Contributions a l'indexation et a la reconnaissance des manuscrits Syriaques

Petra Bilane

N° d'ordre Année 2010

Thèse

Contributions à l'indexation et à la reconnaissance des manuscrits syriaques

Présentée devant
L'Institut National des Sciences Appliquées de Lyon

Pour obtenir
Le grade de docteur

École Doctorale
Informatique et Mathématiques

Par
Pétra BILANE

Soutenue le 23 juin 2010 devant la Commision d'examen

Composition du Jury

Jean-Yves RAMEL	Professeur à l'Université de Tours
Nicole VINCENT	Professeur à l'Université Paris–Descartes, Rapporteur
Mohamed KHOLLADI	Professeur à l'Université de Constantine, Rapporteur
Mohamed HASSOUN	Professeur à l'ENSSIB de Lyon
Robert LAURINI	Professeur à l'INSA de Lyon
Joseph MOUKARZEL	Université Saint-Esprit de Kaslik (Liban)
Hubert EMPTOZ	Professeur à l'INSA de Lyon, Directeur de thèse

Laboratoire d'InfoRmatique en Images et Systèmes d'information-UMR CNRS 5205

رقم التسجيل   سنة 2010

أطروحة

مساهمات في الفهرسة و البحث على المستندات المطبوعة

مقدمة أمام
المعهد الوطني للعلوم التطبيقية في ليون

للحصول على
لقب دكتور

مجال
المعلوماتية و الرياضيات

من قبل
علاء [illegible]

نوقشت في 23 حزيران 2010 أمام لجنة الامتحان المكونة من

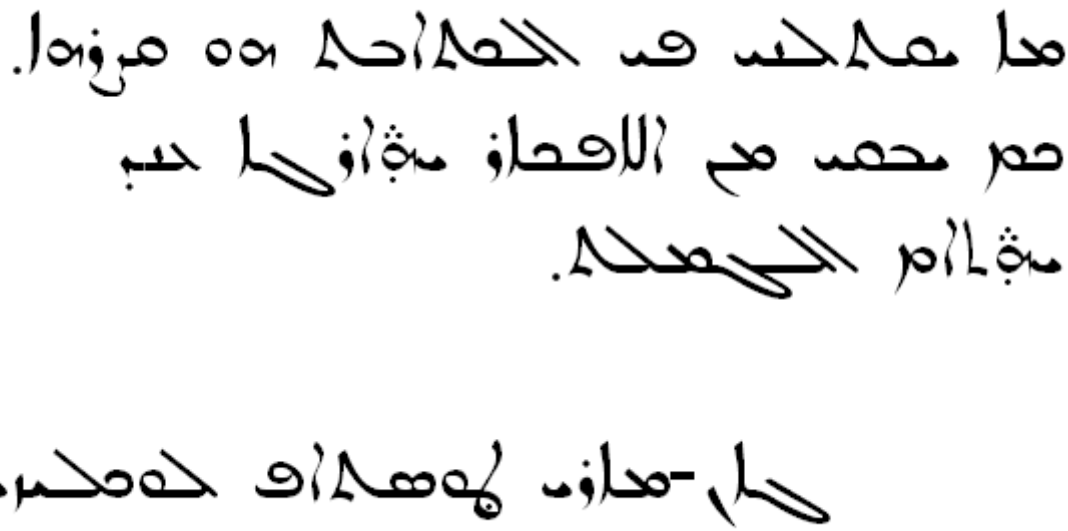
جان-مارك أوجيه	بروفيسور في جامعة لاروشيل، فرنسا
نيكول فنسان	بروفيسور في جامعة باريس-ديكارت، فرنسا
محمد شريت	بروفيسور في جامعة [illegible]، كندا
محمد [illegible]	بروفيسور في [illegible] في ليون
[illegible]	بروفيسور في المعهد الوطني للعلوم التطبيقية في ليون
جورج [illegible]	جامعة الروح القدس الكسليك (لبنان)
هوبير أمبتوز	بروفيسور في المعهد الوطني للعلوم التطبيقية في ليون

مخبر المعلوماتية الصورية و نظم المعلومات-وحدة رقم 5205

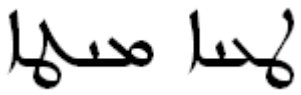
« Ce qui me tue, dans l'écriture, c'est qu'elle est trop courte. Quand la phrase s'achève, que de choses sont restées au-dehors ! »

Jean-Marie Gustave Le Clézio

. . . . . .



Remerciements



Mes remerciements vont tout d'abord à Pr. Hubert Emptoz qui a suivi et dirigé d'une façon continue mes travaux de recherche, pour la confiance qu'il m'a témoignée, pour la patience et la gentillesse qu'il a manifestées à mon égard, pour son encouragement et son soutien moral. Je lui exprime ma profonde gratitude pour son effort à organiser la soutenance de ma thèse avec un Jury d'experts.

Je remercie Pr. Jean-Yves Ramel, Pr. Mohamed Hassoun, Pr. Robert Laurini, et Dr. Joseph Moukarzel qui me font l'honneur de participer au Jury de ma soutenance.

Mes remerciements vont également aux rapporteurs, Pr. Nicole Vincent, et Pr. Mohamed-Khireddine Kholladi, qui ont accepté d'évaluer mon travail, et de participer au Jury de ma soutenance.

Je tiens à remercier Dr. Joël Gardes de France Telecom/Orange pour sa collaboration agréable et son partenariat.

Je voudrais aussi remercier Dr. Stéphane Bres pour l'aide qu'il m'a apportée dans les travaux liés au repérage de mots. Je remercie aussi Dr. Khalil Challita pour sa collaboration notamment dans les travaux effectués lors de mes séjours au Liban.

J'exprime mes remerciements au laboratoire LIRIS et à la Faculté des Sciences et de Génie Informatique de l'Université Saint Esprit de Kaslik au Liban pour m'avoir accueillie et offert un environnement de travail adéquat. Je remercie également l'ensemble des membres du laboratoire, maîtres de conférences, doctorants et docteurs, et stagiaires, pour m'avoir aidée chacun à sa façon.

Je voudrais remercier Dr. Joseph Moukarzel responsable de la Bibliothèque Centrale de l'Université Saint Esprit de Kaslik, qui m'avoir accueillie au sein de sa bibliothèque, et qui a mis à ma disposition des documents Syriaques numérisés de qualité sans lesquels le présent travail n'aurait pu exister.

Je remercie aussi Prof. George Kiraz de Gorgias Press qui m'a aussi fourni des documents Syriaques numérisés de qualité.

Je voudrais également remercier Ing. Pascal Damien pour son soutien moral, ainsi que mes amis Dr. Joseph Rahmé et Mlle. Mira Khoury pour leur aide et leur hospitalité durant mes séjours en France.

Je tiens surtout à remercier mon mari Dr. Charles Yaacoub, l'amour de ma vie, pour son appui infini, pour sa patience et sa compréhension malgré mes longs séjours en France et les distances qui nous ont tellement séparés, et pour avoir toujours été à mon écoute.

Je voudrais exprimer tout mon amour et toute ma reconnaissance à mes parents et mon frère qui m'ont aidée et encouragée durant mes longues années d'études et qui m'ont toujours soutenue moralement et financièrement.

Petra Bilane

Sommaire

Introduction

Le présent mémoire constitue un récapitulatif des travaux de thèse effectués au sein du laboratoire LIRIS en France et de la Bibliothèque Centrale de l'Université Saint-Esprit de Kaslik au Liban, dans un esprit de valorisation des anciens manuscrits, plus précisément de valorisation des anciens manuscrits Syriaques. Ces travaux de thèse ont été financés par l'Agence Universitaire de la Francophonie ; ils ont eu lieu dans le cadre d'une codirection entre le laboratoire LIRIS et l'Université Saint-Esprit de Kaslik.

D'un point de vue scientifique, ce travail de cette thèse consiste en une première exploration des manuscrits Syriaques pour en évaluer les potentialités de traitement et de valorisation par les méthodes relevant de la numérisation.

1 Le Contexte patrimonial

1.1 Le patrimoine Syriaque

Le Syriaque appartient à la branche Armaique des langues Sémitiques, c'est une variante de l'Araméen qui s'est répandue au début de l'ère Chrétienne. Le Syriaque a été faussement considéré comme étant une langue morte ; c'est probablement une des raisons pour lesquelles il a été mis jusqu'à présent de côté par les spécialistes de la reconnaissance d'écriture manuscrite et du traitement de document. Pourtant, c'est une langue assez fascinante qui combine, dans son écriture, d'une façon harmonieuse la structure de mot et la calligraphie Arabe avec cette particularité propre d'être intentionnellement penchée d'un angle d'approximativement 45°.

Les anciens manuscrits Syriaques constituent un trésor d'une valeur inestimable du point de vue textuel et intellectuel et du point de vue artistique. Ces documents constituent les piliers de rites religieux catholiques du Moyen-Orient. Il convient de rappeler que les premiers manuscrits Syriaques enluminés (qui datent du VI^{ième} siècle ap. J. C.) ont joué un rôle important dans le développement de l'iconographie.

Ces documents racontent l'histoire de l'Araméen et des langues Sémitiques, leur évolution au cours des siècles jusqu'à leur déclin. Ils constituent des traces écrites de tous les événements historiques dont ils ont été témoins et qui ont eu lieu dans la région du Moyen-Orient e.

Un tel patrimoine, comme d'ailleurs tout patrimoine manuscrit propre à une région, constitue une archive pour son histoire ; *c'est un héritage précieux à conserver et à transmettre aux générations à venir comme il nous a été légué par nos ancêtres.*

1.2 La conservation d'anciens documents

Les bibliothécaires et les détenteurs de fonds manuscrits sont très soucieux de tout ce qui concerne leurs précieux documents, notamment de la conservation de ces

documents. Malgré leur engagement et leur bonne volonté, les conditions de stockage ne sont pas toujours adaptées aux besoins de ces documents.

Certaines anciennes recettes d'encres requièrent des conditions de stockage assez particulières ; les conditions de température et d'humidité devraient être bien calculées afin de ralentir le processus de corrosion ; un document risque d'être détruit par sa propre encre si ces conditions ne sont pas assurées ; des conditions de conservation spécifiques sont requises pour certains documents déjà très fragilisés. Ces conditions ne peuvent pas toujours être assurées pour diverses raisons liées au contexte historique, social et politique de la région.

Certains documents qui sont rangés dans des boîtes en carton ou des folios, et mis à leur tour dans des placards non entretenus se trouvent exposés aux invasions des conditions extérieures (moisissure, insectes, mauvaise aération, humidité, etc…). D'autres documents, à qui l'on veut éviter d'être rongés par les mites ou par d'autres insectes, sont pulvérisés de sprays insecticides (qui causent des bavures d'encres et des tâches), ou bien encore sont bourrés de boules de naphta, produits pétroliers qui, à leur tour, rongent le papier et laissent des trous.

Ce patrimoine est alors confié à des dits « conservateurs », lesquels, s'avérant parfois assez maladroits en intervenant dans le but de limiter le dégât causé à ouvrage écrit, ne font qu'aggraver son état de détérioration. Des documents se trouvent même avec des pansements ou bien du coton, le tout recouvert de papier adhésif afin de couvrir les trous, ou bien avec une écriture, par-dessus une ancienne qui s'est effacée, effectuée avec un stylo et une couleur d'encre différents de l'originale, le texte n'est parfois pas le même que celui qui a été estompé.

La manipulation manuelle répétitive des documents fragiles risquant de les réduire en poussières, conduit les conservateurs à les séparer des autres documents et à n'autoriser leur accès qu'à de très rares mains expertes ; on se trouve ainsi dans une situation qui contredit le but visé par la création et l'édition du livre, c'est à dire la diffusion de la connaissance.

Pour les raisons mentionnées ci-dessus, et d'autres plus nombreuses, des solutions alternatives à ces méthodes de stockage s'avère alors nécessaire.

Une première alternative adoptée fut le micro-filmage, il y a quelques dizaines d'années. Cette technique consiste en un procédé de photographie du document en niveaux de gris ne tenant pas compte des couleurs ; bien que largement répandue parce qu'apportant un progrès considérable, cette méthode induit plusieurs désavantages que nous évoqueront ultérieurement dans les chapitres.

Une deuxième alternative à ce stockage est la numérisation qui consiste en une prise en image du document en utilisant un capteur et la création de fichier numérique correspondant. Ce dernier pourrait être dupliqué, copié, imprimé et diffusé sans risque de perte de qualité ; la numérisation, au delà de son intérêt pour la sauvegarde apporte une possibilité d'accès dont on n'a pas encore totalement tiré profit.

1.3 La numérisation

La numérisation a d'abord été perçue comme une méthode de sauvegarde du patrimoine écrit, une suite ou une alternative au micro-filmage. Ce n'est qu'ensuite

qu'on a commencé à appréhender la révolution qui y était associée. Numériser une page de document n'est pas en faire une copie comme la photographie (argentique) ou la photocopie ou encore le micro-filmage. *Le résultat de la numérisation est un fichier informatique, c'est une représentation ; cela nous apporte tout le potentiel de l'informatique et du numérique.*

La numérisation avec cette création de fichiers numériques nous permet la manipulation informatique des images des documents, elle ouvre ainsi les portes, au large, vers des applications d'indexation et de reconnaissance d'écriture, d'accès au contenu et finalement vers des *possibilités de réédition virtuelle enrichie.*

Dans les toutes dernières années du vingtième siècle on a entrevu l'avenir que la numérisation allait nous permettre de construire, mais aujourd'hui encore, en 2010, nous ne l'appréhendons pas encore dans sa totalité.

La numérisation est au centre d'une révolution, celle qui va nous faire passer de la culture de l'Ecrit à la culture du Numérique, le défi est considérable ; ce n'est pas celui de la technique mais de l'Homme. Comment va-t-il appréhender la numérisation, quels sont ses besoins et ses usages, évolutifs dans cette nouvelle situation ?

2 Les manuscrits Syriaques

2.1 L'expérience du LIRIS

J'ai souhaité venir faire ma thèse en France au LIRIS parce que j'avais eu connaissance des travaux de ce laboratoire et j'avais vu la diversité des corpus imprimés et manuscrits abordés ; j'ai pensé que ma thèse pourrait alors permettre des développements futurs de coopération avec le Liban pour sauvegarder notre important patrimoine écrit.

L'équipe Numérisation et Documents Ecrits du LIRIS a été constituée il y a plus de dix ans (d'abord dans le cadre du Laboratoire RFV, puis dans le LIRIS à partir de sa création, en 2004) par des chercheurs ayant deux métiers de base le traitement des images et la reconnaissance de formes. Elle est animée par le professeur Emptoz et comprend une douzaine de chercheurs (maîtres de conférences, post-doctorants et doctorants).

Ses travaux recouvrent toute la chaîne informatique qui va de la numérisation (la capture) à l'exploitation des contenus, son expertise relevant essentiellement de l'analyse de l'image et de la reconnaissance des différentes formes présentes dans cette image. Le titre du prochain ouvrage de Hubert Emptoz et Joël Gardes *« Les documents écrits : de la numérisation à la ré-édition virtuelle et augmentée »* résume les grandes lignes de l'activité de cette équipe qui s'est toujours intéressée aux besoins et aux usages des lecteurs.

L'équipe a notamment travaillé sur des grands *corpus patrimoniaux* dans le cadre de projets nationaux ou internationaux et apporté des solutions originales et innovantes.

a) Les premiers ouvrages imprimés de la renaissance

Ce travail s'est déroulé dans le cadre du projet européen DEBORA (pour Digital accEs to the BOoks of RenaissAnce) ; l'équipe a alors élaboré des méthodes de

compression par couches et proposé une nouvelle démarche pour passer du texte en mode image au texte en mode texte, démarche appelée *transcription assistée par ordinateur* (2000-2003)

b) Les grandes gazettes européennes du 17e et du 18e siècle.

Ces gazettes, et en particulier la Gazette de Leyde ont été réunies en collections sous forme de microfilms, il y a quelques dizaines d'années. Le LIRIS a travaillé sur la numérisation de ces microfilms et sur des méthodes de restauration qui permettraient de pouvoir utiliser les logiciels de traitement et de recherche d'information sur les images numérisées de ces microfilms (2005-2009). Un autre chantier a été consacré à la valorisation de ce corpus à partir de sa numérisation directe, Loris Eynard a contribué à ce chantier dans le cadre de sa thèse soutenue en 2009.

c) Les manuscrits du Moyen Age (Français)

L'étude informatique a été initiée dans le cadre du programme du CNRS Société de l'Information, avec l'Institut de Recherche et d'Histoire des textes (IRHT) et l'Ecole des Chartes. Pour accéder aux mots clés (donc indexer !), le LIRIS a conçu et réalisé un moteur de recherche iconique qui utilise des logiciels dits de Word Spotting et de Word Retrievel élaborés initialement pour l'écriture latine, (2004-2006).

Le travail d'exploration initié dans ce projet se poursuit actuellement dans le cadre du projet Graphem financé par l'ANR, l'un des défis de Graphem est de reconstruire une paléographie objective, morphologique et visuelle ; ce sera une étude paléographique assistée par ordinateur des écritures du Moyen Age.

d) Les manuscrits arabes subsahariens.

Ce projet concerne la sauvegarde, numérisation et valorisation de quelques centaines de milliers de manuscrits du 12 au 17e siècle, témoins des grandes civilisations africano-musulmanes et en danger aujourd'hui, notamment dans la région de Tombouctou au Mali. Il a reçu un soutien de la Région Rhône-Alpes dans le contexte des projets MIRA (2004-2007). Ce travail a aujourd'hui pour cadre le projet ANR VECMAS, Valorisation et Edition Critique des Manuscrits Arabes Sahariens (2009-2011)

e) les corpus régionaux,

Ce sont des corpus étudiés et valorisés dans le cadre du cluster recherche Culture Patrimoine et Création de la région Rhône-Alpes (archives de Savoie, de Châtillon sur Chalaronne, manuscrits de Stendhal, manuscrits de Berlioz) ; ce sont aussi des corpus détenus par des laboratoires régionaux.

f) des corpus scientifiques récents

Ils concernent essentiellement les mathématiques, mais aussi les sciences expérimentales.

Le laboratoire a mis en place une plateforme de numérisation conçue autour d'un numériseur Digibook fabriqué par la société I2S et considéré comme une des machines les plus performantes du marché.

Les chercheurs du LIRIS se sont rapidement intéressés au *Syriaque* parce que cette langue présente un certain nombre de caractéristiques au niveau de sa morphologie qui la distingue des langues descendant directement du latin et de l'arabe ; l'appréhension du penché à 45° spécifique de Syriaque a été pour eux un défi…à relever.

2.2 Les manuscrits syriaques

Les travaux du présent mémoire s'attachent particulièrement aux manuscrits Syriaques et font suite aux demandes des conservateurs et des fondateurs des projets de valorisation de ces documents. Ces derniers ont fait l'objet, jusqu'à présent, de très peu de travaux scientifiques bien que présentant une richesse de contenu et une diversité dans l'écriture assez intéressante.

Nous avons eu tendance dans ces travaux, à vouloir comparer les manuscrits Syriaques à leurs confrères Latins, vu le fait que ces derniers ont été l'objet de nombreux travaux de recherche dont une étude similaire à la nôtre, au sein du laboratoire LIRIS, dans le cadre d'une thèse effectuée par Yann Leydier [LEY06]. Les manuscrits médiévaux Latins possèdent une structure assez complexe et une façon de présenter le texte très particulière. La mise en page de ces manuscrits constitue une démarche complexe en tant que telle ; la présence de décorations et d'enluminures ainsi que la non-homogénéité dans la calligraphie et la disposition du texte rendent les outils génériques de traitement de d'analyse de documents développés dans la littérature quasi-inutiles ; c'est ce qui a conduit au développement d'outils spécifiques et adaptés à ces documents.

Les manuscrits Syriaques présentent à leur tour des difficultés similaires. L'écriture Syriaque est tracée en allant de droite à gauche, elle mélange naturellement l'aspect vertical (comme le latin) et un aspect penché (qui lui est propre). De même, la mise en page et la hiérarchie du texte dans ces documents sont présentées d'une façon qui leur est propre. Ces documents possèdent aussi un système de ponctuation et de délimitation de paragraphes qui leur sont spécifiques.

De toutes ces particularités il ressort que les méthodes et outils informatiques développés pour d'autres documents ne seront pas opérationnels pour le Syriaque. D'où le besoin de créer de nouveaux outils spécifiques pour l'analyse de la structure des documents Syriaques, ainsi que des outils de reconnaissance adaptés.

3 Les acteurs dans nos travaux

Comme mentionné précédemment, cette thèse a été effectuée dans le cadre d'une codirection entre le LIRIS (Laboratoire d'InfoRmatique en Images et Systèmes d'information), UMR CNRS 5205 et l'Université Saint-Esprit de Kaslik, dans le but d'apporter des bases pour la valorisation du patrimoine manuscrit Syriaque.

L'Agence Universitaire de la Francophonie (AUF), dans le but de promouvoir l'excellence scientifique et de développer des projets de conservation du patrimoine culturel, a assuré le financement des travaux de recherche dans cette thèse.

L'UMR LIRIS possède en son sein une équipe spécialisée dans le traitement et l'analyse de documents qui a déjà une longue expérience des corpus anciens ; de plus, cette équipe comprend des pionniers et des développeurs de méthodes innovantes dans le domaine de l'indexation d'anciens manuscrits ainsi que dans le domaine de la reconnaissance des structures et des écritures.

L'Université Saint-Esprit de Kaslik au Liban est représentée par sa Bibliothèque Centrale et sa Faculté des Sciences et de Génie Informatique qui ont fondé un projet de numérisation massif de l'archive manuscrit Syriaque de l'Ordre Libanais Maronite ; elles nous ont fourni des échantillons de documents numérisés de cette écriture, lesquels ont constitué la matière première autour de laquelle sont centrés les travaux de cette thèse.

Le département de numérisation de Gorgias Press aux Etats-Unis, a aussi lancé un projet universel de valorisation du patrimoine Syriaque, et nous a volontiers fourni des échantillons de documents Syriaques numérisés, ce qui a apporté une diversité dans la provenance et dans la manière de numériser la base de documents sur laquelle nous avons conduit nos travaux.

Les responsables des Bibliothèques ont aussi exprimé leurs besoins et leur opinion sur les intérêts des pistes de recherches suivies, le but essentiel de cette thèse est de pouvoir répondre au mieux à leurs attentes.

4 Contributions et Organisation du mémoire

Le présent mémoire est divisé en quatre chapitres précédés par la présente introduction.

Le Syriaque un patrimoine à sauvegarder

Le premier chapitre consiste en la description des documents Syriaques. Nous commençons tout d'abord par une vue macroscopique du Syriaque, par une présentation historique et géographique des langues Sémitiques, et de leur évolution à travers le temps arrivant jusqu'à la langue Syriaque, ses différentes calligraphies existantes et les manuscrits Syriaques célèbres. Nous allons aussi parler de la situation temporelle du Syriaque, de sa disparition pour une certaine période de temps puis de sa réapparition sous une autre forme ; où est le Syriaque aujourd'hui ? quels sont les projets en cours qui se rattachent à ce patrimoine ?

Nous allons faire un récapitulatif sur les méthodes de numérisation et une description des paramètres de qualité (résolution, et compression), ainsi que les différentes façons par lesquelles ils peuvent influer sur le résultat d'un algorithme

d'analyse de document. Nous allons aussi discuter du dilemme de la pérennité des supports d'enregistrement numérique.

Des structures du document à la segmentation en lignes et en caractères

Dans le second chapitre nous abordons la description du Syriaque écrit, par une vue microscopique où nous décrirons en détails les documents que nous avons entre les mains, du point de vue qualité visuelle, qualité de numérisation, état de conservation. Puis nous décrirons en détail la structure logique de ces documents en termes de mise en page, hiérarchie du texte, calligraphie, ponctuation, numérotation et pagination, les enluminures, et les réclames et gloses.

Nous développerons la méthode de prétraitement utilisée pour préparer le document pour les applications qui vont suivre. Ce prétraitement consiste en des transformations colorimétriques (binarisation, passage en niveaux de gris, etc…), et en une détection et correction de la penchée des lignes en utilisant la transformée de Hough.

Nous décrivons ensuite l'algorithme de segmentation d'une page de texte en ses lignes constitutives. Cet algorithme est basé sur le tracé du profil des projections horizontales qui détectent les indices de segmentation situés dans les espaces interlignes.

Ceci va conduire à une deuxième segmentation qui, partant d'une ligne de texte permettra de descendre autant que possible à l'élément lettre. Dans une ligne de texte émergent deux directions caractéristiques du Syriaque, une direction saillante oblique, des éléments verticaux. Les espaces inter-mots sont détectables par des bandes blanches verticales, certaines lettres sont aussi disposées verticalement. L'aspect oblique de l'écriture est visible lui aussi à travers des bandes obliques inter-mot et aussi intra-mot.

Afin de pouvoir segmenter une ligne en lettres ou caractères individuels, nous devons tenir compte des deux aspects de l'écriture sans segmenter l'un au détriment de l'autre. Nous avons donc procédé par une première segmentation verticale après laquelle nous nous sommes retrouvés avec des lettres individuelles verticales et avec des groupes de lettres penchées. Nous allons désormais appeler ces groupes de lettres des « n-grammes ». Ce sont ces « n-grammes » qui ont nécessité une deuxième segmentation dite « oblique ».

Il se peut qu'après cette deuxième segmentation il reste encore certains « n-grammes » non segmentés. Leur nombre ainsi que leur dimension « n » sont inférieurs à ceux restant après la segmentation verticale seule. Nous avons constaté que les « n-grammes » persistants sont stables et répétitifs. Ceci prouve que la méthode de segmentation utilisée n'est pas hostile à l'écriture syriaque ; bien qu'elle ne permette pas de segmenter tous les mots en lettres individuelles, elle permet tout de même de la segmenter d'une façon stable et régulière.

Indexation des images par repérages de mots

Dans le chapitre 3 nous abordons l'indexation du texte par une méthode de repérage des mots. Nous avons tout d'abord rappelé l'état de l'art effectué dans cette voie. Le processus commence à partir d'un mot pour lequel nous souhaitons retrouver

les occurrences, ce mot devrait être soumis à un moteur de recherche s'occupant de la fouille des données afin de nous renvoyer aux endroits dans lesquels se trouve notre requête. Le mot souhaité dit « mot requête » peut être saisi soit par pointage direct sur son image dans le document, soit écrit manuellement par l'utilisateur. Le moteur de recherche est basé sur une approche de fenêtres glissantes sélectives desquelles nous extrayons des caractéristiques directionnelles sous la forme de roses de directions pour le mot requête et que nous essayons de retrouver dans le reste du document. L'appariement est effectué par une correspondance entre les caractéristiques extraites du mot et celles extraites du reste du document (essentiellement des composantes de direction). En utilisant cette approche nous étions capables de retrouver toutes les occurrences du mot requête sans en rater aucune ; certains mots dits fausses occurrences ont été retenus, ces derniers bien que parfois différents dans leur sens de celui du mot cherché, présentent une similarité de forme et même certaines lettres en commun avec le mot requête. Nous avons appliqué un test de robustesse sur notre méthode en l'essayant sur des documents dégradés que nous avons synthétisés nous-mêmes.

De la transcription assistée à la transcription collaborative

Au chapitre 4, nous abordons le problème d'indexation par une approche constituant une première contribution à la transcription des manuscrits Syriaques. Nous nous sommes servis des résultats de la segmentation obtenus dans la première partie (graphèmes représentant des consonnes et des « n-grammes »), ces derniers nécessitent une classification selon leur ressemblance. Pour établir cette ressemblance nous utilisons des caractéristiques de profils, propres au Syriaque. Une fois classifiées, ces imagettes seront étiquetées et elles constituent alors une base d'apprentissage.

Pour la reconnaissance nous procéderons en deux passes successives. Dans la première, en utilisant une règle de la famille des k plus proches voisins nous affecterons à chaque imagette une valeur qui peut être un des caractères simples, un trigramme ou une indécision entre les deux consonnes a et l.

La seconde passe permet de remplacer chaque trigramme par ses composantes et permet de lever l'indécision entre les consonnes a et l en analysant la position dans le mot.

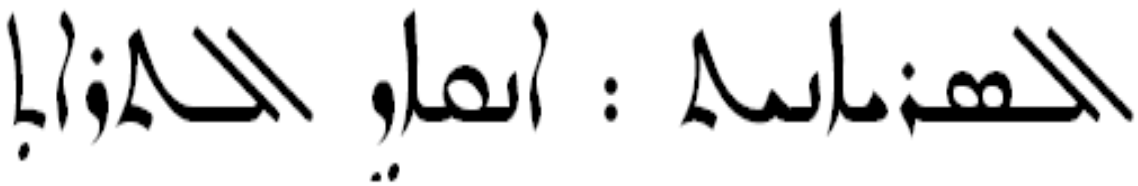
Conclusion

Dans sa simplicité et sa sobriété l'écriture syriaque porte suffisamment d'informations que l'on peut saisir en termes informatiques pour pouvoir mettre en place les techniques d'indexation par repérage de mots et de Transcription collaborative assistée. Notre travail d'exploration n'est qu'une première étude de faisabilité nous incite à vouloir aller plus loin.

Chapitre 1

Le Syriaque :

un patrimoine à sauvegarder.



1 Introduction

Le Syriaque appartient à la branche Armaique des langues Sémitiques, il désigne une famille de variantes de l'Araméen qui se sont répandues au début de l'ère chrétienne. A l'origine, le Syriaque était un dialecte Araméen localement parlé au Nord de la Mésopotamie. Avant la domination de la langue Arabe au Moyen Orient, il constituait la langue principale des communautés Chrétiennes de cette région.

Les documents écrits en Syriaque constituent bien un patrimoine écrit selon la définition de l'Unesco « *Héritage du passé dont nous profitons aujourd'hui (et) que nous transmettons aux générations à venir* ». Cet héritage, constitué au cours des siècles, a subi toutes sortes de circonstances difficiles et de dégradations, certaines dues aux phénomènes naturels, d'autres aux conditions de stockage, et d'autres encore aux problèmes créés par les Hommes.

Ce patrimoine écrit d'une très grande richesse est aujourd'hui en danger, il est menacé de disparition physique, des feuillets risquent d'être réduits en poussière lors d'utilisation par des chercheurs ; peu ou pas de mesures de préservation et de sauvegarde sont mises en œuvre.

Dans la première partie de ce chapitre nous présenterons le Syriaque, l'histoire de cette langue, sa calligraphie et les caractéristiques de son écriture, sa littérature et sa situation actuelle (langue encore pratiquée aujourd'hui, plus de deux mille ans après sa naissance).

Dans la seconde partie de ce chapitre nous aborderons la sauvegarde de ce patrimoine, plus exactement sa sauvegarde numérique et/ou virtuelle. La numérisation est souvent présentée comme la solution idéale, et pourtant, elle ne consiste qu'en une représentation et non une copie. Nous expliquerons quelles en sont les limites aujourd'hui, celles des fichiers numériques ; nous conclurons sur la potentialité de ré-édition enrichie et/ou augmentée, ce qui sera développé dans les chapitres suivants.

2 La Langue Syriaque

2.1 Etendue géographique et origine de la langue

Les langues Sémitiques avaient comme domaine géographique les pays de l'Est de la Méditerranée (le Liban, la Syrie, la Palestine), domaine qui s'étendait jusqu'en Mésopotamie à l'Est, en péninsule Arabique au Sud, et au Nord de l'Afrique à l'Ouest.

Le Syriaque, en tant que tel, a fait son apparition comme un dialecte non écrit de l'ancien Araméen parlé au Nord de la Mésopotamie. Après la conquête de la Syrie et de la Mésopotamie par Alexandre Le Grand, le Syriaque et les autres dialectes Araméens ont commencé à être transcrits par écrit. En l'an 132 avant J. C., le

royaume de Betnovin fut fondé en Edesse (actuellement Urfa, ville située au Sud de la Turquie, juste avant les frontières Syriennes), il adopta le Syriaque comme langue officielle. Les communautés parlant le Syriaque considèrent Edesse comme étant le berceau de leur langue. Le statut de langue officielle a permis au Syriaque d'acquérir une forme cohérente, un style et une grammaire, lesquels sont absents dans les autres anciens dialectes Araméens.

Pour ce qui concerne la zone géographique, le Syriaque était alors parlé en Turquie, Syrie, Irak, Palestine et au Liban. Pendant cette même époque, l'Hébreu était parlé en Palestine, et le Nabatéen en Jordanie et dans certaines régions Syriennes limitrophes de la Jordanie.

Autrefois, les frontières géographiques entre le Liban, la Syrie, la Palestine et la Jordanie n'étaient pas très bien définies. Le Christ est né en Palestine, cependant, il a passé toute sa vie dans la région correspondant à ces quatre pays. Le Syriaque est le dialecte Araméen le plus proche de la langue de Jésus Christ. L'idiome propre du Christ était le Syriaque mêlé à l'Hébreu, et rien ne contredit le fait que le Christ ait pu parler le dialecte Nabatéen, vu le voisinage géographique des deux pays et la proximité temporelle des deux dialectes, comme le montre la figure 1.

2.2 Ancêtres et descendants

Afin de bien situer historiquement le Syriaque par rapport aux autres langues Sémitiques, nous avons dressé un résumé de l'histoire des langues, en remontant dans

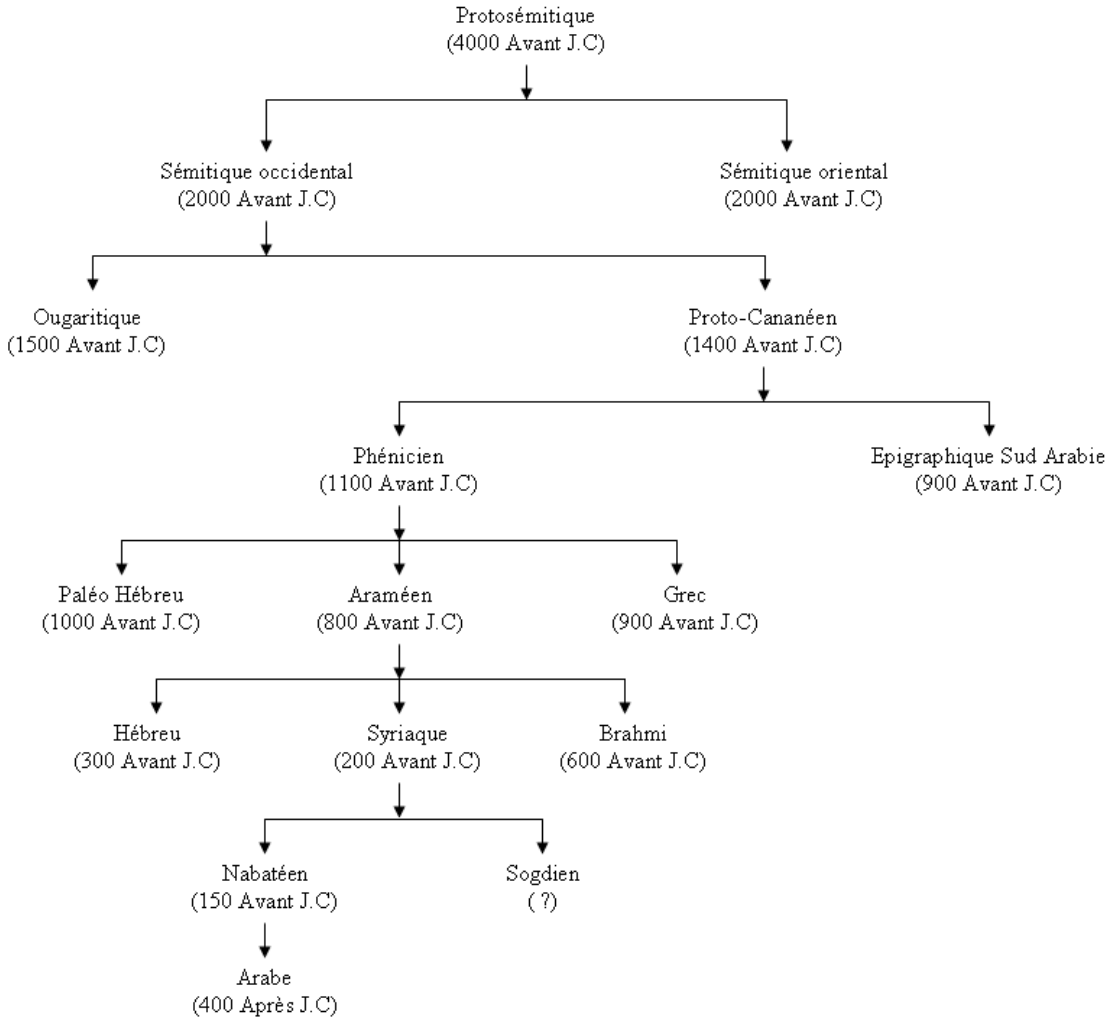


Figure 1: Arbre représentant l'histoire des langues.

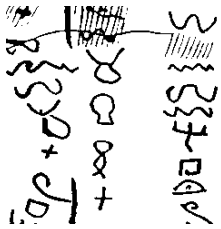
le temps jusqu'au Protosémitique et arrivant le plus récemment jusqu'à l'Arabe, dans un modèle temporel hiérarchique suivant un arbre indiquant les liens de parentés entre les langues Sémitiques. Cet arbre linguistique de l'histoire des langues Sémitiques est représenté dans la figure 1 ci-dessus.

Nous pouvons constater sur cette figure 1, que les langues Sémitiques ne se sont pas limitées aux régions et aux pays du Moyen-Orient ; les communautés parlant ces langues se sont déplacées, elles ont fait des voyages de commerce, transportant non seulement des marchandises mais aussi leurs langues.

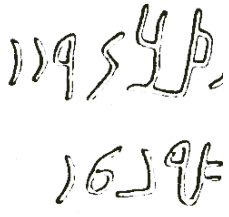
L'Araméen, en particulier, a été adopté comme le script officiel de l'Empire Perse, et de plus, il est considéré comme l'ancêtre de presque tous les alphabets modernes Asiatiques. Le Brahmi, descendant de l'Araméen, a voyagé le long de la route de la soie jusqu'au sous-continent Indien et il s'est répandu vers la Mongolie, le Tibet et l'Indochine. Le Sogdien, descendant du Syriaque, s'est étendu vers ce que nous connaissons aujourd'hui comme l'Ouzbékistan moderne ; ces langues sont les ancêtres de plusieurs autres langues parlées dans les différents pays de l'Asie de l'Est.

Dans la figure ci-dessous, nous présentons des échantillons d'écritures citées dans l'arbre historique précédent, ces illustrations montrent l'évolution de la forme de l'écriture au court du temps. L'écriture a commencé par des représentations picturales

de ce que l'Homme voyait autour de lui, de tels scripts sont qualifiés de pré-alphabétiques ; peu à peu, ces représentations se sont transformées en des formes syllabiques.



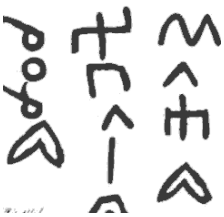
a) Proto-Sémitique.



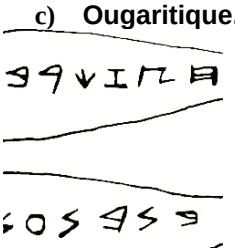
b) Sémitique.

c) Ougaritique.

d) Proto-Cananéen.

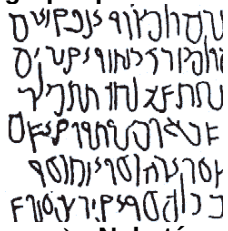


e)Epigraphique Sud-Arabie

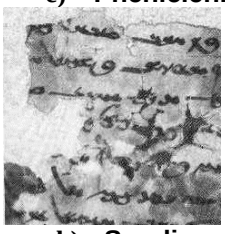


e) Phénicien.

f) Araméen.



g) Nabatéen.



h) Sogdien.

Figure 2: Extraits d'écritures anciennes.

2.3 Relation entre le Syriaque et l'Arabe

L'Arabe est actuellement la langue officielle des pays Arabes, sur une zone du Moyen Orient qui s'étend jusqu'en Irak (Mésopotamie), dans la péninsule Arabique et vers les pays du Nord de l'Afrique.

L'Arabe est une langue descendant du Nabatéen, qui, à son tour, est descendant du Syriaque. Le Nabatéen a fait son apparition vers le II^{ième} siècle avant J.C en Jordanie depuis la ville de Petra puis commença à disparaître vers le VI^{ième} siècle après J. C, date à laquelle la dernière inscription datée en Nabatéen fut retrouvée. La première inscription Arabe remonte au VI^{ième} siècle après J. C., elle fut retrouvée à Zabad en Syrie.

L'alphabet Arabe est dérivé de l'Araméen dans sa variante Nabatéenne ; l'alphabet Nabatéen était lui-même basé sur le Serto (une des trois calligraphies Syriaques desquelles nous allons parler plus loin dans ce chapitre). L'alphabet Arabe est un alphabet consonantique, qui dans sa forme la plus ancienne, suivait le même ordre alphabétique que les autres langues Sémitiques. Cet ordre a été modifié en lui conférant ainsi une certaine originalité.

Petit à petit, le domaine de la langue Arabe s'est étendu, vers le Nord en commençant par la Syrie qui partage une frontière avec la Jordanie, puis vers l'Ouest en Palestine, progressant ensuite vers le Liban et la Syrie depuis le nord de la

Palestine. Les derniers villages Libanais à avoir adopté l'Arabe à la place du Syriaque sont situés à l'extrême Nord du pays, juste avant les frontières Libano-Syrienne. Le peuple Libanais parlant le Syriaque, et ne voulant pas délaisser sa langue d'origine, en a développé une translittération ; c'est donc à ce moment que sont apparus certains textes Arabes écrits avec les caractères Syriaques; les scripts résultants de cette translittération sont dits Karshuni (ou Garshuni).

3 L'écriture

3.1 Calligraphie

Le Syriaque est écrit et lu de droite à gauche. C'est un script cursif écrit intentionnellement penché d'un angle d'approximativement 45°, dans lequel certaines lettres peuvent être connectées à l'intérieur d'un mot par des ligatures.

L'alphabet Syriaque de base est de type consonantique (Abjad) constitué de 22 consonnes écrites dans le même ordre alphabétique que celui des autres langues Sémitiques. A ces lettres peuvent s'ajouter soit 5 voyelles Syriaques, soit 7 voyelles Grecques, soit les voyelles orientales (une combinaison de points pour indiquer la prononciation correcte), soit les voyelles Arabes, soit les voyelles de Jacques d'Edesse qui sont les seules à être écrites comme des lettres séparées. Parfois, trois lettres Persanes, et trois lettres Sogdiennes peuvent se trouver dans un document Syriaque, mais c'est très rare.

Parmi les consonnes de l'alphabet Syriaque, trois peuvent être utilisées comme *mater lectionis* (mère de lecture) pour représenter des voyelles compte tenu de la difficulté dans la lecture d'un texte uniquement écrit avec des consonnes. Ce fait est visible dans les alphabets consonantiques (Abjad) comme étant une première forme d'indication de voyelles. Les trois consonnes en question sont Aleph, Yod, et Waw, et sont dites parfois des voyelles longues.

Nous pouvons distinguer trois calligraphies Syriaques :
- ***l'Estrangelo***, cette dénomination vient de la description Grecque de la forme de cette calligraphie « *στρογγυλη »* prononcé « *strongylê »* qui veut dire arrondie, cette calligraphie est la plus ancienne, et elle est tombée en désuétude,
- le *Syriaque Occidental* dit aussi ***Serto*** (ou bien Jacobite) qui est une écriture cursive et qui est la plus répandue,
- le *Syriaque Oriental* dit aussi ***Nestorien*** (ou bien Madnhaya ou Swadaya) qui est une écriture élégante, relativement carrée et réservée aux documents de valeur.

L'alphabet écrit avec les trois calligraphies est présenté dans le tableau ci-dessous.

Lettre	Estrangelo	Serto	Nestorien
Aleph	ܐ	ܐ	ܐ
Beth	ܒ	ܒ	ܒ
Gamal	ܓ	ܓ	ܓ
Dalath	ܕ	ܕ	ܕ
He	ܗ	ܗ	ܗ

Waw	ܘ	ܘ	ܘ
Zain	ܙ	ܙ	ܙ
Het	ܚ	ܚ	ܚ
Tet	ܛ	ܛ	ܛ
Yod	ܝ	ܝ	ܝ
Kaph	ܟ ܟ	ܟ	ܟ ܟ
Lamad	ܠ	ܠ	ܠ
Mim	ܡ ܡ	ܡ	ܡ ܡ
Nun	ܢ ـܢ	ܢ	ܢ ـܢ
Semkat	ܣ	ܣ	ܣ
Ein	ܥ	ܥ	ܥ
Pe	ܦ	ܦ	ܦ
Sad	ܨ ـܨ	ܨ	ܨ
Qoph	ܩ	ܩ	ܩ
Res	ܪ	ܪ	ܪ
Shin	ܫ	ܫ	ܫ
Taw	ܬ	ܬ	ܬ

Figure 3: Tableau illustrant les trois calligraphies Syriaques.

Les lettres reproduites dans le tableau sont présentées dans la forme qu'elles prennent quand elles sont séparées. Leur forme peut varier selon leur emplacement dans un mot, que ce soit au début, au milieu, ou à la fin, et aussi selon les lettres qui les précédent ou les suivent ; toutefois, le dessin de ces lettres suit un ordre de tracé prédéfini, comme le montre la figure ci-dessous.

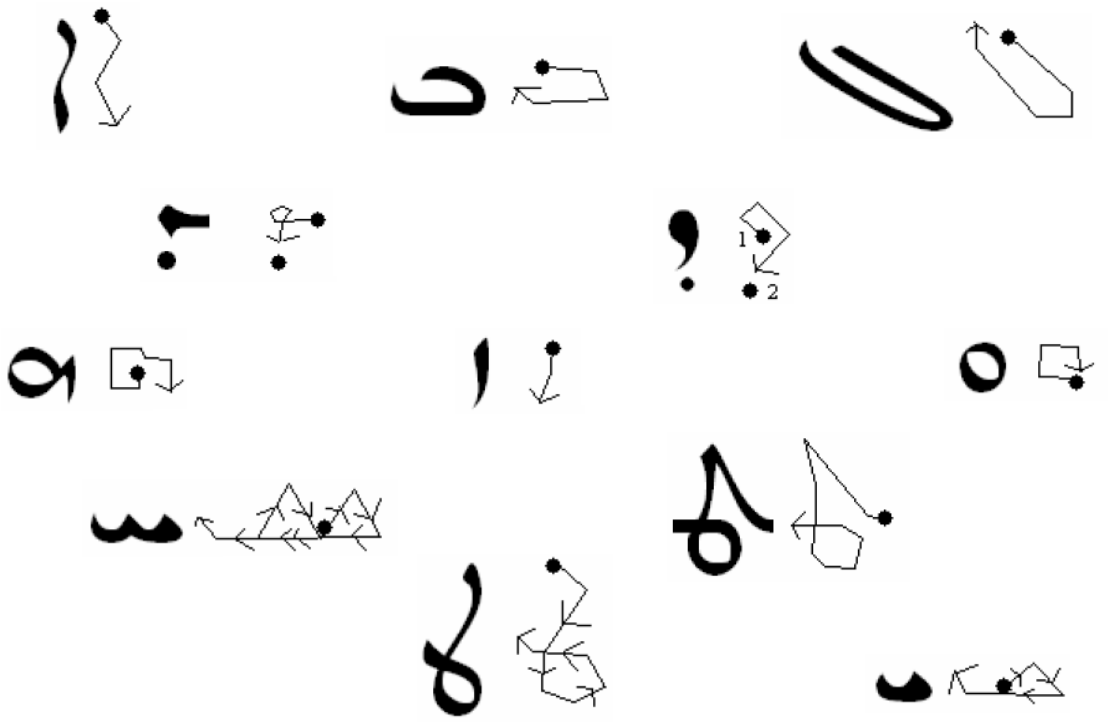


Figure 4: Ordre du tracé de certaines lettres Syriaques.

Les 5 voyelles Syriaques sont dessinées dans la figure 5 ci-dessous.

Figure 5: Les cinq voyelles Syriaques.

Les 7 voyelles Grecques sont représentées dans la figure 6 ci-dessous.

α ε η ι ο υ ω

Figure 6: Les sept voyelles Grecques.

Les voyelles Arabes sont dessinées dans la figure 7 ci-dessous.

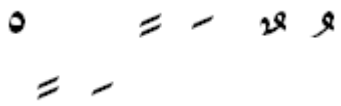


Figure 7: Les voyelles Arabes.

Les voyelles de Jacques d'Edesse sont représentées dans la figure 8 ci-dessous.

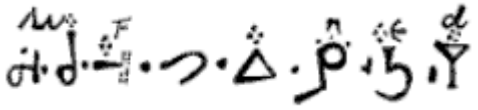


Figure 8: Les voyelles de Jacques d'Edesse.

Les trois lettres Persanes utilisées dans certains manuscrits Syriaques sont Bhath, Ghamal, et Dhalath respectivement représentées dans la figure ci-dessous.

Figure 9: Les trois lettres Persanes.

Les trois lettres Sogdiennes utilisées dans certains manuscrits Syriaques sont Zhain, Khaph, et Fe respectivement représentées dans la figure ci-dessous.

Figure 10: Les trois lettres Sogdiennes.

3.2 Système numérique

Il n'existe pas de chiffres Syriaques. Ce sont les lettres de l'alphabet qui jouent deux rôles, le premier est pour la représentation linguistique, et le deuxième pour la représentation numérique.

Lettre alphabétique	Valeur numérique	Lettre alphabétique	Valeur numérique
ܐ	**1**	ܠ	**30**
ܒ	2	ܡ ܢ	**40**
ܓ	3	ܣ	**50**
ܕ	4	ܥ	**60**
ܗ	5	ܦ	**70**
ܘ	6	ܨ	**80**
ܙ	7	ܩ	**90**
ܚ	8	ܪ	**100**
ܛ	**9**	ܫ	**200**
ܝ	**10**	ܬ	**300**
ܟ ܝ	**20**	ܟ	**400**

Figure 11: Tableau d'équivalence numérique des lettres de l'alphabet Syriaque.

Chaque lettre possède une valeur numérique, et les nombres sont représentés suivant un système numérique semblable au système Grec. La figure 11 ci-dessus montre pour chacune des lettres, la valeur numérique associée.

3.3 Translittération

Il y a 6 lettres de différence entre les alphabets Arabes et Syriaque, puisque l'alphabet Arabe est constitué de 28 consonnes alors que l'alphabet Syriaque est constitué de 22 consonnes. Pour pouvoir écrire le Karshuni, c'est-à-dire écrire l'Arabe

avec des lettres Syriaques, il faut rajouter 6 lettres pour compléter la table de translittération comme la montre la figure ci-dessous.

Caractère Arabe	Caractère équivalent Syriaque
ث	ܒ
ف	ܦ
خ	ܟ
ذ	ܕ
غ	ܓ
ڤ	ܒ

Figure 12: Table de translittération entre le Syriaque et l'Arabe.

4 Le Syriaque : une langue d'hier et d'aujourd'hui

4.1 La Littérature Syriaque

L'Histoire de la littérature Syriaque peut être divisée en trois périodes distinctes ; l'Age d'Or jusqu'au VII^{ième} siècle, la période Arabe jusqu'à l'an 1300, la période moderne de l'an 1300 jusqu'à nos jours. Il n'existe malheureusement pas d'introduction satisfaisante à la littérature Syriaque ; le recueil assez détaillé sur cette littérature écrit par Duval [DUV07], est toujours considéré comme la meilleure référence sur le sujet

L'Age d'Or

La littérature Syriaque en tant que telle n'a commencé à se développer qu'à partir du III^{ème} siècle après J. C.. Depuis le II^{ième} siècle après J. C, des textes de littérature Grecque ont été traduits en Syriaque. Le summum de la littérature Syriaque a été écrit 300 à 400 ans avant l'arrivée de l'Islam et l'expansion de la langue Arabe, cette période fut considérée comme l'Age d'Or de la littérature Syriaque, avec la création de plusieurs œuvres scientifiques, historiques, théologiques, etc…Deux célèbres écrivains se sont fait connaître, Aphrahat et Ephrem, ils sont considérés jusqu'à nos jours comme étant les meilleurs poètes Syriaques.

La période Arabe

Dans cette période, la littérature Syriaque a commencé son déclin face à l'expansion culturelle Arabe. La langue en tant que telle a servi d'intermédiaire pour

la communication et la culture des communautés Arabes quand elles se sont ouvertes aux connaissances Grecques. La traduction directe des textes Grecs en Arabe étant très difficile, ces derniers étaient en premier lieu traduits en Syriaque, puis ce dernier texte était traduit en Arabe. Ce transit linguistique a permis de développer une forme encyclopédique de la littérature Syriaque dans laquelle a excellé le célèbre polymathe Grégory Abul Faraj, connu sous le nom de Barhebraeus.

La période moderne

Pendant la période qui a suivi, le Syriaque a commencé à être réservé pour usage liturgique et à perdre son rôle comme langue d'échange. Toutefois, la littérature ne s'est pas éteinte complètement, des poètes et écrivains Syriaques modernes peuvent toujours être trouvés au Nord de l'Irak. Au début du XX^{ième} siècle, l'intérêt dans ce genre de littérature s'est renouvelé. Même au cours du siècle dernier plusieurs œuvres de littérature Européenne ont été traduites en Syriaque, à citer « The Merchant of Venice » de William Shakespeare, et « A Tale of Two Cities » de Charles Dickens.

4.2 Manuscrits Syriaques célèbres

Les manuscrits Syriaques les plus anciens peuvent dater du I^{er} siècle après J. C. La plus ancienne inscription Syriaque date de l'an 6, le plus ancien parchemin Syriaque connu porte des inscriptions concernant un contrat de vente d'esclaves en l'an 243, il fut trouvé à Doura en Syrie ; le plus ancien manuscrit Syriaque daté a été produit en l'an 411, et il est probablement le plus ancien manuscrit daté de toute langue.

Il existe environ quatre vingt anciennes inscriptions Syriaques datant des trois premiers siècles après J. C, ces inscriptions sont des exemples d'usage de la langue non lié au Christianisme.

Une des plus anciennes versions connues du nouveau testament est écrite en Syriaque, elle est dite « Peshitta ou Peshittô » (mot Araméen qui veut dire pur ou intact), elle date du V^{ième} siècle après J.C, elle a été traduite à partir de la version Grecque écrite en Koinè (une forme commune d'écriture Grecque ancienne vers l'an 300 av. J.C). Deux versions plus anciennes de la « Peshitta » ont été découvertes :

- la version "Syriaque Curetonienne" trouvée par William Cureton en 1842 en Égypte et qui daterait du V^{ième} siècle après J.C.,
- la version "Syriaque Palestinienne" sous la forme d'un palimpseste (un support d'écriture pré-utilisé sur lequel l'écriture a été grattée pour réécrire de nouveau), découverte en 1892 dans la Bibliothèque du Monastère Sainte-Catherine du Sinaï (Egypte) et datant du IV^{ième} siècle après J.C.

La Peshitta est la bible officielle adoptée par les Chrétiens d'Orient. Peshitta est un mot Araméen qui veut dire intact, c'est le nouveau testament, original et pur. La Peshitta est le seul texte authentique et pur qui contient les livres du nouveau testament qui ont été écrits en Araméen, la langue du Christ et de ses disciples.

Par contre, une controverse existe toujours à propos de la langue originale de la Bible entre les Chrétiens d'Orient et les Chrétiens d'Occident. Une partie des spécialistes pensent que la Bible Grecque provient de la traduction de textes Syriaques/Araméens antérieurs. La majorité des spécialistes pensent que la première version écrite de la Bible a directement été rédigée en Grec. À noter que, même dans la version Grecque, il existe des phrases Araméennes éparpillées dans le texte, en

particulier des phrases prononcées par le Christ et conservées dans la version originale pour des raisons religieuses. Il est cependant certain que le Christ a prêché dans la langue parlée par le peuple de l'époque.

4.3 Le Syriaque aujourd'hui

Une langue toujours vivante

Le Syriaque n'est pas une langue morte, le nombre de locuteurs actuels de la langue Syriaque est estimé, aujourd'hui, à 400 000 personnes ; elles résident au Liban, en Syrie, en Irak et en Turquie. Même aujourd'hui, 2000 ans après sa mort, la langue du Christ est bien encore vivante. Nous retrouvons aussi des communautés parlant le Syriaque en Amérique du Nord ou du Sud, et aussi en Europe, ceci est en fait la conséquence du mouvement d'immigration massive qui a touché les Chrétiens d'Orient.

Le Syriaque réservé principalement actuellement pour usage liturgique est la langue parlée d'un certain nombre d'églises d'Orient :

- Église Maronite
- Église Catholique Syriaque
- Église Syriaque Orthodoxe
- Église Catholique Syro-Malankare
- Église Syro-Malankare Orthodoxe
- Église Malabare indépendante
- Église Malankare Orthodoxe
- Église Malankare Mar Thoma

Le Syriaque et Unicode

Les polices de caractères Syriaques sont représentées par la rangée Unicode 0700-074F, et elles sont toujours en usage.

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
700	܀	܁	܂	܃	܄	܅	܆	܇	܈	܉	܊	܋	܌	܍	▯	T
710	ܐ	ܑ	ܒ	ܓ	ܔ	ܕ	ܖ	ܗ	ܘ	ܙ	ܚ	ܛ	ܜ	ܝ	ܞ	ܟ
720	ܠ	ܡ	ܢ	ܣ	ܤ	ܥ	ܦ	ܧ	ܨ	ܩ	ܪ	ܫ	ܬ	▯	▯	▯
730	◌ܰ	◌ܱ	◌ܲ	◌ܳ	◌ܴ	◌ܵ	◌ܶ	◌ܷ	◌ܸ	◌ܹ	◌ܺ	◌ܻ	◌ܼ	◌ܽ	◌ܾ	◌ܿ
740	◌݀	◌݁	◌݂	◌݃	◌݄	◌݅	◌݆	◌݇	◌݈	◌݉	◌݊	▯	▯	▯	▯	▯

Figure 13: Rangée Unicode représentant l'alphabet Syriaque.

Cette rangée comprend les lettres Syriaques, les signes de ponctuation, les voyelles orientales par système de combinaisons de points qui, selon leur disposition, indiquent la prononciation correcte ainsi que trois lettres Persanes (utilisées très rarement), et trois lettres Sogdiennes (utilisées très rarement). Le tableau ci-dessus montre la rangée Unicode en question, la calligraphie montrée est l'Estrangelo

4.4 ASCII, Unicode et UTF-8

Le code **ASCII** (*American Standard Code for Information Interchange*) est un standard américain qui utilise un octet pour définir la correspondance entre symboles et nombres jusqu'au nombre 127. C'est l'un des codes les plus utilisés, en particulier sur la plupart des ordinateurs mais il ne permet pas de représenter les lettres françaises accentuées et à fortiori les caractères cyrilliques ou syriaques. Il nous arrive souvent d'utiliser les codes de 128 à 255 pour les lettres latines accentuées, ces codes sont différents d'un pays à l'autre et donc pas pratiques pour échanger des documents.

Il a donc fallu construire un code plus pratique : ce sera l'**UNICODE**.

Au lieu d'utiliser les codes 0 à 127, il utilise les codes 0 à 65535 (en base 16 : de 0000 à FFFF). Le code UNICODE permet de représenter tous les caractères spécifiques aux différentes langues. De nouveaux codes sont régulièrement attribués pour de nouveaux caractères: caractères latins (accentués ou non), grecs, cyrilliques, arméniens, hébreux,…et syriaques.

Bien que la plupart des systèmes d'exploitation (Windows, Linux, MacOS X...) supportent déjà l'Unicode, ce dernier est encore peu utilisé par rapport à l'ASCII. Mais dans la pratique, c'est une autre paire de manches. Un caractère prend souvent 2 octets, un texte prend deux fois plus de place qu'en ASCII. C'est du gaspillage. Par exemple, si on prend un texte en français, la grande majorité des caractères utilisent seulement le code ASCII. Seuls quelques caractères nécessitent l'Unicode. On a donc trouvé un compromis : **l'UTF-8.**

Un texte en UTF-8 est partout en ASCII, et dès qu'on a besoin d'un caractère appartenant à l'Unicode, on utilise un caractère spécial signalant "*attention, le caractère suivant est en Unicode*".

En UTF-8 chaque caractère est codé par une rangée binaire dont la taille va de 1 à 4 octets (32 bits), selon la rangée Unicode il appartient, cette dernière est déjà représentée en hexadécimal (base 16).

Le remplissage de ces octets n'est pas arbitraire, chaque octet possède de 0 à 4 bits primaires de valeur (1) suivis par un bit de valeur (0) pour indiquer son type. Un nombre n de bits de valeur (1) indique le premier octet d'une séquence de longueur n octets, avec l'exception de 0 bits de valeur (1) qui indique une longueur de 1 octet, et de 1 bit de valeur (1) suivi par 1 bit de valeur (0) qui indique un octet de continuation d'une séquence à plusieurs octets (ceci assure la compatibilité avec le codage ASCII).

Ces bits pré-remplis sont dits les bits de contrôle. La figure ci-dessous permettra d'illustrer ce que nous avons expliqué ci-précédemment, les bits de contrôle sont montrés en gras, les bits « x » représentent les 8 bits de plus faible poids, les bits « y » représentent les 8 bits suivants, et les bits « z » représentent le reste des bits au poids le plus élevé.

Octet 1	**Octet 2**	**Octet 3**	**Octet 4**
0xxxxxxx	-	-	-
110yyyxx	**10**xxxxxx	-	-
1110yyyy	**10**yyyyxx	**10**xxxxxx	-
11110zzz	**10**zzyyyy	**10**yyyyxx	**10**xxxxxx

Figure 14: Table des séquences d'octets en UTF-8.

Codage des caractères Syriaques en UTF-8

Les caractères Syriaques occupent la rangée Unicode 0700-074F, ils ont alors besoin de 2 octets (16 bits) pour être codés en UTF-8, comme le montre la figure ci-dessous.

Octet 1	**Octet 2**
110yyyxx	10xxxxxx

Figure 15: Les deux octets requis pour représenter la rangée Unicode Syriaque.

Chaque caractère Syriaque est représenté par un nombre Unicode hexadécimal, nous allons représenter ce même nombre en binaire, afin de remplacer les bits « x » et « y » par les valeurs (0) et (1) correspondant à la transposition en binaire du nombre hexadécimal Unicode. Prenons par exemple la lettre Syriaque « j », inscrite dans la figure 24.

Figure 16: Lettre Syriaque « j » manuscrite.

Cette lettre est représentée en Unicode par le nombre hexadécimal 0713 lorsqu'elle est tapée sur un clavier, elle est illustrée dans la figure ci-dessous.

Figure 17: Lettre Syriaque « j » dactylographiée.

La transposition en binaire du nombre hexadécimal 0713 est 111 0001 0011, en remplaçant les bits « x » et les bits « y » comme nous l'avons expliqué précédemment la séquence UTF-8 représentant la lettre Syriaque « j » est donnée ci-dessous, les bits de contrôle sont en gras.

Octet 1	**Octet 2**
11011100	**10**010011

Figure 18: Codage en UTF-8 de la lettre Syriaque « j ».

5 Projets Syriaques actuels

5.1 Les projets

La Bibliothèque Centrale de l'Université Saint-Esprit de Kaslik au Liban a initialisé un projet de sauvegarde et de numérisation massive des documents Syriaques provenant des couvents de l'Ordre Libanais Maronite.

L'institut Syriaque Beth Mardutho cherche à favoriser l'étude et la conservation de l'héritage de la langue Syriaque et à faciliter l'accès aux personnes poursuivant des études sur ce patrimoine antique. Elle le fait par l'intermédiaire de plusieurs projets comme celui de la bibliothèque numérique Syriaque et de certaines ressources informatiques comme des polices de caractères Syriaques disponibles pour MS-Word ; elle organise annuellement une conférence sur les études Syriaques.

La maison d'édition Gorgias Press, publie annuellement des articles littéraires Syriaques, ainsi que des livres de grammaire, de vocabulaire et des ouvrages éducatifs pour ceux qui souhaitent apprendre la langue.

5.2 Provenance de nos documents

Dans ce mémoire, nos travaux consistent en l'indexation des images des anciens documents manuscrits. La difficulté se situe dans la possibilité de trouver suffisamment de documents sur lesquels travailler. Les bibliothécaires et les détenteurs de fonds manuscrits sont très réservés et discrets sur tout ce qui concerne leurs précieux documents.

Dans la littérature, les chercheurs ont tendance à choisir les manuscrits situés dans leur entourage, ou bien ceux en conservation ou restoration dans des musées dans leur pays, ou bien même ceux qui sont écrits avec leur langue mère ou une langue qui lui est ancêtre. En d'autres termes, les chercheurs travaillent sur les documents qui sont disponibles à l'endroit où ils se trouvent.

Les documents sur lesquels nous avons conduit nos recherches nous ont été fournis par Père Joseph Moukarzel, responsable de la Bibliothèque Centrale de l'Université Saint-Esprit de Kaslik et par Dr. George Kiraz fondateur de l'Institut Syriaque Beth Mardutho et directeur de la maison d'édition Gorgias Press.

5.3 Nature de nos documents

Nous disposons de documents rédigés avec la calligraphie Serto, et la calligraphie Nestorien. Nous ne disposons malheureusement pas de données avec la calligraphie Estrangelo, étant donné que cette calligraphie est la plus ancienne, et que les documents rédigés avec cette calligraphie sont très rares.

Nos documents proviennent des livres de liturgie, ainsi que de certaines Epîtres issues du Nouveau Testament ; ce sont des documents datant des débuts du XIX^{ième} siècle.

Certains sont numérisés en couleur, avec une résolution de 300 dpi (points par pouce) et sauvegardés en format JPEG ; d'autres sont numérisés en mode binaire inverse (fond noir et texte blanc), avec une résolution de 96 dpi et sauvegardés en format TIFF. Ces documents seront décrits plus en détails dans les chapitres qui vont suivre.

6 Les alternatives au stockage papier pour la sauvegarde

Dans l'introduction de la thèse nous avons décrit les modalités de stockage ainsi que les méthodes usuelles adoptées pour la conservation des documents. Les bibliothécaires, au milieu du XIX^{ième} siècle, sont devenus de plus en plus inquiets sur le sort de leurs précieux documents. Des alternatives aux méthodes traditionnelles de

stockage devraient alors être vite élaborées et mises en usage, sinon les documents seront perdus pour toujours.

Une première méthode pour garder une trace d'un manuscrit en danger est de le recopier, mais cette méthode reste manuelle, elle nécessite beaucoup de temps, et une connaissance approfondie du contenu du document à copier ; de plus, elle est dépendante des erreurs du copiste, ce qui pourrait rendre la copie infidèle au document d'origine.

Une deuxième méthode consiste à photographier le manuscrit qui nous intéresse. Par définition, prendre en photo un objet (quel qu'il soit) c'est *« l'emprisonner dans l'instant »*, comme si son cadre spatio-temporel se figeait soudain. L'objet dans la photographie ne « vieillira » plus jamais, et ne sera plus sensible aux conditions extérieures.

La première approche de photographie de documents fut le *micro-filmage*, la seconde qui est plus populaire est la *numérisation*.

Ces alternatives ont été conçues pour protéger les documents contre les éléments qui les détériorent et aussi pour réduire le contact direct Homme/papier. Les manipulations sont désormais effectuées sur les images des documents, ce qui permet de protéger les originaux. Ces derniers auparavant uniquement accessibles à quelques experts peuvent désormais, et sous forme non corrodable (à débattre), être à la disposition d'un public beaucoup plus large ; de plus, ces alternatives permettent un archivage compact, sans perte de place physique pour le stockage, contrairement au papier qui encombrait et alourdissait les étagères des bibliothèques.

6.1 Le micro-filmage

Le micro-filmage a été adopté comme un moyen économique pour l'archivage à long terme de documents. Il fut au début utilisé pour la préservation d'anciennes collections de journaux qui étaient condamnées à disparition ; ce fut le premier procédé photographique mis en usage pour des fins de préservation de documents. Peu à peu, son usage s'est répandu dans les différentes bibliothèques compte tenu des multiples avantages qu'il présentait.

Définition

Le micro-filmage consiste en un procédé de photographie de documents qui ne tient pas compte des couleurs. L'image résultant est en niveaux de gris, en mode positif ou le plus souvent négatif ; elle est stockée sur une bobine de film argentique ou bien un microfilm. A noter qu'un microfilm est un support de stockage analogique et non pas numérique.

Comme son nom l'indique, c'est une micro représentation de l'image du document. En général, les images sur un microfilm sont 25 fois plus réduites en termes de taille que celles du document d'origine. Cette méthode d'archivage présente de nombreux avantages et autant d'inconvénients.

Avantages

Parmi les avantages du micro-filmage, nous pouvons rappeler le fait qu'il va limiter le contact direct avec les documents fragiles ; toutes les manipulations sont

effectuées sur des copies imprimées à partir du microfilm en laissant les originaux intacts.

C'est aussi un moyen de stockage compact, les archives dans les bibliothèques augmentent en quantités chaque année, une fois ces archives microfilmées elles nécessiteront beaucoup moins d'espace que le papier.

C'est un support de stockage analogique qui ne nécessite donc aucun décodage informatique ni aucun usage de logiciel de lecture, il ne sera en aucun cas affecté par l'évolution technologique (matérielle et logicielle) qui donne sans cesse naissance à de nouveaux standards et formats, laissant les anciens tomber en obsolescence.

De plus, en termes de pérennité, un microfilm conservé dans de bonnes conditions a une durée de vie estimée à plus de 500 ans.

Désavantages

Bien qu'il ait rapporté une première alternative au stockage papier, ce procédé ne va pas sans défauts parmi lesquels le fait qu'il produit des images généralement très petites, ce qui ne permet pas leur lecture à l'œil nu ; un utilisateur a besoin de faire une projection de l'image pour la voir en taille réelle sur un écran, les lecteurs utilisés pour visualiser les images sur microfilm sont difficiles à manipuler de plus ils deviennent de plus en plus chers, et de moins en moins présents sur le marché, par contre, cet obstacle pourrait être franchi en substituant les lecteurs par de simples systèmes de projection et d'agrandissement.

Compte tenu du fait que c'est un support d'enregistrement analogique, les images sur la bobine sont rangées l'une à la suite de l'autre ; un accès direct à une image particulière n'est donc pas possible ; ceci nécessite un embobinage et un dé-bobinage soigneux jusqu'à arriver à l'image cherchée, par conséquent, un lecteur ne pourra pas comparer deux images enregistrées sur une même bobine, une recherche automatisée serait encore plus difficile à réaliser. En effet, pour faciliter la recherche et l'accès à des images particulières, le fonds à conserver devrait passer par une étape de classements et de tris préalables à la prise des images.

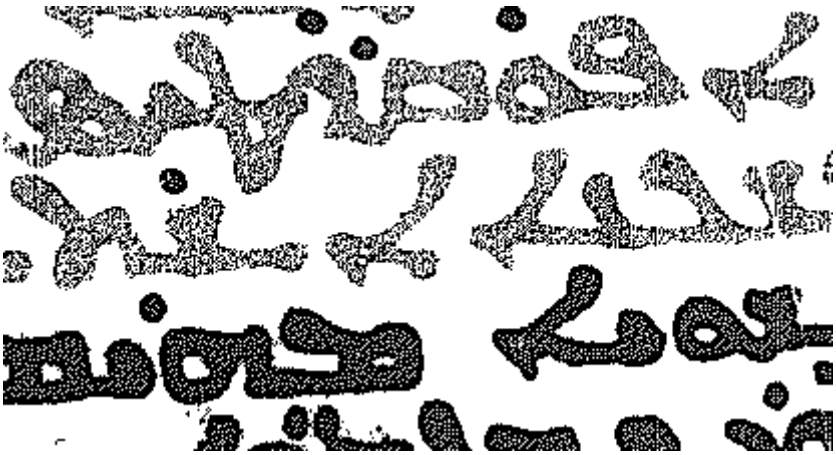


Figure 19: Disparition d'une couleur sur une image de microfilm.

C'est un procédé de photographie en niveaux de gris, la perte au niveau couleur est très importante, parfois même certaines couleurs paraissent en gris très clair que nous n'arrivons plus à les distinguer comme le montre la figure ci-dessus.

Il existe un micro-filmage couleur, mais il est très coûteux, de plus les teintures de photographie couleurs tendent à devenir fades avec le temps, ce qui constitue une perte d'information.

Le résultat d'impression se dégrade au fur et à mesure du copiage et de la reproduction puisque le support de stockage analogique. Le support utilisé noircit alors que l'écriture ternit ou bien devient estompée, ceci est dû au fort éclairage utilisé lors de la prise en photo du document pour augmenter le contraste entre le texte et l'arrière plan ; ceci réduit ainsi considérablement la lisibilité du document, comme le montre la figure ci-dessous.

Figure 20: Lorsque le fond devient plus foncé que l'écriture.

6.2 La numérisation

La numérisation a aussi été adoptée comme un moyen de sauvegarde, une alternative au stockage papier. La numérisation, comme le microfilmage peut contribuer à la préservation des documents manuscrits, et à la « prolongation de leur durée de leur existence ».

On a cru que les images numérisées pourraient témoigner de la présence des livres longtemps après la disparition de leur version papier. Aujourd'hui on s'est aperçu que la durée de vie des documents numériques n'était pas infinie. Ce processus a aussi permis la création de bibliothèques électroniques, facilitant la diffusion des documents anciens et rares à un public plus large, sous une forme imprimable sans limite et sans perte de qualité visuelle.

Définition

La numérisation est la prise en image du document, elle est effectuée par un scanneur ou un numériseur qui convertit les niveaux de lumière et les couleurs en des données binaires, le résultat est stocké dans un fichier informatique.

La numérisation permet de rendre les documents « compréhensibles » par un ordinateur, contrairement à son prédécesseur le micro-filmage qui, lui, produit une représentation analogique du document. A noter que l'image d'un texte n'est pas du texte, c'est une représentation de la réalité, et c'est une agglomération de pixels d'une certaine couleur sur un fond d'une couleur (de préférence) différente.

Il est donc crucial de créer une image de qualité fidèle autant que possible au document d'origine, tout en essayant de conserver ses détails les plus fins. En effet, l'image créée par le numériseur ne possède pas exactement la même « richesse » de contenu que celle du document réel. L'œil humain est facilement trompé par cette « illusion visuelle », puisque même en utilisant un numériseur couleur, certains seuillages sont effectués sur le spectre réel des couleurs. L'œil humain est indifférent par rapport à ces troncatures puisqu'il est incapable de voir toutes lesdites couleurs, ces dernières sont réduites à des mélanges de dosages de trois couleurs ; rouge, verte, et bleue.

Une image pourra être prise en utilisant un capteur, et nécessite alors une certaine qualité de saisie. Plusieurs critères sont alors à envisager ; la compression, la

résolution, et enfin le choix du support d'enregistrement. A noter que ces critères sont étroitement liés à l'usage auquel sera destiné la version numérisée du document.

6.3 La compression

Définition

La compression consiste à réduire l'espace nécessaire à la représentation de l'image, elle a pour utilité de réduire la redondance des données d'une image afin de pouvoir l'emmagasiner sans occuper beaucoup d'espace ou de la transmettre rapidement. Une image peut bien être stockée dans son état brut mais c'est couteux en termes d'espace de stockage.

La compression sans perte

Une première technique de compression dite **RLE** (Run Length Encoding) consiste à remplacer toute suite de bits ou de caractères identiques par un couple de caractères, le premier indiquant le nombre d'occurrences et le deuxième le bit ou le caractère répété ; cette technique est assez bien adaptée aux images binaires ou monochromes.

Une deuxième technique, c'est la compression **LZW** pour Lempel-Ziv-Welch, dite aussi de type dictionnaire. Elle est basée sur le fait que des motifs se retrouvent plus souvent que d'autres et qu'on peut donc les remplacer par un index dans un dictionnaire. Le dictionnaire est construit dynamiquement d'après les motifs rencontrés. Cette technique est utilisée dans les formats **GIF** (Graphics Interchange Format) et **PNG** (Portable Network Graphics). C'est une méthode de compression sans perte. Il y a autant d'information après la compression qu'avant, elle est seulement réécrite d'une manière plus concise.

La compression avec perte JPEG

Une technique de compression avec perte est une technique qui entraîne des modifications sur les données. Le format comprimé avec perte le plus connu est **JPEG** (Joint Photographics Experts Group).

JPEG commence par découper l'image en blocs ou carreaux généralement carrés de 64 (8×8) ou 256 (16×16) pixels, puis effectue sur chaque bloc une transformée DCT (Discrete Cosine Transform), il sauvegarde ensuite les basses fréquences, éliminant ainsi les détails de l'image. Pour l'affichage, les fréquences subissent la transformée inverse et sont affichés comme pixels.

Cette méthode de compression introduit plusieurs dégradations De nombreux artefacts sont visibles sur l'image finale, réduisant fortement la lisibilité d'un document. La compression JPEG offre plusieurs niveaux de perte en fonction du nombre de hautes fréquences rejetées, par conséquent, on peut choisir la qualité désirée pour le fichier final.

Un format modifié du JPEG a été développé pour fournir les mêmes taux de compression, mais avec moins de pertes d'informations de l'image ; ce format est le format JPEG2000.

Ce qui est utilisé

Les bibliothécaires et les détenteurs des fonds manuscrits, soucieux de la préservation de leur patrimoine, voudraient stocker la plus grande quantité de

documents sur le moindre espace possible (dû la plupart du temps en une carence en capacité de stockage). De manière générale, la compression JPEG est choisie bien qu'elle ne soit pas adaptée aux images de documents (des images de traits !)

Malgré la préservation apparente de l'écriture, des artefacts sont introduits réduisant fortement la lisibilité du document, et parfois même dissolvant des parties fines de l'écriture, comme c'est le cas pour les manuscrits Latins, certains caractères Latins étant reliés par des ligatures (fins traits horizontaux).

Les dégradations résultant de cette compression sont irréversibles ; des chercheurs ont tenté une restauration qui consiste par un lissage des artefacts, parfois même des tentatives de restituer les portions d'écriture dissoutes par des approches morphologiques comme dans [ZHE01], ou bien par l'utilisation de contours actifs comme dans [ALL02] [ALL06]. Ces tentatives finissent toujours par confronter le paradigme *« Reconnaître pour restaurer et restaurer pour reconnaître ».* A noter que toutes ces approches tendent à nettoyer les effets JPEG, et à rendre plus agréable la lecture des images bien que l'information perdue ne puisse pas être restituée.

Effets indésirables de la compression avec perte

Afin de montrer les effets néfastes que pourrait introduire ce genre de compression, nous avons choisi une image d'un mot de nos documents que nous avons comprimée en JPEG de qualité 0, et le résultat est visible dans la figure ci-dessous

a) Image d'un mot avant compression.

b) Artefacts introduit par compression avec perte.

Figure 21: Effets indésirables de la compression visibles sur une image comparée avec sa version non comprimée.

Ce qui est remarquable est le fait que, bien que des artefacts soient visibles sur le contour des lettres, l'écriture Syriaque soit suffisamment épaisse pour résister à la dissolution, contrairement à l'écriture Latine qui présente des ligatures et des traits suffisamment fins pour être dissous par une telle compression.

6.4 **La résolution**

Définition

Une image numérique est constituée d'unités de forme carrée dites pixels. Ces unités n'ont pas une taille fixe. La résolution, dite aussi la définition, de l'image indique le nombre de pixels par unité de longueur et elle est exprimée en points par pouce (dpi dots per inch), plus ce nombre est grand, plus la taille des pixels est petite, et plus de détails sont détectés.

Le choix de la résolution adéquate

Comment choisir la résolution convenable ? Plus la résolution est élevée, plus l'image est détaillée, et plus le fichier final est volumineux. Dans l'absolu il n'existe

pas une « bonne résolution » valable pour toutes les images, mais plutôt une « résolution adéquate ».

Le choix de la résolution n'est pas absolu, il faut choisir la résolution convenable qui soit le mieux adaptée au document en question. Un document écrit en Syriaque présente une épaisseur de trait d'écriture de 1mm, par contre un document écrit en Latin présente une écriture plus fine de l'ordre du $10^{\text{ième}}$ de mm, les numériser tous les deux à une même résolution élevée permettrait de bien capter le plus fin détail du document Latin, mais serait considéré comme du gaspillage par rapport au document Syriaque ; si les deux sont numérisés avec une faible résolution, celle-ci pourrait s'avérer suffisante pour le document Syriaque mais insuffisante pour le document Latin.

Le théorème d'échantillonnage « Nyquist-Shannon »

Pour répondre de manière scientifique à cette question, nous avons recours au théorème d'échantillonnage de Nyquist-Shannon : *« la fréquence d'échantillonnage d'un signal doit être égale ou supérieure au double de la fréquence maximale contenue dans ce signal, afin de convertir ce signal d'une forme analogique à une forme numérique ».* Appliqué au traitement d'images ce théorème devient : *« Pour ne pas perdre de détail dans une image, la taille des pixels doit être inférieure à la moitié de la taille du plus petit détail de l'image ».* Ceci est bien justifié par le fait que, dans une image, ce sont les fréquences élevées qui sont porteuses des détails.

Prenons le cas d'un document Latin dans lequel l'épaisseur du trait d'écriture est de l'ordre de 0,1 mm, quelle serait alors la résolution minimale requise pour ne pas risquer de perdre ce trait dans l'image numérique ? En appliquant le théorème de Nyquist-Shannon, la taille du pixel t_{pixel} devrait être :

$$t_{pixel} \le \left(\frac{1}{2} \times \frac{1}{10} = \frac{1}{20} \right)^{\text{ième}} \text{mm} = 50 \text{ microns}$$

Donc la résolution minimale requise est de r :

$$r \ge \left(\frac{25400}{50} = 508 \right) \text{dpi}$$

Par contre pour l'image d'un document Syriaque dans lequel l'épaisseur du trait d'écriture est de 1 mm, en utilisant la même équation que précédemment, la résolution minimale requise serait de 50,8 dpi.

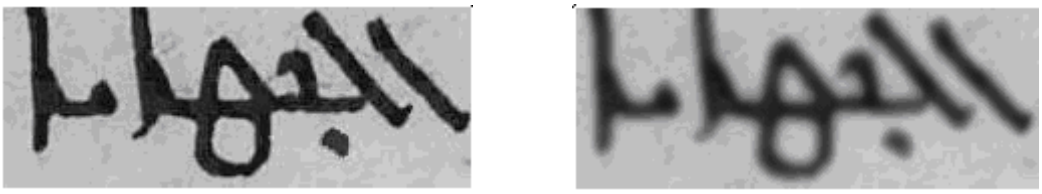
Ce qui est utilisé

Avec la crainte de ne pas avoir suffisamment d'espace de stockage, les chargés de numérisation dans les bibliothèques tendent à baisser la résolution du numériseur. Les tentatives de rétablissement de la faible résolution consistent en un échelonnage forçant ainsi une résolution plus élevée sur l'image d'origine, la plupart des approches se base sur une interpolation linéaire ou bien sur une « cubic spline expansion ». Le résultat est une version floue de l'image d'origine. Thouin et al proposent de restaurer des images à faible résolution en se basant sur la résolution itérative d'un problème non linéaire d'optimisation [THO00].

Effets indésirables de la faible résolution

Afin de montrer l'effet de flou, nous avons imité ce type de dégradation en baissant la résolution d'une image de test puis la restituant de nouveau à sa valeur

initiale par interpolation cubique, l'effet flou introduit est illustré dans la figure ci-dessous.



a) Image d'un mot avant la baisse de la résolution.

b) Effet de floue introduit par la restitution d'une faible résolution.

Figure 22: Effets indésirables de la restitution d'une faible résolution visibles sur une image comparée avec sa version intacte.

Dans ce cas de dégradations, l'écriture Syriaque s'avère suffisamment épaisse pour garder sa lisibilité malgré l'effet de flou ; une écriture Latine aurait été entièrement dissoute dans de telles conditions.

6.5 Le choix du support

Après l'obtention d'un fichier image issu du numériseur, il faudra enregistrer ce dernier sur un support. Un disque compact constitue un support de stockage à prix abordable, facilement lisible et quasiment inaltérable. Les lecteurs de ce type de support sont aussi disponibles et à portée de main de tout utilisateur (pour le moment). Ce n'est pas étonnant qu'il soit le support d'enregistrement numérique le plus populaire (actuellement). C'est un support multi-usage capable de supporter en même temps des fichiers de formats et de natures différents. La duplication répétitive des fichiers stockés sur un disque n'entraîne en aucun cas une dégradation dans l'image de sortie, contrairement à un support analogique tel le microfilm. C'est un support facilement renouvelable, transférable, et peu encombrant (comme son nom « compact » l'indique).

7 La sauvegarde du Numérique

Il serait plus précis de parler de la pérennité et de la durée de vie des supports d'enregistrement.

Les anciens documents manuscrits existent, pour certains, depuis de longs siècles. Leur survie et leur endurance face aux dégradations sont en grande partie dues à la résistance des supports d'écritures (les parchemins issus de peaux animales), et à la qualité des anciennes recettes d'encres, à base de métaux lourds, de teintures végétales, de sèves animales, et de résines qui leur donnent une certaine indélébilité. Le papier et les encres modernes sont loin d'être aussi durables que leurs ancêtres, par exemple, le phénomène de transvision (passage du verso sur le recto) est de plus en plus présent dans les documents récents.

Lorsque ces documents seront numérisés et stockés sur un espace virtuel, quelle sera alors la durée de leur vie que l'on peut espérer ? Est-ce que la mission de sauvegarde du patrimoine manuscrit s'arrête là ? Afin de répondre à ces questions, nous devons introduire la notion de la « préservation numérique », des menaces

pouvant atteindre les supports de stockage numérique, et des stratégies étant nécessaires pour assurer leur longévité.

7.1 La préservation numérique

La préservation numérique consiste en la gérance ou la gestion des données numériques au cours du temps. Dans le contexte de numérisation de documents, c'est un ensemble d'actions, d'outils et de méthodes qui assurent et/ou garantissent la possibilité d'un accès permanent aux versions numériques des manuscrits.

Cette préservation est essentielle pour les documents numérisés pour lesquels l'équivalent analogique ou bien la version originale en papier n'existe plus. La seule trace que nous avons de ces documents se trouve sous forme de son image sur un disque compact ; si cette dernière se trouve corrompue ou bien perdue, le document en question est à jamais disparu.

Il s'avère alors nécessaire de protéger ces données, en protégeant le support de sauvegarde contre les conditions qui peuvent le dégrader. Les menaces pouvant atteindre un support pareil sont de deux natures ; physique et logique.

7.2 Les menaces aux supports numériques

Menaces physiques

Le premier défi auquel est confrontée la préservation numérique est le fait qu'un support de sauvegarde numérique est plus vulnérable aux détériorations et à la perte d'information qu'un support analogique tel que le papier. En effet, un papier subit des détériorations liées à la fragilisation et au jaunissement, ces dégradations évoluent lentement et ne sont visibles qu'après au moins quelques décennies. De plus, il est encore possible de récupérer l'information sur le support après détection des dégradations.

Un support de sauvegarde numérique se détériore à un rythme plus accéléré, et une fois que la corruption commence à l'atteindre, nous pouvons constater qu'il y a déjà une perte de données irrécupérables ultérieurement.

Le disque compact est le support de stockage numérique le plus utilisé. La triste réalité c'est que nous savons bien qu'un disque compact gravé ne durera jamais 100 ans, la durée de vie moyenne d'un disque compact est estimée à une dizaine d'années.

Parmi les détériorations physiques qui peuvent atteindre le disque compact nous pouvons citer ; l'oxydation, les stries résultant d'un accident ou d'une intention nuisible, l'exposition à une source d'ondes électromagnétiques, la poussière, l'humidité, l'exposition à de fortes variations de température, les dégâts résultant d'un mauvais usage (par exemple placer le disque à l'envers dans le lecteur), etc…

Menaces logiques

Un second défi peut être plus dangereux que le premier consiste en le problème d'accès à long terme aux données numérisées ou bien l'obsolescence numérique ; au bout de combien de temps le support de stockage numérique ainsi que le format du fichier utilisé seraient ils considérés comme obsolètes ? Ce problème est rappelé à l'attention des bibliothèques effectuant des projets de numérisation au début des années 1990.

Les technologies informatiques avancent d'une façon très accélérée en synchronisant les deux côtés matériel et logiciel. Les nouveaux supports numériques sont de plus en plus fiables, de moins en moins chers et ils offrent des espaces de

stockage de plus en plus volumineux. Leurs versions plus anciennes sont alors vite dépassées (par exemple, personne n'utilise actuellement une clé USB (Universal Serial Bus) ayant une capacité de 16 Méga Octets).

De nouveaux formats d'enregistrement nécessitent des modifications sur les méthodes de gravure, engendrant ainsi des modifications sur les méthodes de décodage et d'extraction des données, et par la suite la fabrication de nouveaux graveurs et lecteurs. Les anciennes technologies se trouvent éliminées de la chaîne de production. Les enregistrements effectués sous ces technologies courent le risque de ne plus être tangibles, les données devenant désormais inaccessibles.

Il n'existe plus de lecteurs pour un disque compact gravé en 1985 ; de même, il n'existe plus de lecteurs pour les disquettes Zip.

Quel est alors le sort des données qui sont stockées sur des tels supports ? *A quoi sert-il de les avoir pour un siècle si nous ne disposons plus de moyens pour les lire* ?

7.3 Les stratégies de longévité

Afin de pouvoir trouver des solutions aux problèmes mentionnés ci-dessus, une évaluation des risques devrait être effectuée. Le risque primordial est la perte de données, ce que l'on cherche à contourner au mieux. Pour ce faire, plusieurs stratégies ont été élaborées.

Le renouvellement du support d'enregistrement

Ce procédé consiste à recopier les données d'un support d'enregistrement sur un deuxième du même type. Aucun changement ou altération n'est requis au format des données. Cette méthode est nécessaire face à la détérioration et au vieillissement du support.

La conversion

Ce procédé consiste à transférer les données d'un format à un autre, ou bien d'une ancienne version d'un format vers une nouvelle. Les données transférées d'un format à un autre passent à travers un convertisseur, ce dernier peut ne pas être capable de capter et d'interpréter toutes les nuances du format de départ, on parle ici de non fidélité au contenu. Des données transférées d'une ancienne version d'un format vers une nouvelle pourraient perdre certaines de leurs propriétés.

La duplication

Cette méthode consiste à effectuer plusieurs copies des données sur différents supports et différents systèmes d'exploitation, et aussi plusieurs copies du même support. En d'autres termes ne pas ranger les données en un seul endroit, elle permet de garder à l'abri un « backup » des informations au cas où d'autres versions se trouvent endommagées. Si ces dernières se trouvent uniquement en un seul endroit, elles courent le risque d'effondrement logiciel et/ou matériel, et éventuellement la perte.

L'émulation

L'émulation consiste en la duplication des propriétés de formats obsolètes. Cette méthode effectue une analyse des données dans leurs détails de base, elle nécessite une étude de la façon avec laquelle ces données ont été codées, ainsi que la séquence de sauvegarde adoptée. C'est un procédé assez complexe qui est considéré comme

étant une dernière tentative de récupération des données, s'il n'y a pas d'autres alternatives pour le faire.

7.4 Mutation du problème de sauvegarde du patrimoine

Toutes les questions abordées ci-dessus sont sujettes à controverses et à discussions, nous nous trouvons dans une course contre la montre pour savoir quel support survivra à l'autre, le support papier ou bien le support électronique ?

Nous nous trouvons face à une mutation du problème sur lequel nous avons établi notre hypothèse de travail. La problématique en question était de développer des outils afin de sauvegarder les versions en papier des documents du patrimoine manuscrit et ceci en les numérisant. Nous avons vu ensuite qu'une deuxième problématique résulte de la première laquelle consiste à développer des stratégies afin de sauvegarder les versions numérisées des documents du patrimoine manuscrit et ceci en assurant leur conversion numérique. La conclusion à laquelle nous aboutissons et par laquelle nous avons voulu clore ce chapitre c'est que aujourd'hui, ce sont les versions papier des documents manuscrits qui sont en danger et qu'il faut sauver, mais dans un siècle (ou bien peut être même avant) ce seront les versions numérisées des documents manuscrits qui seront en danger et il faut les sauver.

Ce phénomène va s'enchaîner, et à chaque nœud de transition, une alerte va se déclencher en mettant des données en péril de disparition. Si chaque transition résulte en une perte de données même minime, nous allons arriver à un nœud final où nous n'aurons plus de données. C'est ce phénomène là qui augmente l'intérêt accordé par les chercheurs pour le développement d'algorithmes qui permettent des transitions fidèles et sans perte des données.

8 Conclusion

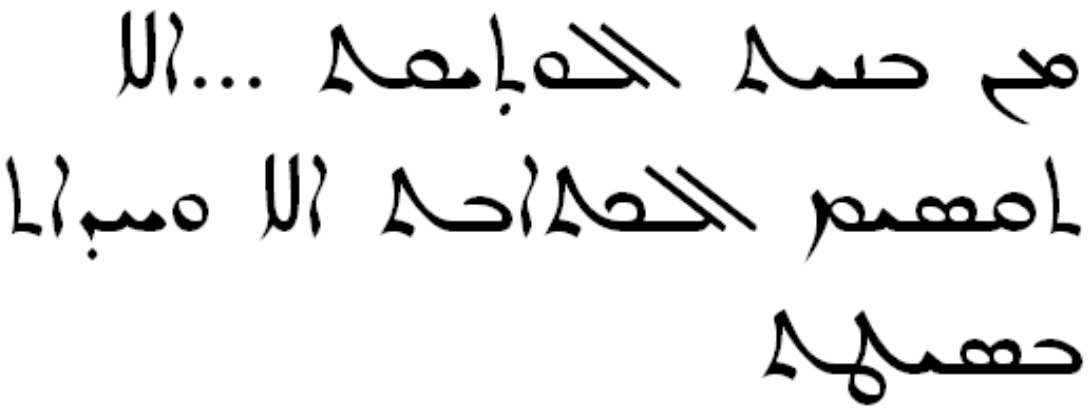
Dans ce chapitre nous avons montré la vie, la vitalité, la persistance et l'intérêt de la langue Syriaque au cours des siècles ; à travers l'histoire des langues qui sont nées dans la région du Moyen-Orient, nous avons vu comment certaines (dont le Syriaque en particulier) ont fortement influé sur les langues modernes.

Le Syriaque a donc joué un rôle considérable au Moyen Orient et les traces écrites qui nous restent constituent un patrimoine de l'Humanité, malheureusement en danger aujourd'hui parce que le support papier sur lesquelles elles se trouvent n'a pas une durée de vie infinie.

Les techniques développées au XXème siècle, photographie puis numérisation ont fait naître de grands espoirs, force est de constater aujourd'hui que la numérisation peut apporter beaucoup en ce qui concerne l'accès et la diffusion mais que nous ne maîtrisons pas encore tous les paramètres de conservation des données et supports numériques. En revanche, la numérisation peut ouvrir la voie à une réédition virtuelle (numérique) enrichie et augmentée. C'est à ce chantier qu'est consacrée la suite de cette thèse.

Chapitre 2

Des structures du document…
à la segmentation
en lignes et en caractères



1 Introduction

Dans ce chapitre nous allons décrire la structure et l'organisation d'un manuscrit Syriaque ; comment sont effectués la ponctuation, la pagination ou la foliotation, la division en paragraphes, la division en chapitres, l'utilisation des couleurs, la décoration et les enluminures, etc... ?

La mise en page du manuscrit Syriaque est plus simple que celle des manuscrits Latins. La composante textuelle est dominante et on cherche à minimiser la place inoccupée.

Dans la seconde partie du chapitre nous allons mettre en place une démarche pour extraire les lignes de textes à partir de cette structure ; nous segmenterons ensuite, par des méthodes algorithmiques ces lignes de textes en composantes élémentaires et nous donnerons un sens aux entités obtenues.

2 Organisation et structuration du document

Tout document manuscrit est une entité complexe dans laquelle de nombreux éléments se rassemblent, plus ou moins harmonieusement, pour lui donner son apparence finale. La structuration et la mise en page d'un texte apportent des informations sur les circonstances liées à sa rédaction, elles permettent aussi de montrer l'hiérarchie du texte ; elles déterminent le rang du manuscrit en question, davantage d'efforts seront mis en action pour un document de valeur supérieure.

Cette mise en page n'est sûrement pas la même pour tous les documents d'une même langue, chaque manuscrit est singulier, mais tous les manuscrits comprennent des éléments constitutifs de base qui leurs sont communs, des paragraphes composés de lignes, elles mêmes composées par des lettres…

Les documents Syriaques ont été rédigés de manière à économiser, autant que possible, le papier, ce dernier étant auparavant une denrée rare et assez coûteuse. Ceci est bien visible dans la mise en page de ces documents, les scribes ont cherché à gérer d'une façon simple la composition textuelle hiérarchique de leurs ouvrages, et ceci en essayant de maintenir la facilité d'un parcours de lecture à plusieurs niveau

2.1 La sobriété de l'écriture syriaque

L'écriture Syriaque fait preuve d'une sobriété assez particulière. Depuis son existence, l'alphabet a gardé l'ordre primitif (dit ordre Levantin) des lettres (abjad) identique pour tous les alphabets sémitiques. Cet ordre est considéré comme étant le premier ordre alphabétique et il est proche du Latin. Pour le Syriaque il est resteé inchangé, contrairement au cas de l'alphabet Arabe moderne dans lequel l'ajout de certains phonèmes a conduit à des modifications dans l'ordre des lettres.

L'écriture Syriaque est dite « monocamérale » (contrairement à l'écriture latine), la distinction entre majuscule et minuscule n'existe pas, il n'y a pas d'opposition de casse (haut de casse pour les majuscules et bas de casse pour les minuscules) ; seule la taille (et non le tracé) d'une lettre peut varier selon qu'elle se trouve dans le titre d'un chapitre ou bien dans le corps du texte ceci a donné une simplicité exemplaire à cette écriture, simplicité transmise à l'écriture Arabe.

Bien que le Syriaque soit une écriture cursive, le dessin des lettres suit un ordre prédéfini comme nous l'avons décrit dans le chapitre 1. Les caractères sont de hauteur inégale, généralement écrasés, petits de corps, avec exagération de certaines hampes. Deux aspects se trouvent présents dans l'écriture, un premier aspect vertical et un second aspect oblique (penché d'environ 45 degrés). Les lettres ne sont pas toutes liées à celles qui les suivent, ceci dépend leur variante contextuelle. Le tracé est courant et rapide, avec davantage un souci de continuité que d'élégance et d'ornementation. Néanmoins, les trois variétés de calligraphies prouvent que les scribes cherchaient à ajouter du style, vu le fait que la calligraphie Serto était réservée pour un usage commun, alors que la calligraphie en Nestorien était réservée pour les documents de valeur.

Ces trois calligraphies distinctes, ont toutefois gardé un aspect assez primaire contrairement aux écritures Latine et Arabe ; dans l'écriture latine on ajoute à certaines lettres des images figuratives, par exemple des lettrines, ces images pouvant varier d'un document à un autre, selon le contexte. Dans l'écriture Arabe, on a développé des styles calligraphiques ayant chacun une valeur symbolique, on a créé des « calligrammes », textes ou poèmes écrits et disposés graphiquement pour former des dessins ont été

2.2 La ponctuation

Les documents Syriaques sont régis par un système simplifié pour la ponctuation :

- la virgule est notée comme un point singulier,
- le point final d'une phrase est noté à l'aide de deux points verticaux.

Jusqu'à présent, outre les symboles utilisés, ceci est aussi vrai pour les documents Latins.

La lettre en majuscule n'existe pas, contrairement à ce qui se passe avec les documents Latins dans lesquels une phrase commence toujours par une lettre en « haut de casse » ou bien par une lettrine.

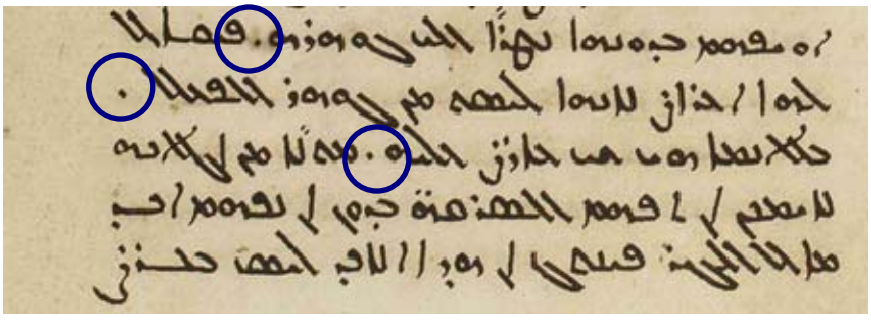


Figure 23: Les virgules dans un extrait d'un document Syriaque.

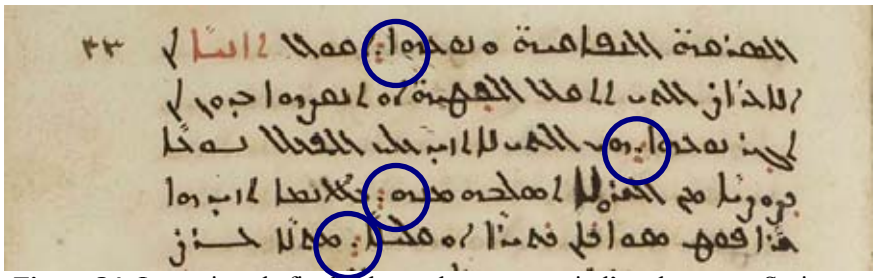


Figure 24: Les points de fin de phrase dans un extrait d'un document Syriaque.

2.3 Délimitation de paragraphes

Les textes Syriaques sont écrits linéairement, il n'y a pas de multicolonnage. Afin de délimiter les paragraphes, on utilise une marque composée par des points disposés en losange

Dans les documents Latins, un paragraphe pourrait commencer par une lettrine plus ou moins sophistiquée, la première ligne d'un paragraphe pourrait être indentée, il y a un retour à la ligne, et même un saut de ligne entre deux paragraphes.

Tous ces éléments de structure de texte sont absents dans les documents Syriaques. Un retour à la ligne est considéré comme une action inutile gaspillant l'espace d'écriture, cette dernière pourrait bien être substituée par une alternative simple et économique qui ne complique pas davantage la lecture et la compréhension de la page en question. Des délimitateurs de paragraphes peuvent se trouver partout dans une ligne de texte.

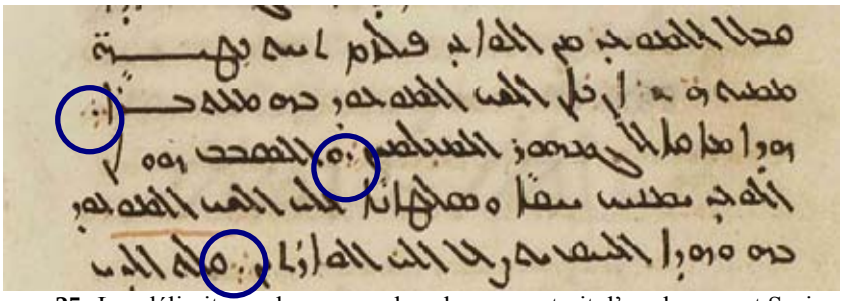


Figure 25: Les délimiteurs de paragraphes dans un extrait d'un document Syriaque.

2.4 Titrage de chapitres

Dans les documents Latins, les chapitres sont séparés par des sauts de pages, et, parfois aussi, par des objets ornementaux tels que des lettrines ou bien encore par des couleurs. Les titres des chapitres sont la plupart du temps écrits en lettres majuscules, ou bien en lettres capitales. Il existe même différents styles de lettrines pour les différents niveaux hiérarchiques du texte

Le saut de page n'est pas présent dans les documents Syriaques, ceci dénoterait, une fois encore, un gaspillage inutile du papier. La notion de majuscule n'existant pas non plus, les titres des chapitres sont tout simplement écrits en taille plus grande et sont placés au milieu de la ligne d'écriture.

Figure 26: Le titre d'un chapitre écrit en taille plus grande que le corps du texte.

2.5 Mise en relief

Les styles typographiques modernes (gras, italique, souligné) ont été inventés pour mettre en relief certains mots, pour attirer l'attention sur certaines expressions dans un texte. Pour les documents Syriaques, cette mise en avant consiste en l'utilisation d'une encre de couleur différente (en général le rouge), à souligner, surligner, ou à encercler la portion de texte en question. La couleur rouge est aussi utilisée pour indiquer une nouvelle section ou rubrique dans une page de texte.

Davantage que pour les documents Latins, la couleur a aussi joué un rôle important dans la mise en avant de certains mots.

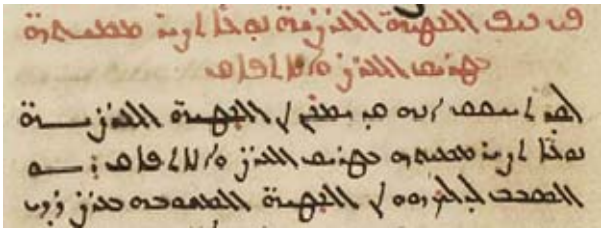
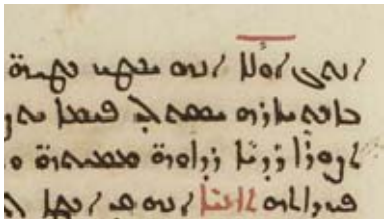
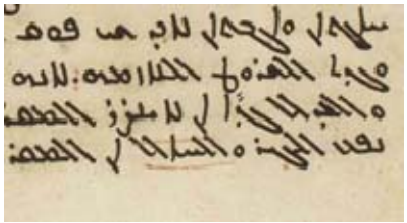


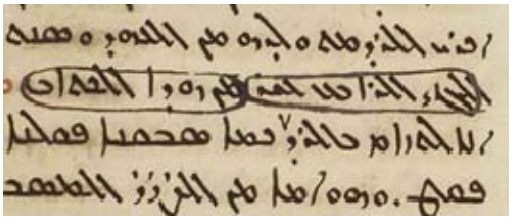
Figure 27: Certains mots mis en relief par la couleur rouge.



a) Mot surligné.



b) Mot souligné.



c) Phrases encerclées.

Figure 28: Différentes façons de mise en relief de certains mots ou phrases.

2.6 Pagination

Les manuscrits Syriaques sont paginés et non pas foliotés. Les numéros des pages sont écrits en chiffres Indiens, ils sont positionnés sur le coin supérieur droit des pages de droite et sur le coin supérieur gauche des pages de gauche.

Figure 29: Pagination d'un document Syriaque.

Les documents sont souvent constitués de cahiers ; un cahier est une grande feuille pliée soit en deux, soit en quatre, soit en huit, ceci dépend de la taille de la feuille au départ et de la taille des pages finales que l'on souhaite obtenir. Un des manuscrits dont nous disposons pour notre étude est dit «in Octavo», c'est un livre composé de cahiers dont chacun est formé par une feuille pliée trois fois de manière à former huit feuillets, c'est-à-dire seize pages. Dans notre document, les cahiers sont repérés (numérotés) avec des lettres qui se répètent toutes les 16 pages.

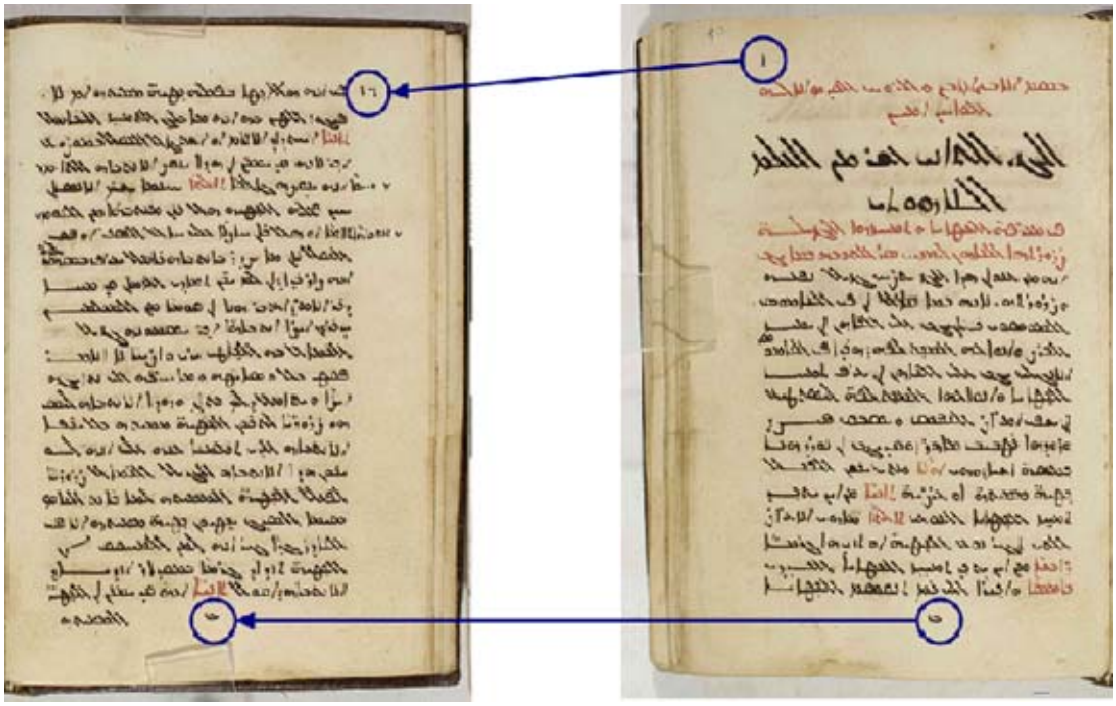


Figure 30: Numérotation des cahiers.

Après le découpage de la grande feuille, et avant de pouvoir passer à l'opération de reliure, on crée le lien appelé « réclame » entre deux pages consécutives, le premier mot d'une page est répété au bas de la page qui précède.

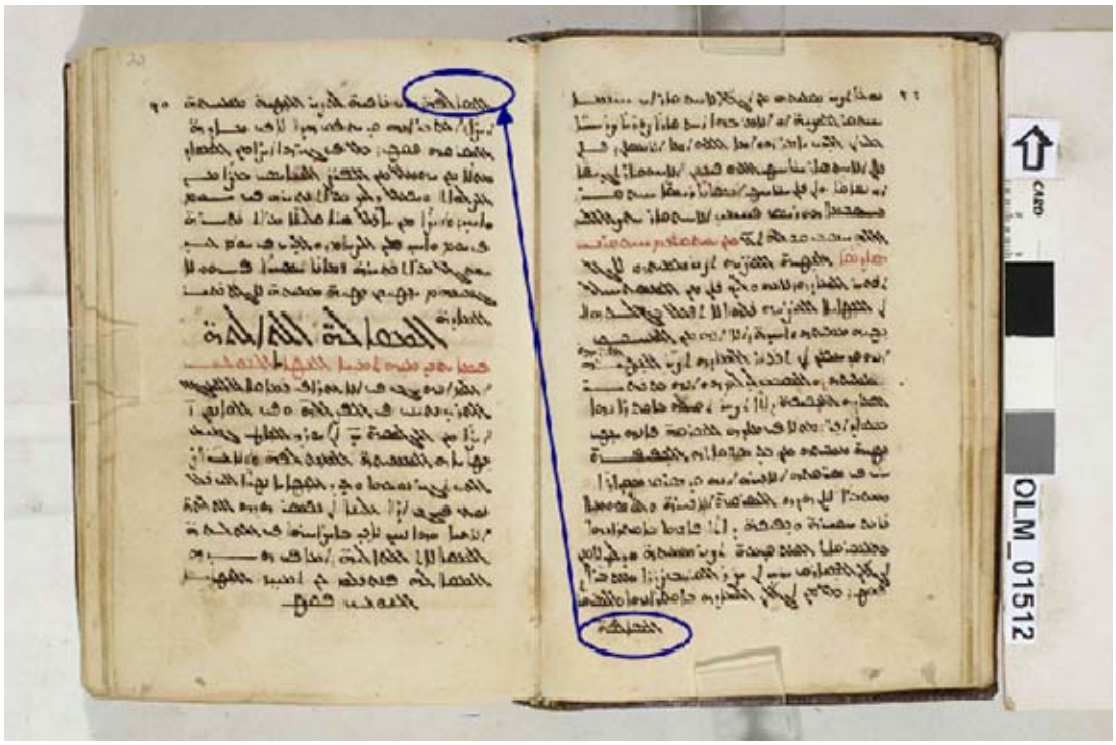


Figure 31: Renvoi d'une réclame d'une page à la page qui suit.

2.7 Annotations

Le texte principal est parfois entouré de gloses c'est à dire de textes écrits dans les marges ; elles sont aussi très présentes dans les manuscrits Arabes et dans les manuscrits Latins. Dans ces derniers, elles peuvent consister en des notes explicatives ou bien des commentaires et définitions ; elles peuvent rendre la structure assez complexe, d'autant qu'elles seront plus nombreuses.

Dans nos manuscrits Syriaques, nous n'avons pas rencontré des gloses aussi importantes, par contre, dans les marges de plusieurs pages du livre se trouve le mot arabe « بلغ » qui veut dire « atteindre » ; nous supposons qu'il est utilisé comme marque page par un lecteur maladroit. Ce qui est remarquable, c'est que ce mot se trouve toujours au milieu de la page de droite. Nous expliquons ceci par le fait que le livre est rédigé de droite à gauche, donc il devrait être lu dans ce même sens, et en le feuilletant, l'œil tombe en premier sur la page de droite.

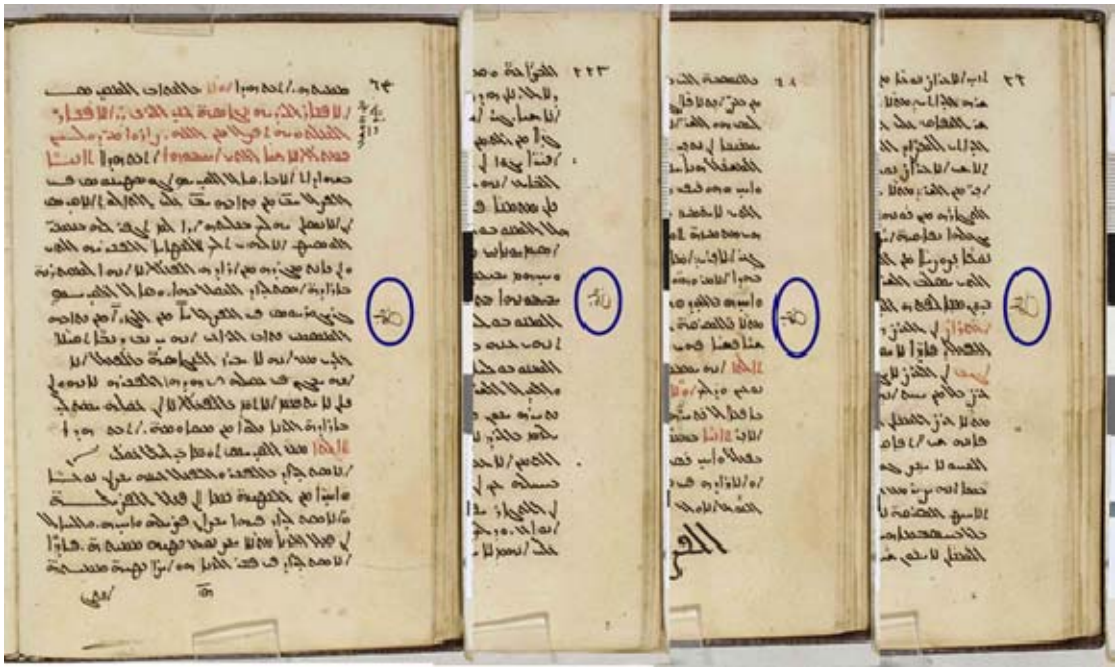


Figure 32: L'ancêtre de la marque page.

2.8 Décoration et enluminures

Comme nous l'avons expliqué précédemment, les documents Syriaques font preuve d'une grande sobriété dans la présentation du texte, contrairement aux textes Latins qui, eux, ont la particularité d'être assez décorés et ornementés (motifs d'enluminures, lettrines, bandeaux, etc…) ; plus un document est décoré et enluminé, plus son rang est élevé. C'est en fait cette décoration qui sert parfois à illustrer l'organisation hiérarchique des textes Latins.

Ceci n'est pas en contradiction avec l'existence de certains manuscrits Syriaques enluminés. Le meilleur des exemples serait sans doute les Evangiles de Rabbula qui se trouvent actuellement à la Biblioteca Mediceo-Laurenziana à Florence en Italie ; ils

datent du VI^{ième} siècle après J. C. Rabbula a accompli son travail de moine copiste au sein du monastère Saint Jean de Beth Zagba, en Mésopotamie, probablement dans le nord de la Syrie actuelle. Le texte de cet Evangéliaire est rédigé avec la calligraphie Estrangelo et fait partie de la Peshitta. Ce manuscrit est enluminé avec un texte encadré de motifs floraux et architecturaux élaborés. Il contient aussi un certain nombre de miniatures, dont une miniature de la Crucifixion, de l'Ascension et de la Pentecôte.

Les Evangiles de Rabbula sont loin de constituer les seuls manuscrits Syriaques à peinture, le recueil de l'Abbé Jules Leroy, "Les manuscrits syriaques à peinture" [LER64] est un excellent ouvrage de référence sur l'étude de ces manuscrits Syriaques enluminés.

De plus, ces manuscrits témoignent du rôle important qu'ont joué les moines Syriaques dans le développement de l'iconographie Chrétienne, que ce soit en Orient ou en Occident. Ci-dessous est une miniature extraite de l'Evangile de Rabbula et qui consiste en la scène de l'Ascension.

Figure 33: Miniature de l'Ascension extraite de l'Evangile de Rabbula.

3 Désordres et Dégradations

3.1 Erreurs et oublis du copiste

Un manuscrit étant le fruit d'un travail manuel humain contiendra inéluctablement des erreurs. Un scribe ou un copiste, aussi exercé soit-il dans son Art, commettra de façon certaine des erreurs dans son manuscrit ; c'est en révisant son œuvre et en relisant le document qu'il pourra se rendre compte de ses erreurs et

essayer de s'en remettre ; cependant, il sera souvent plus judicieux de laisser le texte tel qu'il est plutôt que de le corriger, puisque la correction pourra rendre la situation pire qu'avec l'erreur. Pour le copiste, l'erreur sera simplement réparée, mais cette correction pourrait avoir des effets indésirables pour les traitements que nous souhaitons appliquer au document.

Erreurs de pagination

Pendant la phase de pagination, on peut commettre certaines erreurs, comme par exemple porter le même numéro sur deux pages consécutives, ou bien remplacer

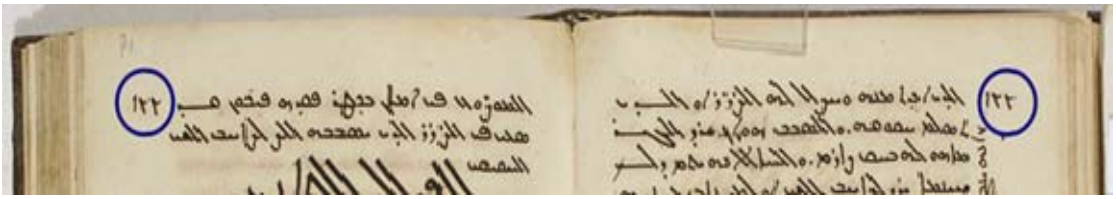


Figure 34: Deux pages successives portant le même numéro.

un chiffre par un autre ; une telle erreur va trainer jusqu'à la fin d'un livre ; le copiste s'en est rendu compte un peu trop tard, mais il l'a corrigée en écrivant le bon chiffre par-dessus l'ancien. Nous ne pouvons pas juger si c'était la meilleure façon de se remettre de cette erreur, étant donné le fait que les ratures qu'il a dû faire seront considérées comme bruits ou dégradations dans la version numérisée du document.

Figure 35: Erreur de pagination et ratures de correction.

Mots oubliés

Il est arrivé qu'en pleine rédaction un copiste, oublie un mot ou une phrase du texte ; dans ce cas, la portion oubliée peut être insérée entre les lignes du texte à l'endroit où elle était sensé exister ; si la partie oubliée est volumineuse et si il n'y a pas suffisamment de place pour la coincer entre les lignes, alors cette partie sera écrite dans la glose avec un petit symbole renvoyant à l'endroit où cette partie devrait être insérée.

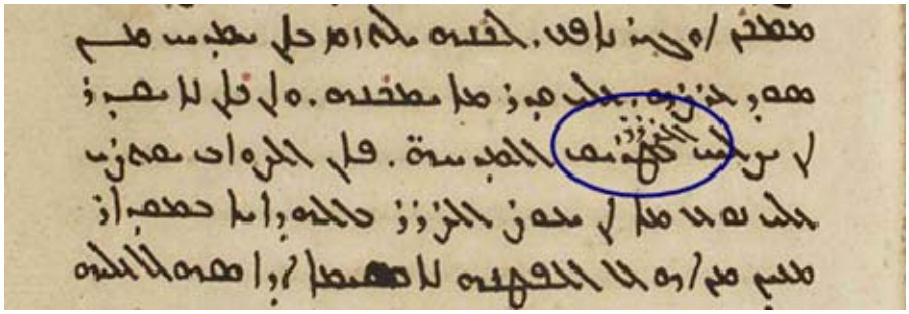


Figure 36: Mot coincé entre deux phrases.

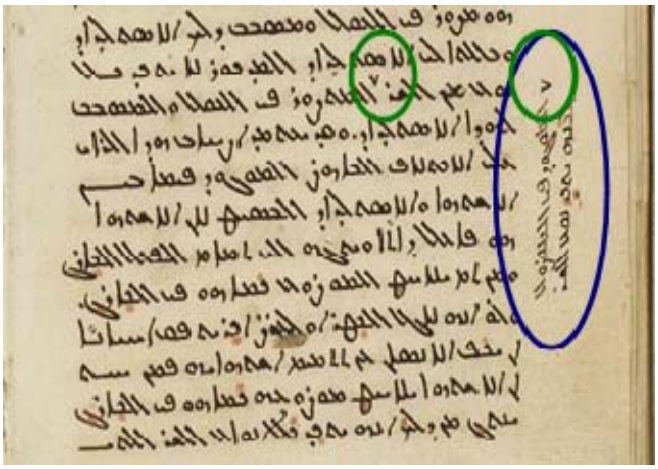


Figure 37: Glose renvoyant à une portion de texte oublié.

Erreurs de modifications

Dans le cas inverse de l'oubli, il peut arriver qu'un copiste ajoute du texte aux mauvais endroits ; au cours d'une relecture, ces parties ajoutées seront raturées. Un exemple de ratures est montré dans la figure ci-dessous.

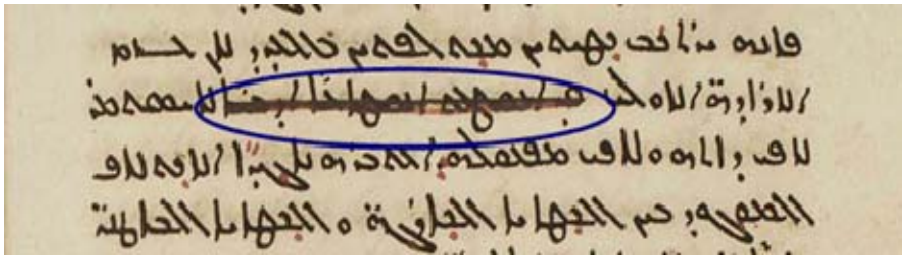


Figure 38: Mots barrés et ratures.

3.2 Qualité visuelle et dégradations

Les documents entre nos mains sont constitués de manuscrits écrits en Serto ou en Nestorien. Ils présentent, comme tout document ancien, diverses détériorations.

Les détériorations pouvant atteindre les versions numérisées des anciens documents manuscrits peuvent avoir plusieurs origines, elles peuvent provenir soit du document lui-même (ratures, mots coincés, taches d'encre, etc…), soit des conditions de stockage dans lesquelles se trouvait le manuscrit avant sa numérisation (humidité et moisissure causant des bavures d'encre et autres taches, trous causés par des insectes, incendies, etc…) ; elles peuvent aussi être dues au processus de numérisation en tant que tel, notamment à des réglages non adéquats du numériseur (faible résolution, forte compression) ou bien encore à un mauvais choix de support de numérisation. Toutes ces détériorations sont considérées comme bruit dans le contexte de traitement d'images.

En numérisant un document, il faut prendre en considération tous ces bruits et régler convenablement les paramètres du numériseur afin de réduire autant que possible leur effets, tout en préservant l'information importante portée par le document.

3.3 Mots grattés

Nous avons remarqué dans nos documents la présence de certains mots grattés, avec un texte de remplacement à côté. Le texte de remplacement peut être le même que celui estompé, comme il peut être différent. Dans ces endroits l'encre a pu être dissoute pour des raisons d'humidité ou a pu être grattée manuellement, par mauvaise intention. On trouve des exemples de mots grattés dans la figure ci-dessous.

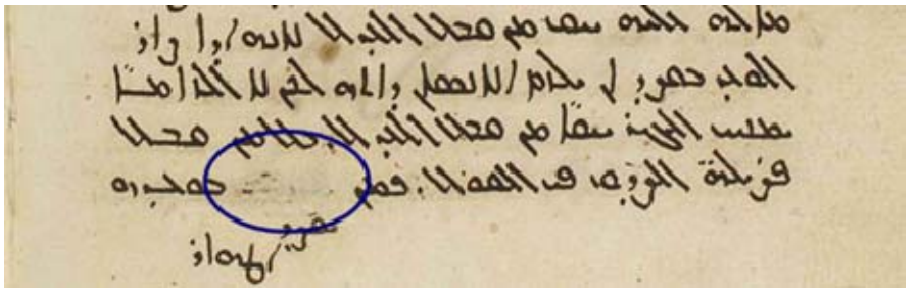


Figure 39: Mots grattés ou estompés.

3.4 Taches

Les anciens documents présentent en général des taches. Ces dernières peuvent avoir plusieurs origines, elles peuvent provenir de l'humidité et de la moisissure, ou bien d'accidents où des liquides ou autres produits se trouvent versés sur le papier, etc… Ces taches peuvent être localisées sur une partie du texte ; en tentant de les supprimer, nous risquons de supprimer aussi le texte.

Ces bruits, dits bruits de fonds, bien que présents dans nos documents influent très peu sur le traitement des images vu le fait qu'un simple passage d'une image couleur en une image en niveaux de gris remonte la contraste entre le texte et le fond réduisant ainsi considérablement leur apparence et effets indésirables.

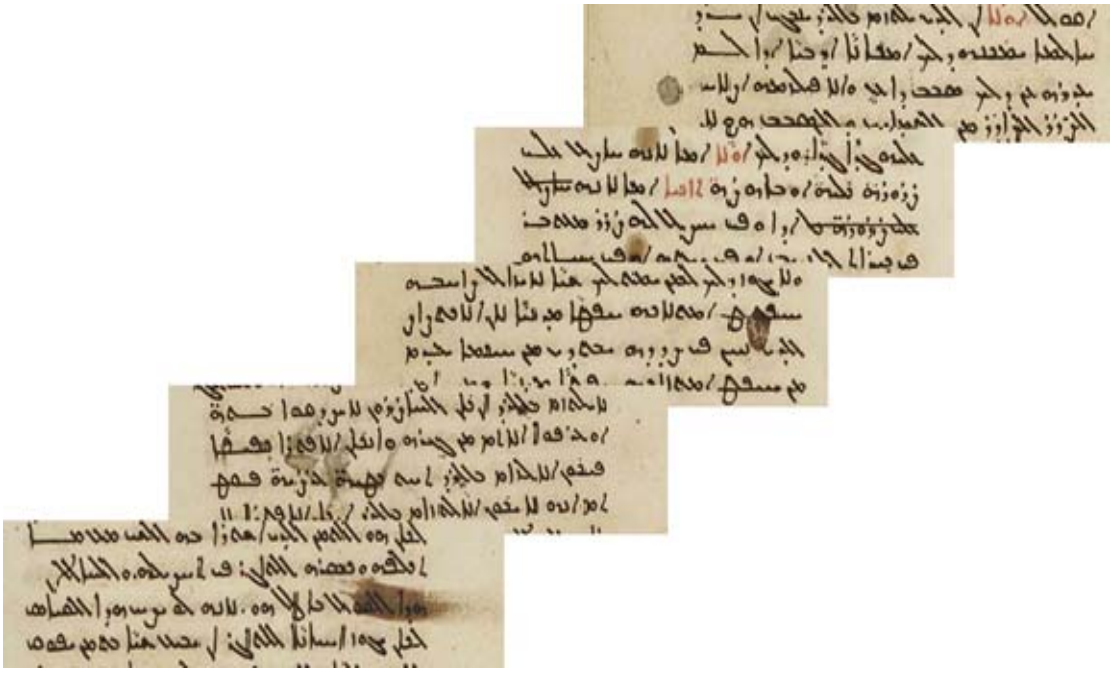


Figure 18: Taches brunâtres multiples visibles sur un corpus en Serto.

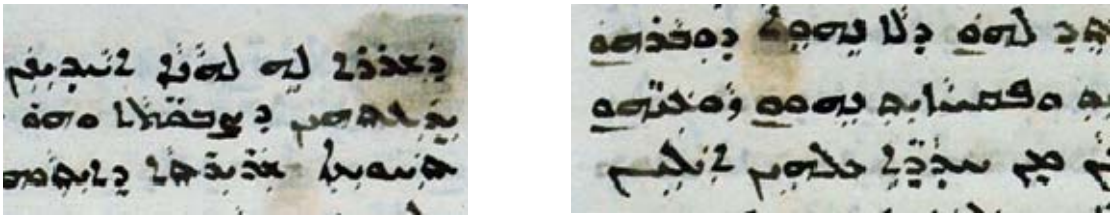


Figure 19: Taches brunâtres multiples visibles sur un corpus en Nestorien.

3.5 Transvision et passage du verso sur le recto

Le texte du verso d'une page peut parfois être visible en transparence sur son recto, c'est le phénomène dit de transvision. Ce fait peut être dû, soit au processus de numérisation par mauvais réglage de la lumière qu'émet le numériseur, soit aux conditions de stockage du document dans un endroit humide ou bien entassé dans des piles lourdes qui appliquent une pression sur le papier, soit à la mauvaise qualité du papier (très buvard) et de l'encre. Des exemples de ces bavures sont illustrés dans les figures ci-dessous.

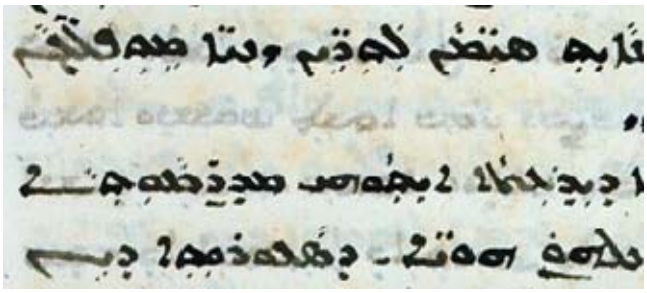


Figure 20: Bavures d'encre du verso visible sur le recto d'un corpus en Nestorien.

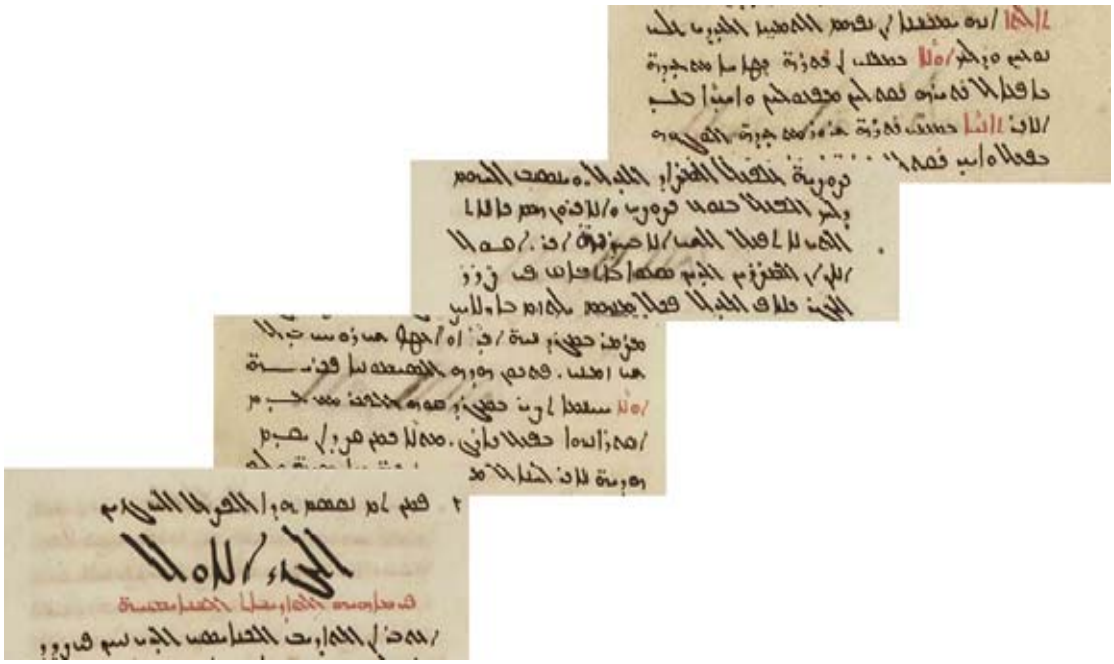


Figure 21: Bavures d'encre du verso visibles sur le recto d'un corpus en Serto.

4 Nos documents d'étude.

Nous disposons de quatre corpus différents, que nous souhaitons indexer. Trois sont rédigés en Serto et le quatrième est rédigé en Nestorien. La résolution a été réglée individuellement pour la numérisation de ces corpus.

A noter que les documents Syriaques présentent une épaisseur du trait de l'écriture de l'ordre de 1 mm. Cette épaisseur permet à l'écriture de résister à la compression avec perte et à la faible résolution lors de la numérisation.

4.1. Le premier corpus

Le premier corpus contient 10 pages et 1813 mots, le texte est écrit en Serto à raison d'une colonne de texte par page. Ce livre a été numérisé en mode binaire inverse avec une résolution de 96 dpi, et les images sont sauvegardées avec le format TIFF (Tagged Image File Format) comprimé sans perte.

Ce document provient du Département de Numérisation de Gorgias Press. La figure ci-dessous montre un extrait de ce corpus.

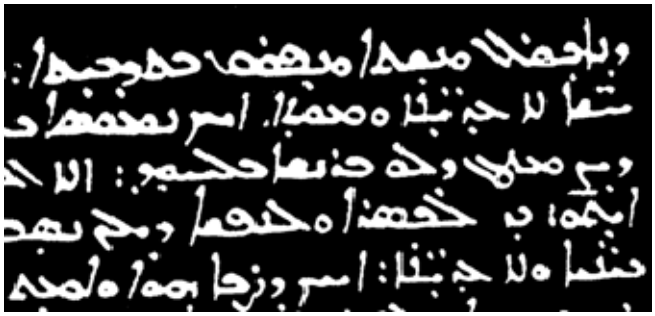


Figure 22: Extrait du premier corpus.

4.2. Le deuxième corpus

Le deuxième corpus consiste en 4 pages et en 1524 mots, le texte est rédigé en Serto à raison de deux colonnes de texte par page. Ce livre a été numérisé en couleur avec une résolution de 300 dpi, et les images sont sauvegardées avec le format JPEG, donc elles ont subi une compression avec perte.

Ce document provient de la Bibliothèque Centrale de l'Université Saint-Esprit de Kaslik. La figure 23 montre un extrait de ce corpus, trois niveaux de couleur sont visibles ; une couleur grisâtre non homogène pour l'arrière plan, une couleur noire pour le texte et une couleur rouge pour la délimitation des phrases et des paragraphes.

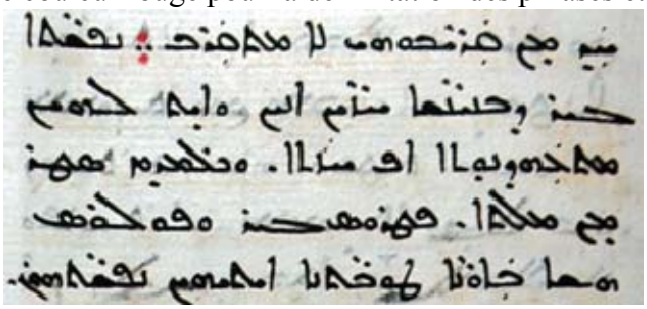


Figure 23: Extrait du deuxième corpus**.**

4.3. Le troisième corpus

Le troisième corpus a 377 pages et environ 50 000 mots, le texte est rédigé en Serto à raison d'une colonne de texte par page. Ce livre a été numérisé en couleur avec une résolution de 300 dpi, et les images sont sauvegardées avec le format JPEG.

Ce document provient de la Bibliothèque Centrale de l'Université Saint-Esprit de Kaslik, au Liban. La figure ci-après montre un extrait de ce corpus, trois niveaux de couleur sont visibles ; une couleur beige non homogène pour l'arrière plan, une couleur brunâtre pour le texte, et une couleur rouge pour les symboles diacritiques, les délimitations des phrases et des paragraphes, et les mots mis en relief.

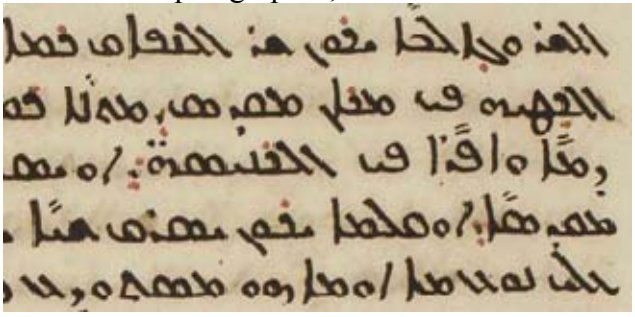


Figure24 : Extrait du troisième corpus.

4.4. Le quatrième corpus

Le quatrième corpus comporte 4 pages et 576 mots, le texte est rédigé en Nestorien à raison d'une colonne de texte par page. Ce livre a été numérisé en couleur avec une résolution de 71 dpi et ses images sont sauvegardées avec le format JPEG.

Ce document provient de la Bibliothèque Centrale de l'Université Saint-Esprit de Kaslik, au Liban. La figure ci-après montre un extrait de ce corpus, deux niveaux de couleur sont visibles : une couleur grise non homogène pour l'arrière plan, une couleur noirâtre pour le texte.

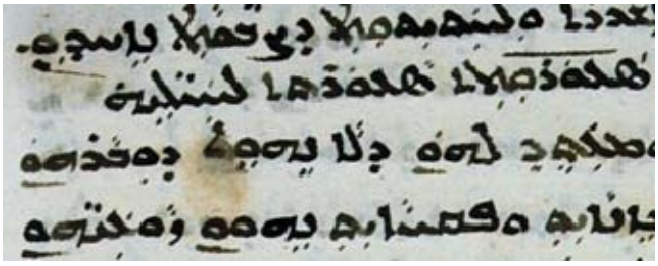


Figure 25: Extrait du quatrième corpus.

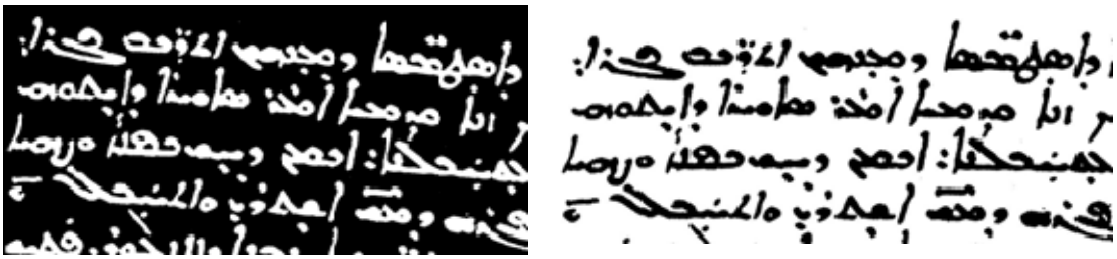
5 Extraction des lignes de texte

5.1. Le prétraitement

Elle consiste à préparer le document pour les étapes qui vont suivre. Cette phase dépend de l'état initial dans lequel se trouve le document, ainsi de ce que nous désirons y appliquer comme traitement. Ce pourra être une simple binarisation [GAT04] [GAT06], un passage d'une image couleur en image en niveaux de gris, ou bien un nettoyage des bruits de fonds (élimination des bavures recto/verso, et des taches, etc…). Plusieurs approches ont été élaborées et testées dans [DRI07] sur ce propos. Cette phase servira à remonter le contraste entre le fond et le texte, à mettre ce dernier en avant et à diminuer autant que possible l'influence du bruit sur les traitements futurs.

5.2 Conversion colorimétrique

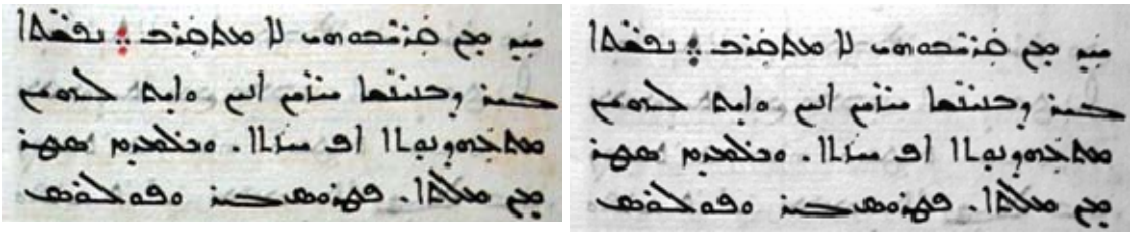
Pour le premier corpus, nous avons tout simplement permuté le noir et le blanc puisqu'une écriture noire sur un fond blanc est plus agréable à l'œil du lecteur. La figure ci-dessous montre un extrait du premier corpus avant et après la conversion.



a) Image en mode binaire inverse. b) Image en mode binaire normal.

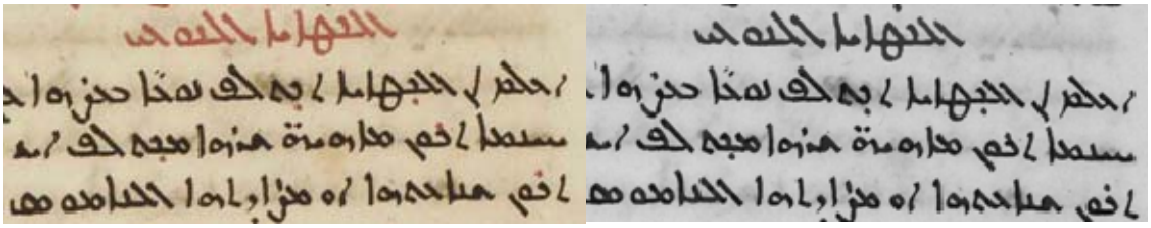
Figure 26: Conversion colorimétrique effectuée sur le premier corpus.

Pour les deuxième et troisième corpus, certains mots sont écrits en rouge, si cette couleur est claire, elle risque de disparaître si nous passons immédiatement en niveaux de gris, d'autre part nous remarquons la présence de certaines taches et de bavures recto/verso qui pourraient nous déranger. Avant d'attaquer ces problèmes par des méthodes de restauration complexe, nous avons tout d'abord tenté les approches simples. Nous avons commencé par assombrir la composante rouge dénotant les mots et les expressions en relief, puis nous avons appliqué une transformation de passage d'une image couleur en une image en niveaux de gris.



c) Image en mode couleur. d) Image en niveaux de gris.

Figure 27: Conversion colorimétrique effectuée sur le deuxième corpus.

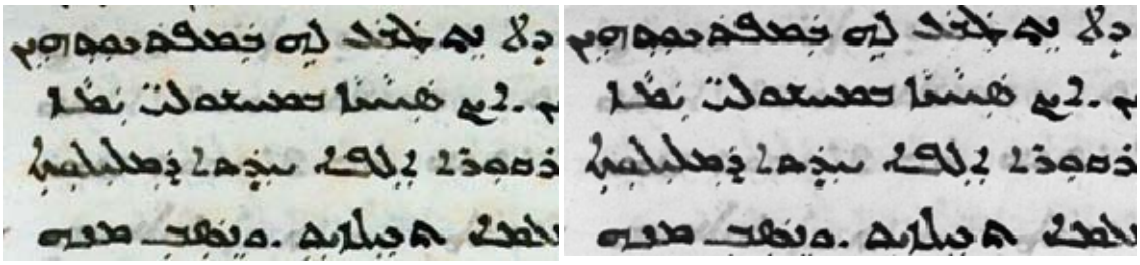


e) Image en mode couleur. f) Image en niveaux de gris.

Figure 28: Conversion colorimétrique effectuée sur le troisième corpus.

Ce traitement a été suffisant dans le cas de nos documents pour éviter l'effacement des mots en rouge, et pour estomper les bruits de fonds. L'écriture étant suffisamment sombre et épaisse pour résister à ces influences. Les figures ci-dessus montrent respectivement des extraits du deuxième et troisième corpus avant et après la conversion.

Pour le quatrième corpus, nous avons appliqué une transformation d'une image couleur en image en niveaux de gris. La figure ci-dessous montre un extrait du quatrième corpus avant et après la conversion.



g) Image en mode couleur. h) Image en niveaux de gris.

Figure29: Conversion colorimétrique effectuée sur le quatrième corpus**.**

5.3 Détection et correction du «pencher» des lignes

Même si l'écriture dans une page de texte est justifiée et si les lignes sont bien parallèles, le texte n'est pas toujours exactement horizontal. Ceci est dû soit au

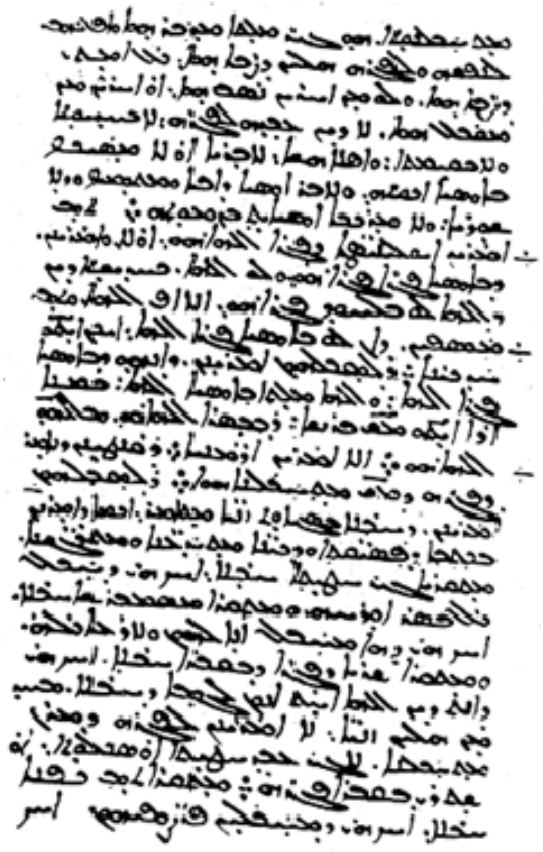
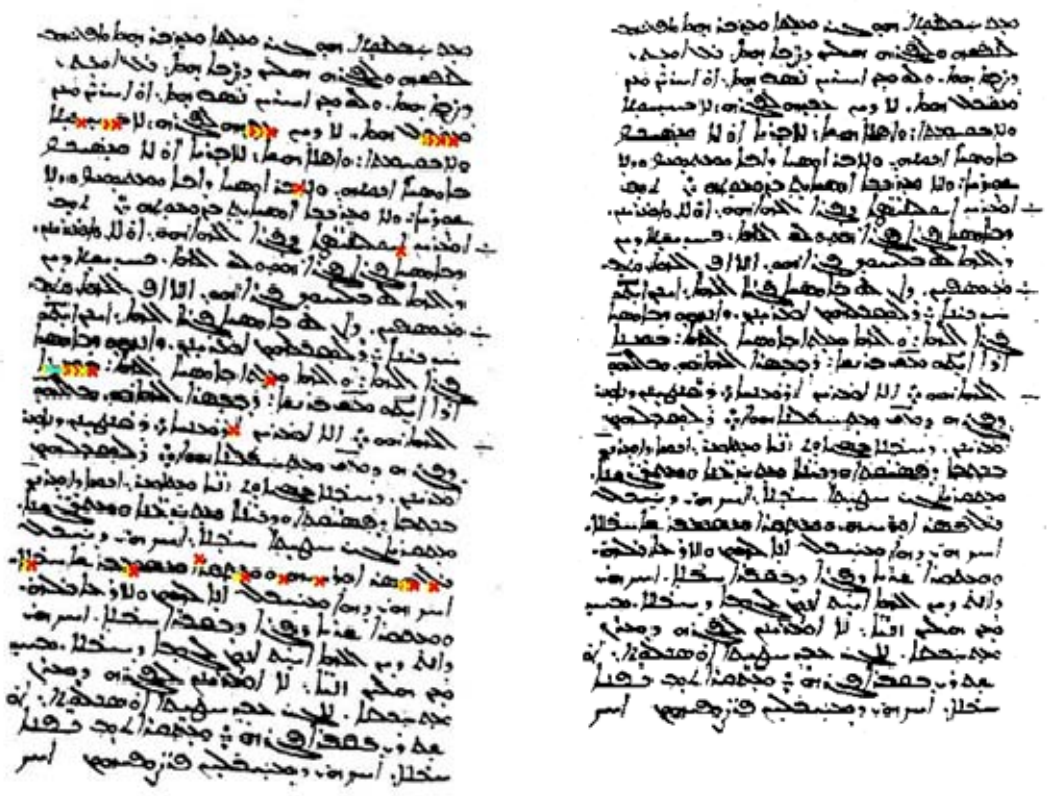


Figure 30: Exemple d'une page de texte montrant une écriture penchée.

scripteur lui-même, soit à un découpage mal fait du cahier avant la reliure, ou soit au fait que la personne qui s'occupe de la numérisation a tourné un peu la page avant de la scanner. La figure 30 ci-dessus montre une page de texte qui a été mal coupée avant la reliure, les lignes sont bien parallèles mais elles sont toutes penchées vers la droite.

Afin de redresser l'écriture, une première étape consiste en l'estimation de l'angle de pencher. Pour ce faire, nous avons choisi la transformée de Hough [DUD72], qui détecte la ligne de plus grande pente et par la suite la direction saillante de l'écriture. Une fois le pencher estimé, la correction de cette dernière est faite par cisaillement dans le sens inverse à l'angle de penchée. La figure ci-dessous montre la direction saillante de la page de texte précédente en couleur Cyan ainsi que la même page après correction du pencher.



a) Direction saillante du pencher. b) Correction du pencher des lignes.

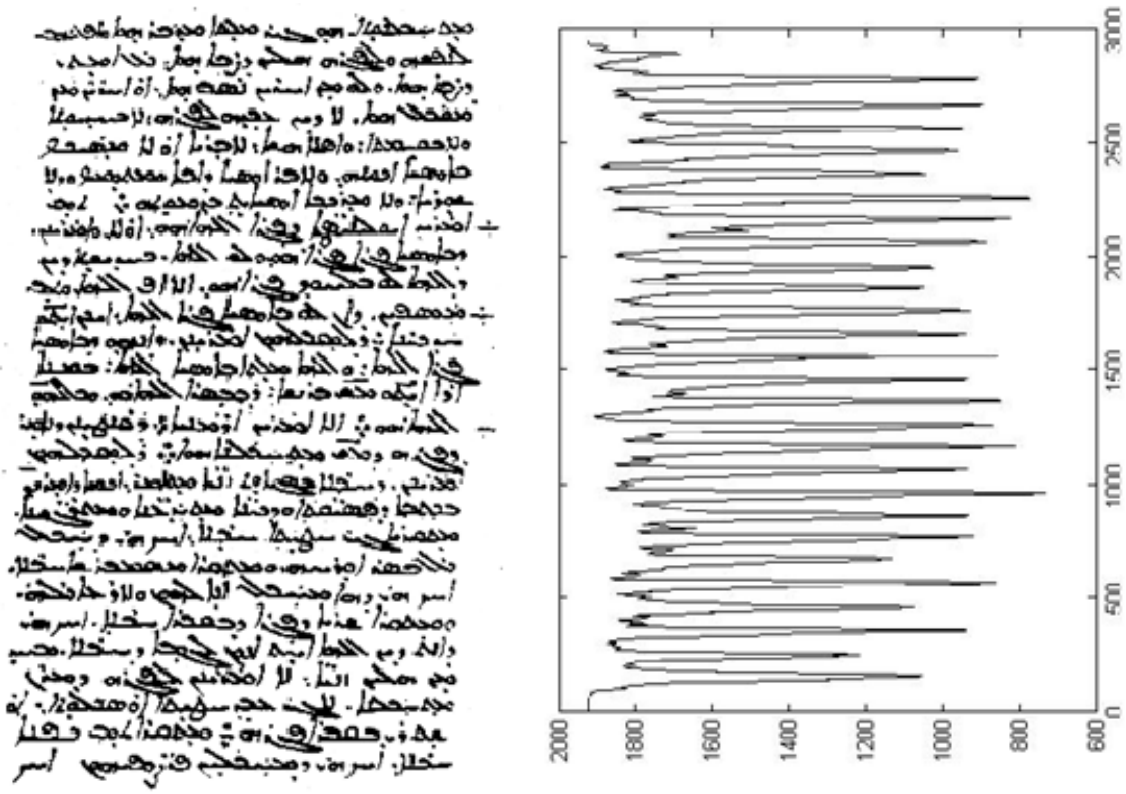
Figure31: Détection et correction du pencher des lignes dans une page de texte.

Ce prétraitement s'est avéré suffisant pour les images de nos documents. Aucun nettoyage supplémentaire des traces du verso sur le recto ou bien des taches brunes n'a été nécessaire, les conversions colorimétriques que nous avons appliquées ont suffi pour hausser le contraste entre le texte et le fonds, les bruits mentionnés ont désormais des effets minimes.

5.4. Segmentation d'une page en lignes constitutives

Afin de pouvoir segmenter une page de texte en ses lignes constitutives, nous commençons par tracer son histogramme horizontal. Ce dernier va montrer la silhouette de la page.

Dans la figure ci-dessous, le tracé d'un histogramme horizontal représente une page de texte, cet histogramme montre 27 sommets, donc cette page contient 27 lignes distinctes. Les positions des sommets indiquent les coordonnées qui situent où chaque ligne commence dans la page.



a) Image d'une page de texte. b) Histogramme horizontal associé.

Figure 32: Une page de texte et son histogramme horizontal associé.

6 La Segmentation des lignes en graphèmes

La lettre est le plus petit élément dans l'écriture d'une langue. Dans le traitement des images de textes manuscrits, et pour aboutir à des fins de transcription, nous voulons revenir à cet élément de base, puis reconstituer la chaîne linguistique en remontant dans l'hiérarchie jusqu'aux mots, puis aux lignes, et enfin aux pages de texte.

Pour pouvoir « descendre » au niveau de l'élément « lettre » de l'alphabet, nous devons effectuer une série de segmentations successives commençant par une segmentation d'une page de texte en ses lignes constitutives, puis pour chaque ligne de texte une segmentation en lettres autant que possible.

Cette tâche n'est pas évidente, surtout lorsqu'il s'agit de segmenter un document manuscrit. Que dire alors lorsque l'écriture est cursive ? Plusieurs approches de reconnaissance d'écriture manuscrite ont tenté d'éviter ou de contourner cette étape, compte tenu des difficultés qu'elle présente, surtout lorsqu'il s'agit de segmenter des lettres connectées à l'intérieur d'un mot. Dans la plupart des cas, l'indice correct de segmentation est indéterminé ; il peut apparaître très ambigu sans un minimum de connaissances linguistiques.

Une écriture penchée telle le Syriaque ajoute des contraintes supplémentaires à un algorithme de segmentation déjà assez compliqué. L'aspect oblique conduit au fait que certaines lettres peuvent se trouver cachées sous d'autres, le segment horizontal les reliant ou bien l'espace blanc intra-mot les séparant ne pouvant plus être détecté par des projections verticales.

6.1. Segmentation d'une ligne de texte en caractères individuels

Une ligne particulière de texte en Serto montre la direction saillante oblique de l'écriture, cependant cette caractéristique n'est pas présente pour toutes les lettres, l'élément vertical de l'écriture est toujours préservé. Les espaces inter-mots sont détectables par des bandes blanches verticales, certaines lettres sont aussi disposées verticalement. L'aspect oblique de l'écriture est visible intra-mot pour certaines de ses lettres constitutives. Ce que montre bien la figure ci-dessous.

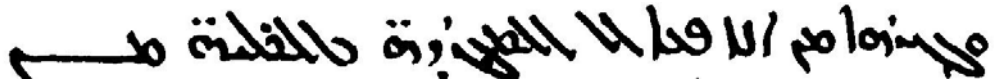


Figure 33: Une ligne de texte écrite en Serto.

Donc afin de pouvoir segmenter une ligne en Serto en caractères individuels, nous devons prendre en compte des deux aspects de l'écriture sans segmenter l'un au dépit de l'autre ; en d'autres termes, nous devons éviter de finir avec des fragments de lettres insignifiants. La segmentation devra se faire en étapes successives et d'une façon hiérarchique, c'est-à-dire commencer par capter un aspect puis en cas de besoin essayer de capter le deuxième.

6.2. Capture de l'aspect vertical de l'écriture

Pour pouvoir capter l'aspect vertical de l'écriture dans une ligne, nous traçons son profil de projection vertical associé. Avec ce profil nous sommes non seulement capables de repérer les bandes verticales inter-mots, découpant ainsi une ligne en ses mots constitutifs, mais aussi capter les espaces blancs verticaux intra-mot séparant ainsi les lettres à aspect vertical des autres lettres penchées.

Figure 34: Indices de segmentation verticale de la ligne de texte.

Après cette segmentation, nous nous retrouvons avec des lettres individuelles verticales et avec des groupes de lettres penchées. Nous allons désormais désigner ces groupes de lettres par l'expression « n-grammes » ; dans cette nomenclature la lettre « n » indique la dimension du groupe, c'est-à-dire le nombre de caractères qui le constitue. Ce nombre n'est pas fixe mais il est nécessairement supérieur ou égal à 2 :
→ pour un groupe de deux caractères, $n = 2$ et nous avons alors un bi-gramme.
→ pour un groupe de trois caractères, $n = 3$ et nous avons alors un tri-gramme.
→ etc…

La ligne segmentée ci-dessus dans la figure 2, contient des exemples de lettres individuelles, de bi-gramme, de tri-gramme et même de quadri-gramme. D'autres échantillons sont montrés dans la figure 35 ci-dessous.

a) Lettre individuelle.

b) Bi-gramme.

c) Tri-gramme.

d) Quadri-gramme.

Figure 35: Echantillons restant après la segmentation verticale.

Comme nous l'avons mentionné précédemment, l'aspect penché de l'écriture fait que certaines des lettres se trouvent dissimulées par d'autres, cela rend inefficace l'emploi d'un histogramme vertical pour les séparer. Ceci est bien visible dans les images des « n-grammes » présentées ci-dessus.

6.3. Capture de l'aspect oblique de l'écriture

Ce sont ces circonstances qui nécessitent une intervention pour la deuxième phase de segmentation à laquelle nous allons référer par « segmentation oblique ».

C'est essentiellement cette « segmentation oblique » qui est spécifiquement conçue pour les manuscrits Syriaques. Ne pas capter cet aspect distinct de l'écriture voudrait dire la traiter comme toutes les autres ; il en résulterait une perte conséquente d'informations qui sont propres à l'écriture syriaque et qui la décrivent dans ses caractéristiques qui ne sont observables qu'au niveau de ses éléments linguistiques de base i.e. les lettres de son alphabet.

L'objectif auquel nous souhaitons d'aboutir, autant que faire se peut lors de la segmentation des lignes en lettres individuelles est donc de réduire le nombre des « n-grammes » tout en minimisant la valeur de « n ». L'aspect oblique est présent surtout dans ces « n-grammes », ces derniers ne pouvant être subdivisés en lettres individuelles en utilisant uniquement une segmentation suivant la direction verticale. Ce qui nous conduit à une « segmentation oblique » nécessaire.

En résumé, la « segmentation oblique » partant d'une image d'un « n-gramme » consiste par détecter les indices corrects de segmentation puis récupérer les imagettes résultants, ce processus est illustré dans la figure ci-après en prenant comme exemple l'image initiale d'un bi-gramme.

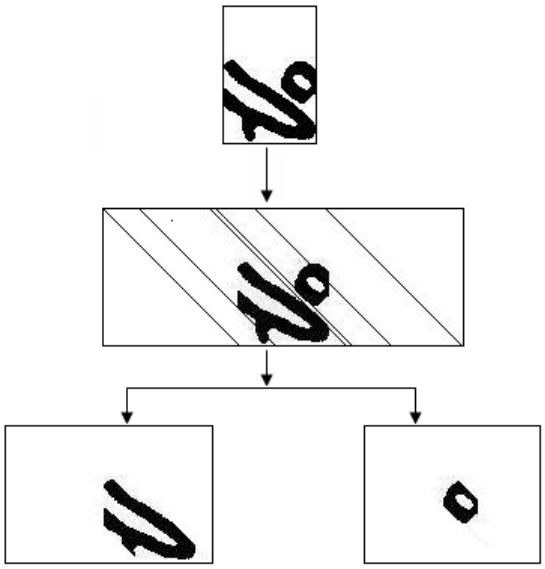


Figure 36: Schéma résumé illustrant en résumé la segmentation oblique.

Le schéma ci-dessus est bien compact, et montre une « descente » quasi-immédiate du niveau « n-gramme » au niveau lettres. En réalité, le schéma est plus compliqué que ça, les étapes mentionnées précédemment, surtout l'étape intermédiaire aux deux niveaux « n-grammes » et lettres, sont déployées en une série de sous-étapes consécutives que nous allons développer par la suite.

Estimation de l'angle d'inclinaison des lettres

Cette estimation est effectuée en utilisant la transformée de Hough [DUD72], une transformée dont l'usage est très classique pour ce type d'applications. Il est important de capter cet aspect penché, individuellement pour chaque « n-gramme », parce qu'il y a des fluctuations dans l'écriture, souvent due à la fatigue du scripteur. Tous les « n-grammes » ne sont pas penchés suivant le même angle, un scribe aussi habile soit-il ne peut pas toujours réussir à réaliser un parallélisme parfait dans l'écriture.

Une autre méthode pour estimer l'angle saillant de la direction d'une écriture est basée sur les roses de directions, cette méthode permet une estimation grossière de l'inclinaison. Pour le cas de nos « n-grammes », cette méthode conduirait à un manque de précision puisqu'elle estime suivant des intervalles et non pas des valeurs. Certaines valeurs peuvent tomber dans le même intervalle bien qu'elles soient différentes. Afin d'augmenter la précision des roses de direction, il faut subdiviser davantage le cadran de l'image (représenter davantage des directions), augmentant ainsi le nombre d'intervalles requis pour le couvrir, et diminuant l'angle de couverture. Dans notre cas, nous avons besoin de la valeur précise de l'angle d'inclinaison afin de pouvoir parfaitement redresser l'écriture verticalement.

Redressement de l'écriture

Quand l'angle d'inclinaison exact est calculé, nous appliquons un cisaillement, « shear transformation », dans le sens inverse de cet angle. Ceci permettra de redresser les lettres inclinées, et de repérer les lettres qui étaient cachées en dessous d'autres, ces dernières se trouvent ainsi en position verticale. Les espaces blancs inter-

lettres ainsi que les ligatures non détectables auparavant à cause de l'oblique sont désormais visibles. Ce sont ces éléments-là qui constituent les indices corrects de segmentation intra-mots pour pouvoir aboutir aux lettres individuelles.

Remplissage des trous

Certaines lettres présentent des trous dans leur dessin, ces trous peuvent être confondus avec des ligatures ce qui pourrait conduire à une subdivision de la lettre elle-même. Afin d'éviter cet incident, nous procédons par un remplissage de ces trous par une application d'un algorithme de reconstruction morphologique, qui change la valeur des pixels de l'arrière plan (1) en celle du texte (0). La condition d'arrêt de cet algorithme est obtenue quand le remplissage atteint les bords du trou. Cette étape est indépendante du résultat de l'étape de redressement de l'écriture qui lui précède, elle pourrait bien être effectuée avant. Par contre, elle devrait nécessairement devancer celle de la détection des indices de segmentation.

Détection des indices corrects de segmentation

Cette étape est effectuée en utilisant les projections verticales sur l'image du « n-gramme » à trous remplis. Une fois l'écriture redressée et les trous remplis, l'image du « n-gramme » montre des bandes blanches désormais verticales, ainsi que les ligatures qui sont visibles, et qui constituent les zones de segmentation intéressantes, le profil de projection vertical d'une telle image permet de repérer les indices corrects de segmentation.

Segmentation de l'image du « n-gramme »

Quand les indices de segmentation sont repérés sur l'image du « n-gramme » redressée et à trous remplis, nous pouvons nous en servir pour la segmentation de l'image du « n-gramme » redressée avec les trous d'origine. Nous récupérons ainsi les sous-images des lettres, avec leurs trous si elles sont trouées.

6.4. Stabilité des résultats

Les lettres

Après les deux segmentations successives, nous nous retrouvons (autant que possible) avec des imagettes de lettres. Ces dernières sont toujours redressées suivant la verticale, ce qui n'est pas leur forme d'origine. Nous appliquons alors un cisaillement suivant l'angle d'inclinaison original afin de retrouver la forme initiale des caractères.

Le schéma détaillé et complet du processus de segmentation oblique, ainsi que le développement de l'étape intermédiaire de la figure 4 en ses sous-étapes, sont alors illustrés dans la figure ci-après.

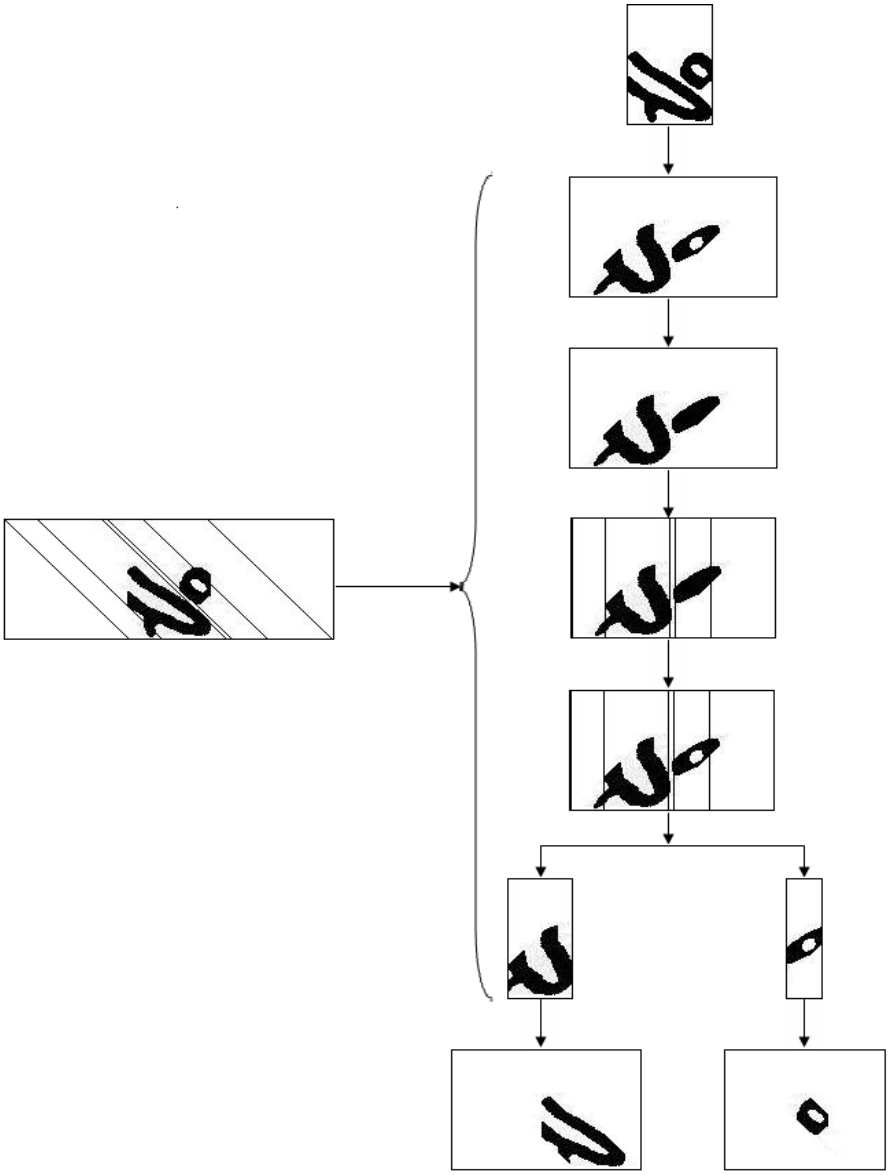


Figure 37: Schéma détaillé illustrant la segmentation oblique.

Persistance des « n-grammes »

Il se peut qu'après ces segmentations successives il reste certains « n-grammes » non segmentés. Leur nombre ainsi que leur dimension « n » sont bien évidemment inférieurs à ceux restant après la segmentation verticale seule.

En parcourant et en analysant visuellement les imagettes de n-grammes résultant des segmentations successives, nous constatons que les « n-grammes » persistants sont répétitifs, nous dirons qu'ils sont stables ou réguliers. Ceci prouve que la méthode de segmentation que nous avons utilisée est bien adaptée à ce genre d'écriture ; bien que cette méthode n'ait pas pu segmenter tous les mots en lettres individuelles, elle permet tout de même de segmenter l'écriture d'une façon assez régulière et consistante.

Nous n'avons pas voulu faire des tentatives de segmentation pour les « n-grammes » restants. Nous craignons, en fait, que de tels essais supplémentaires n'aboutissent pas à une séparation en lettres individuelles, mais conduisent à fragmenter l'image en miettes dérisoires ne portant aucune information significative pour des fins de reconnaissance.

A titre d'illustration nous donnons les modèles de formes de chacune des lettres et leur fréquence d'apparition dans notre base de travail.

Etiquette lettre	**Forme lettre**	**Pourcentage de présence**	**Etiquette lettre**	**Forme lettre**	**Pourcentage de presence**
a		10,25%	m		7,625%
a_ou_l		10,875%	n		1,5%
b		3,25%	o		8%
ch		0,125%	p		6,5%
d		3%	q		3,625%
ein		5%	r		1,375%
h		0,25%	s		4,625%
hh		1,875%	sa		3%
i		2,625%	t		4,75%
j		2,125%	ta		3,625%
k		3,125%	z		2,5%
l		3,125%			

Figure 38: Liste des lettres de l'alphabet ainsi que leur taux d'occurrence.

La régularité dans la présence des « n-grammes » persistants justifie le calcul leur taux d'occurrence dans notre base d'imagettes.

N-gramme	Forme	Pourcentage de présence	N-gramme	Forme	Pourcentage de presence
al		0,5%	alt		0,375%
alj		1,125%	li		0,875%
alk		1,625%	lt		2,5%
aln		0,25%			

Figure 39: Liste des « n-grammes » ainsi que leur taux de présence.

7 Comparaison des calligraphies et leur racine

Nous rappelons que nous disposons d'échantillons de deux des trois calligraphies Syriaques, à savoir de trois corpus rédigés en Serto et d'un corpus moins volumineux rédigé en Nestorien.

La racine linguistique d'une langue est son alphabet, la racine d'une écriture est sa calligraphie, c'est l'art dans le dessin des lettres. C'est au niveau des lettres que nous pouvons signaler les différences entre les calligraphies.

Nous étions capables de signaler la différence et de séparer les deux calligraphies Serto et Nestorien en trois niveaux ; visuel, algorithmique, et arithmétique.

7.1. Distinction visuelle

La première distinction consiste à marquer visuellement les traits particuliers pour chacune des deux calligraphies. Par un simple parcours à l'œil nu, nous pouvons constater une différence au niveau de l'apparence entre les deux calligraphies Serto et Nestorien. Nous pouvons dire que le Serto comporte un élément angulaire à 45° assez conséquent au niveau des lettres, par contre la calligraphie Nestorien s'avère plus plate et plus carrée, les lettres ont une apparence plus compacte.

En Serto il existe plus de lettres écrites penchées qu'en Nestorien, en outre, les mêmes lettres peuvent s'écrire de façon plus penchées avec la première calligraphie qu'avec la deuxième.

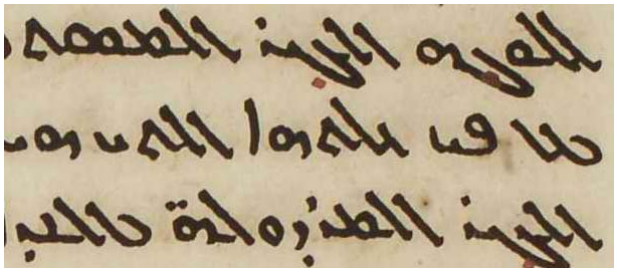


Figure 40: Document écrit avec la calligraphie Serto montrant le penché dominant.

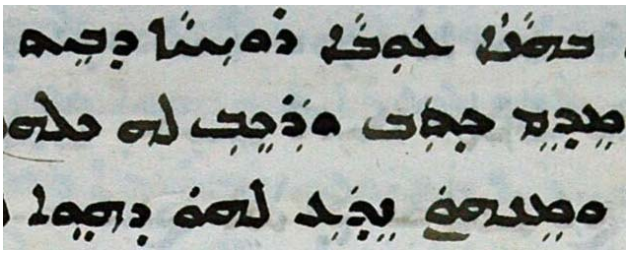


Figure 41: Document écrit avec la calligraphie Nestorien montrant une écriture plus plate.

7.2. Distinction algorithmique

Le processus de segmentations successives décrit ci-précédemment a été appliqué sur les documents rédigés en Serto plutôt que sur celui rédigé en Nestorien. Il s'est avéré qu'une segmentation suivant la direction verticale basée sur un profil de projection était suffisante pour les documents en Nestorien. Par la suite nous avons obtenu moins de « n-grammes », et nous tendons à une segmentation qui converge vite vers les lettres individuelles sans avoir à passer par une segmentation supplémentaire oblique.

Ceci concorde bien avec l'impression que nous avons eue, en première vue sur les deux calligraphies, et que l'aspect oblique de l'écriture est moins présent dans la calligraphie Nestorien.

7.3 Distinction arithmétique

La conclusion à laquelle nous sommes parvenus théoriquement ci-précédemment devrait être aussi valable au niveau calcul. Pour ce faire, et pour nos deux bases d'imagettes Serto et Nestorien, nous avons voulu montrer arithmétiquement la distinction calligraphique au niveau des lettres.

En utilisant la transformée de Hough [DUD72], nous avons calculé l'angle moyen de penchée des imagettes en Serto et celui des imagettes en Nestorien. L'angle moyen de penchée de l'écriture en Serto est d'environ 42°, et celui de l'écriture en Nestorien est d'environ 29°. Ce qui convient aussi bien avec la première impression que l'écriture en Serto est plus penchée vers un angle de 45° alors que l'écriture en Nestorien est plus plate et compacte.

Cette opération de calcul d'angle moyen d'inclinaison a permis une première séparation automatisée des deux calligraphies présentes entre nos mains. Bien que ce soit une approche grossière, elle permet une première classification et distinction des

deux calligraphies qui est assez satisfaisante avec une marge de différence angulaire d'environ 15°.

Afin d'illustrer ces calculs par un exemple, nous avons choisi de montrer une imagette de la lettre « b » écrite en Serto et une écrite en Nestorien dans la figure ci-dessous. Pour la lettre « b » en Serto l'angle de penchée calculé par la transformée de Hough est de 35°, quand cette même lettre est écrite en Nestorien, son angle de penchée calculé est de seulement 3°. Cet angle de penchée est montré par la ligne de plus grande pente tracée en couleur Cyan pour chacune des deux versions de la lettre « b » dans la figure ci-dessous.

a) « b » en Serto penchée de 35° b) « b » en Nestorien penchée de 3°

Figure 42: Différence de penchée entre la calligraphie Serto et la calligraphie Nestorien.

8 Conclusion

Nous avons dans ce chapitre décrit l'information qui était visible dans les documents Syriaques. Avant d'aborder un traitement d'images de documents du point de vue logiciel et informatique, il faut tout d'abord effectuer un dépistage visuel, et noter tout ce qui est perceptible à l'œil, cette connaissance nous servira pour mettre en place des outils capables d'extraire informatiquement ce que nous voyons.

Il est patent que le Syriaque est écrit avec beaucoup de sobriété, notamment parce que son écriture obéit à un principe d'économie de papier ; les espaces entre paragraphes n'existent pas, il n'y a pas de retour à la ligne, de plus les débuts de chapitres ne sont distinguables que par le fait que leurs titres sont écrits en une typographie plus grande et non pas en majuscules (trait inexistant dans cette écriture). Afin de faciliter la lisibilité de ces textes, les scribes ont développés un système particulier pour la ponctuation. L'écriture présente aussi une épaisseur du trait d'écriture, la rendant plus résistante face aux détériorations et à une numérisation à forte compression ou à faible résolution, ce sont des cas assez problématiques pour une écriture telle le Latin qui peuvent entraîner une dissolution du trait d'écriture.

L'inclinaison d'une partie des caractères à 45° est sa caractéristique la plus remarquable et va constituer un défi les chercheurs voulant de segmenter l'écriture en lettres individuelles. Pour ce faire, nous avons développé un algorithme adapté à ces documents consistant en une double segmentation qui nous a permis de décomposer l'écriture en lettres, d'une part, en « 3-grammes » stables ou réguliers, d'autre part.

Chapitre 3

Indexation des images de texte

…….par repérage de mots

1 Introduction

L'indexation d'un texte consiste à repérer en son sein, certains mots ou expressions particulièrement significatifs (*des termes*) dans un contexte donné, à créer ensuite un lien entre ces termes et le texte original ; par exemple, les pages d'index d'un livre reprennent les termes significatifs apparaissant dans le livre, et les relient aux pages du livre où ces termes (ou leurs synonymes) apparaissent. Ceci facilite, pour le lecteur, la localisation des pages ou des sections où l'on mentionnne un sujet particulier.

Rappelons aussi que, étymologiquement, *indexer* signifie montrer du doigt quelque chose qu'on veut identifier à telle ou telle fin.

L'indexation des documents manuscrits anciens requiert une recherche suivant leur contenu textuel. Cette tâche nécessite un temps et un effort considérables lorsqu'elle est effectuée manuellement. Avec l'expansion des projets de numérisation et le développement des bibliothèques électroniques, elle tend à devenir semi-automatisée, et cela, d'autant mieux que les méthodes de calcul et d'extraction automatiques des méta-données progressent.

Notre contribution à l'indexation de ces manuscrits se situe à deux niveaux du texte, nous allons élaborer :

- une méthode de reconnaissance collaborative de caractères que nous appellerons plus loin *transcription*
- et un principe de repérage des mots (clés !) dans les versions numérisées de ces documents.

Rappelons que l'image d'un texte n'est pas du texte, c'est une représentation numérique de ce texte, un fichier. On peut bien sûr repasser à un dessin, ce n'est en fait rien d'autre qu'une agglomération de pixels d'une certaine couleur sur un fond d'une couleur (de préférence) différente, le tout étant « décrit » en interne dans la représentation.

Après avoir introduit les deux catégories de méthodes que nous proposons, dans cette thèse, pour indexer les versions numérisées de documents manuscrits, nous décrirons en détails, dans ce chapitre, la première catégorie d'approches dites holistiques.

Deux axes complémentaires constituent l'essentiel de ces méthodes holistiques :
- le premier consiste en la formulation de la requête et de son mode de saisie,
- le deuxième consiste à l'élaboration du *moteur de recherche* approprié, lequel va faire des comparaisons entre les représentations des mots ; ce moteur de recherche peut être qualifié d'*iconique* voire de sémiotique à la différence des moteurs de recherches que l'on utilise habituellement et qui sont de nature symbolique.

L'approche heuristique sera l'objet du chapitre suivant intitulée «de la transcription assistée à la transcription collaborative ».

2 Approches pour l'Indexation

Comme nous l'avons rappelé dans l'introduction, il existe plusieurs approches pour l'indexation textuelle des images de documents manuscrits, elles relèvent d'une des deux catégories suivantes :
- les approches heuristiques, au niveau des caractères,
- les approches holistiques par repérage de mots.

La plupart des auteurs préfère une approche de niveau de mot plutôt qu'une approche de niveau de caractères (sauf dans le cas où le texte est en mode texte et où il existe des outils d'analyse du plain texte très performants).

2.1 Approches heuristiques

Les approches heuristiques reposent sur des méthodes de segmentation successives, qui partant de l'image d'une page de texte, permettent, autant que faire se peut, d'arriver :

→ jusqu'aux caractères ; dans ce cas-là, si nous sommes en situation de reconnaître ces caractères, nous parlons alors de Reconnaissance Optique des Caractères ou de son équivalent anglais Optical Character Recognition dont l'acronyme est OCR.

→jusqu'aux graphèmes, c'est-à-dire à un ensemble de composantes de base du tracé, déterminées par un principe algorithmique ; ces graphèmes peuvent correspondre à des morceaux de caractères ou à des groupes de caractères spécifiques à une langue.

→jusqu'au niveau des traits de base (comme c'est le cas des écritures Chinoise, Japonaise, et d'autres alphabets asiatiques).

A partir des résultats de cette segmentation nous pouvons passer à la reconnaissance, donc à la construction d'une version en mode texte du contenu textuel de l'image, c'est-à-dire en ASCII ou en Unicode. Ce qui sera commun à toutes les situations, c'est l'élaboration d'un apprentissage que nous permettra d'inférer des règles de reconnaissances avec un dictionnaire de formes ou de caractères ou un code book. Il est clair que ces méthodes de reconnaissance dépendent fortement de la robustesse de l'algorithme de segmentation,

2.2 Approches holistiques

Les approches holistiques abordent le problème d'indexation au niveau des mots, le mot devenant alors la « plus petite entité » que l'on veut reconnaître ou repérer.

Les premiers travaux

L'équipe de Manmatha, à l'University of Massashusetts à Amherst, a été la pionnière dans l'élaboration de ces méthodes et ses travaux sur la collection de Georges Washington numérisée en niveaux de gris sont toujours cités [MAN05], [RAT07], [RAT03], et [MAN03] ; les chercheurs de cette équipe ont simulé un moteur de recherche basé sur une approche de repérage de mots dite de « Word Spotting ».

Les caractéristiques qu'ils ont utilisées étaient des caractéristiques multidimensionnelles de profil, tels que le profil de projections, le profil de mot et les transitions de fond/encre. Leur algorithme d'appariement est basé sur du DTW (Dynamic Time Warping) qui permet de faire correspondre les caractéristiques extraites des images des mots.

Dans certains de leurs autres travaux récents [LAV04], ils ont combiné les caractéristiques scalaires avec les caractéristiques de profil pour la reconnaissance holistique de mots et la correspondance est faite en utilisant les chaînes de Markov cachées HMMs (Hidden Markov Models). Ils ont également comparé une approche de SVM (Support Vector Machine) à un modèle de Bayes Naïf combiné avec des évaluations gaussiennes de densité pour la reconnaissance holistique de mots [FEN05].

Les travaux de Terasawa et al. ont été effectués sur les manuscrits Japonais anciens, cette équipe a aussi tenté une approche de « Word Spotting » basée sur l'espace propre (Eigen Space) [TER05], et [TER07]. Les signatures sont extraites à partir des fenêtres glissantes. Les niveaux relatifs du gradient dans les 8 directions principales sont calculés. Ceci mène à des caractéristiques qui sont invariantes à l'échelle.

Les différences morphologiques entre les mots sont surmontées en employant un algorithme de correspondance basé sur le DTW. Le DTW est aussi utilisé dans [BAL06] pour la correspondance entre mots. Une autre approche sans-segmentation basée sur les Modèles de Markov Cachés et d'autres modèles statistiques est décrite dans [VIN04].

Les travaux sur le Word Spotting semblent émerger dans le contexte de la recherche internationale, c'est ce que notait le professeur Guy Lorette lors de la dernière conférence ICDAR à Barcelone en juillet 2009.

Les recherches du LIRIS

Depuis près de 10 ans l'équipe lyonnaise qui m'a accueillie a entrepris des investigations sur le Word Spotting, tout en essayant de lui donner un statut, alternative à l'OCR, moteur de recherche ; elle propose aujourd'hui plusieurs voies.

Leydier et al. ont travaillé sur les manuscrits latins médiévaux numérisés en niveaux de gris, le but de leurs travaux était l'extraction et l'identification des mots choisis. Pour ce faire ils ont construit des caractéristiques basées sur l'orientation du gradient [LEY07], [LEY05]. Ces caractéristiques sont invariantes aux transformations géométriques et stables par rapport aux variations d'échelles. Après avoir représenté le mot requête par un vecteur de caractéristiques, une recherche dans les différentes pages du document est effectuée afin de retrouver toutes ses occurrences. L'algorithme de correspondance repose sur un balayage de surface qui permet de trouver une superposition correcte entre les caractéristiques.

Trois méthodes d'appariement ont été testées.

La première est une correspondance naïve et rigide dans laquelle l'image du mot requête effectue un balayage sur la surface d'une page de texte, chaque pixel est

considéré comme une zone d'intérêt, et la comparaison s'effectue pixel par pixel, cette méthode d'appariement ne tolère aucune variation spatiale et nécessite une superposition parfaite de l'image du mot requête et celle du texte à chercher.

La deuxième consiste en un algorithme de correspondance élastique dans laquelle un pixel de l'image du mot requête est apparié au gradient le plus proche de son voisinage dans l'image de texte, cette méthode est plus précise que la première mais présente une forte complexité.

Dans la troisième méthode, ils cherchent à trouver l'information la plus représentative, dans les documents Latins, cette information est située dans les traits verticaux. Comme la distance entre ces traits pourrait varier, ce qui pourrait entraîner un resserrement ou bien un espacement entre les lettres qui pourrait différer entre les occurrences, il est plus approprié de comparer les pixels autour de ces traits que de comparer les formes entières. La morphologie mathématique est appliquée aux niveaux gris des images afin de calculer des guides. Une ouverture avec un élément structurant vertical est effectuée, et les boîtes délimitant les guides sont agrandies pour obtenir les zones d'intérêt.
Le mot requête est ensuite fragmenté suivant les zones d'intérêt, et la recherche de correspondance est effectuée par fragment. En utilisant cette méthode, ils ont pu retrouver toutes les occurrences de mots requêtes sur les documents Latins, ils ont même testé cette méthode sur la base utilisée par Rath et al. [RAT03], et ils ont obtenu un taux de récupération de 85.18% (23 occurrences retrouvées parmi 27) alors que celui obtenu par Rath et al. était de 77.64% (21 occurrences retrouvées parmi 27).

Le travail que je présente ici s'appuie sur les « roses des directions «.

En relation directe avec les travaux présentés dans le cadre de cette thèse, nous citerons les travaux de **Brès et Wagan** qui font l'objet d'une thèse en cours et qui portent sur un nouveau descripteur baptisé MSIO pour Multi Scale Integral Orientations. L'objectif de ce descripteur est de caractériser des détails importants sur les images à partir de l'information liée à l'orientation locale des gradients, et cela à plusieurs échelles. Par ailleurs, une optimisation de l'extraction de cette information permet de l'obtenir de façon très rapide ce qui nous donne la capacité de trouver le descripteur de n'importe quelle région dans l'image en temps constant.

Les caractérisations envisagées a priori par ce descripteur sont les suivantes, donc différentes de celles que nous allons présentées dans le cadre de cette thèse :
- les MSIO s'attachent à réaliser une description « exhaustive » de la surface (2D) de l'image
- nous proposons une description sélective du tracé (1D) de l'écriture.

Ces deux descripteurs partagent donc la même information de base, à savoir l'information d'orientation locale calculée par l'intermédiaire de gradients, mais nous proposons une approche plus directement spécifique et adaptée à des applications liées aux écritures alors que les MSIO ont une vocation plus généraliste de caractérisation d'images.
Ils ont déjà été testés pour le Word Spotting et pour la classification de scripteurs, mais aussi dans des domaines plus variés comme la localisation ou la reconnaissance de visages ou l'indexation de formes 3D.

3 Repérage de mots et saisie du mot requête.

Par définition le repérage de mots consiste en un parcours rapide d'un support enregistré afin de localiser un mot et/ou un passage déterminé. Le processus commence à partir du choix du mot (la requête) que nous souhaitons retrouver, ce mot sera proposé ou soumis à un moteur de recherche s'occupant de la fouille des données, lequel nous retournera les positions dans lesquelles se trouve notre requête.

Deux aspects sont à évoquer :

- le premier consiste en la forme que pourrait prendre la requête ainsi que sa procédure de nomination ou de saisie,
- le deuxième consiste à l'élaboration du moteur de recherche approprié.

Pour le repérage de mots nous pouvons distinguer trois modes de saisie du mot requête : la saisie par pointage sur l'image, la saisie manuscrite et la saisie par l'intermédiaire du clavier. Ces modes correspondent respectivement à ce que l'on trouve dans la littérature sous les vocables de word spotting, word retrieval.

3.1. La saisie par pointage

Ce mode de saisie consiste à sélectionner l'image du mot requête à partir de l'image du document lui-même. L'avantage de cette méthode est essentiellement dû au très haut niveau d'homogénéité de l'image du mot requête et de celle du document, puisque le mot cherché fait physiquement partie du support enregistré ; cette cohérence repose sur le fait que l'image du document et l'image de la requête :

- ont été produites par le même scribe (généralement),
- ont subi les mêmes conditions de numérisation, contiennent les mêmes bruits,
- et, par voies de conséquences, requièrent le même prétraitement (sans qu'il soit nécessaire d'avoir recours à des interventions supplémentaires au niveau de l'une ou l'autre des images).

L'inconvénient majeur de cette méthode de saisie est le fait que l'utilisateur doit pointer lui-même une première occurrence du mot qu'il cherche ; il doit donc faire un premier balayage visuel du document en question. Si le mot est fréquent, c'est aisé, au contraire, si mot rare (voire absent), il faut imaginer d'autres stratégies.

3.2. Saisie manuscrite directe

Pour cette approche, l'utilisateur est muni d'un périphérique de saisie. Ce périphérique pourra être une tablette graphique, un crayon optique, ou bien un simple instrument d'écriture sur papier et un scanner. Plusieurs difficultés sont inhérentes à ce mode de saisie.

La première vient du fait que l'utilisateur est un scribe étranger par rapport au document, il a un style d'écriture et une « touche » qui lui sont propres. C'est un scripteur contemporain, alors que le document pourrait être un manuscrit ancien ; même si la police d'écriture est la même que celle du document, cette dernière a subi des évolutions et des changements depuis la date de la création du manuscrit jusqu'au temps présent.

De fait, l'utilisateur crée une image étrangère par rapport au document, ce qui nécessitera une homogénéisation entre la requête et l'image du manuscrit ; une cohérence colorimétrique devra être présente, vu le fait que l'arrière plan du document présente souvent un aspect non homogène, lequel ne pourra pas être

reprodui sur le fond de la requête. Une phase de passage à un mode bicolore (binaire) ou bien un mode encre noir sur fond papier blanc est nécessaire, mais cette binarisation entraîne une perte conséquente de l'information, surtout si le document en question est très dégradé et/ou que le tracé consiste en un fin trait d'écriture qui pourrait se trouver effacé.

La troisième difficulté apparaît lorsque la saisie est effectuée par l'intermédiaire d'une tablette graphique ou d'un crayon optique. Deux types de sauvegardes sont à envisager ; la sauvegarde sous la forme d'une image, et la sauvegarde sous la forme de coordonnées des points correspondant au mouvement de la souris ou du stylo lorsqu'il est appuyé sur la tablette, ces coordonnées sont stockées dans le même ordre que leur création.

Dans le cas d'une sauvegarde sous la forme de coordonnées, la recherche du mot requête dans l'image de document nécessite une comparaison entre des données dynamiques (qui sont les coordonnées), et des données statiques sous la forme d'agglomérations de pixels. Cette comparaison requiert une reconstitution du parcours du tracé manuscrit statique, en d'autres termes un besoin de chercher à retrouver la trajectoire qu'a empruntée le stylo pour créer le texte, d'où le besoin de squelettiser le tracé de l'écriture afin d'extraire la ligne fine correspondant à cette trajectoire, une démarche qui résulte en une perte d'informations significatives pour la reconnaissance. Cette approche fut d'abord adoptée par Rousseau et al. pour la reconnaissance de caractères isolés [ROU04], elle fut appliquée ultérieurement dans des buts d'indexation de documents numérisés [ROU07].

Dans le cas d'une sauvegarde sous la forme d'une image, il est important de retrouver la même épaisseur de tracé de trait entre le mot requête et l'image du document, sinon une squelettisation est là aussi nécessaire. De plus, il est important de retrouver les mêmes paramètres de compression et de résolution entre l'image requête et le document intégral.

Enfin, dans le cas d'utilisation d'un simple instrument d'écriture et d'un papier pour reproduire le texte, il faudra scanner pour obtenir la version numérique de la requête. Les instruments d'écriture modernes sont très différents des instruments anciens et ne donnent pas le même effet final ; chaque instrument d'écriture produit une épaisseur de trait différente. Un spécialiste des écritures anciennes pourrait, peut-être, reproduire la même écriture avec un instrument moderne, mais un utilisateur classique ayant besoin d'une information ne dispose pas de tous les outils anciens d'écriture ; ceci va créer des différences et augmenter la dissemblance entre le mot requête et le document.

3.3. La saisie au clavier

Dans cette méthode de saisie, un utilisateur tape au clavier le mot qu'il souhaite chercher. Cette méthode est la plus intuitive et la moins encombrante à l'utilisateur. Par contre, elle devra être soutenue par une base de données conséquente contenant des échantillons d'images de lettres sous toutes leurs formes. Cette base de données devra être étiquetée, chaque groupe d'images de la même lettre manuscrite sera libellé par son équivalent typographié. Quand l'utilisateur saisira le mot, chacune des lettres nous conduira vers l'étiquette d'une classe manuscrite ; le mot manuscrit sera reconstitué à partir des combinaisons des lettres. L'image du mot reconstitué sera

envoyé vers le moteur de recherche qui parcourra le document afin de retrouver les occurrences possibles du mot requête. Cette approche porte alors le nom de « Word Retrieval ».

3.4 Le moteur de recherche

Le moteur de recherche est le même, quelle que soit la méthode de saisie de la requête, il constitue le noyau du processus de repérage ; c'est l'outil le plus complexe à construire. Les différences de stratégies entre les moteurs de recherche se situent au niveau :
- des caractéristiques extraites,
- de l'algorithme d'appariement,
- des permissions de recherche (filtrage des mots vides, éliminations des terminaisons et des conjugaisons, etc…).

4 Notre moteur de recherche

Notre méthode de repérage de mots « Word Spotting » est basée sur une sélection de fenêtres glissantes desquelles nous extrayons des caractéristiques (directionnelles) des objets à comparer.

4.1 Le contexte

Dans ce paragraphe, nous n'utiliserons que des mots requêtes présents dans le document de recherche. Suivant la définition donnée au début de ce chapitre, indexer consiste à repérer des mots particuliers ou significatifs, des mots qui correspondent au mieux au contenu informationnel du texte. Un mot qui apparaît souvent dans un texte représente un contexte important. Donc nous avons tendance à cibler des mots fréquents comme choix pour nos mots requêtes. Un mot est dit fréquent s'il figure plusieurs fois dans un corpus et dans des pages différentes. D'après cette définition, nous nous apercevons que les mots les plus fréquents sont des mots fonctionnels ou mots outils, les prépositions pourraient être considérées comme des mots fréquents par la nature de leur utilisation, par contre elles ne sont pas choisies comme mots requêtes puisqu'elles ne rapportent aucune information significative pour l'indexation.

Notre but c'est de pouvoir repérer et retrouver toutes les occurrences d'un mot fréquent « significatif ». Ce repérage est effectué en utilisant une seule image d'un mot fréquent à partir de laquelle nous devons retrouver toutes les autres occurrences.

Les corpus dont nous disposons sont tous mono-scripteurs, sauf un seul corpus qui a été rédigé par trois scripteurs différents. Nous sommes donc dans des conditions favorables bien que la façon d'écrire un mot donné puisse différer d'une page à une autre, même si le scribe, aussi habile soit il, est inchangé ; plus le corpus est long, plus les différences peuvent être considérables. Ces fluctuations dans l'écriture sont dues à plusieurs causes :
- la fréquence de trempage de la plume ou de l'instrument d'écriture dans l'encrier,

- la capacité d'absorption de l'encre par le papier,
- l'affutage ou bien au contraire l'usure de l'instrument d'écriture,
- les variations des conditions de luminosité,
- les tentatives de justifier l'écriture,
- la fatigue et l'humeur du scribe,
- les soucis de consommation excessive de papier, etc...

Les variations seraient alors visibles au niveau de :

- L'épaisseur de l'écriture.
- L'espacement inter-mots.
- L'espacement interligne.
- L'extension ou bien le resserrement des mots.
- La disposition du l'écriture par rapport à la ligne médiane.
- La penchée des caractères.

Ces différences peuvent réduire le degré de similarité entre l'image du mot requête et celles de ses occurrences. Ci-dessous l'image d'un des mots que nous avons choisi comme mot requête, ainsi que certaines des ses occurrences, les différences mentionnées ci-précédemment sont visibles entre les images. Il faudra faire en sorte que notre caractérisation ne soit pas trop influencée par ces variations !

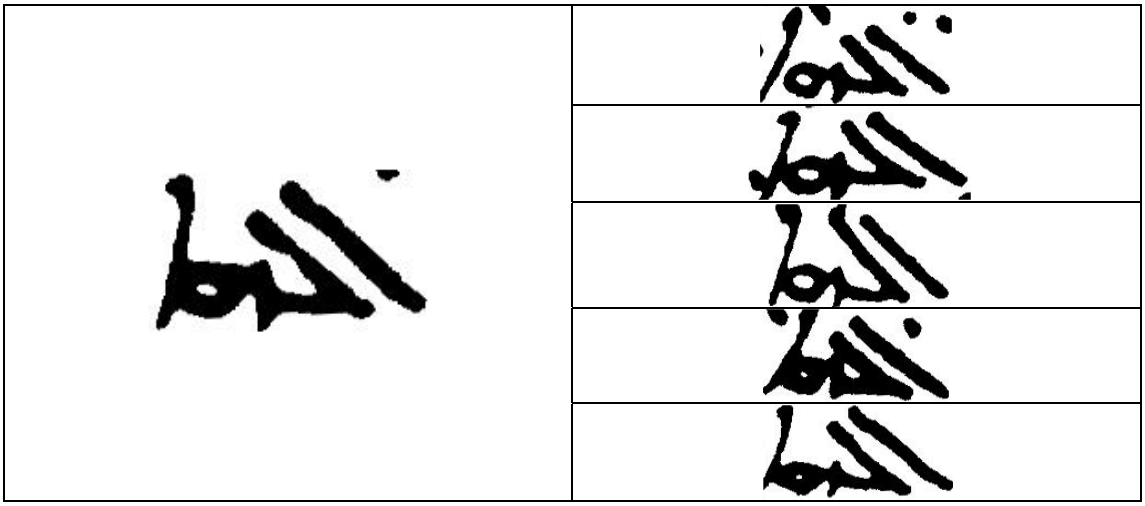


Figure 40: Un mot fréquent et les différences visibles avec ses occurrences.

4.2 Les principes de notre moteur

Le principe de notre démarche va être de décomposer les lignes d'écritures en petits morceaux (des fenêtres de taille assez petites) dans lesquels on « mesurera » la direction, plus précisément les directions, avec les techniques d'auto-corrélation dont les résultats seront résumés par une rose des directions

4.3 Les fenêtres glissantes : élaboration et sélection

La méthode que nous proposons est basée sur une technique de fenêtres glissantes sélectives. Contrairement à Terasawa et al. qui ont pris en compte toutes les fenêtres glissantes qu'ils pourraient extraire, nous avons éliminé les fenêtres qui ne répondent pas à nos critères de choix..

Nous avons introduit deux formes de fenêtres glissantes :
- la première forme est une fenêtre carrée de taille 32×32 pixels,
- la deuxième est une fenêtre rectangulaire de taille 96×32 pixels.

Nous justifierons les choix de tailles plus tard.

La hauteur moyenne des lignes de texte est de 96 pixels, quand une ligne de texte est divisée en trois zones d'écriture (supérieure, centrale, et inférieure), délimitées par des lignes de bases « upper and lower baseline ». La hauteur de chaque zone est de l'ordre de 32 pixels ; si nous souhaitons glisser une fenêtre carrée dans chaque zone, la largeur de cette fenêtre doit être 32 pixels. Nous avons repéré les lignes de bases en utilisant les projections horizontales. Les lignes de bases d'une ligne de texte particulières sont montrées dans la figure ci-dessous.

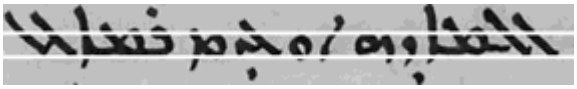


Figure 41: Détection des lignes de base supérieure et inférieure d'une ligne de texte.

Dans chacune de ces trois zones, nous allons faire glisser une fenêtre carrée de taille 32×32 pixels de gauche à droite avec un pas de 1 pixel. A chaque pas, une analyse du contenu de la fenêtre est effectuée et nous éliminons les fenêtres qui ne répondent pas aux critères suivants :

a) Tout d'abord, les fenêtres ayant une densité de pixels noirs inférieure à un seuil particulier seront rejetées. Nous estimons qu'une fenêtre glissante ne contient pas d'information si il y a moins de 1/5 de sa surface en pixels noirs. Puisque la largeur moyenne de trait d'écriture dans nos documents est de 12 pixels, une lettre de petite taille, ou bien une ligature occupera en moyenne 16 pixels, 12*16 pixels est environ le 1/5 des 32*32 pixels.

b) Nous étudions le mouvement du centre de gravité des pixels noirs, et nous gardons les fenêtres ayant un centre de gravité qui a bougé suffisamment, c'est-à-dire de plus de 16 pixels qui est la moitié de la fenêtre, suivant l'axe des abscisses, par rapport à son prédécesseur. Si le centre de gravité de la zone noire ne s'est pas trop déplacé, nous estimons que les deux fenêtres contiennent des formes presque identiques

c) Une fenêtre qui couvre plus que la moitié de la surface de celle qui la précède sera rejetée car on considère qu'elle ne contient pas d'information différente ; cette dernière étape consiste donc à éliminer la redondance.

4.4 Extraction de caractéristiques

L'aspect particulier et propre à l'écriture Syriaque est le pencher de certaines lettres dans une direction à 45 degrés alors que d'autres sont verticales ou horizontales. Ces directions sont ce qui caractérise le mieux cette écriture. C'est la raison pour laquelle pour la décrire le plus fidèlement, nous avons choisi d'extraire des caractéristiques directionnelles sous la forme de roses à 8 directions. Cette méthode avait été utilisée pour d'autres applications telles que l'identification de scripteurs dans des anciens manuscrits [EGL07], ainsi que pour la description et l'analyse de la mise en page d'anciens documents imprimés [JOU05]. Afin d'extraire ces caractéristiques nous procédons de la façon suivante :

a) Les fenêtres que nous avons choisi de garder sont divisées en quatre sous-fenêtres de taille 16×16 pixels chacune, cela nous permettra de capter un trait ou une ligature dont l'épaisseur moyenne est de 12 pixels.

b) La fonction d'auto-corrélation est calculée dans chacune de ces quatre sous-fenêtres. Les motifs résultants de cette auto-corrélation représentent les directions principales dans les quatre quadrants d'une fenêtre.

Fonction d'auto-corrélation

Soit l'image f, de taille $M\times N$, dans notre cas 16×16 pixels, la fonction d'auto-corrélation de cette image est évaluée comme suit, tout d'abord nous calculons la transformée de Fourrier de cette image, la fonction d'auto-corrélation dans le domaine fréquentiel serait le module de la transformée de Fourrier élevé au carré, pour revenir au domaine spatial nous calculons la transformée de Fourrier inverse du résultat précédent qui nous permet d'obtenir la fonction d'auto-corrélation:

$$F(u,v)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y)e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)}$$

$$f(x,y)\circ f(x,y)=\frac{1}{MN}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}f^{*}(m,n)f(x+m,y+n)$$

$$f(x,y)\circ f(x,y)\Leftrightarrow F^{*}(u,v)F(u,v)=\left|F(u,v)\right|^{2}$$

$$f(x,y)\circ f(x,y)=\frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}\left|F(u,v)\right|^{2}e^{j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)}$$

c) Cette information est résumée sous la forme d'une rose à 8 directions.
Chaque sous-fenêtre est représentée par une signature de 8 valeurs résultant en un total de $8\times4=32$ valeurs représentant chaque fenêtre. La longueur d'une direction est obtenue en faisant la somme des niveaux de gris de la fonction d'auto-corrélation dans cette direction.

Pour garder l'information la plus discriminante, nous avons choisi de garder les variations des directions qui sont au dessus de la direction la moins représentée qui, elle, sera mise à zéro. La direction dominante sera mise à 1, et ceci pour réduire l'influence de la dynamique de l'image d'origine.

De plus, l'utilisation de la fonction d'auto-corrélation pour définir les signatures locales réduit l'influence du bruit et des dégradations puisque c'est la structure de la sous-fenêtre qui va influer sur le résultat.

La figure ci-dessous montre les quatre sous-fenêtres associées à une fenêtre glissante sélectionnée, ainsi que leurs fonctions d'auto-corrélation respectives.

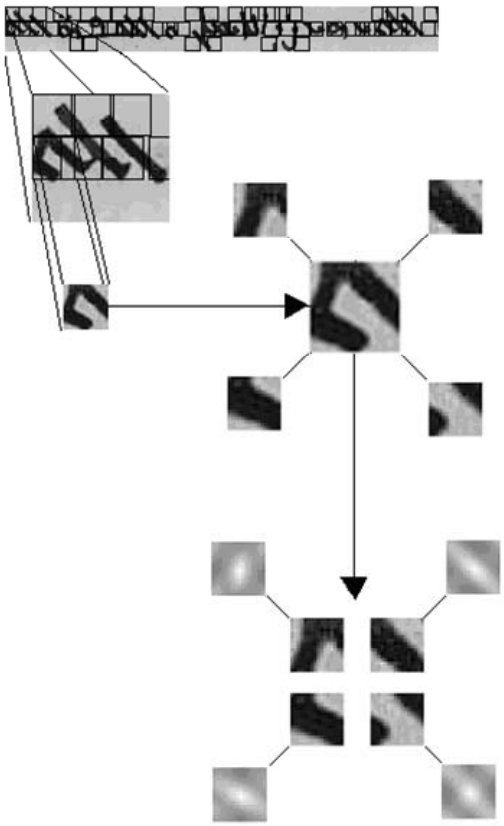


Figure 42: Les fonctions d'auto-corrélation des quatre sous-fenêtres associées à une fenêtre glissante.

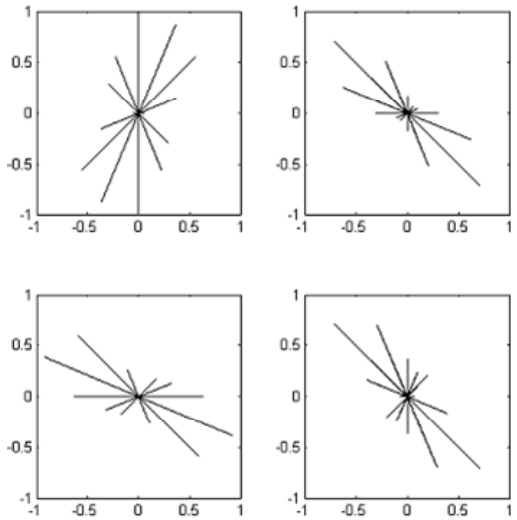


Figure 43: Roses de directions associées aux sous-fenêtres de la figure 14.

La figure ci-dessus laisse apparaître les roses de directions associées aux quatre sous-fenêtres de la figure précédente. Une fenêtre glissante sélectionnée est représentée par les 4 roses chacune à 8 directions.

Les roses de direction mettent bien en relief la direction saillante dans chaque sous-fenêtre. Malgré que cette décomposition soit invariante à l'échelle, nous ne prenons pas cette propriété en considération puisque nos sous-fenêtres ont une taille

fixe. Par contre une normalisation pourrait être effectuée en phase de prétraitement afin de bien situer le texte entre les lignes de base.

La réduction de taille ne résulte en aucun cas en une perte de précision vu que la taille que nous avons choisie pour les caractères et éventuellement les mots contient suffisamment d'information pour notre application, une augmentation de taille résultera en un effet de flou sur l'image, ce cas sera évoqué en détails plus tard.

4.5 Algorithme d'appariement

Quand toutes les signatures de toutes les sous-fenêtres de toutes les fenêtres sélectionnées des trois zones de l'image du mot requête ont été extraites, on peut alors rechercher et retrouver toutes leurs occurrences. Elles seront comparées à toutes celles extraites des trois zones de chaque ligne de texte de l'image de la page de test, celles qui leur sont les plus similaires sont détectées et la région ayant l'agglomération de sous fenêtres similaires à celles de l'image du mot requête la plus importante est considérée comme étant une occurrence possible du mot requête.

Par contre, une décision basée seulement sur ce critère de choix n'est pas très fiable étant donné qu'une superposition de candidats de correspondance possibles peut avoir lieu dans les trois zones d'une ligne de texte résultant en des choix incorrects mais possibles d'occurrences du mot requête.

Pour surmonter ce problème de confusion, nous avons pensé à estimer grossièrement les endroits où les occurrences correctes du mot sont susceptibles d'être situées, nous allons référer à ces endroits là en les nommant « régions d'intérêt ». Afin de détecter ces régions, nous traçons le graphe des positions des centres de gravité des pixels noirs des fenêtres glissantes dans chacune des trois zones du mot requête, puis nous traçons aussi le même graphe pour chacune des trois zones pour toutes les lignes de la page de test.

Les portions de graphes issus de la page de texte les plus semblables à celles du mot requête sont repérées, une région d'intérêt sera celle où se superposent trois graphes jugés semblables à ceux du mot requête. L'évaluation de la similitude est basée sur le minimum de la distance euclidienne. Une région d'intérêt bien situé peut être considérée comme une occurrence du mot requête, par contre une région d'intérêt mal placée sera une fausse occurrence.

En couplant l'appariement direct des fenêtres glissantes avec la pré-détection des régions d'intérêt, nous avons effectué une double élimination. La première consiste à écarter les fenêtres qui ne sont pas situées dans une région d'intérêt, et la deuxième consiste à éliminer les régions d'intérêt qui ne contiennent pas des fenêtres.

4.6 Remarque

Les caractéristiques directionnelles que nous utilisons pour le repérage de mots ont été utilisées dans la littérature pour de l'identification de scripteur. En effet Sidiqqi et al [SID08] utilisent des descripteurs directionnels pour faire l'identification de scripteur

Ils utilisent des fenêtres glissantes de différentes tailles, qu'ils divisent « implicitement » en quatre sous-régions afin de déterminer dans chacune la direction saillante du tracé de l'écriture. Parmi les caractéristiques qu'ils extraient ; le profil

vertical, le profil horizontal, et l'orientation générale de l'écriture, ce que nous nous appelons la direction saillante du tracé.

5 Résultats

Nous allons illustrer cette approche dans trois cas différents où nous cherchons les occurrences du mot requête sur une page du document numérisé, celle qui le contient :
- la premier concerne une image de texte normal à compression et résolution raisonnables (que nous allons utiliser comme référence d'évaluation),
- le deuxième concerne la même image de texte comprimée en JPEG qualité 0,
- le troisième concerne encore la même image de texte pour laquelle nous avons réduit la résolution puis restituée de nouveau à sa valeur initiale par interpolation cubique.

5.1 Résultats obtenus sur une image normale

Appariement direct des fenêtres glissantes

Dans la figure ci-dessous, les fenêtres retenues dans les trois zones de texte pour un mot requête, ainsi que celles qui leurs sont les plus similaires dans les lignes de texte sont détectées. Il y a trois endroits où les superpositions correspondent aux occurrences correctes du mot, par contre nous remarquons un endroit où il y a des superpositions semblables à celles du mot requête mais qui ne correspondent pas à une occurrence correcte.

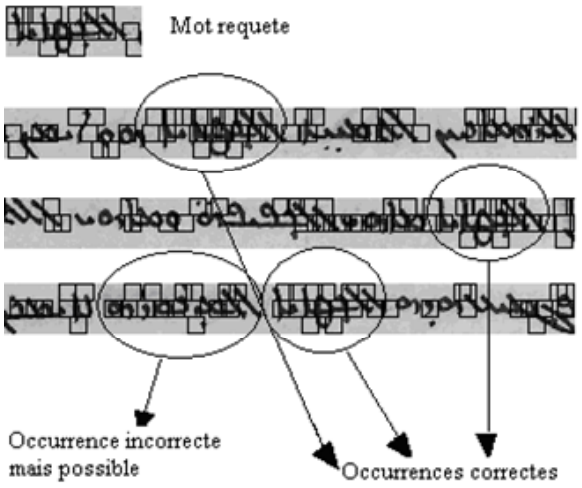


Figure 44 : Mauvaise occurrence détectée sur la similarité des fenêtres.

Détection des régions d'intérêt

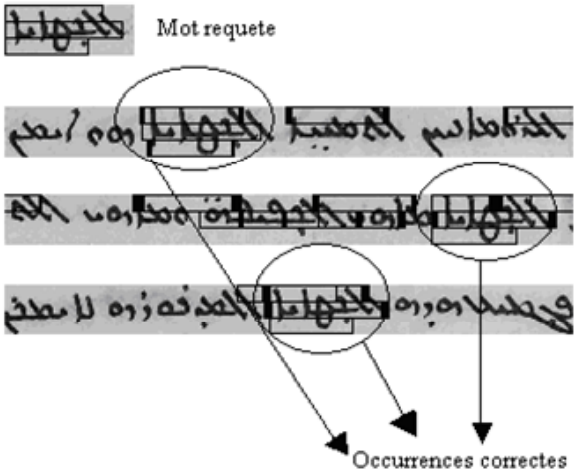


Figure 45: Régions d'intérêt associées au mot requête.

Dans la figure ci-dessus, nous détectons les régions d'intérêt associées au mot requête, les régions d'intérêt sont la superposition dans les trois zones de graphes semblables à ceux du mot requête. Nous remarquons que cette superposition a eu lieu sur les bonnes occurrences. Donc les régions d'intérêt sont bien situées.

Double élimination

Désormais, les coefficients des roses de directions extraits à partir de l'image du mot requête seront comparés à ceux extraits des régions d'intérêt. La correspondance est basée sur le minimum de distance euclidienne, et les occurrences correctes sont celles ayant le plus grand nombre de coefficients correspondants dans les trois zones.

Nous voyons bien dans la figure ci-dessous que la fausse occurrence qui était retenue à la première étape se trouve éliminée, et les occurrences correctes sont restées intactes. Aucune occurrence correcte n'a été ratée.

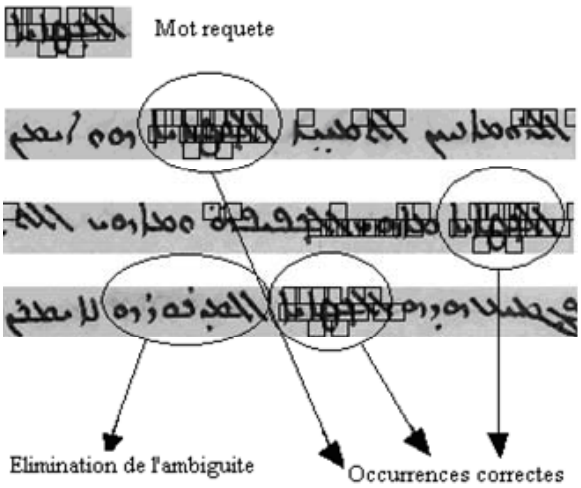


Figure 46: Elimination de l'ambigüité et repérage des occurrences correctes.

5.2 Résultats obtenus sur une image fortement comprimée

La compression JPEG avec perte excessive introduit des artéfacts visibles autour du contour des caractères, ce qui réduit la lisibilité du document et ajoute des contraintes et des difficultés à tout procédé de traitement d'images.

Les figures ci-après montrent une image de texte avant et après la compression excessive, les artefacts sont bien visibles sur le contour des caractères. En termes de lisibilité et de continuité du tracé, l'écriture Syriaque s'avère assez résistante face à ces artéfacts.

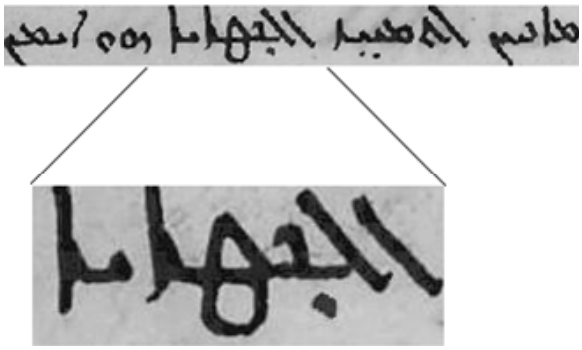


Figure 47: Image à compression raisonnable.

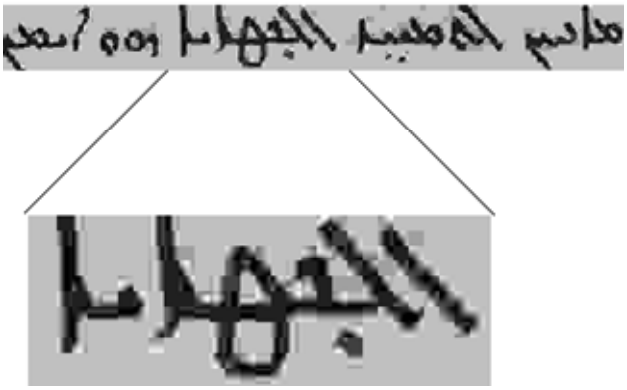


Figure 48: Image ayant des artefacts résultant d'une compression excessive.

Appariement direct des fenêtres glissantes

Nous remarquons qu'en essayant de faire correspondre des fenêtres sur les trois zones, nous pourrons obtenir des fausses occurrences, comme le montre la figure ci-dessous. Les trois occurrences correctes sont repérées avec une fausse occurrence supplémentaire.

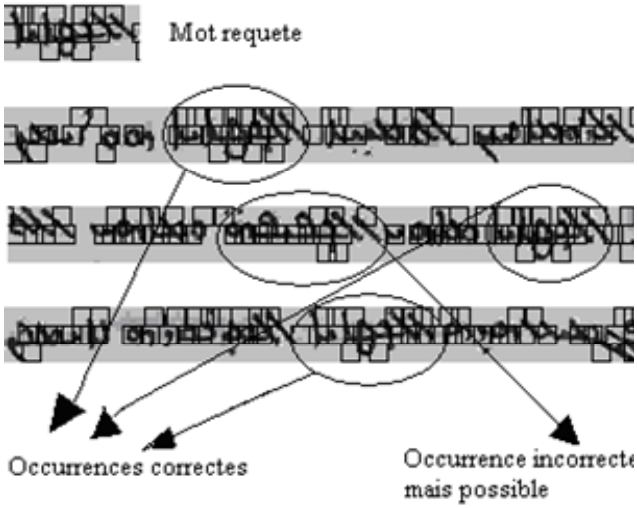


Figure 49: Occurrence incorrecte détectée sur la similarité des fenêtres.

Détection des régions d'intérêt

La pré-détection des régions d'intérêt résiste à cette dégradation et met en avant les

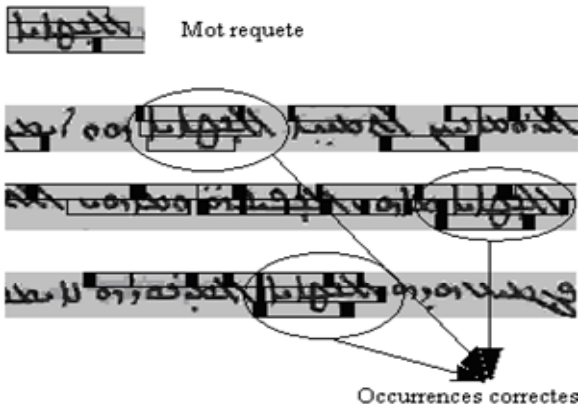


Figure 50: Régions d'intérêt associées au mot requête.

endroits dans lesquels se trouvent les bonnes occurrences. Nous voyons bien que les régions d'intérêt sont bien situées.

La double élimination

La combinaison des deux étapes précédentes a entrainé l'élimination de l'occurrence incorrecte puisque les fenêtres constituant cette fausse occurrence ne sont situées dans aucune des régions d'intérêt, les trois occurrences correctes ont été correctement repérées.

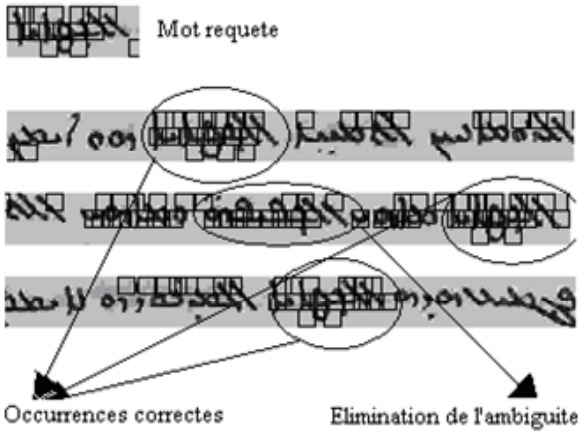


Figure 51: Elimination de l'ambigüité et repérage des occurrences correctes.

5.3 Résultats obtenus sur une image à faible résolution

Un numériseur réglé avec une faible résolution ne permettra pas de capter tous les détails de l'image en cours de numérisation. Une tentative de hausser ultérieurement la résolution d'une image résultera en une version floue de cette dernière.

Nous avons réduit la résolution de l'image de test d'origine, puis nous l'avons restituée de nouveau à sa valeur initiale par interpolation cubique. L'effet de flou est bien visible comme la figure ci-dessous le montre.

Tout comme le cas des dégradations JPEG, l'écriture Syriaque se montre assez résistante et garde son tracé clair et sa lisibilité lorsque confrontée à une faible résolution.

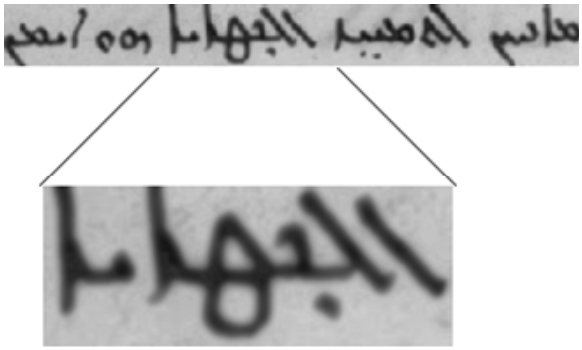


Figure 52: Effet de flou obtenu par tentative de hausser la résolution d'une image.

Appariement direct des fenêtres glissantes

Nous remarquons de même dans le cas d'une image à faible résolution qu'en essayant de correspondre des fenêtres sur les trois zones, nous pourrons obtenir des fausses occurrences, comme le montre la figure ci-dessous. Les trois occurrences correctes sont repérées avec une fausse occurrence supplémentaire.

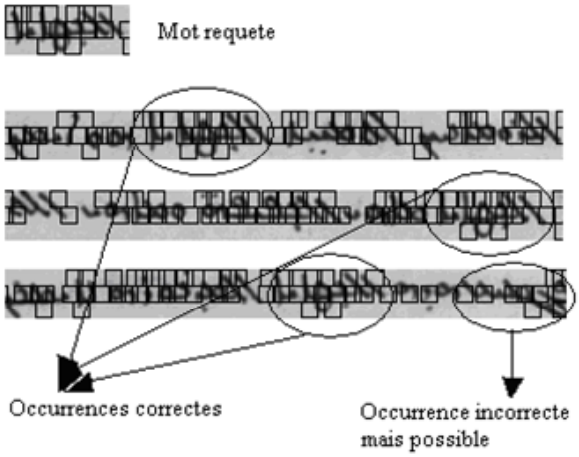


Figure 53: Occurrence incorrecte détectée sur la similarité des fenêtres.

Détection des régions d'intérêt

La pré-détection des régions d'intérêt résiste de même à ce genre de bruit et met en relief les endroits dans lesquels se trouvent les bonnes occurrences. Nous voyons que ces régions sont aussi bien situées.

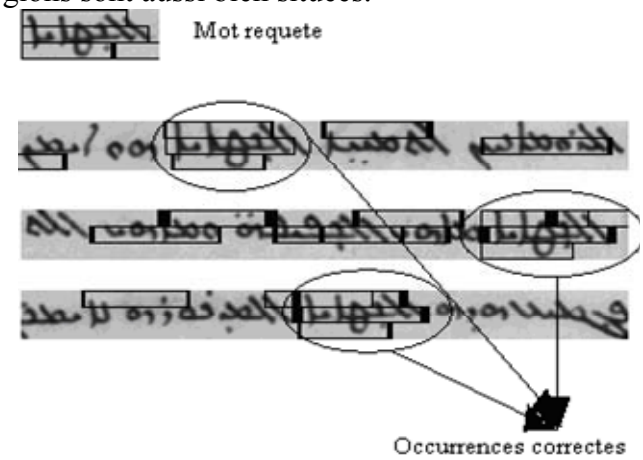


Figure 54: Régions d'intérêt associées au mot requête.

La double élimination

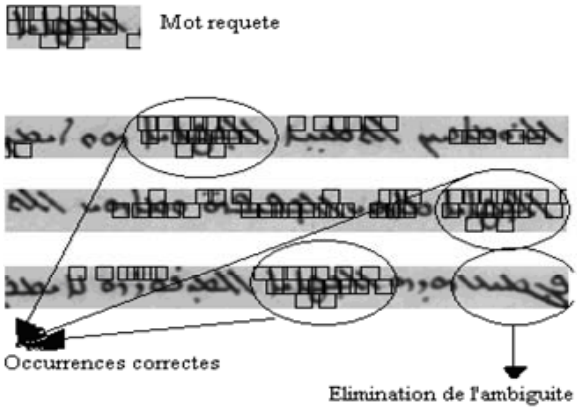


Figure 55: Elimination de l'ambigüité et repérage des occurrences correctes.

A la fin, la combinaison des deux étapes précédentes élimine les ambigüités ce qui permet de repérer et de garder uniquement les bonnes occurrences du mot requête.

La fausse occurrence a été rejetée puisque ses fenêtres constitutives sont en dehors des régions d'intérêt.

5.4 Conclusion

Tous les tests mentionnés ci-dessus ont été effectués sur une seule page de texte contenant 3 ou 4 occurrences d'un mot requête. Le mot recherché et toutes ses occurrences figurent dans la même page. Ces tests modestes ont prouvé que la première méthode proposée est efficace sur une petite base de tests. Toutes les occurrences du mot requête ont pu être repérées. De plus, sur les exemples proposés, la méthode a pu résister aux deux dégradations les plus courantes pouvant atteindre les versions numérisées des documents manuscrits.

6 Améliorations

6.1 Choix de la fenêtre rectangulaire 96×32 pixels

L'efficacité de la méthode ne peut être prouvée sans son application sur une base de test importante. C'est ce qui a constitué les travaux que nous décrirons par la suite, dans lesquels nous allons utiliser la deuxième forme de fenêtre glissante qui est une fenêtre rectangulaire de taille 96×32 pixels, la fenêtre va donc recouvrir la totalité de la ligne d'écriture, elle pourra permettre de décrire la forme d'une lettre pareille dans son intégralité.

Les fenêtres glissantes constituent une première alternative à un algorithme de segmentation en lettres individuelles.

Les mêmes critères de sélection que ceux utilisés pour la fenêtre de taille 32×32 pixels sont adoptés afin de ne garder que les fenêtres intéressantes.

Un mot requête ainsi qu'une page entière de texte se trouvent réduits à un nombre de fenêtres. Pour l'image d'un mot les fenêtres retenues sont dans la figure ci-dessous.

a) Un mot requête. b) Les fenêtres glissantes sélectionnées.

Figure 56: Un mot requête avec ses fenêtres glissantes associées.

6.2 Extraction de caractéristiques

Les fenêtres 96x32 sont divisées en douze sous-fenêtres de taille 16×16 pixels chacune. La fonction d'auto-corrélation est calculée dans chacune de ces douze sous-

fenêtres. Les motifs résultants de cette auto-corrélation représentent les directions principales dans les quatre quadrants d'une fenêtre.

Cette information est résumée sous la forme d'une rose à 8 directions. Chaque sous-fenêtre pixels est représentée par une signature de 8 valeurs résultant en un total de 96 valeurs représentant chaque fenêtre. La longueur d'une direction est obtenue en faisant la somme des niveaux de gris de la fonction d'auto-corrélation dans cette direction.

La figure ci-dessous montre la fonction d'auto-corrélation pour chacune des douze sous-fenêtres constituant une fenêtre glissante sélectionnée.

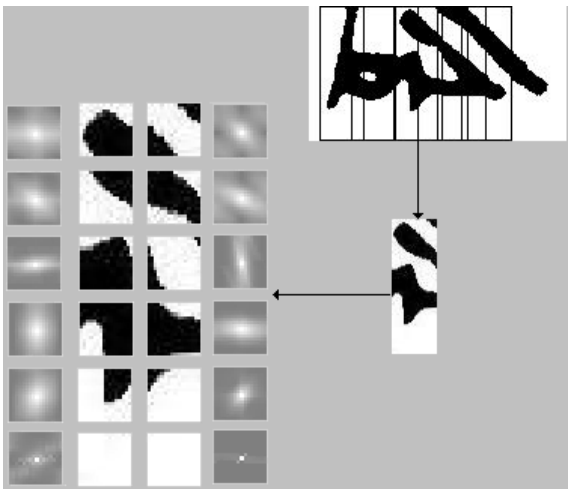


Figure 57: Les fonctions d'auto-corrélation des douze sous-fenêtres constituant une fenêtre sélectionnée.

Les roses de direction associées aux fonctions d'auto-corrélation de la figure précédente sont illustrées dans la figure ci-après.

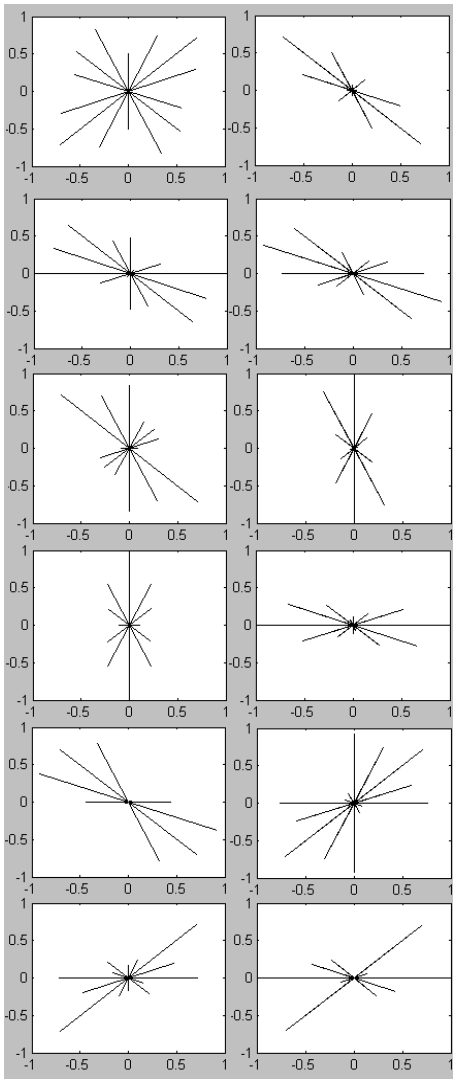


Figure 58: Les roses de directions représentant respectivement les douze sous-fenêtres.

6.3 Algorithme d'appariement

Une fois les signatures de toutes les fenêtres choisies de toutes les lignes des pages de texte sont extraites, elles sont comparées à celles extraites à partir de l'image du mot requête Les plus semblables sont détectées et les régions ayant une agglomération des fenêtres glissantes semblables à celles de l'image du mot requête sont considérées une occurrence possible à ce dernier. Cependant une décision basée uniquement sur ce critère n'est pas très fiable, en augmentant la taille de la base de test nous remarquons la présence d'agglomérations de fenêtres semblables à celles du mot requête sans tout de même en être des occurrences.

6.4 Détection des régions d'intérêt

Pour surmonter ce problème, nous aurons donc recours à une pré-détection des régions d'intérêt où les occurrences possibles du mot requête peuvent être situées. Pour ce faire, nous avons étudié le mouvement des centres de gravité des fenêtres

retenues pour le mot requête. Les positions de ces centres sont tracées sur un graphe, et le but est de trouver les portions de graphes dans les lignes de texte qui sont les plus similaires au graphe requête. La mesure de similarité est basée sur le calcul de la distance euclidienne minimale. Une région d'intérêt bien repérée est considérée comme une occurrence possible du mot requête, par contre une région d'intérêt mal-retenue sera une fausse occurrence.

Sur la figure 20 nous montrons les régions d'intérêt associées au mot requête des figures précédentes. Nous voyons sur la figure que les régions d'intérêt sont bien situées sur les onze occurrences correctes, par contre elles se placent aussi dans trois endroits incorrects.

Les critères de décision mentionnés ci-dessus, chacun pris à part, ne sont pas suffisants pour décider si une bonne occurrence est trouvée ou non, mais la combinaison de ces deux critères résulte, comme expliqué précédemment, en deux phases d'élimination :
- dans la première phase, les fenêtres glissantes incorrectes qui ont été retenues seront éliminées si elles ne sont pas situées dans une région d'intérêt,
- dans la deuxième phase, les régions d'intérêt mal placées seront éliminées si elles ne contiennent pas de fenêtres semblables à celles du mot requête.

Si après cette double élimination, certaines fausses occurrences persistent, leur image est très semblable à celle du mot cherché même si leur sens en est différent, ces dernières seront dites les fausses bonnes réponses associées au mot requête.

Une évaluation quantitative de nos résultats sera difficile à présenter pour le moment. Une telle évaluation requiert une vérité de terrain de laquelle nous ne disposons pas pour nos documents. Une vérité de terrain n'est pas facile à construire non plus. Nous comptons manuellement les occurrences du mot requête, et nous vérifions visuellement si elles ont été toutes retrouvées ou non. Si le dénombrement des occurrences est à faire automatiquement, ceci nécessite un seuil minimum de reconnaissance, donc nous devrions faire la reconnaissance afin de compter. Or pour reconnaître les mots que nous recherchons, nous devrions savoir où ils sont situés et combien ils sont nombreux, donc ici nous devrions compter afin de reconnaître. Ainsi nous tombons dans le dilemme *« Compter afin de reconnaître et reconnaître afin de compter »*.

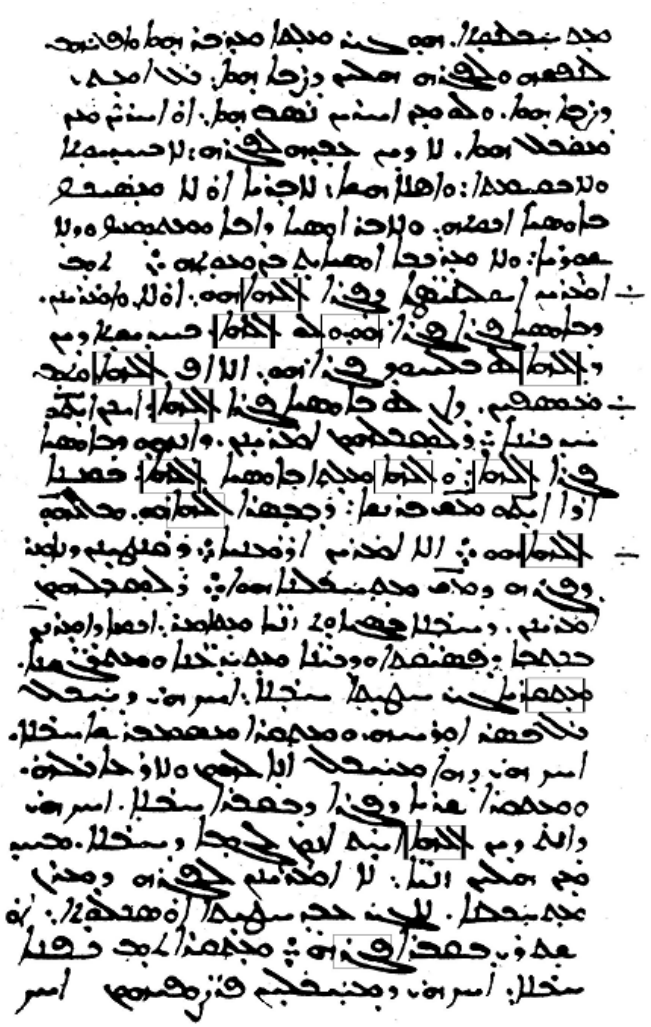


Figure 59: Régions d'intérêt associées au mot requête de la figure 30.

7 Expérimentations sur nos corpus de références

7.1 Résultats obtenus sur les corpus en Serto

Premier corpus

Nous allons montrer les résultats que nous avons obtenus avec trois mots fréquents dans ce corpus de 10 pages, ces trois mots sont écrits dans la figure 21.

Mot 1	Mot 2	Mot 3
		

Figure 60: Images des trois mots requêtes fréquents dans le premier corpus.

Nous avons résumé les résultats dans le tableau statistique de la figure 22.

Mot	Nb total des occurrences	Nb d'occurrences retrouvées	Nb d'occurrences ratées	Nb de fausses bonnes réponses
Mot 1	**14**	**14**	**0**	**1**
Mot 2	**13**	**13**	**0**	**3**
Mot 3	**7**	**7**	**0**	**1**

Figure 61: Tableau des statistiques évaluées pour les trois mots fréquents.

La méthode que nous avons proposée a prouvé son efficacité dans le repérage de toutes les occurrences des mots requêtes. Aucune bonne occurrence n'a été ratée, par contre, certaines fausses bonnes réponses ont été retenues. Leur nombre est très réduit comparé au nombre total de mots existants dans le corpus.

La figure ci-après montre le mot requête 1 ainsi que sa fausse occurrence associée ; nous constatons sur cette image montre que la fausse occurrence possède deux lettres parmi quatre en commun avec le mot requête.

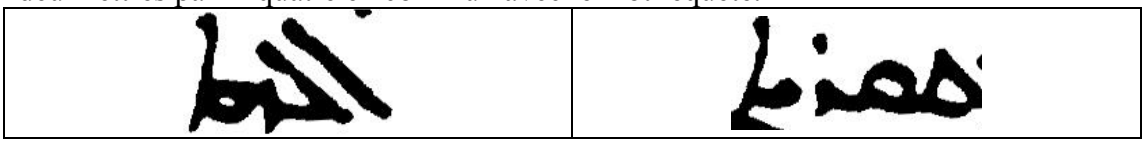


Figure 62: Le mot requête 1 et sa fausse occurrence retenue.

La figure ci-après montre le mot requête 2 ainsi que ses trois fausses occurrences associées. Nous pouvons remarquer que la première fausse occurrence a trois lettres parmi quatre en commun avec le mot requête, la troisième fausse occurrence possède aussi trois lettres parmi quatre en commun avec le mot requête, et la deuxième fausse occurrence diffère du mot requête uniquement par son suffixe. Nous pouvons voir que les trois occurrences présentent une similarité d'allure avec le mot cherché.

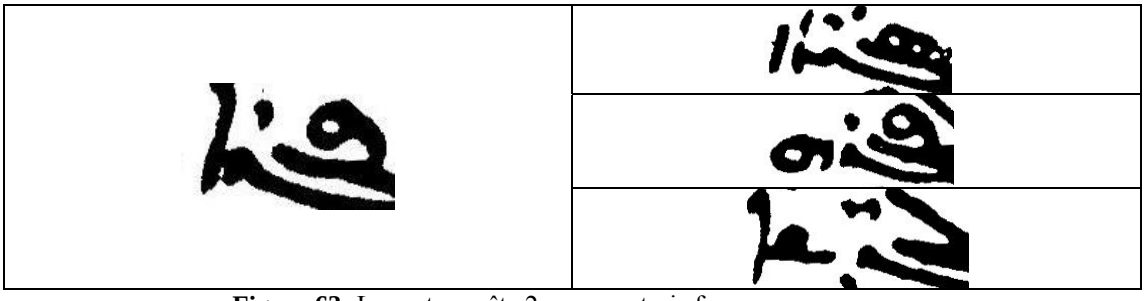


Figure 63: Le mot requête 2 avec ses trois fausses occurrences.

La figure ci-dessous montre le mot requête 3 ainsi que sa fausse occurrence associée. Ce qui est remarquable c'est que la fausse occurrence possède une préposition détachée du mot initial alors que le mot requête c'est ce même mot mais avec un préfixe à la place de la préposition. Le mot requête 3 et sa fausse occurrence associée partagent une racine commune.

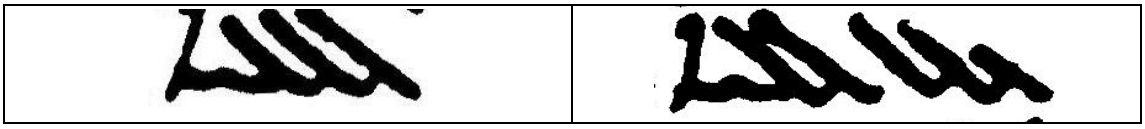


Figure 64: Le mot requête 3 et sa fausse occurrence.

Deuxième corpus

Dans ce deuxième corpus comportant 4 pages soit 1524 mots, nous avons choisi trois mots fréquents (Figure 26) dont nous allons chercher les occurrences.

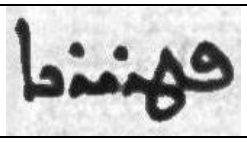
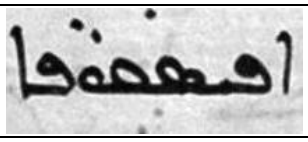
Mot 1	Mot 2	Mot 3
		

Figure 65: Images des trois mots requêtes fréquents dans le deuxième corpus.

Les statistiques mentionnées ci-dessus ont aussi été évaluées pour chacun de ces mots fréquents, ces dernières sont résumées dans le tableau ci-après.

Mot	Nb total des occurrences	Nb d'occurrences retrouvées	Nb d'occurrences ratées	Nb de fausses bonnes réponses
Mot 1	**15**	**15**	**0**	**3**
Mot 2	**9**	**9**	**0**	**1**
Mot 3	**16**	**16**	**0**	**3**

Figure 66: Tableau des statistiques évaluées pour les trois mots fréquents.

La méthode prouve son efficacité dans le repérage de toutes les occurrences des mots requêtes dans le deuxième corpus. Aucune bonne occurrence n'a été ratée, par contre certaines fausses bonnes réponses ont été retenues. Leur nombre est aussi très réduit comparé au nombre total de mots existants dans le corpus.

La figure ci-après montre le mot requête 1 ainsi que ses trois fausses occurrences associées. Toutes les trois fausses occurrences ont des lettres en communs avec le mot requête.

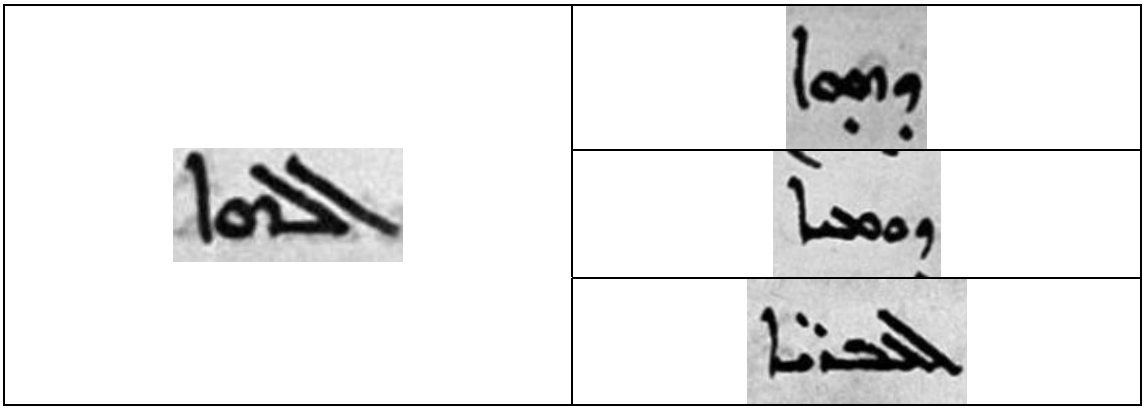


Figure 67: Le mot requête 1 avec ses trois fausses occurrences.

La figure ci-après montre le mot requête 2 ainsi que sa fausse occurrence associée. Nous pouvons remarquer que cette fausse occurrence ressemble beaucoup au mot requête et possède plusieurs lettres en commun avec

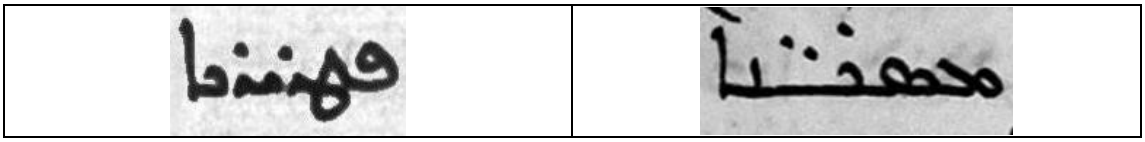


Figure 68: Le mot requête 2 et sa fausse occurrence associée

La figure ci-après montre le mot requête 3 ainsi que ses trois fausses occurrences associées. La première fausse occurrence possède 5 lettres parmi 7 en commun avec le mot requête, la deuxième fausse occurrence possède 4 lettres parmi 7 en commun avec le mot cherché, et la troisième fausse occurrence possède 3 lettres parmi 7 en commun avec le mot requête. Toutes les trois fausses occurrences présentent une similarité d'apparence avec le mot cherché.

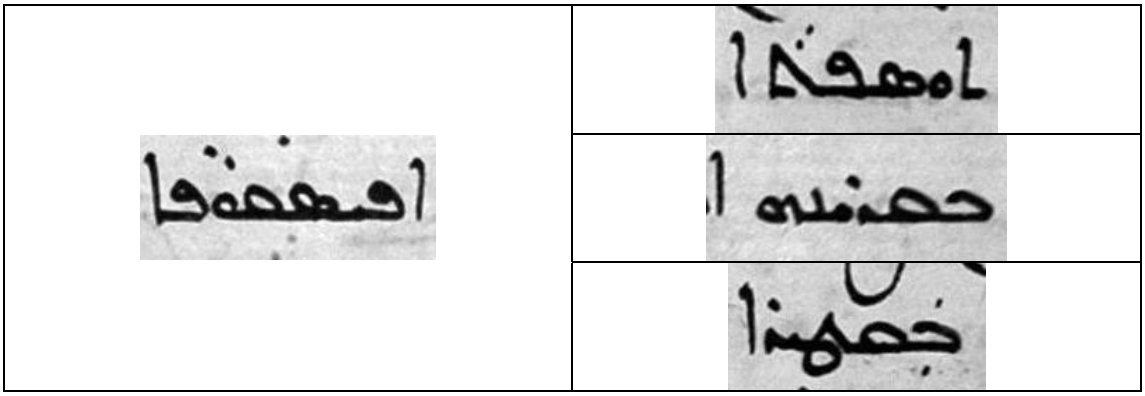


Figure 69: Le mot requête 3 avec ses trois fausses occurrences.

Troisième corpus

Le troisième corpus est le plus volumineux parmi ceux que nous avons entre les mains, 377 pages et 50 000 mots, environ. Des mots fréquents peuvent s'y trouver des centaines de fois. Des différences de forme entre un mot cherché et ses occurrences peuvent être de plus en plus importantes, de même, le risque de tomber sur des fausses occurrences augmente. Dans ce corpus nous avons sélectionné deux mots fréquents (Figure 31) pour lesquels nous allons montrer nos résultats.

Mot 1	Mot 2
	

Figure 70: Images des deux mots requêtes fréquents dans le troisième corpus.

L'évaluation des statistiques pour chacun de ces mots fréquents, est résumée dans le tableau ci-après.

Mot	Nb total des occurrences	Nb d'occurrences retrouvées	Nb d'occurrences ratées	Nb de fausses bonnes réponses
Mot 1	**240**	**240**	**0**	**6**
Mot 2	**101**	**101**	**0**	**4**

Figure 71: Tableau des statistiques évaluées pour les deux mots fréquents.

Nous parvenons aussi dans ce troisième corpus à retrouver toutes les bonnes occurrences du mot requête aussi fréquentes soient elles, aucune occurrence correcte n'a été ratée. Certaines fausses occurrences ont été retenues, nous remarquons qu'elles sont plus nombreuses que celles pour les deux corpus précédents, ceci est justifié par le volume de ce troisième corpus et aussi par le fait qu'il a été rédigé par trois scribes différents.

Dans la figure ci-dessous, nous avons recopié les ses six fausses occurrences associées au premier mot requête.

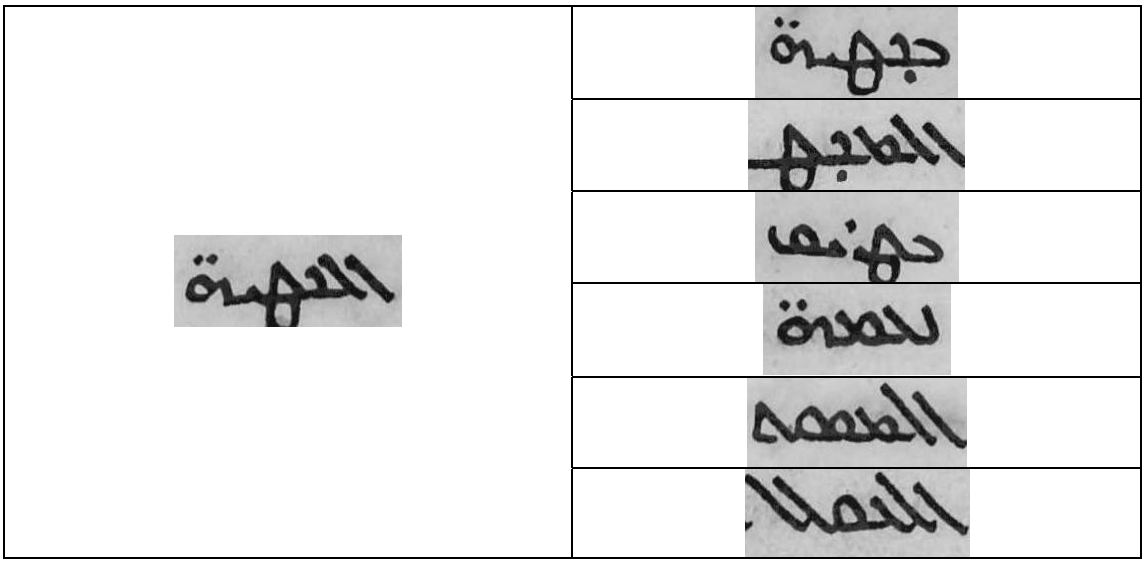


Figure 72: Le mot requête 1 avec ses six fausses occurrences associées.

Nous remarquons que la première «fausse occurrence» est exactement le mot cherché, mais sans le préfixe, la deuxième fausse occurrence possède 4 lettres parmi 5 en commun avec le mot requête, les troisième, quatrième, cinquième, et sixième fausses occurrences possèdent toutes 2 lettres parmi 5 en commun avec le mot cherché. Toutes les six fausses occurrences montrent une similarité d'allure avec le mot requête.

Dans la figure ci-dessous nous montrons le second mot requête avec ses quatre fausses occurrences associées.

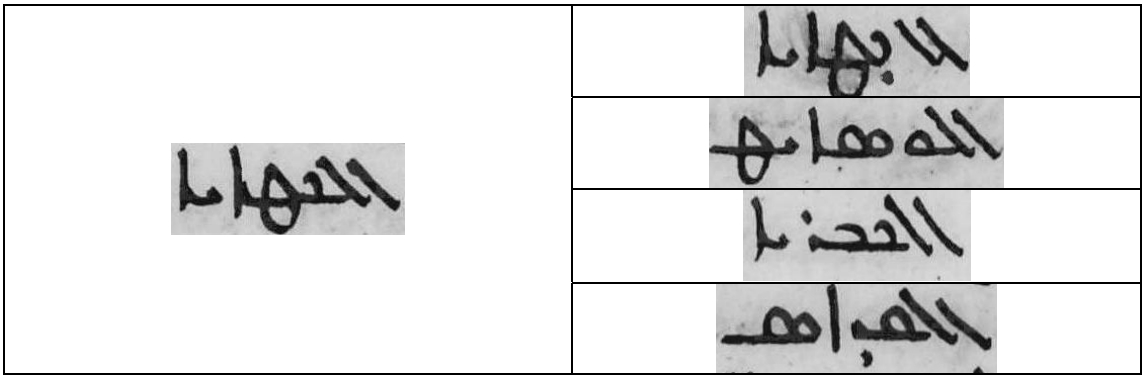


Figure 73: Le mot requête 2 avec ses quatre fausses occurrences.

Nous remarquons que la première fausse occurrence possède exactement la même racine que le mot cherché, la différence est uniquement au niveau du préfixe remplacé par une préposition détachée dans la fausse occurrence ; les deuxième et troisième fausses occurrences possèdent 5 lettres parmi 7 en commun avec le mot requête, et la quatrième fausse occurrence possède 4 lettres parmi 7 en commun avec le mot cherché.

7.2 Résultats obtenus sur le corpus en Nestorien

Dans ce corpus de 4 pages, nous avons repéré deux mots fréquents (Figure 35).

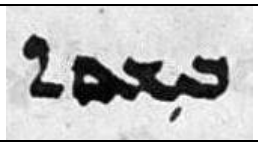
Mot 1	Mot 2
	

Figure 74: Images des deux mots fréquents choisis pour le quatrième corpus.

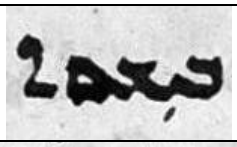
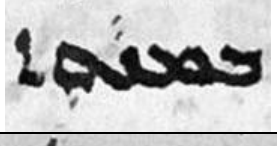
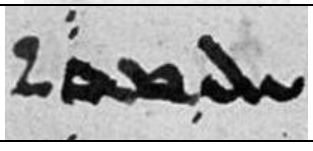
L'évaluation des statistiques pour chacun de ces mots fréquents, est résumée dans le tableau ci-après.

Mot	Nb total des occurrences	Nb d'occurrences retrouvées	Nb d'occurrences ratées	Nb de fausses bonnes réponses
Mot 1	**10**	**10**	**0**	**4**
Mot 2	**6**	**6**	**0**	**3**

Figure 75: Tableau des statistiques évaluées pour les deux mots fréquents.

De même pour le Nestorien, nous sommes parvenus à retrouver toutes les occurrences des mots requêtes ; comme pour les autres essais certaines fausses occurrences ont été retenues ; leur nombre est assez élevé par rapport au nombre total d'occurrences à trouver, elles présentent toutes une similarité de forme avec les images cherchées.

Dans la figure ci-dessous nous montrons le premier mot requête avec ses quatre fausses occurrences associées.

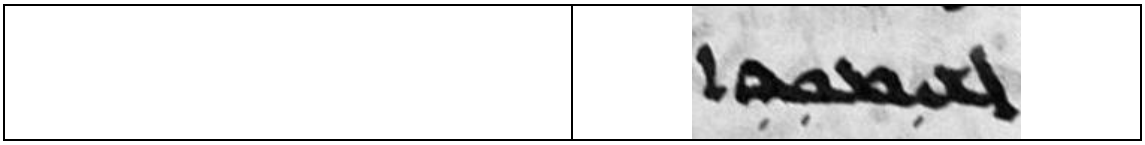


Figure 76: Le mot requête 2 avec ses quatre fausses occurrences.

Nous remarquons que la première et la deuxième fausse occurrence ont 3 lettres parmi 4 en commun avec le mot requête, que la troisième et la quatrième fausse occurrence ont 2 lettres parmi 4 en commun avec le mot cherché.

La figure ci-dessous montre le deuxième mot requête avec ses fausses occurrences associées.

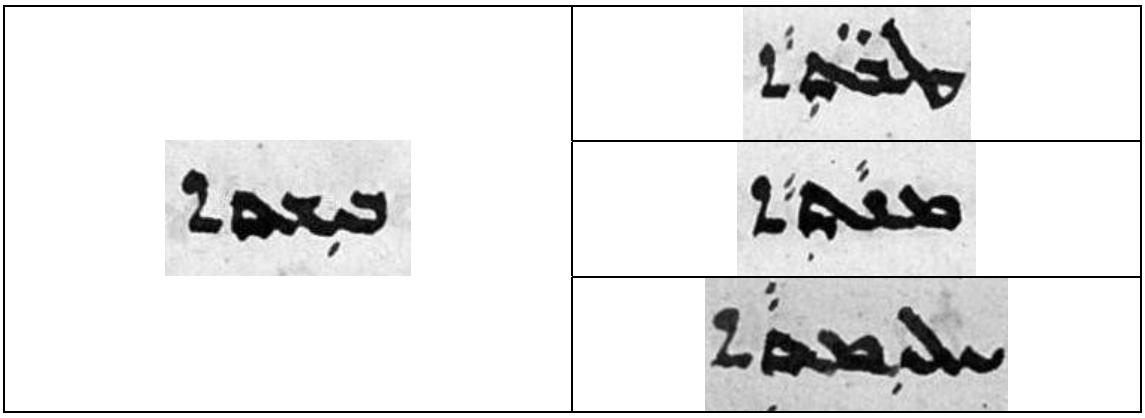


Figure 77: Le mot requête 2 avec ses trois fausses occurrences.

Nous remarquons que la première fausse occurrence possède 3 lettres parmi 4 en commun avec le mot requête, et que la deuxième et la troisième fausse occurrence ont 2 lettres parmi 4 en commun avec le mot cherché.

7.3 La saisie Manuscrite

Une dernière tentative de notre part était de jouer le rôle de scribe et d'écrire nous-mêmes nos mots requêtes, nous avons écrit les trois mots requêtes du deuxième corpus en Serto que nous avons testé ci-précédemment.

Nous avons choisi comme instrument d'écriture un marqueur de couleur noire qui puisse reproduire la même épaisseur de trait que celle dans le document, puis nous les avons scannés à la même résolution que les images de texte, mais en mode binaire (écriture noire sur fond blanc), et sauvegardés avec le même format que celui du document, nous avons ensuite binarisé les images des pages du texte avant de lancer la requête.

La binarisation était importante puisque le fond des images d'origine était d'une couleur assez irrégulière, un passage en niveau de gris résultait en un fond gris non homogène, nous ne pouvons pas reproduire ce même fond sur les mots que nous avons nous-mêmes écrits, donc nous avons choisi de retourner en mode encre sur papier et de travailler en mode binaire.

Les images des mots que nous avons écrits sont présentées dans la figure ci-dessous.

Mot 1	Mot 2	Mot 3

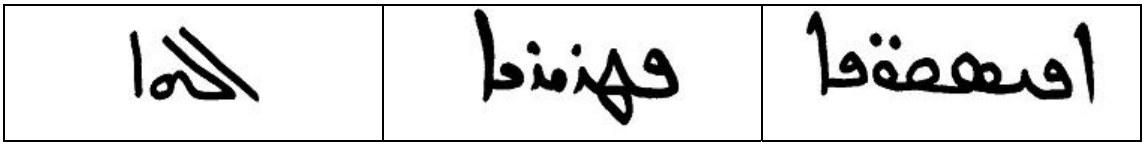


Figure 78: Images de nos trois mots fréquents manuscrits.

L'évaluation qualitative des résultats obtenus pour cette méthode de saisie s'effectue en utilisant les mêmes statistiques que celles utilisées précédemment. Ces statistiques pour chacun des trois mots sont montrées dans le tableau ci-après.

Mot	Nb total des occurrences	Nb d'occurrences retrouvées	Nb d'occurrences ratées	Nb de fausses bonnes réponses
Mot 1	**15**	**15**	**0**	**5**
Mot 2	**9**	**9**	**0**	**1**
Mot 3	**16**	**16**	**0**	**1**

Figure 79: Tableau des statistiques évaluées pour nos trois mots fréquents manuscrits.

Dans la figure ci-dessous, nous montrons le premier mot requête avec ses cinq fausses occurrences associées.

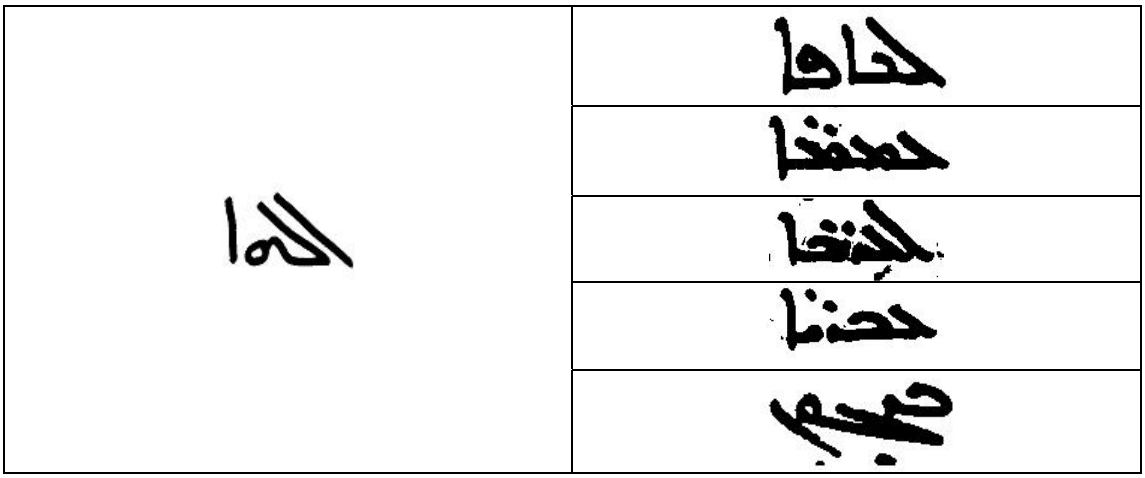


Figure 80: Le mot requête 1 avec ses cinq fausses occurrences.

La première fausse occurrence possède 3 lettres parmi 4 en commun avec le mot requête, les deuxième, troisième, quatrième et cinquième occurrences possèdent une lettre parmi 4 en commun avec le mot cherché.

Dans la figure ci-dessous, nous montrons le deuxième mot requête avec sa fausse occurrence associée.

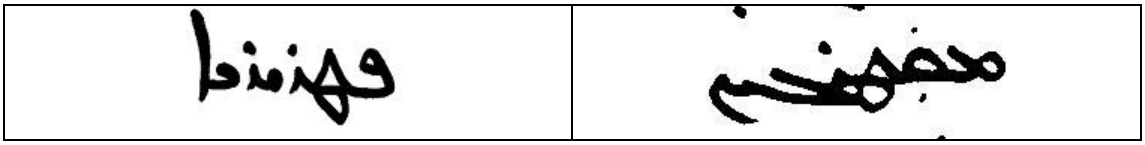


Figure 81: Le mot requête 2 et sa fausse occurrence associée.

Cette unique fausse occurrence pour le mot requête 2 possède 3 lettres parmi 6 en commun avec lui.

Dans la figure ci-dessous, nous montrons le troisième mot requête avec sa fausse occurrence associée.

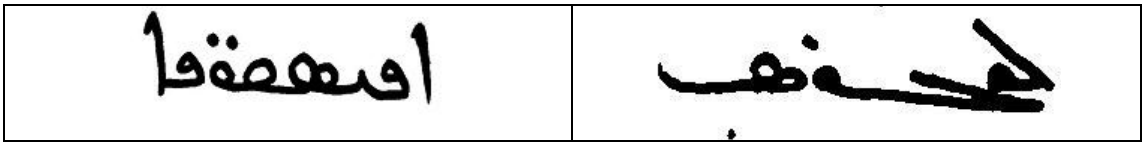


Figure 82: Le mot requête 3 et sa fausse occurrence associée.

Cette unique fausse occurrence pour le mot requête 3 possède 3 lettres parmi 7 en commun avec lui.

7.4 La saisie par le clavier

Nous n'avons pas pu tester cette approche de saisie, puisque elle nécessite une certaine vérité de terrain de laquelle nous ne disposons pas pour nos documents. Cette vérité de terrain afin d'être construite nécessite un nombre important de documents desquels nous ne disposons pas, ainsi qu'un algorithme de segmentation de texte en caractères associé à un classifieur afin de créer une base de données conséquente et qui soit étiquetée. Cette base de données devrait être couplée aussi avec un dictionnaire de la langue, de plus elle nécessite l'intervention et la présence continue d'experts en la langue.

7.5 Evaluation globale des résultats obtenus

Les résultats obtenus sur le corpus en Nestorien n'étaient pas aussi satisfaisants que ceux obtenus pour les corpus en Serto. Il nous semble que notre approche était mieux adaptée aux documents en Serto qu'à ceux en Nestorien

Pour les deux calligraphies, notre approche a retenu des fausses occurrences supplémentaires aux occurrences correctes du mot requête. Par contre, le nombre de ces fausses réponses était relativement bas par rapport au nombre total d'occurrences pour les documents en Serto, les résultats les moins bons ont été obtenus avec le Nestorien pour lequel nous disposons que d'un seul corpus de petite taille numérisé avec une résolution de 71 dpi. L'écriture en Nestorien étant assez compacte, de nombreux mots ont tendance à se ressembler par la forme même s'ils sont assez différents par leur contenu, par contre la cursivité dans l'écriture Serto pourrait augmenter la différence d'allure entre les différents mots.

Il est important de noter que, pour tous les tests que nous avons conduits, notre algorithme bien qu'il a retenu de fausse occurrences, n'en a raté aucune des correctes, quelles que soient la taille du corpus, son état de numérisation et la calligraphie adoptée.

Il est vrai que certaines fausses occurrences gardées par notre approche, diffèrent en terme de sens du mot que l'on recherche, mais elles présentent une similarité dans le dessin des lettres, elles ont même certaines lettres en commun avec le mot requête.

Dans certains cas, une fausse occurrence pourrait être une dérivée grammaticale du mot requête, puisqu'en Syriaque presque toutes les fonctions grammaticales sont écrites sous forme de préfixes et de suffixes au lieu d'être des mots séparés. Lesdites fausses occurrences partagent une racine commune avec le mot requête. Partant de cette idée, une autre opération devrait être ensuite couramment appliquée lors de l'indexation. Elle consiste à effacer les terminaisons, la conjugaison, et les déclinaisons, afin de retrouver les racines des mots. Cette opération est appelée « *stemming* », et nécessite un dictionnaire linguistique de la langue ainsi qu'un dictionnaire grammatical. Ce procédé permet de relever les fréquences en cumulant les nombres d'occurrence des variations des mêmes mots. Dans ce cas, si nous avons pu repérer le mot lui-même ou bien une de ses dérivées, le sens initial est capturé et l'indexation est effectuée. La fausse occurrence serait alors considérée comme correcte.

8 Expérimentation sur un corpus Arabe

Nous avons voulu teste notre algorithme de repérage de mots sur une autre langue sémitique, l'Arabe. Pour ce faire nous avons conduit des tests sur un manuscrit Arabe, en provenance de Tombouctou, du Mali en collaboration avec Alfadoulou Abdoulahi. Ce document a subi les mêmes étapes de prétraitement, de sélection de mots requête, et d'application algorithmique que ses précédents Syriaques.

Ce manuscrit consiste en un traité du droit Musulman, il est formé de 6 pages et de 2166 mots, le texte est rédigé en écriture Maghrébine à raison d'une colonne par page de texte, et a été numérisé en couleur avec une résolution de 500 dpi, et les images sont sauvegardées avec le format TIF.

La figure ci dessous montre un extrait de ce corpus, trois niveaux de couleur sont visibles ; une couleur beige non homogène pour l'arrière plan, une couleur brunâtre pour le texte, et une couleur rouge pour les symboles diacritiques, les délimitations des phrases et des paragraphes, et les mots mis en relief.

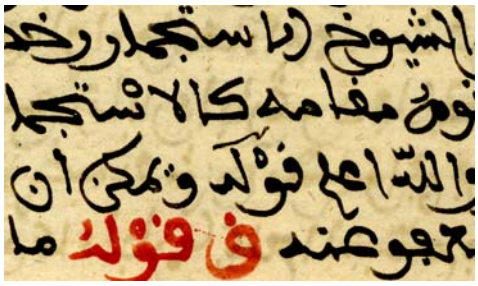


Figure : Extrait du corpus Arabe**.**

Dans ce corpus nous avons sélectionné deux mots fréquents pour lesquels nous allons montrer nos résultats.

Mot 1	**Mot 2**

Figure : Images des deux mots requêtes fréquents dans le corpus Arabe**.**

L'évaluation des statistiques pour chacun de ces mots fréquents, est résumée dans le tableau ci-après.

Mot	Nb total des occurrences	Nb d'occurrences retrouvées	Nb d'occurrences ratées	Nb de fausses bonnes réponses
Mot 1	**8**	**8**	**0**	**2**
Mot 2	**16**	**16**	**0**	**4**

Figure 83: Tableau des statistiques évaluées pour les deux mots fréquents.

Nous parvenons aussi dans ce corpus à retrouver toutes les bonnes occurrences du mot requête aussi fréquentes soient elles, aucune occurrence correcte n'a été ratée. Certaines fausses occurrences ont été retenues, bien que parfois ces dernières diffèrent dans leur sens du mot cherché, elles présentent une similarité dans le dessin des lettres.

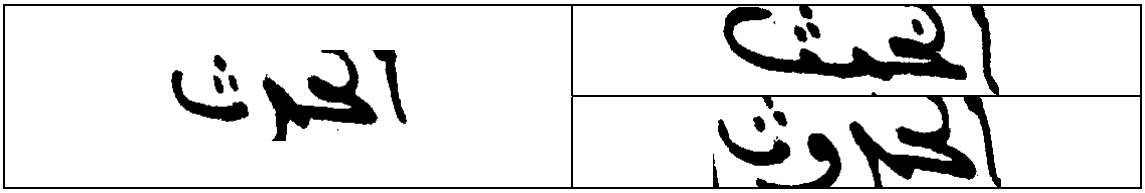


Figure 84: Le mot requête 1 avec ses deux fausses occurrences.

Nous remarquons que la première fausse occurrence possède 4 lettres parmi 5 en commun avec le mot requête ; nous remarquons aussi que la deuxième fausse occurrence possède exactement la même racine que le mot cherché, en effet c'est le substantif du verbe associé au mot requête, cette dernière fausse occurrence pourra alors bien être considérée comme correcte.

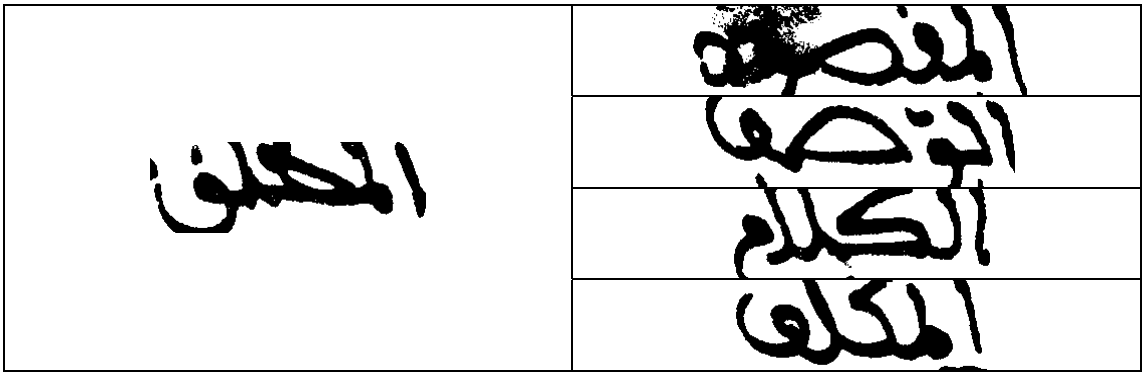


Figure : Le mot requête 2 avec ses quatre fausses occurrences.

Nous remarquons que la première fausse occurrence possède 4 lettres parmi 6 en commun avec le mot requête ; nous remarquons aussi que les deuxième et troisième fausses occurrences possèdent 3 lettres parmi 6 en commun avec le mot requête ; et la dernière fausse occurrence possède 5 lettres parmi 6 en commun avec le mot cherché.

Nous pouvons conclure, que l'algorithme que nous avons développé pour le Syriaque pourra bien être prometteur pour d'autres langues, notamment de la même racine, et précisément l'Arabe.

9 Repérage de mots par appariement de caractères

Les approches dites classiques de repérage de mots, sont basées sur un algorithme de recherche et de reconnaissance holistique des différentes occurrences d'un mot dit « requête ». L'approche holistique est adoptée dans des cas où une segmentation en caractères pour des fins de transcription ou bien d'OCR est jugée inapplicable, le mot en tant que tel devient la plus petite entité à rechercher.

Ces deux façons d'aborder le problème d'indexation ont été utilisées sans convergence directe de l'une vers l'autre, hormis dans l'approche suggérée dans des travaux de Khurshid et al. [KHU09], dans laquelle une segmentation de mots en caractères est effectuée pour des fins de « Word Spotting ». En effet, la recherche d'occurrences d'un certain mot requête peut être faite suivant deux phases d'élimination. Dans la première phase, une mesure de grandeur telle le rapport entre le nombre des « caractères » du mot requête et celui d'un mot pris pour requête, permettra d'éliminer plus que la moitié du nombre total des mots.

Dans la deuxième phase, un affinage des premiers résultats peut être effectué, ceci en descendant aux détails heuristiques du mot en question, suivant une correspondance basée sur un calcul de distance de séparation/groupement (Split /Merge), entre le résultat de la segmentation du mot requête et celui des mots candidats suivant le cas en question (si le rapport des nombres de caractères est inférieur ou bien supérieur à 1). Ceci résultera en un calcul des coûts de correspondance entre les caractères du mot requête et celui des mots candidats, disposés suivant une matrice (en cas idéal carrée) afin de calculer la distance finale entre les deux mots. Ce procédé permet de réduire le nombre de fausses occurrences retenues, ainsi que le nombre des bonnes occurrences ratées par rapport à des travaux dans la littérature [FAU08].

Cette approche de repérage de mots par appariement de caractères nécessite une vérité de terrain sous la forme d'une base d'apprentissage classifiée et étiquetée

qui soit volumineuse et consistante. Une base de laquelle nous ne disposons pas à une échelle pareille pour les documents Syriaques.

Un autre moteur de repérage de mot procède par appariement d'entité plus petite que le mot, celui élaboré par Sari et al. [SAR08] pour l'Arabe ; ils proposent une approche d'indexation d'anciens manuscrits Arabes dégradés, par appariement de chaînes de caractères (pas nécessairement des mots) sur des images de document ; un utilisateur souhaitant chercher les occurrences d'un mot pointe sa requête.

Les caractéristiques extraites sont sous forme de composantes connexes et similarité est basée sur un calcul de la distance de Levenshtein ou distance d'édition) entre l'image du mot requête et l'image d'un mot du document. Les résultats préliminaires étaient prometteurs compte tenu de la complexité des documents Arabes.

10 Conclusion

Ce chapitre présente un aspect assez applicatif, dans lequel nous présentons et analysons les résultats de notre algorithme de repérage de mots appliqué sur les différents documents qui se trouvent entre nos mains.

Le repérage de mots consiste en une première approche d'indexation d'images de documents que nous avons testé sur le Syriaque ; cette démarche ne requiert aucune connaissance à priori et, sur le plan technique aucune segmentation de la ligne d'écriture en graphèmes ou caractères individuels.

La méthode proposée repose sur une sélection de fenêtres glissant sur les lignes d'écriture ; dans les fenêtres portant suffisamment d'informations nouvelles on « mesurera » la direction, plus précisément les directions, avec les techniques d'auto-corrélation dont les résultats seront résumés par une rose des directions

Il est important de remarque que notre méthode n'a raté aucune occurrence des mots requêtes proposé, la sobriété de l'écriture syriaque et le fait que nous avons fait une quantité de tests trop limitée sont à opposer à ce résultat.

Nous avons constaté que notre méthode a davantage tendance à retenir des fausses occurrences pour des mots courts que pour des mots longs. En effet plus un mot est long, plus il contient de l'information et donc des caractéristiques pertinents. Cette « lacune » est commune à toutes les approches de repérage de mots et c'est certainement à ce niveau que les techniques de repérages de mots et les techniques de reconnaissance de caractères diffèrent

Chapitre 4

De la transcription assistée à …
la transcription collaborative

1 Introduction

Le terme Transcription est polysémique ; dans le contexte de l'écriture on s'accorde sur le fait que la transcription est l'opération qui consiste à substituer à chaque phonème (on parle alors de transcription phonologique) ou à chaque son (transcription phonétique) d'une langue un graphème ou un groupe de graphèmes d'un système d'écriture. Plus simplement, c'est l'acte de traduire un signal en caractères d'écriture.

Par extension, on a appelé transcription l'acte permettant de passer d'un écrit en mode image à sa forme texte (ASCII ou Unicode) et cette démarche se fait autrement que par une simple saisie, en se faisant assister par un ordinateur. Les premières recherches que nous avons appréhendées ont été faites à Lyon, il y a une dizaines d'années, elles portaient sur des corpus particuliers, les premiers ouvrages de la Renaissance. On a d'abord perçu la transcription assistée comme une alternative aux méthodes d'O.C.R. (Optical Character Recognition). C'était l'élaboration d'un O.C.R. dédié à un corpus particulier et l'apprentissage allait nous permettre d'extraire et de formaliser les spécificités de chaque composante de l'écriture ce corpus que nous pourrions ensuite utiliser dans le module de reconnaissance.

Les travaux faits ces dernières années tendent à faire évoluer la transcription de simple alternative à l'OCR vers une véritable démarche collaborative (homme-machine).

Les travaux faits jusqu'à présent ont porté presqu'exclusivement sur des textes imprimés complexes où il y avait un grand nombre de caractères et d'abréviations. On est loin du Syriaque manuscrit avec sa sobriété et son nombre restreint de caractères.

Notre premier objectif sera de démontrer que la démarche transcription est transposable au Syriaque, la sobriété de l'écriture syriaque ne simplifiera pas nécessairement la tâche de reconnaissance car il y a souvent peu de signes visuellement distinctifs dans les différents caractères composant le syriaque. La décomposition de l'écriture en graphèmes de la fin du chapitre 2 servira de point de départ à notre étude de faisabilité. Il faut noter que la transcription d'un manuscrit nous permettra de pouvoir disposer d'une copie imprimée de ce manuscrit !

Clocksin et al. [CLO03] ont déjà fait une tentative dans cette direction sur la calligraphie Estrangelo, leurs meilleurs résultats ont été obtenus sur les documents imprimés.

2 Etat de l'Art sur La Transcription assistée

2.1 La transcription dans DEBORA

De la compression à la transcription

DEBORA est l'acronyme de Digital accEs to Books of RenaissAnce ; ce projet de recherche/développement a été soutenu par l'Union Européenne dans le cadre du 4ième PCRD.

Il avait pour objectif la mise en place d'une stratégie pour la numérisation et la diffusion des ouvrages de la Renaissance. Le consortium qui avait été constitué comprenait des bibliothèques européennes riches en ouvrages de la Renaissance et de nombreuses équipes de recherches dont le laboratoire lyonnais RFV ; RFV avait la charge de traiter les images de document pour en extraire des métadonnées et faciliter le stockage et la transmission. Dans ce contexte, la compression était un sujet important pour deux raisons :

- le stockage des documents numérisés est couteux,
- la transmission était plus lente qu'aujourd'hui ; une façon d'atténuer l'attente pour l'utilisateur était de faire de l'affichage progressif

L'équipe lyonnaise avait mis en place une décomposition l'image considérée comme une superposition de couches. Ces couches n'ont pas été choisies a priori, mais par leur potentialité à pouvoir être compressée sans perte au moindre coût.

Le choix des couches n'est pas fait à priori, mais plutôt « a postériori » : des couches correspondant à une potentialité de compression sans perte élevée, couches de fonds, couches de graphisme et desseins, couches textuelles. Les couches textuelles sont composées par des juxtapositions de tracés d'écritures, des composantes connexes,

c'est à dire les morceaux de traits d'un seul tenant. Comme on travaille sur de l'imprimé, ces composantes correspondent la plupart du temps à des lettres ou encore à des fragments de lettres (quand celles-ci sont cassées ou rompues). Ces composantes peuvent dans certains cas rares correspondre à plusieurs lettres (collées accidentellement ou abréviations).

La première idée fut de coder ces composantes connexes en utilisant une méthode de compression entropique (ou avec Huffmann), le blanc et le e étant codés sur un bit…Ce qui intéressait en priorité les chercheurs c'était « la redondance des formes » qui leur permettait de compresser efficacement et de faire une importante économie en termes de volume de stockage. Le codage semblait simple à réaliser, on balayait la page dans le sens de la lecture et on étudiait chaque composante connexe relativement à la compression, il y a donc deux éventualités :
- la composante a déjà été rencontrée et elle est donc déjà associée avec un code, on lui attribue alors ce code,
- elle n'a jamais été rencontrée, alors on lui attribue un nouveau code.

Cette démarche naturelle ressemble à celle qui sous-tend les méthodes de clustering par propagation. Quand un nombre suffisant de page a été codé on peut reformuler le codage pour attribuer des codes de longueur inversement proportionnelle à la fréquence d'apparition de l'élément.
C'est alors que les chercheurs ont vu que le meilleur code était probablement celui qui consistait à associer le caractère ASCII à son occurrence.

La transcription DEBORA

Le principe de la transcription repose sur l'identification des formes (les caractères) similaires dans une image, quand toutes les formes similaires sont collectées, l'ordinateur demande à un opérateur de les identifier manuellement, et il propage alors la transcription à tout le document.

Cependant, la transcription assistée pose le problème fondamental qui se pose concerne l'évaluation de la ressemblance entre deux formes entre deux formes. Compte-tenu du bruit des images binaires, un même caractère imprimé est représenté par plusieurs images légèrement différentes. Si la comparaison est trop rigoureuse alors le nombre de formes nouvelles à saisir va augmenter. Inversement, si la comparaison n'est pas assez rigoureuse, l'algorithme provoquera des erreurs de substitution et considérera deux caractères différents comme une seule et même forme. Pour réaliser une transcription il faut définir un algorithme de calcul de similarité à la fois précis et performant, condition nécessaire pour obtenir une collecte d'un nombre minimal de formes différentes sans erreur de substitution.

La figure 1 montre la liste des 243 formes différentes de caractères identifiées dans la première page de document du 16^{e} comportant 2091 formes de caractères (figure). Le taux de redondance de est de 88 si l'on poursuit sur les pages suivantes alors ce taux de redondance croit fortement jusqu'à atteindre 99%. Les difficultés viennent du fait que beaucoup de ces formes correspondent à des caractères dégradés ou cassés

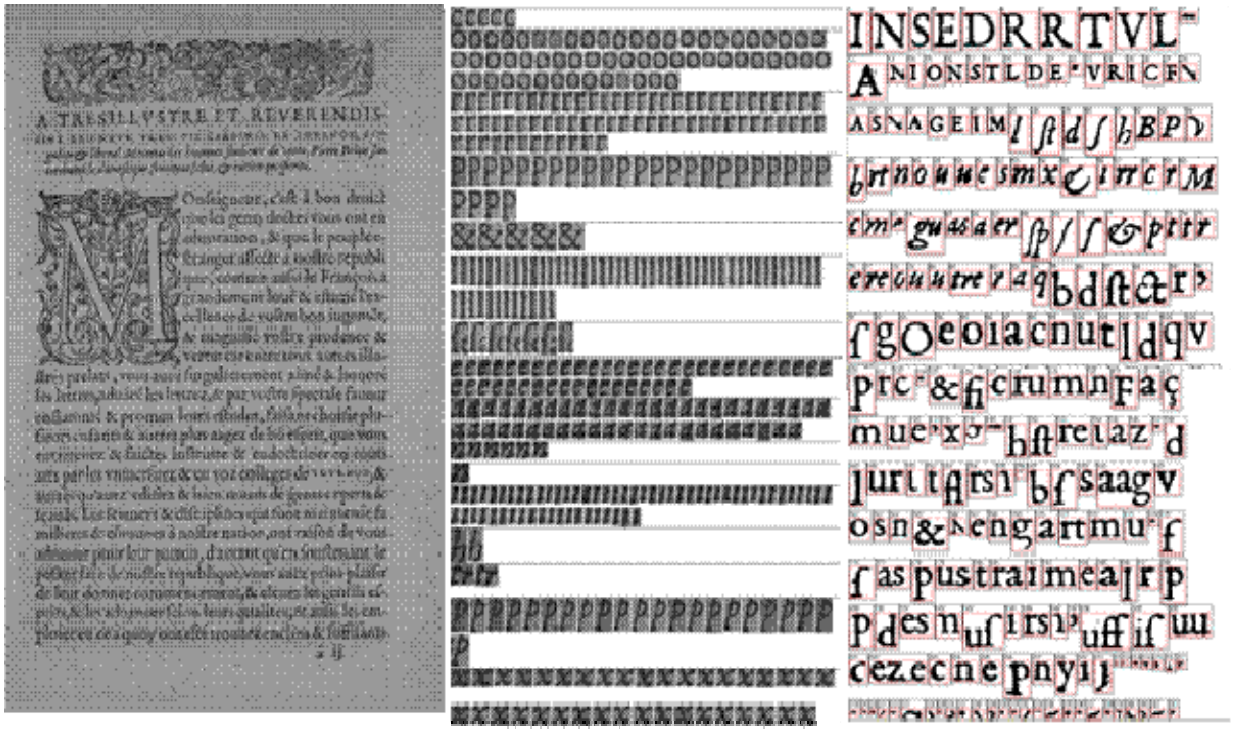


Image originale | collections de redondances | 243 formes différentes

Figure 1 : illustration de la transcription

La poursuite des recherches sur la transcription assistée a tardé après la fin de projet DEBORA ; peu de gens voyaient l'intérêt de cette démarche eu égard à la croyance que l'OCR pouvait tout résoudre !

Une suite du travail a été entreprise à Tours, au Centre d'Etudes Supérieures de la Renaissance et au laboratoire d'Informatique de l'Université Rabelais.
A Lyon les travaux ont repris dans le cadre des Thèses de Loris Eynard, d'une part, et de Joël Gardes d'autre part. Ces trois reprises ont exploré des développements très complémentaires que nous allons présenter sommairement.

2.2 Les développements faits à Tours

A Tours, en relation avec le Centre supérieur d'Etudes sur la Renaissance l'équipe du professeur Ramel a mis en oeuvre deux projets complémentaires pour donner une suite à DEBORA :
- AGORA, acronyme de Analyseur Graphique pour OuvRages Anciens,
- PIVOAN, pour Plateforme Informatique de Valorisation des Ouvrages ANciens
Le premier concerne la chaîne recouvrant le traitement et l'analyse jusqu'à sa segmentation en composantes élémentaires précédant la transcription, le deuxième propose une étude et des stratégies pour la transcription proprement dite

Nous retiendrons de ce travail le fait qu'après le constat sur les difficultés rencontrées pour mettre en place une classification des caractères résultant de la segmentation Ramel s'éloigne de la transcription simple alternative à l'OCR ; il

introduit une première idée d'interaction homme/machine et parle alors de transcription semi automatique.

Ainsi il propose que l'on fasse appel à l'utilisateur bien avant la phase contextuelle de correction des erreurs et de post-traitement.
On commencera par l'étiquetage manuel de quelques classes (les plus volumineuses).
On fusionnera les ensuite les caractères coupées (par des approches contextuelles).
On fusionnera aussi des classes par étude de voisinages, on s'intéressera à la position des caractères dans les mots si nécessaires
Ramel évoque aussi une possibilité de va et vient entre la machine et l'utilisateur

2.3 Approche « analyse et classification » dans le projet Gazette de Leyde.

L'ambition affichée au départ du projet de thèse de Eynard était de faire de la transcription assistée une véritable alternative à l'OCR, en d'autres termes, de créer un logiciel d'OCR dédié à la gazette. Les exemplaires de la Gazette de Leyde ont en moyenne deux siècles de moins que les ouvrages phares de DEBORA, aussi on pensait que ce serait aisé que de faire ce logiciel dédié. Mais la gazette est un journal, produit dans des conditions matérielles souvent ardues, caractérisées par
- du papier difficile à trouver et de qualité irrégulière,
- des lettres d'imprimerie en plomb dont les dimensions à l'origine sont un peu « aléatoires », et ensuite plus ou moins usées et écrasées par l'usage,
- une irrégularité dans la façon de serrer les presses

Eynard s'est attaché à reprendre les différentes composantes techniques de DEBORA et à les améliorer et à les rendre plus robustes, en commençant par proposer des méthodes d'extraction de caractères (composantes connexes) plus fiables
Il a proposé deux nouvelles approches de caractérisations de ces composantes et de façon corrélée des nouvelles évaluations des ressemblances ainsi que des méthodes de regroupement-classification des composantes basées sur les densités donc très souples et paramétrables (mean shift). Eynard a enfin proposé l'élaboration d'un indice (de qualité) de la TAO en prenant en considération le rapport du nombre de classes de caractères donné par la machine (donc à étiqueter) et celui de caractères à reconnaitre.

2.4 La transcription avec intervention humaine

L'approche développée dans la thèse de Joël Gardes a été mise en œuvre par Leydier et Gardes, dans un prototype de service dénommé "loupe numérique hypertexte". L'objectif principal de ce prototype est de pouvoir effectuer une reconnaissance de mots dans un mode supervisé, c'est-à-dire dans une mise en situation où l'homme supervise la machine.
Ce prototype a pour vocation de permettre à l'utilisateur de demander la transcription d'un fragment d'image contenant du texte, quels que soient les conditions de prise de vue de l'image, de typographie utilisée et de langue ; le fragment de texte retranscrit devant servir de requête à destination, par exemple, d'un moteur de recherche, d'un traducteur automatique ou d'un service d'encyclopédie en ligne.

Les contraintes d'utilisation de ce prototype ont conduit à écarter d'emblée l'utilisation d'un OCR comme composant logiciel :
- le prototype doit pouvoir fonctionner sur un terminal mobile limité en ressources de calcul et en mémoire

- les images sont prises sur le vif,
- il n'y a pas de possibilité d'entraînement a priori d'un OCR du fait de la très grande variété de formes à reconnaître pour la transcription,
- une erreur de transcription ne doit pas interrompre la boucle d'interaction homme machine,
- inversement, l'utilisateur peut à tout moment corriger la reconnaissance faite par la machine en ajustant la segmentation en composantes connexes et en corrigeant ou ajoutant la valeur du caractère.

A partir de ces contraintes, Gardes à conçu un système de transcription supervisé dans lequel l'utilisateur retrace au stylet la forme du caractère mal ou non reconnu et la reconnaissance du tracé donne à la fois une hypothèse de segmentation et une hypothèse de valeur du caractère en cours de traitement.
- L'hypothèse de segmentation est obtenue par le rectangle englobant du tracé fait par l'utilisateur ; elle en comparée à la segmentation faite dans l'image. Cela permet de détecter les composantes connexes correspondant à des caractères tronqués ou ligaturés. Si cette détection est négative, l'information de segmentation n'est pas remise en cause, dans le cas contraire, la ou les composantes connexes incriminées dans l'image dont re-segmentées en se basant sur le rectangle engobant du tracé.
- l'hypothèse de valeur du caractère est obtenue par la reconnaissance du tracé effectué par l'utilisateur. Le principe pragmatique est que si la valeur reconnue du tracé est jugée bonne par l'utilisateur, c'est que cette valeur correspond au caractère en cours de reconnaissance ; par conséquent, la machine affecte cette valeur à la composante connexe et reprend la poursuite du traitement.

Ce mode de fonctionnement résout le problème de la validation supervisée d'une reconnaissance de forme que l'on qualifiera de "tout terrain". Cependant, il s'agissait également de compenser l'absence d'entraînement a priori du système par un apprentissage à la volée.

Cet apprentissage se base actuellement sur une heuristique simple : à chaque composante connexe correspond un caractère. Les limites de cette heuristique concernent le traitement des caractères accentués ; limites partiellement compensées par un traitement allant vérifier la présence d'une composante connexe voisine de celle en cours de traitement et située au dessus de cette dernière. Cette heuristique permet de mettre en œuvre un test de décision très simple : si la segmentation des composantes connexes n'est pas remise en cause par une action de l'utilisateur alors, la composante connexe est insérée dans la base de référence pour la classification. L'entrainement de cet "OCR supervisé" devient effectif très rapidement (deux ou trois exemples suffisent à automatiser la reconnaissance). En outre, Gardes n'a pas constaté de phénomène de sur-apprentissage qui se traduirait par l'apparition d'ambiguïtés dans la reconnaissance.
Globalement, il s'agit d'une méthode d'apprentissage incrémentale dont la mise en œuvre est simple sur le plan logiciel et compatible avec les contraintes initiales du prototype : l'entraînement d'une reconnaissance de tracés manuscrits (selon une approche markovienne incrémentale), est répercuté progressivement dans "l'OCR supervisé" à partir des actions de l'utilisateur en situation de correction. Il s'agit d'une véritable mise en œuvre de l'approche ICT de Caelen (Interaction, Contexte, Trace) expliqué plus loin.

2.5 Un nouveau statut pour la transcription

On a d'abord perçu la transcription comme une simple alternative à l'OCR donc une démarche de même nature, c'était la vision des années 2000. Pour chaque corpus on va élaborer un algorithme de transcription. De fait, cela nous ramenait aux démarches mises en oeuvre au début de l'OCR où on travaillait sur une police ; on a commencé par élaborer des OCR mono-police, puis multi-police avant de passer aux omni-polices d'aujourd'hui ; on avait même eu l'idée de fabriquer des polices simples à reconnaître, l'OCR A et à l'OCRB.

Les OCR font, de plus, aujourd'hui un appel systématique à des dictionnaires, ce qui pourrait dans des documents patrimoniaux nous conduire à un non respect de l'original (encore faudrait-il qu'il existât des dictionnaires numérisés pour chaque période du passé).

L'analyse des orientations prises dans les trois chantiers récents montrent qu'on cherche d'autres voies que la démarche OCR ; Ramel veut introduire un va et vient entre le transcripteur et l'ordinateur. Eynard veut introduire une maitrise de la qualité liée à la technicité des algorithmes et au travail demandé au travail d'alignement demandé à l'humain. Gardes et Leydier ont déjà introduit une interaction en phase de reconnaissance.

Des réflexions sont en cours pour aller plus loin, et pour faire de la transcription assistée une véritable stratégie de rupture, ce ne sera ni de l'OCR, ni de la ressaisie manuelle. Le cadre de travail pourra être qualifié par le sigle ICT pour Interaction, Contexte et Trace. Dans le cadre d'une véritable interaction, l'ordinateur coopérera avec des spécialistes humains du corpus à transcrire ; dans un premier temps on initialisera le processus en utilisant les travaux rapportés ci-dessus, dans une seconde phase interactive et coopérative, on éliminera progressivement les points critiques.
On sera dans ce contexte dans une situation de réédition coopérative, annoncée pour un horizon de deux à trois ans en France.

3 Stratégie pour le Syriaque

Nous nous interrogeons quant à savoir si, la sobriété et la simplicité du Syriaque qui, jusqu'à présent, nous ont plutôt aidés dans notre démarche, vont laisser suffisamment d'information à nos procédures algorithmiques de reconnaissance, et constituent pas une garantie de bon fonctionnement de la transcription sur cette langue.

Comme point positif on a obtenu une segmentation des lignes d'écriture en trente imagettes ou graphèmes et à 22 de ces graphèmes on sait associer les 22 consonnes.

On va développer une stratégie avec l'objectif de démontrer qu'une forme de transcription est applicable au Syriaque, on est dans une phase d'exploration et on vise dans un premier temps une étude de faisabilité. Nous nous positionnons dans une démarche de faisabilité notamment pour le fait que les travaux concernant la transcription ont porté essentiellement sur des imprimés.

4 Distance entre nos graphèmes

4.1 La distance de compression

Notre attention s'est tournée vers une approche de classification utilisant une distance dite de compression jusqu'à présent pas utilisée dans le domaine des images de documents excepté par Gardes que vient de lui consacré des pages dans sa thèse [GAR09].

Gardes ont fait usage de cette méthode pour classifier des pages de journaux anciens, des fichiers informatiques de formats différents, et des ensembles de tracés graphiques. Les résultats qu'ils ont obtenus montrent des regroupements des pages de journaux similaires selon qu'elles contiennent des illustrations ou non, selon leur provenance (si elles sont du même journal), et selon la mise en page et le nombre de colonnes. Cette méthode leur a permis aussi de regrouper ensemble les fichiers informatiques de même format, ainsi que les tracés graphiques similaires.

Cette méthode ne requiert aucune connaissance à priori, et ne fait pas usage de caractéristiques spécifiques dépendant de la nature des objets à classifier. Elle a été développée par Cilibrasi et al [CIL05] et fut appliquée initialement dans le domaine de la bioinformatique pour le traitement des séquences d'ADN (acide désoxyribonucléique), et opère comme suit : tout d'abord une distance de similarité la Distance Normalisée de Compression (Normalized Compression Distance) est déterminée, son calcul est effectué à partir des longueurs des fichiers de données comprimés, sur lesquels une méthode de classification hiérarchique est appliquée.

Pour évaluer la distance NCD entre deux objets x et y, on procède comme suit : Une compression sans perte est d'abord appliquée sur x et sur y séparément, laquelle conduit à C_x et C_y respectivement ; la concaténation de x et de y est ensuite comprimée et conduit à $C_{(xy)}$; à cette dernière valeur la plus petite des valeurs C_x et C_y, et on divise ce résultat par le maximum de C_x et C_y :

$$NCD_{(x,y)} = \frac{C_{(xy)} - \min\{C_x, C_y\}}{\max\{C_x, C_y\}}$$

Application pour la classification

Soit N le nombre total des objets à classer. En choisissant un objet parmi ce nombre, et en se servant de l'équation ci-dessus un calcul de vecteur de différence entre cet objet et tous les autres est effectué, les valeurs dans ce vecteur seront rangées par ordre croissant. Ce vecteur a comme taille $1 \times N$.

En choisissant à chaque fois un objet différent, toutes les combinaisons possibles des vecteurs de différences seront alors calculées. Une matrice dite de distance est alors construite, c'est une matrice carrée de taille $N \times N$.

A partir de cette matrice, un arbre binaire sans racine est construit. Chaque nœud interne de cet arbre constitue la racine d'un sous-arbre, ce qui facilite l'identification des classes.

Application sur nos imagettes

Comme première tentative, nous avons essayé d'appliquer cette méthode afin de classifier les imagettes en Serto résultant des segmentations successives. Nous avons choisi de nous concentrer sur la base Serto vu le fait que nous disposons de données en cette calligraphie plus volumineuses que celle en Nestorien, de plus l'usage de calligraphie Serto est plus répandu que celui de la calligraphie Nestorien.

Les imagettes sont de tailles différentes, elles ont toutes la même hauteur (96 pixels qui est la hauteur moyenne d'une ligne de texte), la différence apparaît alors au niveau de la largeur. L'imagette d'un « n-gramme » est plus large que celle d'une lettre individuelle, et les imagettes de certaines lettres s'avèrent plus larges que d'autres. Les imagettes des occurrences d'une même lettre peuvent aussi varier en largeur. En calculant la moyenne des largeurs des imagettes, et en se servant de cette valeur pour seuillage, toute imagette de largeur supérieure est considérée comme celle d'un « n-gramme », sinon ce serait celle d'une lettre individuelle.

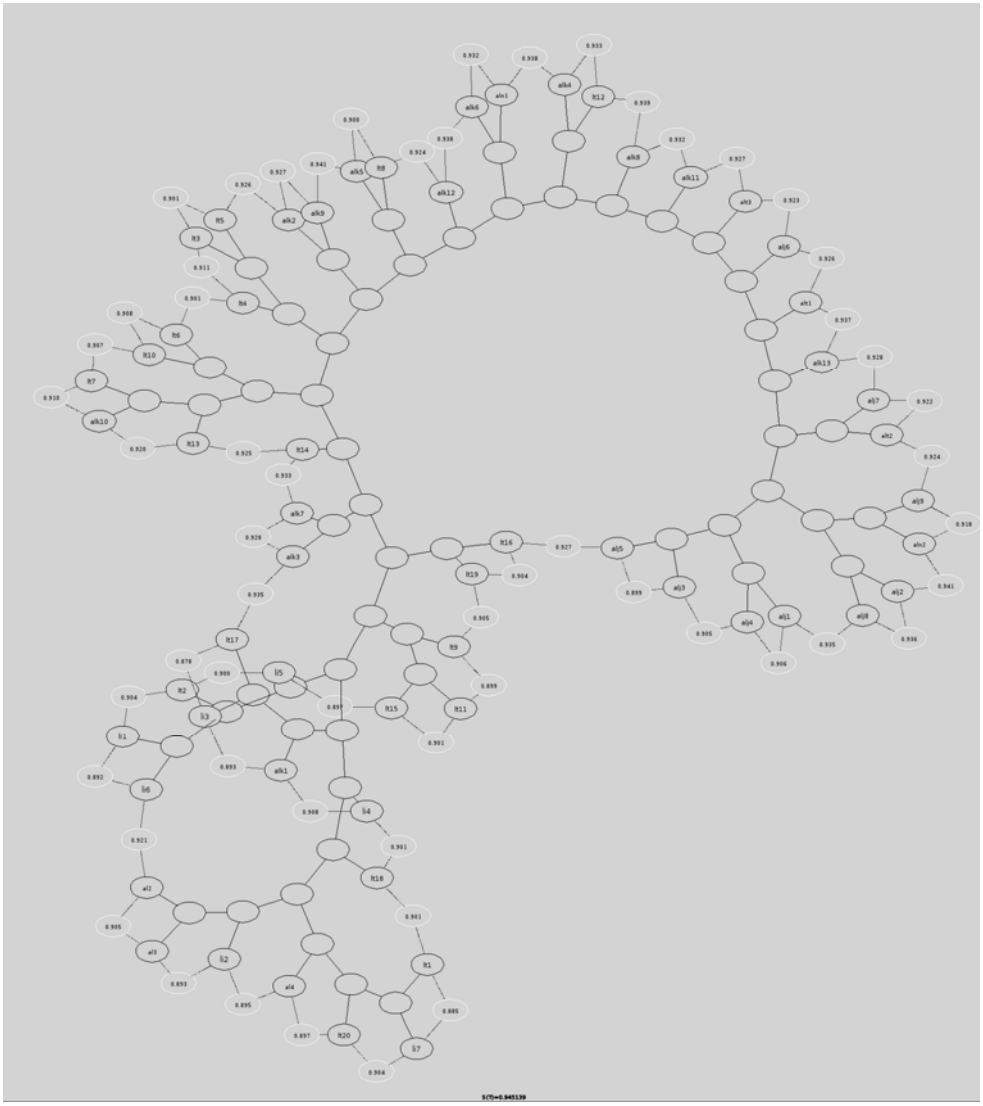


Figure 2: Arbre de distances obtenu pour les « n-grammes ».

Nous avons donc tenté de classer en utilisant cette méthode les « n-grammes » qui se ressemblent, et les lettres qui se ressemblent. Le graphe obtenu pour les images des lettres étant trop large pour inclure dans ces pages, nous allons montrer dans la figure ci-dessous le graphe obtenu pour les « n-grammes ».

A première vue, nous remarquons la dispersion du graphe ainsi qu'un nombre assez important de nœuds indiquant les classes. Ce résultat montre que ce classifieur n'a pas réussi à grouper ensemble sous un même nœud toutes les imagettes d'un même « n-gramme », il les a plutôt distribuées sur plusieurs classes (nœuds). De plus, il a groupé sous de mêmes nœuds des imagettes correspondant à des « n-grammes » différents.

Nous allons agrandir certaines portions de ce graphe afin de montrer les zones dans lesquelles les résultats sont cohérents avec la similarité réelle des imagettes, et les zones où ils ne le sont pas.

La figure ci-dessous montre une zone du graphe dans laquelle plusieurs versions du même « n-gramme » sont groupées proches les unes des autres sous un même nœud principal, mais distribués sur plusieurs sous-nœuds.

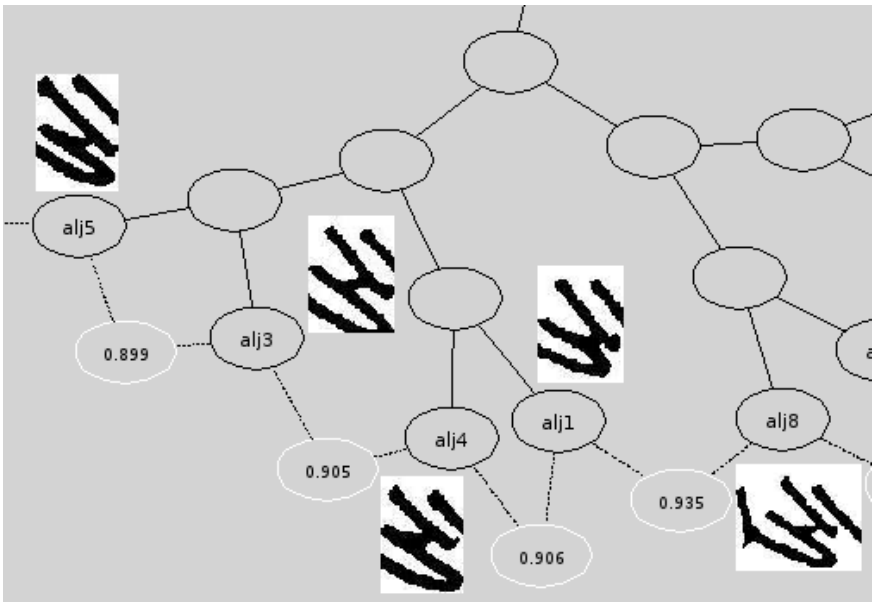


Figure 3: « N-grammes » similaires groupés ensembles.

La figure ci-dessous montre une zone du graphe dans laquelle des « n-grammes » différents sont groupés ensembles. Par contre, bien que différents, il existe une certaine ressemblance entre leurs images.

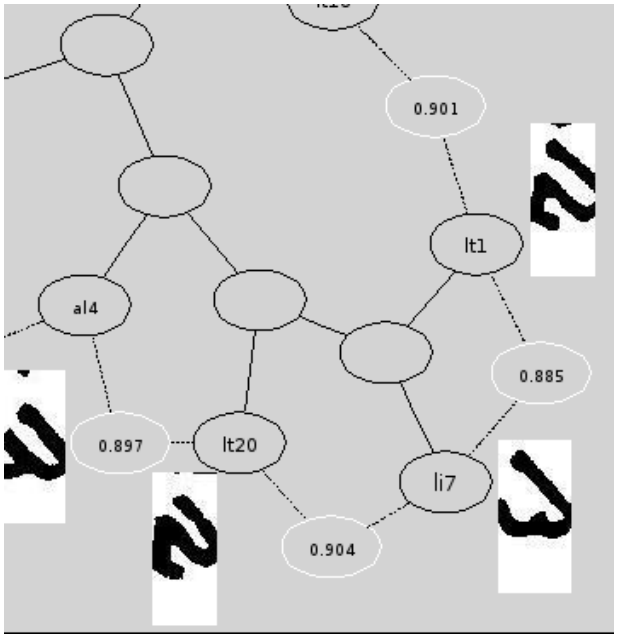


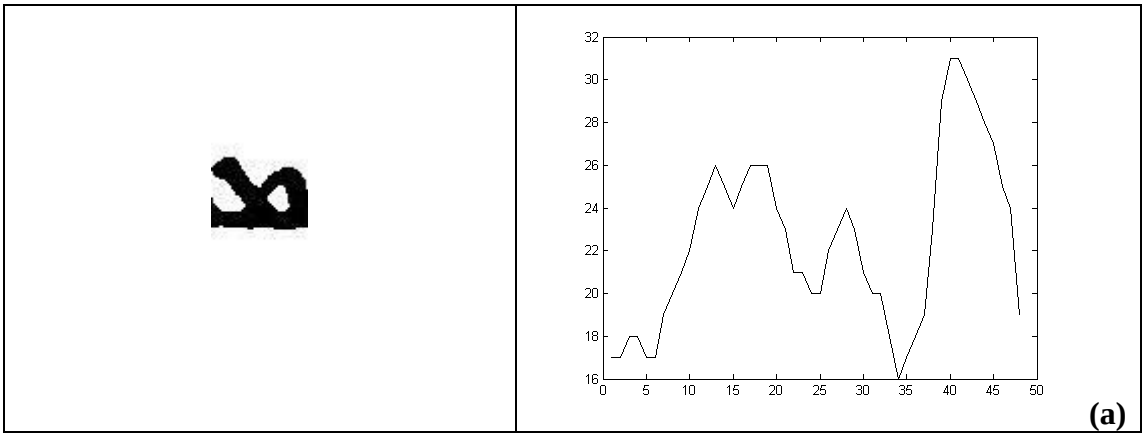
Figure 4: « N-grammes » différents groupés sous le même nœud.

A noter que les résultats de l'application de ce classifieur sur les imagettes des lettres individuelles n'étaient pas meilleurs que ceux obtenus pour les imagettes des « n-grammes ».

4.2 Distance entre les graphèmes syriaques

Caractérisation des caractères et des graphèmes

Les caractéristiques que nous avons choisies d'extraire sont des caractéristiques de profil ; horizontal, vertical, et diagonal. Les imagettes étant en mode binaire, nous comptons le nombre de pixels noirs suivant les deux axes vertical et horizontal, et suivant la diagonale de l'image. La figure montre les trois profils : vertical (a), horizontal (b) et diagonal (c) pour une imagette de la lettre « m ».



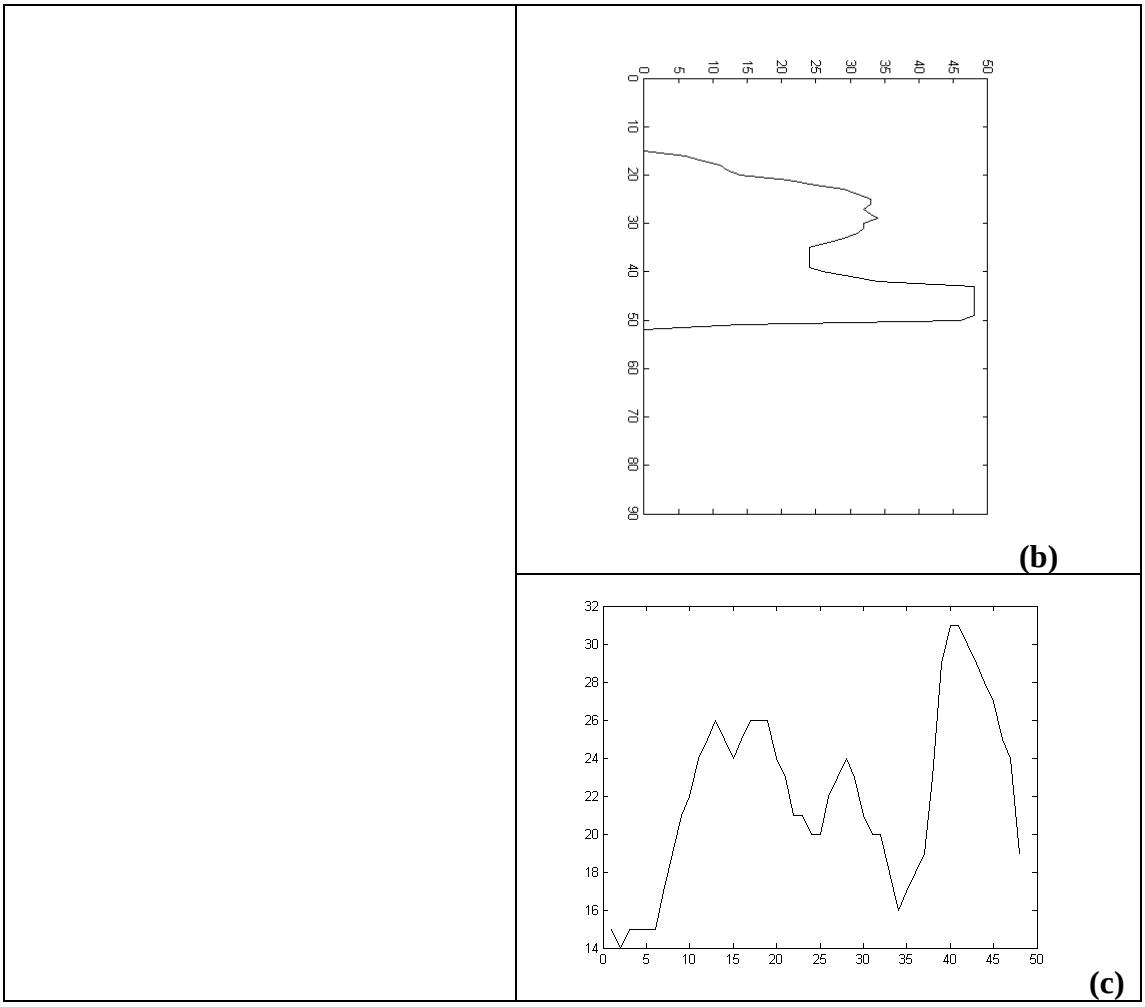


Figure 5 : Les trois profils pour une imagette de la lettre « m ».

Pour chacun de ces trois profils nous avons choisi de retenir cinq valeurs, ce qui nous conduira à caractériser les caractères par un vecteur ayant quinze composantes positives ou nulles. Pour chaque profil, ces cinq valeurs consistent en la valeur milieu du profil, la valeur située au 1/6^{ième} du profil, la valeur située au tiers du profil, la valeur située aux 2 tiers du profil, et la valeur située au 5/6^{ième} du profil.

Ce choix de ces valeurs n'est pas anodin, nous avons voulu capter la distribution de la densité des pixels noirs sur la surface de l'imagette, sans toutefois négliger l'information centrale, l'aspect diagonal, ainsi que la présence ou non de symétrie par rapport à cette dernière.

Les cinq valeurs respectivement gardées pour chacun des profils sont représentées dans la figure 18 ci-dessous

5 valeurs pour le profil vertical	5 valeurs pour le profil horizontal	5 valeurs pour le profil diagonal
		

Figure 6.

Distances utilisées

Pour évaluer une dissemblance entre deux imagettes (deux graphèmes) représentées par deux vecteurs de dimension15 caractéristiques on peut envisager plusieurs types de distances. Désignons par p et q deux vecteurs de caractéristiques de taille n.

La distance euclidienne, dite aussi distance ordinaire, est la métrique de distance la plus intuitive à utiliser, elle est calculée comme suit :

$$d_{pq} = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

La distance City Block qui est utilisée en général comme une approximation de la Distance Euclidienne afin de réduire la charge opérationnelle est calculée comme suit :

$$d_{pq} = \sum_{k=1}^{n} |p_k - q_k|$$

Nos vecteurs de caractéristiques n'ont que 15 composantes et notre base d'apprentissage est relativement petite, l'usage d'une telle distance ne se justifie pas

La distance du Sup, dite aussi distance de Chebychev entre les deux vecteurs est donnée par :

$$d_{pq} := \max_{k=1 \to n} (|p_k - q_k|)$$

Le désavantage de cette distance c'est qu'elle néglige les faibles fluctuations qui pourraient apparaître entre deux lettres différentes au dépit d'une caractéristique qui pourrait être très différentes entre deux imagettes d'une même lettre, tandis que les autres distances pénalisent les différences globales entre les caractéristiques.

La distance de Hamming permet de quantifier la différence entre les deux vecteurs de caractéristiques en comptant le nombre des valeurs différentes entre les deux. Cette distance s'est avérée très sensible et très binaire face aux variations des pixels. Pour différentes occurrences d'une même lettre, nous pouvons parfois avoir un tracé plus fin ou plus épais, cette variation se traduirait par une différence de l'ordre de quelques pixels noirs entre le vecteur de caractéristiques cherché et celui d'une imagette de la base étiquetée ; pour la distance de Hamming une différence de 1 pixel donnerai le même résultat qu'une différence de 30 pixels, ce qui ramènerai à des résultats erronés.

Notre démarche est à rapprocher de celle que Eynard a choisie pour construire une distance entre les caractères extraits de la Gazette de Leyde ; il représente les caractères issus de la segmentation par des vecteurs de projection ayant un total de 40 composantes,

5 Apprentissage

Création d'une base d'apprentissage

La base d'apprentissage créée après les deux segmentations successives d'une dizaine de pages de texte en Serto, contient environ 1100 imagettes de lettres individuelles et de « n-grammes » stables. L'étiquetage (donc la classification) de cette base a été effectué manuellement en regroupant sous un même nom les différentes imagettes représentant une même lettre ainsi que les tri-grammes stables.

Les groupes d'imagettes ne sont pas tous de la même taille, ceci dépend de la fréquence des lettres dans la langue. Nous ne connaissons pas ces fréquences pour le Syriaque, nous avons donc évaluées ces fréquences dans le contexte de notre base

Nous avons donc étiqueté les 22 consonnes de l'alphabet Syriaque, et comme nous l'avons déjà mentionné, nous avons eu une persistance de 7 « n-grammes ». Ces derniers sont réguliers, nous avons pu compter 7 formes de « n-grammes » différentes, ainsi qu'une forme de lettre particulière assez présente qui pourrait être un « a » ou un « l », elle constitue une classe séparée à qui nous attribuons l'étiquette « a_ou_l ». Donc au total nous avons 30 classes.

Il est important de remarquer que cet étiquetage manuel n'associe pas toujours un nom de lettre à chaque classe : il y a la classe qu'on a appelé « a_ou_l ». Au début d'un mot la lettre « a » est toujours verticale sauf si elle est suivie de la lettre « l » alors le « a » va suivre la penchée du « l », et le « l » est toujours penché quelque soit sa position dans un mot. En l'absence du contexte nous ne pouvons pas décider si c'est un « a » ou un « l ».

Nous donnons un aperçu de cette base d'imagettes étiquetées dans la figure 7. Ce n'est pas la base d'imagettes en entier, ce sont quelques échantillons que nous avons choisis pour donner une idée des classes que nous avons obtenues.

Figure7 : Aperçu de la base d'imagettes de lettres et de « n-grammes ».

Fréquence des lettres

Le tableau ci-dessous montre pour chaque lettre le pourcentage de présence de ses occurrences dans notre base. Nous allons référer à la forme de la lettre par l'étiquette que nous lui avons attribuée.

Etiquette lettre	**Forme lettre**	**Pourcentage de présence**	**Etiquette lettre**	**Forme lettre**	**Pourcentage de presence**
a		10,25%	m		7,625%
a_ou_l		10,875%	n		1,5%
b		3,25%	o		8%
ch		0,125%	p		6,5%
d		3%	q		3,625%
ein		5%	r		1,375%

h		0,25%	s		4,625%
hh		1,875%	sa		3%
i		2,625%	t		4,75%
j		2,125%	ta		3,625%
k		3,125%	z		2,5%
l		3,125%			

Figure 8 : Liste des lettres de l'alphabet ainsi que leur taux d'occurrence.

Fréquence des « n-grammes »

Afin de montrer la régularité dans la présence des « n-grammes » persistants, nous avons aussi calculé leur taux d'occurrence dans notre base d'imagettes.

N-gramme	**Forme**	**Pourcentage de présence**	**N-gramme**	**Forme**	**Pourcentage de presence**
al		0,5%	alt		0,375%
alj		1,125%	li		0,875%
alk		1,625%	lt		2,5%
aln		0,25%			

Figure 9 : Liste des « n-grammes » ainsi que leur taux de présence.

Classification automatique

Les différents essais de classification automatique que nous avons tenté de mettre en œuvre sur la base d'apprentissage n'ont pas donné des résultats encourageant, et ce, quels que soient les procédures de classification utilisées

6 Essai de Transcription-Reconnaissance

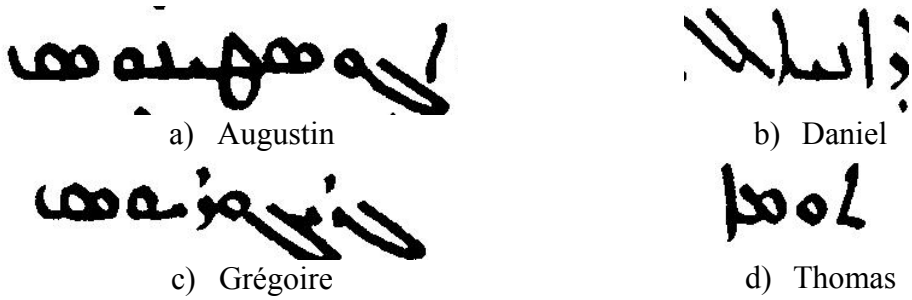
6.1 Reconnaissance de caractères

Une base d'apprentissage correctement étiquetée constitue le support de tout algorithme de reconnaissance de caractères. Il est clair que nous ne sommes pas dans cette situation puisque au moins une classe n'est pas étiquetée de façon univoque, la classe qu'on a appelée « a_ou_l ». Nous allons devoir introduire des informations contextuelles pour pouvoir décider entre le « a » et le « l ».

6.2 Constitution d'une base de tests

La base

Nous avons constitué une petite base de tests sur laquelle nous allons tester notre méthode de transcription ; selon les résultats nous pourrons l'évaluer et l'améliorer au fur et à mesure. Nous avons extrait 50 mots, dont des noms propres, dans un de nos documents Serto, nous les segmentons, d'une part, et les étiquetons, d'autre part ; plusieurs de ces mots apparaissent sur la figure 10.



a) Augustin b) Daniel

c) Grégoire d) Thomas

Figure 10 : Quelques noms propres choisis du dictionnaire.

Segmentation des mots de la base de tests

Nous avons adopté la même démarche de segmentations successives pour segmenter les mots du dictionnaire en lettres individuelles et « n-grammes » (puisque nous travaillons sur la calligraphie Serto).

Nous montrons ci-dessous le résultat de la segmentation du nom « Augustin ». La figure montre une division en lettres individuelles, le nom est constitué de 9 lettres au total.

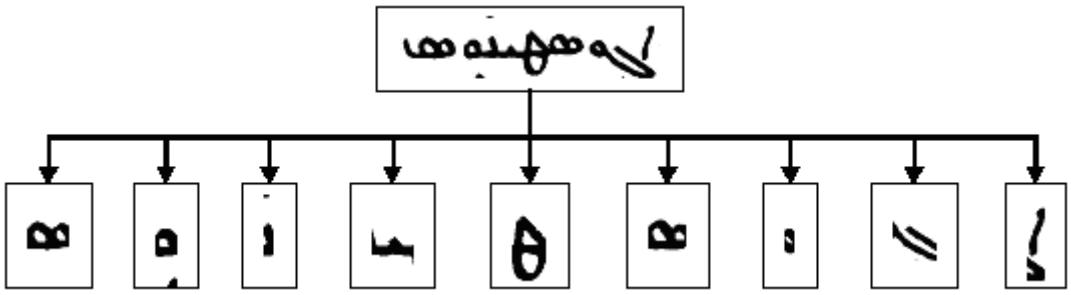


Figure 11 : Segmentation du mot « Augustin ».

6.3 Choix d'une méthode en cascade

Nous avons mis en place une procédure de classification procédant en cascade, selon deux phases distinctes, en accord avec les résultats de la segmentation qui nous propose trente formes assez stable et récurrentes

- 22 consonnes,
- 7 trigrammes,
- une forme associée à une indécision entre un « a « et un « l «

Première Phase

Pour cette phase nous avons choisi le k-plus proche voisin ou kppv, c'est l'une des méthodes les plus simples utilisées dans des applications de reconnaissance de séquences telles la reconnaissance d'écriture manuscrite ; cette méthode est assez fiable et facile à mettre en oeuvre

La méthode kppv cherche les k imagettes les plus proches d'une observation inconnue que l'on veut reconnaître et lui attribue l'étiquette de la classe la plus représentée parmi ses k plus proches voisins ; cette étiquette peut selon le cas être

- un nom de lettre
- l'indicateur d'un des 7 trigrammes,
- l'étiquette « a_ou_l ».

Seconde phase

Elle ne concernera que certains résultats de la première phase, celles ooù le résultat était indicateur d'un des tri-gramme ou « a_ou_l ».

- pour le cas « a_ou_l » on décidera de l'étiquette a ou de l'étiquette l, c'est-à-dire du code UTF-8, après une étude du contexte (lettre précédant et lettre suivant),
- pour les trigrammes on donnera les 3 codes UTF-8 correspondants.

Après avoir associée une étiquette aux lettres et aux n-grammes, la transcription se termine par leur remplacement par leur représentation en Unicode. Une fois « codés », les lettres ou les « n-grammes » reconnus peuvent désormais être affichés sur l'écran en leur forme dactylographiée.

6.4 Les résultats selon la valeur k.

L'évaluation des résultats pour le kppv se fait en termes de calcul d'un taux de reconnaissance, qui dans notre cas consiste en le pourcentage des lettres correctement reconnues parmi le nombre total des lettres. Les taux de reconnaissance que nous allons montrer ci-dessous ont été évalués pour l'ensemble total des mots de notre base de tests.

Pour une première évaluation, nous avons choisi k=1, une démarche classique consiste à se référer en premier au plus proche voisin, et attribuer l'étiquette de sa classe à la lettre inconnue. Ce choix a résulté en un taux de reconnaissance sur l'ensemble du mini-dictionnaire de 61%.

Pour une deuxième tentative avec k = 3 le taux de reconnaissance croit à 65%.

Pour la troisième tentative avec k=5 le taux de reconnaissance monte à 69%.

On serait tenté de croire qu'en augmentant la valeur k le taux de reconnaissance continue à s'améliorer. Il n'en est rien, avec k=7 ce taux redescend à 54%. Cette régression pourrait être justifiée par le fait que dans notre base étiquetée, les classes sont de tailles différentes. Nous avons des classes de lettres et de « n-grammes » représentées par un cardinal d'imagettes faible comparé à d'autres classes qui elles sont de taille plus conséquente ; de plus, certaines lettres bien que différentes en sens sont assez ressemblantes ; aussi, plus nous augmentons la valeur de k plus nous risquons de dissoudre une classe de faible taille dans une autre différente et plus volumineuse ayant des images ressemblantes.

Augmenter la valeur de k permet de donner de meilleurs taux de reconnaissance lorsque la base d'apprentissage est volumineuse et que les classes dans cette base sont équitablement représentées, nous allons nous contenter pour le moment de k = 5

7 Deux cas de transcription

« Augustin »

Pour le nom propre « Augustin » prononcé « Ajostinos » (à rappeler que le Syriaque est écrit et lu de droite à gauche), nous allons illustrer les étapes par lesquelles est passé notre algorithme afin d'afficher sur l'écran la version dactylographiée de l'image du mot manuscrit.

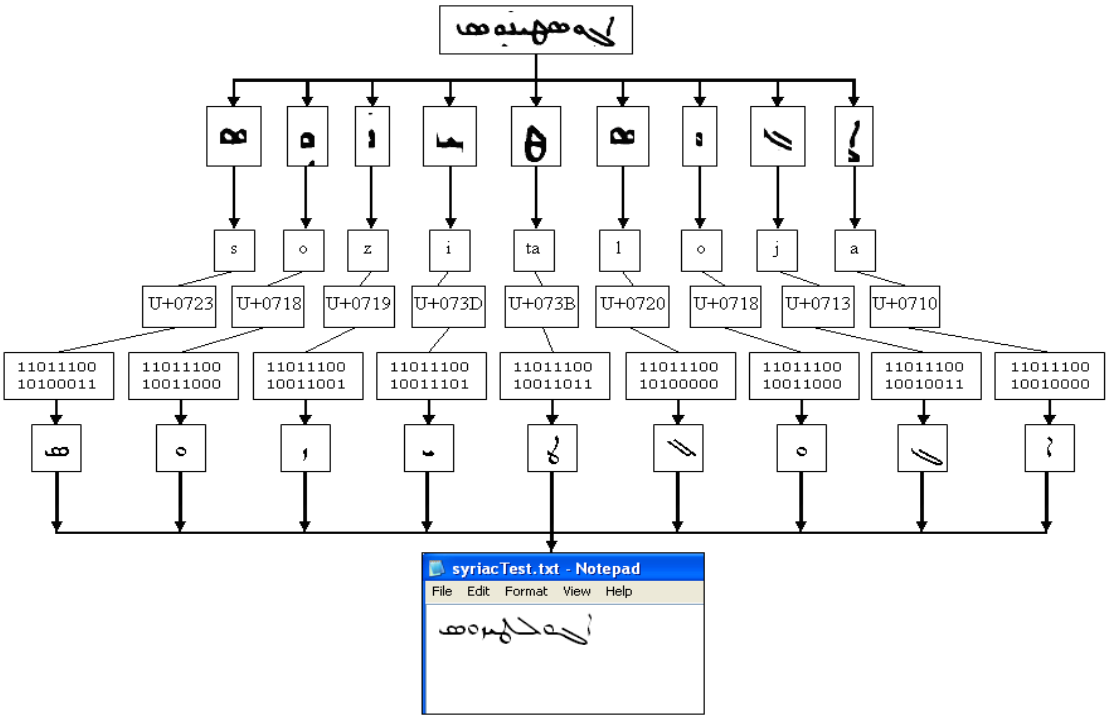


Figure12 : Schéma récapitulatif sur la reconnaissance du mot « Augustin » et sa transcription.

La quatrième lettre était censée être reconnue comme étant un « s » à la place elle a été jugée comme un « l », et la septième lettre était censée être reconnue comme étant un « n » , en fait elle a été reconnue comme un « z ». Pour ce mot, avec la valeur k = 5 qui maximise le taux de reconnaissance, notre algorithme a reconnu

correctement 7 lettres parmi 9, les erreurs de reconnaissance ont eu lieu au niveau de la quatrième et la septième lettre depuis la droite.

« jour »

Nous allons montrer un mot pour lequel toutes les lettres ont été correctement reconnues, c'est mot « jour » qu'on prononce « iom », un mot court de trois lettres.

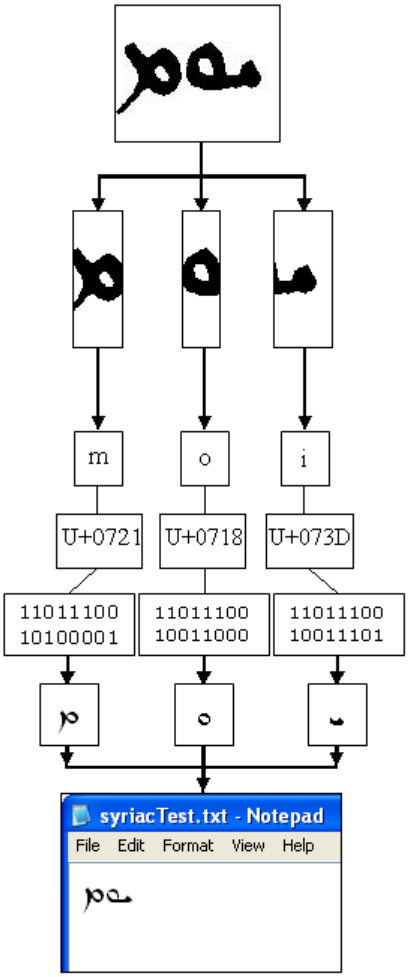


Figure 85: Le mot « jour » pour lequel toutes les lettres ont été correctement reconnues.

8 Autres travaux sur la reconnaissance du Syriaque

Très rares sont les gens qui se sont lancés dans l'étude des écrits Syriaques. Clocksin et al. [CLO03], ont effectué une tentative vers une transcription automatique de manuscrits Syriaques. Pour ce faire, les caractéristiques qu'ils ont utilisées sont des moments géométriques qu'ils ont extraits à partir de l'image d'une lettre et de celle de sa transformée polaire, leur vecteurs de caractéristiques varient en longueur de 5 à 175 caractéristiques. Pour un nombre n de classes ils ont utilisé n(n-1)/2 classifieurs basés chacun sur un SVM (Support Vector Machine) nécessitant une phase préalable d'entraînement. Ils ont conduit leurs tests sur des documents manuscrits et des documents imprimés avec la calligraphie Estrangelo, les taux de reconnaissance qu'ils

ont obtenus varient de 61% pour les documents manuscrits à 100% pour les documents imprimés.

Dans d'autres travaux [CLO04], ils se sont concentrés sur la reconnaissance de caractères Syriaques Estrangelo isolés, les caractéristiques qu'ils ont utilisées sont des points basés sur la «order structure invariance » de Carlsson [CAR98] extraites à partir des contours des lettres, l'appariement entre deux images d'un même caractère est effectué suivant une correspondance de point à point. Cette approche n'a pas abouti à des résultats prometteurs mais a constitué quand même une piste de recherche sur l'écriture Syriaque.

9 Proposition : correction à la volée issue de la loupe

Ce paragraphe consiste en la présentation d'une démonstration de transcription dans laquelle l'opérateur peut aider l'ordinateur

10 Conclusion

La transcription assistée semble une voie possible pour le Syriaque mais, pour progresser en identifiant les difficultés il faudra travailler sur des corpus importants en taille.

La transcription vue comme une simple alternative à l'OCR, c'est-à-dire comme un OCR dédié à l'écriture Syriaque n'est pas possible. Il faut s'orienter dans la voie appelée ICT pour Interaction, Contexte et Traces ; le cas du graphème a_ou _l où on ne peut décider qu'après que les lettres voisines soient identifiées. C'est en ce sens que nous pensons qu'il faudra développer une véritable « coopération « entre la machine et l'opérateur.

Conclusion et prospectives

1. Bilan

Dans ce mémoire nous avons voulu attirer l'attention sur le Syriaque et contribuer à la rendre accessible grâce à la numérisation.

Le Syriaque est une langue fascinante, par son histoire, mais aussi par son écriture qui est d'une sobriété exemplaire, simple et concise, dépouillée sans être triste ; la structure logique est élémentaire, les textes n'ont pas eu besoin de toutes règles de mise en page développées en Occident pour le Latin et ses descendants. Ses fragments de lettres tracés avec un angle de 45° par rapport à la ligne d'écriture la font identifier instantanément parmi n'importe quelle autre écriture

Aussi dans ce mémoire, nous nous sommes attaché à développer des méthodes et des algorithmes conçus spécifiquement pour satisfaire aux caractéristiques propres et aux exigences de la langue et de l'écriture en tant que tels. La simplicité et la concision du Syriaque font que dans le trait écriture il y a très peu d'information « redondante ». Aussi, il ne faut pas trop en perdre lors de la numérisation et des diverses traitements que l'on fait pour définir et évaluer les caractéristiques ; le choix des méthodes d'appariement dans le chapitre 3 nous semblent bien répondre à cette exigence.

Dans le Chapitre 1, nous avons présenté la langue Syriaque et nous avons tenté de la situer historiquement, géographiquement, et scientifiquement par rapport aux

autres langues. Notre démarche se voulant aussi pédagogique, nous avons rappelé les difficultés inhérentes à la préservation du patrimoine écrit et nous avons aussi rappelé que la numérisation doit être abordée de façon professionnelle.

Dans le Chapitre 2, nous avons abordé l'organisation et la structure des documents Syriaques, ils ont une présentation particulière, notamment de la hiérarchie du texte. L'information visuelle apparente de ces documents rapporte des détails propres à cette écriture. La sobriété de l'écriture Syriaque, sa mise en page élémentaire, son système de ponctuation rudimentaire attirent immédiatement l'attention vers le contenu de la page de texte sans aucune distraction introduite par des éléments ornementaux ; dans les documents latins, un tel contenu se trouve dissimulé derrière des lettrines, des enluminures et autres éléments décoratifs. Nous avons choisi de faire une comparaison avec les documents Latins, ces derniers ayant fait l'objet d'une étude exhaustive similaire. La description visuelle apparente des documents Syriaques se trouve désormais plus aisée, la structure plus accessible aux algorithmes d'analyse et d'extraction. Le développement de ces algorithmes notamment d'algorithmes de segmentation d'une page de texte en ses lignes constitutives et des lignes de texte en graphèmes ont constitués une partie importante de ce chapitre.

Dans le Chapitre 3, nous avons effectué une première démarche d'indexation des manuscrits Syriaques, laquelle consistait en un repérage de mots ; plus précisément, partant de l'image du mot requête, nous cherchons à retrouver toutes les occurrences dans le document. L'élément directionnel nous a paru être la clef dans la distinction entre les formes, c'est la raison pour laquelle nous avons choisi des caractéristiques conduisant à la rose des directions. Une telle approche nous a permis de retrouver toutes les occurrences du mot cherché même dans des conditions simulées de faible résolution et de forte compression avec perte ; pour l'heure, ce bon repérage se fait au prix du repérage de quelques fausses occurrences ; ces dernières bien qu'elles aient un sens différent de celui du mot requête, montrent des traits de ressemblance et partage des lettres en commun avec le mot cherché.

Les résultats de la segmentation décrits dans le Chapitre 2 ont servi à développer une approche de transcription assistée des manuscrits Syriaques, travail faisant l'objet du Chapitre 4 ; chaque mots est décomposé en à ses graphèmes constitutifs, des lettres individuelles et des « n-grammes ».

Cette approche est jugée comme étant risquée vu l'aspect cursif de l'écriture. Nous avons pu dégager, à partir de ces résultats une première vérité de terrain pour la langue, qui a servi de base d'apprentissage pour notre algorithme de reconnaissance testé sur une sélection de mots.

Nous avons introduit une méthode de transcription selon deux étapes. Dans la première nous cherchons à apparier ces graphèmes avec notre base classée et étiquetée et à attribuer au graphème inconnu l'étiquette de la classe (qui peut être une étiquette de lettre), dans la seconde nous terminons l'étiquetage en lettres.

Les travaux développés dans ce mémoire constituent une première exploration informatique conçue pour les manuscrits Syriaques numérisés, ainsi que des épreuves de faisabilité vu le fait que ces documents n'ont jusqu'à présent jamais été objet d'une recherche scientifique. Ceci nous a conduits à travailler sur tous les aspects liés au traitement des images de ces documents, que ce soit l'aspect descriptif et visuel,

l'aspect indexation et recherche textuelle, l'aspect reconnaissance d'écriture et transcription et à expérimenter dans des approches jugées auparavant trop difficiles ou même irréalisables. De premiers outils informatiques spécifiques à la langue ont été développés, ils ont réintroduit la langue Syriaque dans la recherche scientifique et montré que cette langue constitue un champ prolifique de recherche. L'expérience acquise lors de ce travail pourrait aider dans des travaux concernant d'autres langues qu'elles soient Sémitiques ou pas.

2. Prospectives

Comme pistes ultérieures de recherche nous pouvons proposer des tentatives de constitution d'une vérité de terrain pour les documents Syriaques, ainsi que la création d'un dictionnaire électronique de cette langue ; ce dernier constituera un support sur lequel nous pouvons nous baser afin d'effectuer des tentatives de « Word Retrieval » (décrit dans le Chapitre 3) pour du repérage de mots. Nous n'avons pas pu aborder actuellement cette recherche par faute de moyens et de matières premières, pour cela nous avons besoin d'une base d'images de lettres segmentées assez volumineuse et qui soient classées et étiquetées le plus équitablement possible, cette étape nécessite un apprentissage qui dans le cas traité dans ce mémoire pour notre base préliminaire était supervisé, ainsi qu'un dictionnaire électronique non seulement vocabulaire mais aussi grammatical englobant les abréviations et les dérivées grammaticales des mots, ainsi qu'une étude approfondie sur la constitution des mots Syriaques.

En ayant cette vérité de terrain, nous envisageons la mise en place d'un moteur de recherche textuelle dans les images de documents dans lequel le mot requête est saisi au clavier par l'utilisateur ; des images virtuelles de ce mot sont alors constituées en concaténant ensemble différentes images des lettres individuelles, une recherche de similarité de ces images synthétisées peut alors être effectuée dans les différentes pages du corpus afin de retrouver les occurrences du mot requête.

Une autre piste dans laquelle nous souhaitons continuer, est effectivement la transcription de l'écriture Syriaque. Ce que nous avons proposé dans le Chapitre 4 constitue une première tentative de transcription assistée qui s'est avérée assez prometteuse. Les taux de reconnaissance obtenus ne soient pas pour le moment très élevés, mais ils pourraient être considérablement améliorés

- dès lors que la vérité de terrain serait mieux appréhendée et que l'on aurait une base d'apprentissage suffisante,
- en prenant davantage en compte le contexte et les interactions possibles entre le lecteur et la machine.

Liste des publications

[1] **P. Bilane**, S. Bres, K. Challita, H. Emptoz, "Indexation of Syriac manuscripts using directional features", *in Proceedings of the IEEE International Conference on Image Processing (ICIP 09)*, Cairo, Egypt, Nov. 2009, pp. 1841-1844.

[2] **P. Bilane,** S. Bres, K. Challita, H. Emptoz, "A segmentation free approach for indexing digitized Syriac manuscripts", *in Proceedings of the 17th European Signal Processing Conference (EUSIPCO 09)*, Glaskow, Scotland, Aug. 2009, pp. 303-307.

[3] **P. Bilane,** S. Bres, H. Emptoz, "Word Spotting dans des manuscrits Syriaques dégradés en utilisant des caractéristiques directionnelles", *in Proceedings of the 10ème Colloque International Francophone sur l'Ecrit et le Document (CIFED 08)*, Rouen, France, Oct. 2008, pp. 201-202.

[4] **P. Bilane,** S. Bres, H. Emptoz, "Local orientation extraction for word spotting in Syriac manuscripts", *in Proceedings of the 3rd International Conference on Image and Signal Processing (ICISP 08)*, Normandy, France, Jul. 2008, pp. 481-489.

[5] **P. Bilane**, S. Bres, H. Emptoz, "Robust directional features for word spotting in degraded Syriac manuscripts", *in Proceedings of the 6th International Workshop on Content-Based Multimedia Indexing (CBMI 08)*, London, UK, Jun. 2008, pp. 526-533.

[6] **P. Bilane,** E. Youssef, B. Eter, C. Sarraf, J. Constantin, "Simplified point to point correspondence of the Euclidean Distance for online handwriting recognition", *in Proceedings of the IEEE International Conference on Signal Processing and Communications (ICSPC 07)*, Dubai, UAE, Nov. 2007, pp. 1011- 1014.

Bibliographie

[ALL02] B. Allier, and H. Emptoz, « Degraded character image restoration using active contours: A first approach ». *In Proceedings of the 2nd ACM Symposium on Document Engineering*, McLean Virginia, United States of America, p. 142-148, Nov. 2002.

[ALL06] B. Allier, N. Bali, and H. Emptoz, « Automatic accurate broken character restoration for patrimonial documents ». *International Journal on Document Analysis and Recognition (IJDAR'06)*, Springer-Verlag, Berlin Heidelberg, Vol. 8, No. 4, p. 246-261, Sep. 2006.

[BAL06] A. Balasubramanian, M. Meshesha, C. V. Jawahar, « Retrieval from document image collections ». *In Proceedings of the 7th International Workshop on Document Analysis Systems (DAS'06)*, Nelson, New Zealand, p. 1-12, Feb. 2006.

[CAR98] S. Carlsson. « Order structure, correspondence and shape based categories ». *In Lecture Notes in Computer Science*, Springer Verlag, Berlin Heidelberg, Vol. 1681, p. 58–71, 1998.

[CIL05] R. Cilibrasi, L. Van Iersel, S. Kelk, and J. Tromp, « On the complexity of several haplotyping problems ». *In Proceedings of the 5th International Workshop on Algorithms in Bioinformatics (WABI'05)*, Mallorca, Spain, p. 128-139, Oct. 2005.

[CLO03] W. F. Clocksin, and P.P.J. Fernando, « Towards automatic transcription of Syriac handwriting ». *In IEEE Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP'03)*, Mantova, Italy, p. 664-669, Sept. 2003.

[CLO04] W. F. Clocksin, « Handwritten Syriac character recognition using order structure invariance ». *In IEEE Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, Cambridge, United Kingdom, p. 562-565, Aug. 2004.

[DEB00] DEBORA Projet européen n° LB 5608A, coordination Prof. R. Bouche, Edition ENSSIB Lyon, juin 2000, 192 p.

[DRI07] F. Drira, « Contribution à la restauration des images de documents anciens », 221 p. Thèse : Informatique : INSA De Lyon : 2007.

[DUD72] R. O. Duda, and P. E. Hart, « Use of the Hough transformation to detect lines and curves in pictures ». *Published in the ACM Communications*, Vol. 15, p. 11-15, Jan. 1972.

[DUV07] R. Duval, « La littérature syriaque », 3ème édition. Paris : Librairie Victor Lecoffre, 1907, XVIII-430 p.

[EGL07] V. Eglin, S. Bres, and C. Rivero, « Hermite and Gabor transforms for noise reduction and handwriting classification in ancient manuscripts ». *In International Journal on Document Analysis and Recognition (IJDAR'07)*, Springer-Verlag, Berlin Heidelberg, Vol. 9, No. 2-4, p. 101-122, Apr. 2007.

[EMP03] H. Emptoz, F. Lebourgeois, V. Eglin, and Y. Leydier. « La reconnaissance dans les images numérisées : OCR et transcription, reconnaissance des structures fonctionnelles et des méta-données ».*In La numérisation des textes et des images*, Editions du CNRS, p. 105-130, Lille, Jan. 2003

[EYN09] L. Eynard, « Contribution à la numérisation de documents imprimés du XVIIIème siècle : Application au cas de la Gazette de Leyde ».Thèse : Informatique : INSA de Lyon : 2009.

[FAU08] K. Khurshid, C. Faure, and N. Vincent, « Feature bases Word Spotting in ancient printed documents », *In Proceedings of the 8th International Workshop on Pattern Recognition in Information Systems (PRIS'08)*, Barcelona, Spain, p. 193-198, Jun. 2008.

[FEN05] S. L. Feng, and R. Manmatha, « Classification models for historical manuscript recognition ». *In IEEE Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05)*, Seoul, Korea, p. 528-532, Aug. 2005.

[GAC09] D. Gaceb. « Contributions au tri automatique de documents et de courriers d'entreprise ». Thèse : Informatique : INSA de Lyon : 2009.

[GAR09]. J. Gardes, « Le document numerique: la complexite des formes et les formes de la complexite ». Thèse : Informatique : INSA de Lyon : 2009.

[GAT04] B. Gatos, I. Pratikakis, and S. J. Perantonis, « An adaptative binarization technique for low quality historical documents ». *In Proceedings of the 6th International Workshop on Document Analysis Systems (DAS'04)*, Florence, Italy, p. 102-113, Sep. 2004.

[GAT06] B. Gatos, I. Pratikakis, and S. J. Perantonis, « Adaptive degraded document image binarization ». *In Pattern Recognition, The Journal of the Pattern Recognition Society*, Elsevier Science, Vol. 39, No. 3, p. 317-327, Mar. 2006.

[GOV04] V. Govindaraju, and H. Xue, « Fast handwriting recognition for indexing historical documents ». *In IEEE Proceedings of the 1st International Workshop on Document Image Analysis for Libraries (DIAL'04)*, Palo Alto, California, United States of America, p. 314-320, Jan. 2004.

[HOC10] S. Hoquet, and J. Y. Ramel, « Analyse de formes pour la transcription de textes imprimés anciens», *in Proceedings of the 12ème Colloque International Francophone sur l'Ecrit et le Document (CIFED 10)*, Sousse, Tunisie, Mar. 2010, (à paraître).

[JOU05] N. Journet, R. Mullot, J. Ramel, and V. Eglin. « Ancient Printed Documents Indexation: A New Approach ». *In Proceedings of the 3rd International Conference on Advances in Pattern Recognition (ICAPR'05)*, Springer-Verlag, Berlin Heidelberg, p. 580-589, Aug. 2005.

[KHU08] K. Khurshid, C. Faure, and N. Vincent. « Recherche de mots dans les images de documents par appariements de caractères », *in Proceedings of the 10ème Colloque International Francophone sur l'Ecrit et le Document (CIFED 08)*, Rouen, France, Oct. 2008, p. 91-96.

[KHU09] K. Khurshid, C. Faure, and N. Vincent, « A novel approach for Word Spotting using merge-split edit distance ». *In Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns (CAIP'09)*, Munster, Germany, p. 213-220, Sep. 2009.

[LAV04] V. Lavrenko, T. M. Rath, and R. Manmatha, « Holistic word recognition for handwritten historical documents ». *In IEEE Proceedings of the 1st International Workshop on Document Image Analysis for Libraries (DIAL'04)*, Palo Alto, California, United States of America, p. 278-287, Jan. 2004.

[LER64] J. Leroy, « Les manuscrits syriaques à peintures conservés dans les bibliothèques d'Europe et d'Orient. Contribution à l'étude de l'iconographie des églises de langue syriaque », 2 Volumes. Paris: Librairie orientaliste Paul Geuthner, Bibliothèque archéologique et historique 77, 1964.

[LEY05] Y. Leydier, F. Lebourgeois, and H. Emptoz, « Omnilingual segmentation-free word spotting for ancient manuscripts indexation ». *In IEEE Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05)*, Seoul, Korea, p. 533-537, Aug. 2005.

[LEY06] Y. Leydier, « Numérisation et exploration des manuscrits médiévaux », 202 p. Thèse : Informatique : INSA De Lyon : 2006.

[LEY07] Y. Leydier, F. Lebourgeois, and H. Emptoz, « Text search for medieval manuscript images ». *In Pattern Recognition, The Journal of the Pattern Recognition Society*, Elsevier Science, Vol. 40, No. 12, p. 3552-3567, Dec. 2007.

[LEY09] Y. Leydier, A. Ouji, F Lebourgeois, and H. Emptoz, « Towards an omnilingual word retrieval system for ancient manuscripts ». *In Pattern Recognition, The Journal of Pattern Recognition*, Elsevier Science, Vol. 42, No. 9, p. 2089-2105, Jan. 2009.

[MAN03] T. M. Rath, and R. Manmatha, « Features for word spotting in historical manuscripts ». *In IEEE Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR'03)*, Edinburgh, Scotland, p. 218-222, Aug. 2003.

[MAN05] R. Manmatha, and J. L. Rothfeder, « A scale space approach for automatically segmenting words from historical handwritten documents ». *In IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI'05)*, Vol. 27, No. 8, p. 1212-1225, Aug. 2005.

[RAT03] T. M. Rath, and R. Manmatha, « Word image matching using Dynamic Time Warping ». *In IEEE Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'03)*, Madison, Wisconsin, United States of America, p. 521-527, Jun. 2003.

[RAT07] T. M. Rath, and R. Manmatha, « Word spotting for historical documents ». *In International Journal on Document Analysis and Recognition (IJDAR'07)*, Springer-Verlag, Berlin Heidelberg, Vol. 9, No. 2-4, p. 139-152, Apr. 2007.

[ROU04] L. Rousseau, E. Anquetil, and J. Camillerapp, « Reconstitution du parcours du trace manuscrit hors-ligne de caractères isoles ». *In 8ème Colloque International Francophone sur l''ecrit et le Document, (CIFED'04)*, La Rochelle, France, p. 123–127, Jun. 2004.

[ROU07] L. Rousseau, « Reconnaissance d'écriture manuscrite hors-ligne par reconstruction de l'ordre du tracé en vue de l'indexation de document d'archives », 172 p. Thèse : Informatique : INSA de Rennes : 2007.

[SAR08] T. Sari, and A. Kefali, « A search engine for Arabic documents », *in Proceedings of the 10ème Colloque International Francophone sur l'Ecrit et le Document (CIFED 08)*, Rouen, France, Oct. 2008, p. 97-102.

[SID08] I. Siddiqi, and N. Vincent, « Descripteurs locaux de forme pour la reconnaissance de scripteur», *in Proceedings of the 10ème Colloque International Francophone sur l'Ecrit et le Document (CIFED 08)*, Rouen, France, Oct. 2008, p. 157-162.

[TER05] K. Terasawa, T. Nagasaki, and T. Kawashima, « Eigenspace method for text retrieval in historical document images ». *In IEEE Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR'05)*, Seoul, Korea, p. 437-441, Aug. 2005.

[TER07] K. Terasawa, and Y. Tanaka, « Locality sensitive pseudo-code for document images ». *In IEEE Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR'07)*, Curitiba, Brazil, p. 73-77, Sep. 2007.

[THO00] P. D. Thouin, and C. I. Chang, « A method for restoration of low-resolution document images ». *In International Journal on Document Analysis and Recognition (IJDAR'00)*, Springer-Verlag, Berlin Heidelberg, Vol. 2, No. 4, p. 200-210, Jun. 2000.

[VIN04] A. Vinciarelli, S. Bengio, and H. Bunke, « Offline recognition of unconstrained handwritten texts using hmms and statistical language models ». *In IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI'04)*, Vol. 26, No. 6, p. 709-720, Jun. 2004.

[WEI01] H. Weihua, C. L. Tan, S. Y. Sung, and Y. Xu, « Word shape recognition for image-based document retrieval ». *In IEEE Proceedings of the International Conference on Image Processing (ICIP'01)*, Thessaloniki, Greece, p. 1114-1117, Oct. 2001.

[ZHE01] Q. J. Zheng, and T. Kanungo, « Morphological degradation models and their use in document image restoration ». *In IEEE Proceedings of the International Conference on Image Processing (ICIP'01)*, Thessaloniki, Greece, p. 193-196, Oct. 2001.