texte ; en revanche, si l'on veut connaître le nombre d'occurrences du lemme « avoir » , on compte le nombre de fois où l'on retrouve le verbe « avoir » dans le texte sous toutes ses formes (a, avons, ayez, eut...).

La lemmatisation est donc le processus qui consiste à réduire un mot à sa forme canonique, c'est-à-dire en supprimant les traits fonctionnels du mot : nombre, pluriel, marques de temps et de personne pour les verbes. Chaque mot est ainsi ramené à sa forme « canonique » appelée lemme.

Le *padapatha* est une version lemmatisé du texte réalisé par l'éditeur qui indique les séparations entre les mots, les racines, les préfixes etc. A cause des difficultés liées à sa réalisation, un seul de nos manuscrits est lemmatisé.

L'absence de caractère séparateur complique énormément la reconnaissance des mots dans les différents manuscrits translittérés. Afin de rendre possible l'identification de ces mots, nous allons utiliser l'alignement de notre texte lemmatisé avec les autres manuscrits.

## 4 Alignement, score et distance

### 4.1 Introduction à l'alignement et aux problèmes de « granularité »

Pour comparer des manuscrits, il convient tout d'abord de décider de la granularité de l'alignement, c'est-à-dire de la nature des éléments à mettre en correspondance. L'absence de segmentation explicite et systématique dans les manuscrits sanskrits représente une difficulté supplémentaire. La segmentation peut être effectuée à plusieurs niveaux : les caractères, les syllabes, les mots, les lemmes, etc. Par expérience, plus la segmentation est riche en sens (des lemmes plutôt que des mots), plus les informations sont pertinentes.

Dans un premier temps, nous nous contenterons de 2 niveaux de base : les caractères et les mots à cause des difficultés liées au sanskrit.

### 4.2 Alignement

C'est un problème complexe dans le cas qui nous occupe car les spécificités du sanskrit et de la translittération gênent l'utilisation des alignements classiques liés à la distance d'édition. Une série de problèmes est en effet à constater :

- Quels poids doit-on affecter aux opérations élémentaires pour que le score minimum corresponde au meilleur alignement (au niveau de sanskrit, de la copie,...) ?

Considérons les deux alignements suivants :

1) A C G T A A T A T T G A - -                      2) A C G T A A T A T T G A  
A C G T A - - - T T G A T A                      A C G T A T T G A T - A

Le premier alignement est celui dans lequel le plus grand nombre de symboles alignés a été conservé, alors que le deuxième alignement est celui dont le nombre d'erreurs d'alignement est minimal. Ces deux alignements peuvent donc être considérés comme optimaux compte-tenu de la mesure adoptée

- Le nombre d'opérations élémentaires est-il correct ?

Certains caractères sanskrits peuvent correspondre à plusieurs caractères latins du fait de la translittération ; une comparaison au niveau des caractères latins ne correspond pas à une comparaison au niveau des caractères sanskrits qui est préférable.

Caractères sanskrits	अ	औ	भ
Lettres latines correspondantes	a	au	bha
Nbre de Lettres	1	2	3

FIG. 5 – Problème de nombre de lettres lors de la translittération

- L'alignement doit essayer de conserver les mots dans la mesure du possible. Dans l'exemple suivant le mot « combien » se trouve éclaté par l'alignement.

- COMPTE BIEN

- COM - - - BIEN

En plus, comme le sanskrit n'a pas de séparateurs entre les mots, il faut impérativement utiliser le padapatha pour optimiser cette contrainte.

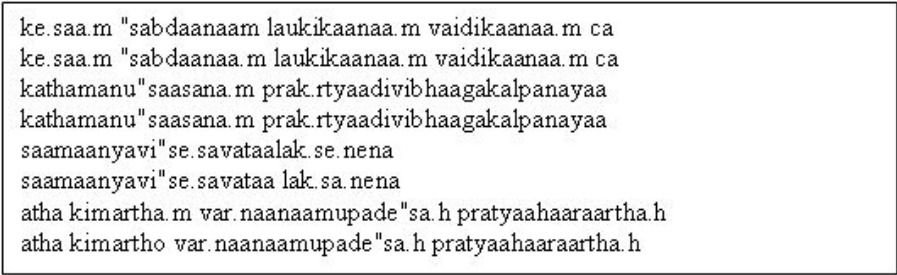
## Filiation de manuscrits sanskrits et arbres phylogénétiques

- Enfin, une autre difficulté provient des sandhis qui n'existent que dans le cas de liaison entre deux mots. Si un des mots disparaît dans une version, le sandhi aussi et l'alignement du mot restant devient délicat.

Voyons maintenant les méthodes proposées pour calculer nos « distances », l'une au niveau des caractères, l'autre au niveau des mots :

**1<sup>re</sup> méthode : Alignement de corpus multilingues + distance d'édition** Pour calculer une distance entre nos manuscrits, nous procédons en deux étapes. La première consiste à aligner les phrases des différents manuscrits puis pour chaque phrase, on aligne les caractères et on détermine alors une distance d'édition.

Afin d'aligner les phrases, nous utilisons les techniques de l'alignement de corpus multilingues qui sont utilisables même dans le cas d'une langue unique. Les algorithmes d'alignement utilisent tantôt une approche statistique, tantôt une approche lexicale ou un mélange des deux approches. Dans notre cas, l'approche lexicale n'est pas exploitable, c'est donc une méthode statistique qui a été utilisée, celle de Gale et Church (1991). Ce type d'alignement utilise uniquement les longueurs de phrase et aucune information sur leur contenu lexical. Les auteurs partent de la constatation que la longueur des phrases du texte source est fortement corrélée à celle des phrases du texte cible. Gale et Church ont montré que la corrélation entre les longueurs de deux phrases suit une loi probabilisée normale centrée réduite. Elle nous donne d'excellents résultats pour l'alignement de nos phrases permettant également de repérer les commentaires et les phrases non « comparables ».



```
ke.saa.m "sabdaanaam laukikaanaa.m vaidikaanaa.m ca
ke.saa.m "sabdaanaa.m laukikaanaa.m vaidikaanaa.m ca
kathamānu"saasana.m prak.rtyaativibhaagakalpanayaa
kathamānu"saasana.m prak.rtyaativibhaagakalpanayaa
saamaanyavi"se.savataalak.se.nena
saamaanyavi"se.savataalak.sa.nena
atha kimartha.m var.naanaamupade"sa.h pratyaahaaraartha.h
atha kimartha.m var.naanaamupade"sa.h pratyaahaaraartha.h
```

FIG. 6 – Résultats de l'algorithme de Gale et Church

Au niveau de l'alignement des caractères, nous avons opté pour la distance de Levenshtein (1966) qui est couramment utilisée dans de nombreuses applications où il faut mesurer la similarité entre deux séquences, ici nos phrases. Elle permet de déterminer quelle est la longueur d'une séquence minimale d'opérations élémentaires sur les caractères (suppression, insertion et substitution) pour transformer une phrase en une autre.

**2<sup>e</sup> méthode : Alignement de corpus à l'aide du padapatha + distance lexicale** Ici, pour mesurer la ressemblance entre nos manuscrits, nous procédons encore en deux étapes. La première consiste à aligner les différents manuscrits sur le padapatha afin d'identifier les mots, puis on construit une distance lexicale (ou presque) entre nos manuscrits.

Une méthode intéressante actuellement développée par Csernel et Bertrand (2005) permet d'aligner les manuscrits par rapport au padapatha. Cet alignement doit permettre en utilisant le padapatha comme lexique d'établir la liste des mots du padapatha qui sont reconnus ou non dans un manuscrit avec leur position. On n'a pas d'informations sur les mots du manuscrit qui ne sont pas dans le padapatha ; en effet, ceux-ci ne sont pas identifiables de façon automatique.

Au niveau des mots, l'usage d'indices tels que, parmi beaucoup d'autres, le célèbre indice de Jaccard (1912) ou la connexion lexicale de Muller (1977) permet de calculer le rapport entre les mots qui sont communs aux deux textes et ceux qui n'appartiennent qu'à l'un des deux. C'est une méthode similaire qui est utilisée pour construire une mesure de ressemblance avec le padapatha comme lexique.

La procédure de comparaison s'effectue séquentiellement selon l'ordre des mots du padapatha qui est vu ici, comme un lexique. L'algorithme de comparaison nous délivre 4 types d'informations :

$n$ =Nombre de mots du padapatha présents dans les 2 manuscrits.

$n_1$ =Nombre de mots du padapatha présents dans le manuscrit 1 et pas dans l'autre.

$n_2$ =Nombre de mots du padapatha présents dans le manuscrit 2 et pas dans l'autre.

$n_{12}$ =Nombre de mots du padapatha absents dans les 2 manuscrits.

On peut alors construire notre mesure de ressemblance entre les 2 manuscrits :

Mot $i$ du padapatha	Absence(0) ou présence(1) dans le manuscrit 1	Absence(0) ou présence(1) dans le manuscrit 2	$d_i$
viprakiir	1	1	1
kriyate	1	0	0
naama	0	0	1
kathamamu	0	1	0

TAB. 1 – Tableau de construction des  $d_i$  selon la présence où l'absence

$$\text{on a donc } d_i = \begin{cases} 1 & \text{si le mot est présent ou absent dans les 2 manuscrits} \\ 0 & \text{si le mot n'est présent que dans un des manuscrits} \end{cases}$$

$$\text{On pose alors } d = \frac{\sum d_i}{\text{Nb de mots}} = \frac{n+n_{12}}{n+n_1+n_2+n_{12}}$$

## 5 Les Arbres

### 5.1 La phylogénie

La phylogénie peut être considérée comme la construction de l'histoire évolutive d'un ensemble d'espèces. On choisit alors de représenter les relations qui existent entre elles sous forme d'un arbre phylogénétique (le plus souvent binaire).

Un arbre phylogénétique est une représentation graphique de la phylogenèse d'un groupe de *taxa*. C.-à-d. les feuilles représentent les espèces et les branches définissent les relations entre les taxa en terme de descendance. Les nœuds représentent des ancêtres hypothétiques.

Il peut paraître intéressant de comparer cette théorie de l'évolution avec la filiation de manuscrits comme l'a suggéré Buneman (1971). En effet, les séquences « génétiques » de

## Filiation de manuscrits sanskrits et arbres phylogénétiques

différentes espèces sont comparées et selon leur ressemblance ou dissemblance on détermine alors un arbre de l'évolution. Une situation similaire se retrouve si l'on compare les écrits de différents témoins les uns avec les autres afin de déterminer une sorte d'arbre de filiation des différents manuscrits.

Cependant, il existe une différence entre notre problème et celui de la théorie de l'évolution. Dans cette dernière, les espèces dont on compare les séquences d'ADN sont toutes à l'instant t de l'évolution donc, dans l'arbre, elles sont toutes des feuilles. Dans la filiation de manuscrit, rien n'empêche de trouver un manuscrit qui ne soit pas une feuille mais un nœud c.-à-d. un manuscrit « père » (cf. fig. 7). L'algorithme de reconstruction doit en tenir compte.

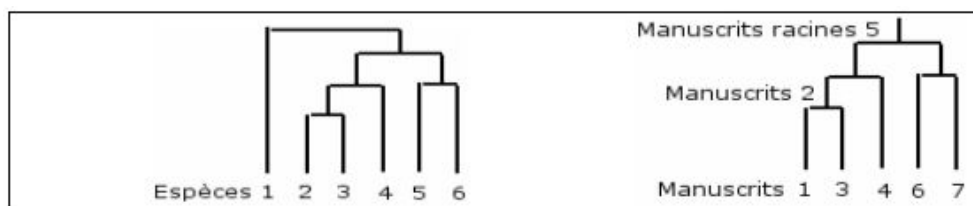


FIG. 7 – Différence entre phylogénie et filiation

Il existe différentes méthodes permettant de construire des arbres phylogénétiques :

- La méthode du maximum de parcimonie,
- Les méthodes probabilistes,
- Les méthodes basées sur les distances.

Au vu du nombre important de manuscrits, on privilégie une méthode basée sur les distances. Il peut en revanche être intéressant d'utiliser les autres méthodes pour confirmer les hypothèses que l'on a pu faire avec la méthode des distances sur une partie de l'arbre (sous-arbre).

## 5.2 Méthodes basées sur les distances

**Préliminaires** Il s'agit d'obtenir, sous forme de graphe planaire, la meilleure représentation possible des distances de chacun des textes à tous les autres. Chaque texte est représenté par une feuille ou un nœud de l'arbre. La distance qui le sépare d'un autre est matérialisée par la longueur du chemin à parcourir sur l'arbre pour unir ces deux textes. Les textes qui sont rattachés à un même nœud forment des groupes plus ou moins homogènes en fonction de la longueur des distances.

Cet ajustement étant NP-difficile (cf. Day (1987)) , la plupart des méthodes d'approximation utilisées sont heuristiques.

Nous utiliserons par la suite deux des nombreuses méthodes, la méthode des Groupements de Luong (1988) et NJ de Saitou et Nei (1987) :

**Méthodes des groupements** Les distances ou dissimilarités sont regroupées dans une matrice carrée où les textes sont rangés, en ligne et en colonne et appelées matrice de dissimilarité. L'algorithme procède à la construction de groupe en classant ensemble les textes séparés par la



distance inférieure à un score. Il recalcule alors les distances entre textes et groupes de textes et réitère le procédé jusqu'à la constitution d'un groupe unique.

La méthode permet d'obtenir des regroupements de plus de deux sommets, c.-à-d des arbres non binaires. Elle a une complexité en  $O(n^4)$  et propose une racine par construction.

**Méthode NJ** C'est une méthode qui combine une approche de distance avec le principe de parcimonie. On part d'un arbre en étoile dans lequel on recherche le couple de sommet  $i$  et  $j$  qui une fois rassemblé minimise la longueur totale de l'arbre. On recalcule la matrice de distances en considérant les sommets  $i$  et  $j$  comme un groupe  $(i,j)$  indissociable. On réitère les étapes jusqu'à ce qu'il ne reste plus de sommet dans la matrice. Deux différences avec la méthode précédente sont à noter : Elle a une complexité en  $O(n^3)$  et ne permet d'obtenir que des arbres binaires.

## 6 Résultats actuels

### 6.1 Résumé des méthodes développées ou en cours de développement

Précédemment, nous avons donc envisagées plusieurs méthodes de reconstruction d'arbre à partir de corpus. Le schéma (cf. fig 8) nous résume les différentes possibilités. Trois types d'arbres sont proposés dans le schéma et décrits dans Barthélemy et Guénoche (1988). Nous utilisons surtout les arbres hiérarchiques pour leur ressemblances avec des stemmas déjà réalisés

### 6.2 Corpus de test pour les méthodes

Pour les premiers tests, plusieurs corpus « fictifs » dont on connaît le stemma ont été réalisés. Chacun de ces corpus permet de tester un problème particulier (ex : arbre déséquilibré par perte de manuscrits dans une branche). Les méthodes se comportent en général très bien et l'expérimentation permet de les affiner.

Par la suite, nous avons aussi essayé les programmes sur un corpus de « Chain Letter » collecté par Bennett et al.. Les résultats sont en cours d'analyse mais permettent actuellement de retrouver globalement la classification des lettres.

### 6.3 Application sur des Manuscrits

La reconnaissance des mots à travers le Padapatha n'étant actuellement pas terminée, seule la méthode travaillant avec les distances d'édition a été testée sur les manuscrits sanskrits.

Pour les comparer, on a réalisé 3 corpus d'une cinquantaine de manuscrits :

- Un corpus n° 1 contenant les 34 premiers paragraphes du premier chapitre.
- Un corpus n° 2 contenant les 17 premiers paragraphes.
- Un corpus n° 3 contenant les 17 derniers paragraphes.

L'expérimentation fait rapidement apparaître des manuscrits qui sont très loin des autres dans le sens vertical de la figure 9 (exemple mns IO5 ). Ce sont des manuscrits très délabrés. Ceci confirme que la distance construite mesure bien les dégradations des manuscrits. Dans

## Filiation de manuscrits sanskrits et arbres phylogénétiques

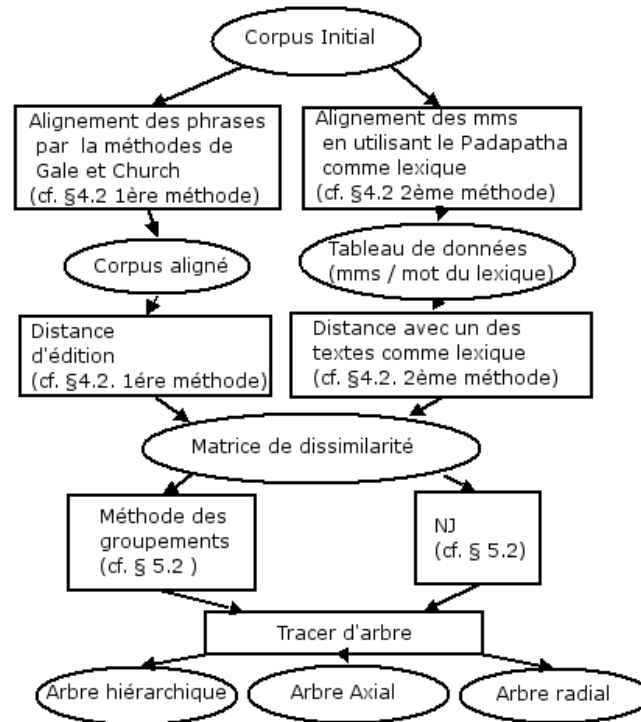
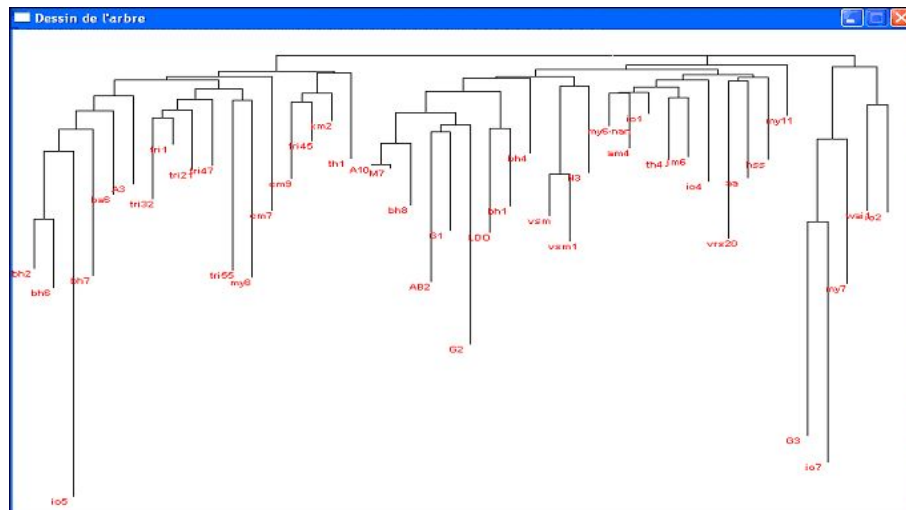


FIG. 8 – Schéma récapitulatif des différentes méthodes



notre construction, on considère que le temps est « proportionnel » aux dégradations par hypothèse ; plus le temps passe, plus le manuscrit est modifié.

Le plantage de l'arbre sur la figure 9 est celui proposée par la méthode NJ. Cela correspond à la recherche du manuscrit original dans notre problème. L'utilisation d' « outgroup » (cf. Watrous et Wheeler (1981)) ou de la médiane d'un graphe (cf. Harary (1969)) sont des méthodes similaires couramment utilisées en phylogénétique. Elles ne permettent pas non plus dans le cas d'un arbre déséquilibré d'obtenir la racine. Actuellement le problème d'orientation de l'arbre reste à résoudre et le plantage proposé est arbitraire.

Au niveau de la validation des résultats, les différents arbres obtenus sur les trois corpus convergent si l'on impose la même racine. Les résultats semblent en effet peu sensible aux changements de corpus ce qui laisse à penser que la classification reste constante quelques soient les parties des manuscrits sélectionnées. Ensuite les analyses des sanskritistes montrent que les classifications obtenues par les arbres sont conformes à leurs attentes. Ils retrouvent rassemblés des manuscrits qu'ils savaient proches. Ces deux validations, ainsi que le bon comportement de nos algorithmes sur des corpus fictifs tentent à prouver l'intérêt de la méthode.

Enfin, Pour faciliter la lecture des arbres et visualiser plus facilement les filiations, des techniques de consensus (cf. Adams (1972)) sont actuellement mise en oeuvre.

## 7 Conclusions et perspectives

On peut avant tout se demander si, après le nombre important de « prétraitements » effectués sur le corpus, on classe toujours la filiation des manuscrits et non pas les écoles de copistes, les écritures, etc.

Un autre problème est de savoir déterminer au mieux la racine. Actuellement les techniques utilisées n'ont pas donné satisfaction. Pour cela il faut sûrement intégrer des informations extérieures comme la paléographie et l'ecdotique ou la datation. Il faudra alors inférer notre arbre en y incorporant des observations supplémentaires.

Une étude des règles de l'acte de copie peut permettre d'orienter le graphe et de développer des méthodes d'intermédiation entre les textes pour reconstruire le stemma. Cette étude peut aussi résoudre le problème de la contamination (hybridation) des textes.

Enfin des connaissances supplémentaires sur le sanskrit peuvent être utilisées pour passer d'une analyse au niveau des mots à une analyse au niveau des lemmes sans doute plus riche de sens.

## Références

- Adams, E. (1972). Consensus techniques and the comparison of taxonomic trees. *Syst. Zool.* 21, 390–397.
- Barthélemy, J. P. et A. Guénoche (1988). *Les Arbres et les Représentations des Proximités*. Masson.
- Bédier, J. (1928). La tradition manuscrite du lai de l'ombre. réflexions sur l'art d'éditer les anciens textes. *Romania* 54, 162–86, 321–56.
- Bennett, C., M. Li, et B. Ma. Linking chain letters.

## Filiation de manuscrits sanskrits et arbres phylogénétiques

- Buneman, P. (1971). Filiations of manuscripts mathematics in archaeological and historical sciences. *Edinburgh University Press*.
- Csernel, M. et P. Bertrand (2005). Comparaison de manuscrits sanskrits(édition critique et classification). *Modulad 33*, 1–20.
- Day, W. H. E. (1987). Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology 49*, 461–467.
- Filliozat, J. (1941). *Catalogue du fonds sanskrits* (Paris, Adrien Maisonneuve ed.).
- Gale, W. A. et K. W. Church (1991). A program for aligning sentences in bilingual corpora. *Computational Linguistics 19*, 75–102.
- Harary, F. (1969). *Graph Theory*. Reading, Mass. : Addison-Wesley.
- Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New Phytol. 11*, 37–50.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics - Doklady 10*(8), 707–710.
- Luong, X. (1988). *Méthodes d'analyse arborée. Algorithmes. Applications*. Ph. D. thesis, Thèse de doctorat, Paris V.
- Muller, C. (1977). *Principes et méthodes de statistique textuelle*. Hachette Paris.
- Quentin, H. (1926). *Essais de critique textuelle*. Picard.
- Saitou, N. et M. Nei (1987). The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Mol Biol Evol 4*, 406–425.
- Salemnans, B. J. P. (1999). *Building Stemmas with the Computer in a Cladistic, Neo-Lachmannian Way : the Genealogy of the Fourteen Versions of Lanseloet van Denemerken*. Ph. D. thesis, Nimega.
- Veltuis, F. (1991). Package devenagari pour tex latex.
- Watrous, L. E. et Q. D. Wheeler (1981). The outgroup comparison method of character analysis. *Systematic Zoology 30*, 1–11.

## Summary

The establishment of a stemma codicum is one of the most rigorous approaches of textual criticism. It requires the rebuilding of the history of the text by classifying the corpus to decide if a group of manuscripts is generated by a lost intermediary. To cluster our corpus, we use methods of the computerized textual analysis and phylogenetic reconstruction in order to establish the tree of filiation or pedigree. The method employed is dedicated to Sanskrits manuscripts taking into account all the specificities of this language.