Hypertextes ou documents structurés ? études de cas en critique génétique

Jacques André et Hélène Richy

IRISA, campus universitaire de Beaulieu,
F-35042 Rennes cedex, France
{jandre, richy}@irisa.fr

RÉSUMÉ. *L'expression « documents structurés » est très réductrice. Nous montrons, dans le cadre d'un travail spécifique en critique génétique, que les notions de graphe (et non simplement d'arbre), de vue, de documents actifs ajoutées à celles de liens hypertextuels permettent d'utiliser des documents structurés pour bien autre chose que leur édition.*

ABSTRACT. *This paper shows that graphs (and not only trees) are available within structured documents. This concept plus concepts such as views, active documents and hypertext links make structured documents very powerfull tools. Experiments on Genetic Criticism case studies are exhibited.*

MOTS-CLEFS : *Document structuré, vues, documents actifs, comparaison de chaînes, critique génétique.*
KEY WORDS : *Structured documents, views, active documents, string comparison, genetic criticism.*

1. Introduction

Depuis longtemps, linguistes, philologues, spécialistes de l'histoire des textes, etc. travaillent sur des données textuelles ou des images de textes informatisées. Selon la nature du travail à faire, ils utilisent des systèmes de base de données et de recherche documentaire, des outils linguistiques (lemmatisation, syntaxe, dictionnaires, etc.), des systèmes de traitement de textes pour l'édition et, depuis quelques années, la notion d'hypertexte pour relier divers documents ou concepts entre eux. Mais il semble que ces outils ne soient pas toujours compatibles et qu'il faille beaucoup d'énergie pour, par exemple, publier un cartulaire depuis un hypertexte. Diverses études [AFR95, AND95, STI94, VIR93] ont analysé la notion de poste de travail, comme le poste de lecture de la TGB. Nous voudrions ici montrer l'intérêt de la notion de document structuré comme moteur central d'un tel poste, en donnant à document structuré non pas le sens restreint de structure organisationnelle hiérarchique qu'il peut avoir parfois (par exemple avec SGML) mais celui plus général incluant notamment la notion de document actif (en donnant à ce mot le sens de « non-passif » ce qui et plus général que l'attribut « interactif » donné aux formulaires du Web) [QUI94].

Notre étude est basée sur diverses expérimentations qui ont été faites dans le cadre du projet Philectre que nous allons d'abord présenter.

2. Le projet Philectre

Philectre (PHILologie élECTRoniquE) est un projet [1] d'étude de faisabilité d'emploi de l'informatique dans le cadre de la critique génétique. La critique génétique (voir notamment [GRE95]) est l'étude de la genèse des œuvres littéraires (telles que celles de Flaubert, Proust, etc.[HAY93]), scientifiques (par exemple les écrits de Pasteur [BPR95]) ou musicales (par exemple les chansonniers des troubadours).

L'informatisation d'un travail de critique génétique concerne plusieurs de ses phases.

1. La constitution du dossier génétique, c'est-à-dire de l'inventaire de toutes les sources (écrits, images, enregistrements, etc.) et leur concordance (par exemple, tel folio de la troisième version de *Féerie pour une autre fois* de Céline correspond à tel autre d'une autre version et à telle page de son édition finale) : tout ceci constitue une base de donnée avec beaucoup de pointeurs d'un élément à l'autre que de nombreux systèmes « hypertextes » permettent de gérer.

2. La constitution d'une base de données informatisée des sources. Celles-ci sont essentiellement graphiques : manuscrits d'auteurs (par exemple les carnets ayant servi de cahier des charges à *La vie mode d'emploi* de Perec [HMN93]) mais aussi les manuscrits médiévaux. Il ne suffit pas de « scanner » ces documents, encore faut-il le faire sans perte d'information (il est difficile par exemple aujourd'hui de traiter la texture, l'or, etc. [AFR95]) et sans prendre trop de place ; et surtout il faut pouvoir s'y référer, ou plutôt se référer à telle ou telle partie d'un document. En particulier, dans le cadre du projet Philectre, ont été étudiées des méthodes permettant de détecter les lettrines dans des chansonniers en vue du repérage des strophes (IUT de Reims) ou de reconnaître (et donc indexer) les lignes d'un manuscrit de Flaubert[LSF96, LRL97, RLL97].
   Un problème connexe évident (mais dont les solutions ne le sont pas) est celui de l'espace vraiment utilisable sur un écran : il est difficile de concevoir un système ergonomique présentant « à la fois » dix versions d'un chansonnier de

1. Dans le cadre du GIS Sciences de la cognition du CNRS (appel d'offre « Lives électroniques »). Il est animé par Bernard Cerquiglini (Paris VII) et Jean-Louis Lebrave (Institut des Textes et Manuscrits Modernes) et comprend des équipes de l'ENST, l'IRHT, l'Inria, etc. En particulier, deux œuvres sont à l'étude : un corpus en ancien provençal de *Can vei la lauzeta* (du troubadour Bernard de Ventadour, XIII^e siècle) et un corpus moderne, la *Légende de Saint Julien l'Hospitalier* de Flaubert. Par ailleurs dans le cadre du projet Gecri (GIS patrimoine écrit, CNRS), nous avions aussi abordé certains aspects des manuscrits de Céline (*Féerie pour une autre fois*) en collaboration avec Corinne Chuhat et les *Maximes* de La Rochefoucault avec Roger Laufer. Les opinions émises ici sont celles des auteurs et pas nécessairement celles de tout le groupe Philectre.

troubadours ; en revanche il est possible d'en appeler à la demande deux ou trois par exemple, ou une version et sa traduction diplomatique (figure 2).

3. La publication de travaux de critique génétique a une spécificité qui est « l'édition diplomatique »[GRE95], copie fidèle des manuscrits où l'écriture manuscrite est remplacée par du texte tapé à la machine pour permettre au non-spécialiste de tel auteur d'en lire les manuscrits (voir par exemple figure 2). Cette édition peut se faire sous forme de pages PostScript (affichables, éditables mais pas réutilisables), toutefois on a là confusion entre édition et outil de travail. Nous verrons en 3.1. qu'il y a d'autres structures de données plus utilisables que la vision classique de la forme diplomatique.

4. La notion d'hypertexte n'est pas uniquement, comme l'emploie la majorité des chercheurs, un moyen de mettre en relation des parties physiques de documents (on clique ici pour montrer telle image ou là pour montrer tel passage, ce qui n'est qu'une copie électronique des « voir » des encyclopédies papier). Elle permet aussi des relations plus sémantiques que quelques chercheurs ont commencé à utiliser. Citons par exemple les travaux de Lebrave [LEB92, LEB96], Laufer[LAU94], Balpe[BLS95], Clément[CLE95], etc.

5. Enfin, tout comme on fait des études statistiques ou de linguistique « computationelle » sur des textes linéaires, il faut aussi pouvoir en faire sur des hypertextes.

Par ailleurs, une spécificité des recherches en critique génétique est que chaque auteur étudié est un cas particulier. Pas question donc d'offrir un modèle universel de poste de travail pour généticiens, mais bien plutôt de concevoir un poste générique qui puisse être « progammé », adpaté à chaque cas spécifique. Certains systèmes de manipulation de documents structurés permettant cette approche programmable (c'est le cas de Thot), nous pensons qu'ils sont de particulièrement bons candidats pour être le moteur (c'est-à-dire le cœur ou le système) permettant de relier entre elles toutes les différentes taches énumérées ci-dessus [AND95]. Les expériences que nous allons décrire ont été faites en utilisant Thot, un éditeur structuré issu de Grif [BQR96].

3. Quelques possibilités des documents structurés

3.1. *Il n'y a pas que des arbres*

Une erreur des auteurs des synthèses sur les documents structurés (par exemple [AFQ89] !) est probablement d'avoir tellement insisté sur la notion de structures hiérarchiques (un article est formé de sections, elles mêmes formées de sous-sections, elles mêmes formées de sous-sous-sections, etc.) qu'on a fini par occulter la possibilité de définir d'autres structures, notamment des graphes. Or, les documents structurés ne sont pas limités à l'emploi de ces arbres ! Voici un exemple.

La figure1-haut montre un manuscrit de Proust. Sa version diplomatique n'apportant rien d'utilisable informatiquement, Jean-Louis Lebrave [GLV92] a cherché

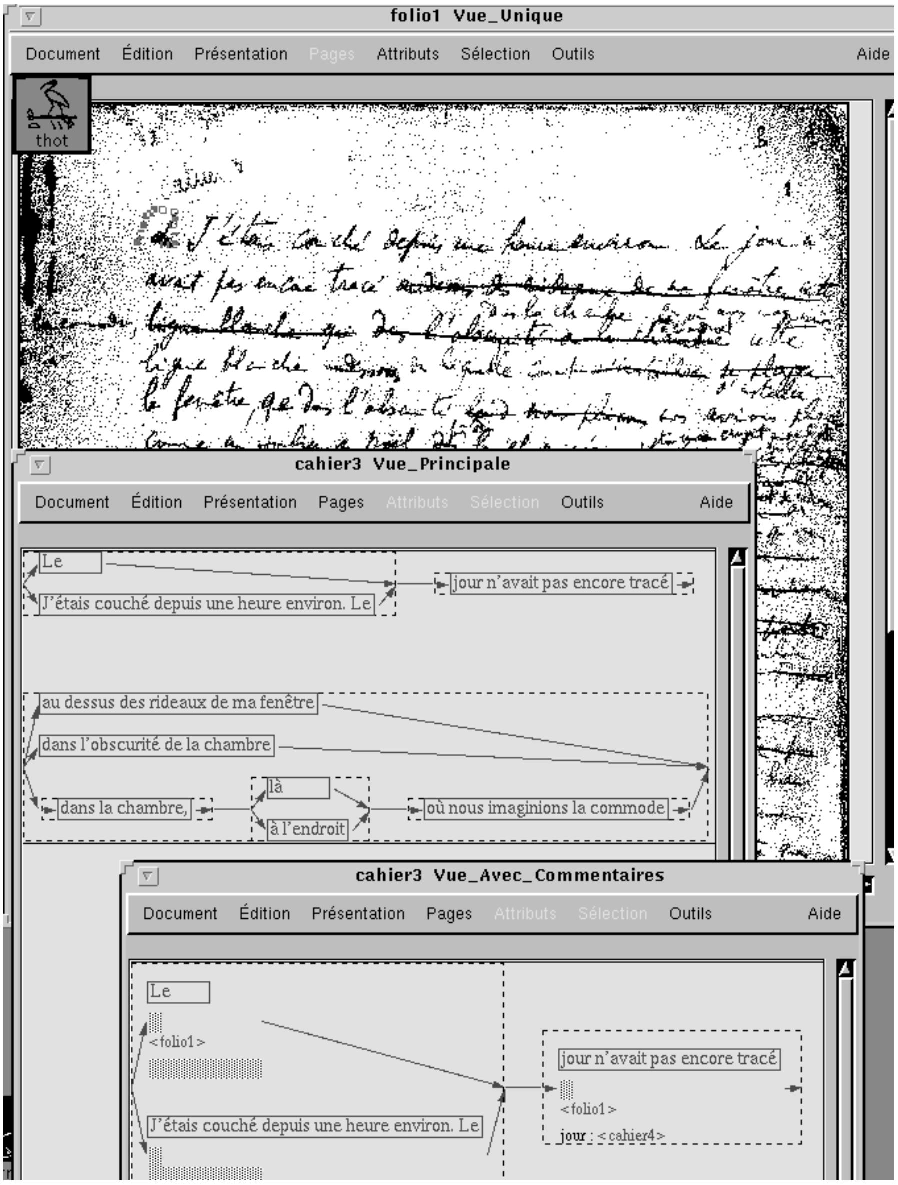

FIG. 1 – *De haut en bas : manuscrit de Proust, graphe génétique correspondant et détails des nœuds du graphe.*

diverses représentations de cette page plus propices au travail informatique. Il nous a finalement proposé une mise en page comme celle de la figure 1-milieu.

En cliquant dans un rectangle, le généticien peut faire toutes les opérations usuelles de la PAO (couper, coller, etc.) mais au niveau de ce graphe, c'est-à-dire par exemple inclure un rectangle parallèlement à un autre, ou après un autre, etc. C'est bien sûr à lui de décider des variantes et de leur synchronie (en figures 1 et 2, on suppose que de deux rectangles superposés, celui contenant le tout premier texte est le plus haut).

Nous voudrions donc insister sur le fait que de telles structures sont tout à fait compatibles avec les DTD des documents structurés et adaptables à chaque dossier génétique spécifique.

3.2. *Notion de vue*

Le graphe génétique de la figure 1-milieu a été construit en Thot. En fait les rectangles de ce graphe sont des « vues » partielles d'une structure de donnée plus grande et dont l'utilisateur peut cacher à volonté une partie (un exemple typique est celle d'un livre dont on peut cacher tout ce qui est « en-dessous » de la notion de section : ce qui reste visible est la table des matières !).

Dans la vue globale (figure 1-bas), chaque rectangle fait partie d'un ensemble de données où le généticien peut, à son choix, mettre un pointeur sur l'image du manuscrit (sur laquelle il suffit donc de cliquer pour faire apparaître sur l'écran le « folio 1 »), des indications topologiques (« dans la marge » par exemple), graphiques (« biffé », « encre rouge »), diachroniques (« probablement le 3 janvier »), etc.

Par ailleurs, la notion de vue est très probablement une réponse à la question ergonomique d'accéder à un moment donné à n manuscrits différents sans pour autant ouvrir n fenêtres en même temps (n pouvant parfois être supérieur à dix). La figure 2 montre le graphe génétique de neuf versions d'un même chansonnier (ici la granularité choisie est celle du vers, mais nous aurions pu prendre aussi bien une strophe entière).

3.3. *Liens hypertextuels*

Les documents structurés ne sont pas incompatibles avec la notion de lien hypertexte. Dans Thot ceux-ci sont typés (il y a donc une sémantique attachée à ce lien) [QVA92]. Dans l'exemple de la figure 1-b, depuis le rectangle « le jour n'avait pas encore tracé » on peut mettre un lien du type « antinomie » ou « jour » vers le rectangle contenant « il faisait nuit » d'un autre folio voire même d'un autre cahier. Toutefois, nombre de problèmes sont encore à résoudre : d'une part trouver un moyen d'accepter n'importe quelle granularité (mot, syntagme, phrase, paragraphe, etc.) sans pour autant pénaliser la saisie des variantes et, d'autre part, permettre de travailler (informatiquement) sur ces liens.

4. Documents actifs

La figure 2 utilise la même représentation en graphe que la figure 1 mais pour représenter « en parallèle » les n versions textuelles d'une même œuvre, ici un chansonnier provençal du XIII[e] siècle : *Can vei la lauzeta* du troubadour Bernard de Ventadour. Outre les manuscrits, on dispose aussi des n versions translitérées et codées,

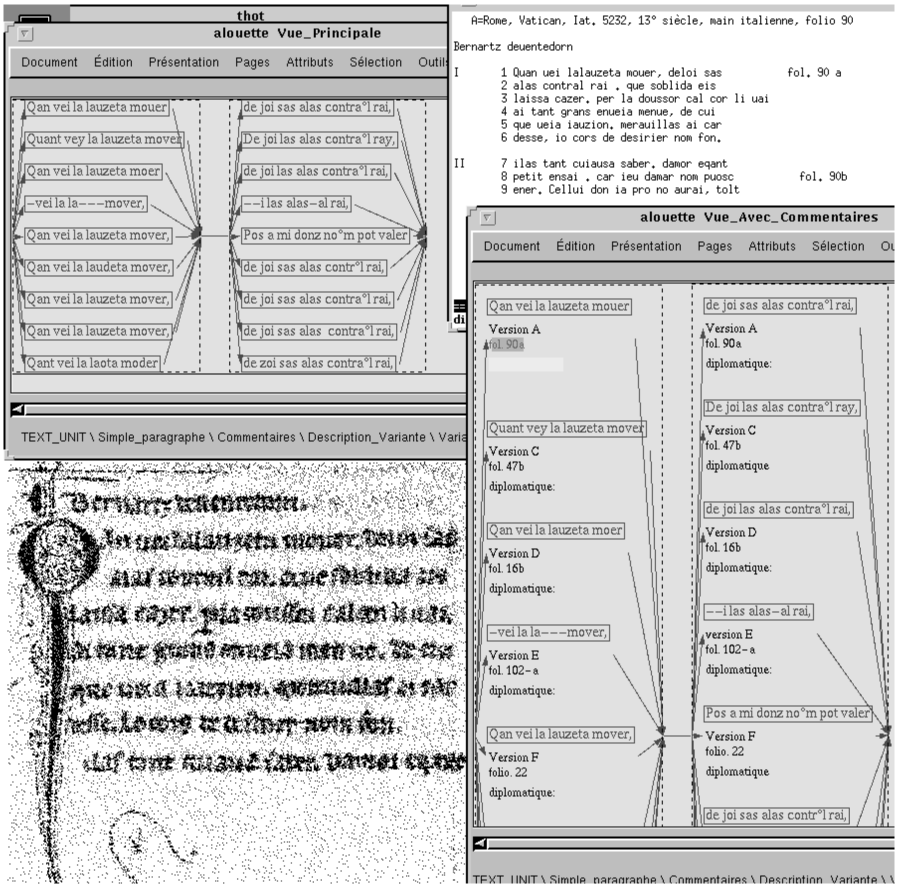

FIG. 2 – *Étude d'édition hypertextuelle d'un chansonnier de troubadour provençal (*Can vei la lauzeta *de Bernard de ventadour). Graphe génétique (en haut à gauche et en bas à droite, avec détail des nœuds) ; en haut à droite, version diplomatique de la version A (folio 90 du manuscrit en bas à gauche). Édition de Anne Wanono (Institut de Recherche sur l'Histoire des Textes) qui nous a aimablement confié ces textes ; le folio 90 a été scanné et reproduit ici avec l'autorisation de Madame Hasenohr (IRHT).*

probablement avec une classe de documents proche de la TEI [BUR95]. Le graphe que présente cette figure n'est en fait qu'un document de travail — il ne sera sans doute pas inclus dans une édition finale — qui peut être construit automatiquement à partir de ces n documents primaires. C'est déjà une forme de document actif. De façon plus générale[QVA94], on entend ainsi des documents qui se comportent comme des programmes et qui peuvent produire à leur tour des documents ou d'autres actions. Typiquement, rentrent dans cette catégorie les formulaires interactifs du Web. Montrons quelques possibilités de ce concept sur un problème spécifique.

Restons dans le cadre des variantes de manuscrits ou de versions d'une même œuvre et considérons le cas où nous disposons du texte source linéaire, soit qu'il existe dans quelque base de données (c'est par exemple le cas des *Quatre évangiles* qui ont d'ailleurs fait l'objet de nombreuses études de comparaisons et de présentation « synoptique » des versions [BEN90]), soit qu'on l'ait saisi à partir des éditions connues d'une œuvre (cas des *Maximes* de La Rochefoucault dans les classiques Garnier ou l'œuvre de Goldoni [TOS96]) soit que l'on ait une édition reconstituée des manuscrits (par exemple les diverses versions de *Féerie pour une autre fois* de Céline dans la Pléiade). Proposons nous donc de prendre les versions A et B de *Féerie* et de montrer comment Céline a pu passer de l'une à l'autre non pas en construisant manuellement le graphe (comme en figure 1) mais en créant automatiquement un hyperdocument dans lequel les passages différents et les passages semblables des deux versions apparaissent clairement (et par là ce qui a été fait : suppression, insertion, remplacement, déplacement, etc.). Cette manipulation se fait en trois phases principales :

1. **Lancement de l'opération —** un généticien ayant affiché et sélectionné dans deux fenêtres les parties des versions qu'il veut comparer appelle par le biais d'un menu l'action désirée, ici la comparaison de deux chaînes. Grâce à la notion d'interface programmable, Thot appelle ainsi un script (écrit par exemple en C, TCL, Java, etc.) fournissant à une application quelconque le moyen de travailler sur la représentation interne des documents. L'écriture de ces scripts et leur inclusion dans le menu d'appel est à la charge des utilisateurs, ce qui ouvre à notre généticien accès à toutes les applications dont il a besoin.

2. **Travail de comparaison —** le principe de la comparaison [AHO90, STE94] est de construire à partir de chaque mot A_i de la version A et B_j de la version B, une matrice $M_{i,j}$ telle que $M_{i,j} = 1$ si $A_i = B_j$, 0 sinon. Signalons dès à présent que la proximité de deux mots pourrait être enrichie sémantiquement (par exemple accepter que « commode » soit proche de « console »). Outre certains points de bruit (identité aléatoire de certains mots), cette matrice présente des diagonales correspondant aux parties identiques de textes (figure 4-droite). En reliant celles-ci on peut trouver ce qui a été ajouté ou supprimé par rapport à la première version (figure 4-gauche). En fait, pour travailler sur des données aussi volumineuses et pour donner une certaine tolérance à la comparaison, nous utilisons des méthodes anciennes de comparaison de chaînes, méthodes réactualisées en reconnaissance de la parole ou en biologie moléculaire pour les recherches sur l'alignement génétique [LAV96]. On trouvera dans [RIA97] les détails de notre implémentation. Il est possible de donner au généticien le

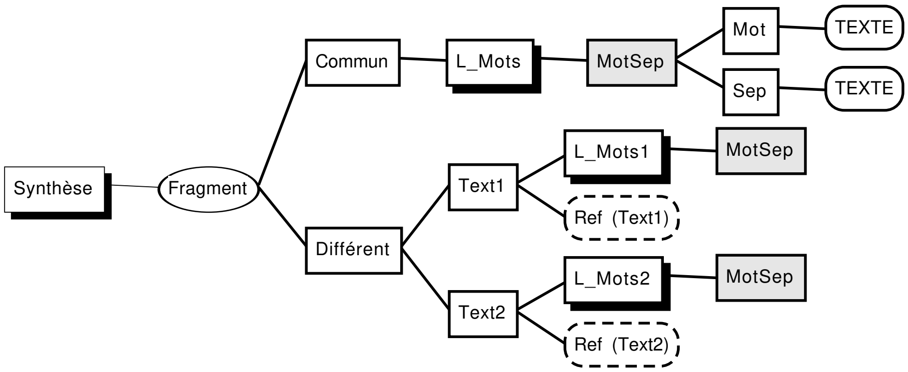

FIG. 3 – *Modèle logique de comparaison de textes — un rectangle arrondi figure une unité de base (texte ou référence), un rectangle ombré une liste d'éléments de même type, un rectangle un agrégat d'éléments de types différents et une ellipse un choix entre deux types d'éléments.*

moyen de visualiser cette matrice et de zoomer sur telle ou telle partie pour suivre cette comparaison en fonction du texte (figure 5).

3. Création du document de synthèse — pour que le document résultat soit un document structuré, nous avons défini un modèle logique et un modèle visuel. Le premier permet de considérer un document comme une liste de fragments qui sont soit communs aux deux textes, soit propres à un seul. Des liens entre ces deux versions et une présentation synchronisée des deux textes fait du document un hyperdocument. La figure 3 donne une représentation de ce modèle logique. L'indépendance de ce modèle logique avec le modèle visuel facilite la production de vues individualisées au gré de chaque utilisateur. Thot permet de spécifier plusieurs vues du même document ; dans cet exemple on peut avoir une vue globale (les deux textes) ou des vues décorées du premier ou du second texte. L'utilisateur peut alors analyser les résultats de la comparaison et en contrôler les choix (segmentation ; contraction des passages longs ; mises en valeur par des couleurs, des différences ou ressemblances, etc.). La figure 6 montre un exemple d'affichage de comparaison des deux premières versions de *Féerie pour une autre fois*. Le modèle visuel étant facilement extensible, d'autres vues peuvent être définies afin de faciliter la perception des « lieux variants » des textes comparés : index alphabétique des mots, présentation d'une liste de passages communs classés par taille décroissante, numérotations des occurrences de fragments, présentation comme en figure 1-b, etc. Il faut pour cela créer de nouvelles feuilles de style adaptées au modèle logique défini précédemment. Ce n'est pas facile de manipuler des langages de feuilles de style, aussi faut-il étudier la possibilité d'en définir de façon plus ergonomique [HER96, RIC97].

Précisons bien que ce que nous avons montré ici est très spécifique à cette appli-

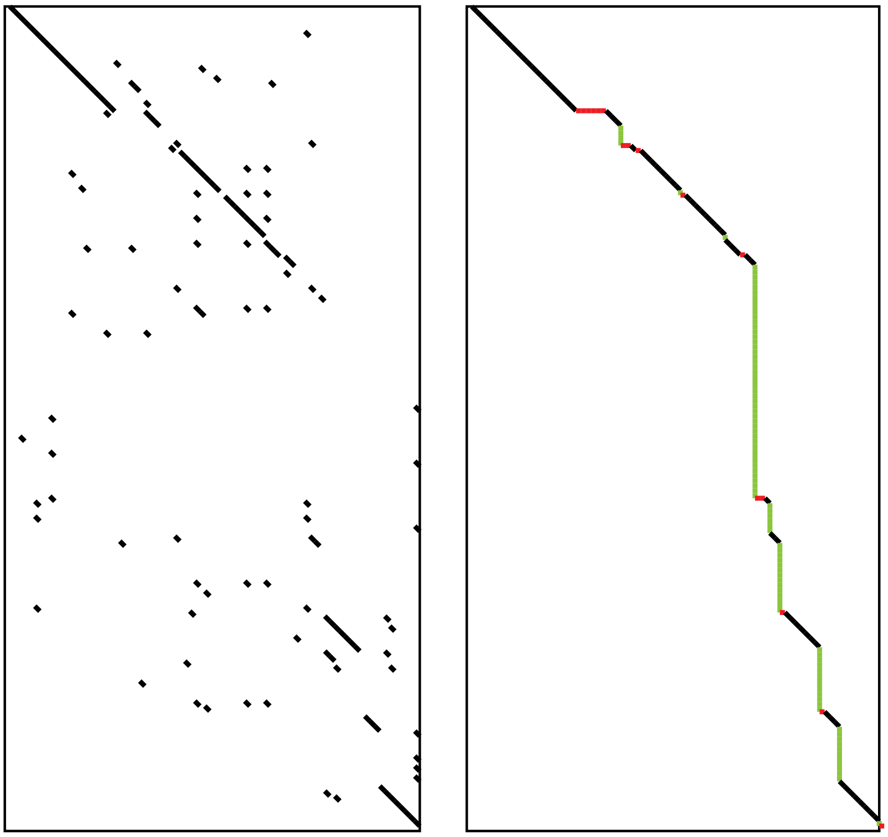

FIG. 4 – *Comparaison mot à mot du début de deux versions de* Féerie pour une autre fois *de Céline ; à gauche, comparaison brute, à droite après élimination du bruit.*

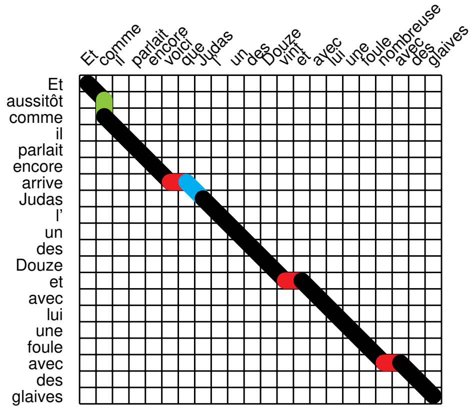

FIG. 5 – *Comparaison de deux versions du* Nouveau testament *(hor. Mt26/47 et vert. Mc14/43 ; textes d'après [BEN90]).*

cation. Mais ce serait applicable à d'autres travaux (par exemple de statistique) avec d'autre modèles de documents et d'autres scripts.

5. Conclusion

En intégrant des concepts d'hypertextes (liens) aux documents structurés, surtout s'ils sont manipulables en tant que documents actifs, on peut définir toute une classe d'applications utilisables pour l'informatisation des dossiers génétiques en amont bien de leur simple « édition » c'est-à-dire pour la recherche.

Remerciements : Nous tenons à remercier Chrystelle Hérault et les étudiants stagiaires de l'Insa et de l'Ifsic qui nous ont aidé à préparer certaines de ces expériences. Merci aussi bien sûr à nos collègues des projets Gecri et Philectre sans qui ce travail n'aurait pu avoir lieu.

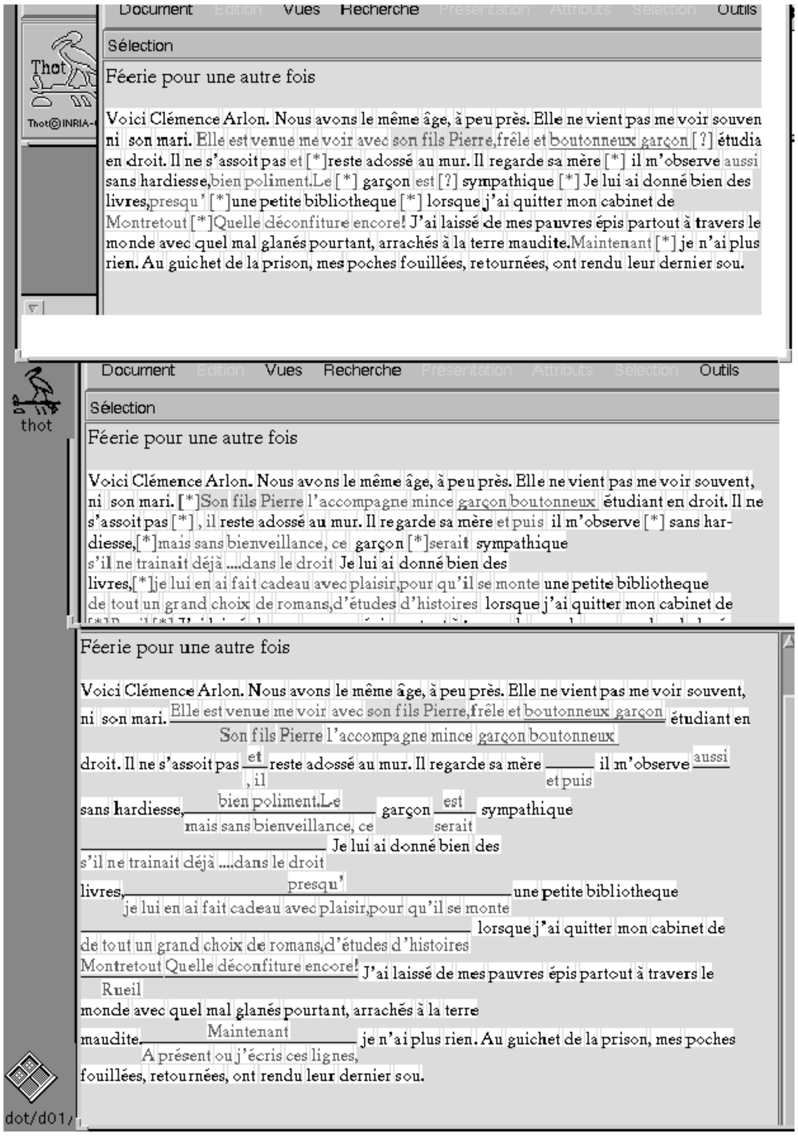

FIG. 6 – *De haut en bas : vues du début des versions A et B de* Féerie pour autre fois *(d'après la collection de la Pléade) puis vue de synthèse*

Références

[AFQ89] Jacques André, Richard Furuta and Vincent Quint, *Structured documents*, Cambridge University Press, 1989.

[AFR95] Jacques André, Jean-Daniel Fekete et Hélène Richy, « Traitement mixte image/texte de documents anciens », *Cahiers Gutenberg*, n^{o} 21, juin 1995, pages 75-85.

[AHO90] A. Aho, « Algorithms for finding pattern in strings », *Handbook of Theoritical Computer Science*, J. van Leeuwen ed., MIT Press, 1990, p. 257-300.

[AND95] Jacques André, « Vers un poste de travail sur l'écrit », in *Hypertextes et hypermédias*, J.P. Balpe, Alain Lelu et Imad Saleh eds., Hermes, 1995, pages 119-126.

[BEN90] P. Benoit et M.-E. Boismard, *Synopse des Quatre Évangiles*, Les éditions du Cerf, 1990.

[BLS95] J.P. Balpe, Alain Lelu et Imad Saleh eds., *Hypertextes et hypermédias*, Hermes, 1995

[BPR95] Françoise Balibar et Marie Laure Prévost, *Pasteur, cahiers d'un savant*, CNRS éditions, BNF et Zulma eds., 1995.

[BQR96] Stephane Bonhomme, Vincent Quint, Hélène Richy et Cécile Roisin, et Irène Vatton, *Manuel Thot*, Inria-Grenoble, 1996
`http://opera.inrialpes.fr/thot/doc/Thotman-F.html`

[BUR95] Lou Burnard, and C.M. Sperberg-McQueen, *TEI Lite : An Introduction to Text Encoding for Interchange*, http ://www.uic.edu/org/tei/intros/, juin 1995 ; traduction française dans *Cahiers Gutenberg*, n^{o} 24, pages 23-152, 1996.

[CLE95] Jean Clément, « Fiction interactive et modernité », *Littérature*, numéro spécial « Informatique et littérature », n^{o} 96, Larousse, décembre 1994.

[GLV92] A. Grésillon, J.-L. Lebrave et Catherine Viollet,*Proust à la lettre — les intermittences de l'écriture*, Du Lérot éd., Tusson, 1990.

[GRE95] Almuth Gresillon, *Éléments de critique génétique*, PUF 1994.

[HAY93] Louis Hay, *Les manuscrits des écrivains*, Hachette et CNRS éditions, 1993.

[HER96] Chrystelle Herault, « Définition incrémentale de feuilles de styles », *Cned*, Nantes, 1996.

[HMN93] Hans Hartje, Bernard Magné et Jacques Neefs, *Cahier des charges de* La Vie mode d'emploi *de Georges Perec*, CNRS éditions et Zulma, 1993.

[LAU94] Roger Laufer, « Hypertexte : visualisation comparative et explicitation », *Hypermédias, éducation et formation* (édités par E. Bruillard, G.-L. Baron et B. de La Passardière), Masi, Université de Paris VI, 1994, pages 55-73.

[LAV96] Dominique Lavenier, « Dedicated hardware for Biological Sequence Comparison », *Journal of Universal Computer Science*, 2(2), février 1996. Voir aussi *Rapport de recherche Inria* n° 2845, mars 1996.

[LEB92] Jean-Louis Lebrave, « L'hypertexte et l'avant-texte », *EP92*, (C. Vanoirbeek et G. Coray eds.), Cambridge University Press, 1992, p. 233-246.

[LEB96] Jean-Louis Lebrave, « Hypertexte, édition savante, édition génétique », *Actes du séminaire Hypermédias, Education et Formations* (édités par E. Bruillard, G.-L. Baron et B. de La Passardière), IUFM Créteil, LIP ParisVI et INRP, 1996, p. 51-65.

[LRL97] Laurence Likforman-Sulem, Laurent Robert, Eric Lecolinet, Jean-Louis Lebrave et Bernard Cerquiglini, « Édition hypertextuelle et consultation de manuscrits : le projet Philectre », *4e conférence Hypertextes et hypermédias* (ces actes) (Lelu et Saled eds.), à paraître, Hermes, septembre 1997.

[LSF96] Laurence Likforman-Sulem et Claudie Faure, « Structuration des manuscrits pour l'édition électronique », *Colloque CNED*, Nantes, 1996, p.267-273.

[QUI94] Vincent Quint (sous la direction de), *Electronic Publishing - Origination, Dissemination and Design*, special issue on Active Documents, 7(2), June 1994.

[QVA92] Vincent Quint and Irène Vatton, « Combining Hypertext and structured document », *ECHT'92 Proceedings*, Lucarella ed., ACM Press, Milan, décembre 1992.

[QVA94] Vincent Quint and Irène Vatton, « Making Structured documents active », *Electronic Pubishing – Origination, Dissemination and Design*, 7(2), pp. 55-74, June 1994.

[RIA97] Hélène Richy et Jacques André, « Comparaison de chaînes, documents structurés et critique génétique », Publication interne Irisa, à paraître, 1997.

[RIC97] Hélène Richy, « Feuilles de style pour le Web », *Cahiers Gutenberg*, n° 26, mai 1997, p.133-145.

[RLL97] Laurent Robert, Laurence Likforman-Sulem and Éric Le Colinet, « Image and Text coupling for Creating Electronic Books for Manuscripts », *Colloque ICDAR'97*, Ulm, 1997 (à paraître).

[STE94] Graham A. Stephen, *String Searching Algorithms*, Lectures Notes Series on Computing - Vol. 3, World Scientific Publishing Co., Singapour, 1994.

[STI94] Bernard Stiegel, « Lecture et édition savante assistée par ordinateur : l'hypertraitement de texte », *Actes du congrès Afcet 1993*, tome 4, p. 37-45.

[TOS96] Luca Toschi. « L'ipertesto d'autore. "La famiglia dell'antiquario" di Carlo Goldoni », in *Edizione elettronica su CD-ROM*. Venezia, Marsilio, 1996

[VIR93] Jacques Virbel, « Reading and Managing texts on the "Bibliothèque Nationale de France" Station », *The digital word : text based computing in the Humanities* (P. Delany and G. Landau eds.), MIT press, 1993, p. 31-52.