Schriften des Instituts für Dokumentologie und Editorik — Band 2

Kodikologie und Paläographie im digitalen Zeitalter

Codicology and Palaeography in the Digital Age

herausgegeben von | edited by

Malte Rehbein, Patrick Sahle, Torsten Schaßan

unter Mitarbeit von | in collaboration with

Bernhard Assmann, Franz Fischer, Christiane Fritze

2009

BoD, Norderstedt

Bibliografische Information der Deutschen Nationalbibliothek:
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über **http://dnb.d-nb.de/** abrufbar.

Leicht veränderte Fassung für die digitale Publikation (siehe Vorwort).

Slightly modified version to be published digitally (see preface).

Publication réalisée avec le soutien d'Apices
Association Paléographique Internationale
Culture – Écriture – Société
Dotation J.M.M. Hermans.
http://www.palaeographia.org/apices/

Herstellung und Verlag: Books on Demand GmbH, Norderstedt
ISBN: 978-3-8370-9842-6
Einbandgestaltung: Katharina Weber
Satz: XeTeX und Bernhard Assmann

Approche informatique du document manuscrit*

Gilbert Tomasi, Roland Tomasi

Résumé

Les technologies mises en œuvre avec le logiciel BIT-Alpha, sont la base d'un outil informatique d'aide à l'analyse de l'écriture manuscrite naturelle, pour transcription et identification, pour une utilisation en paléographie. Dans l'image numérique, la capture du contenu du document en permet une analyse, puis une interprétation et une valorisation. La binarisation, la capture des lignes et la segmentation de l'image sont exposées et commentées. L'identification des mots, puis des lettres permet une première reconnaissance de l'écriture, basée sur l'analyse du graphisme. La transcription du texte s'appuie en complément sur des considérations linguistiques. Les critères d'analyse du graphisme permettent aussi une aide à l'identification du scribe. Une idée de mesure et de normalisation de la différentiation entre écritures est envisagée. Les éléments graphiques peuvent être édités. En exemple, le traitement d'un texte manuscrit est détaillé.

Zusammenfassung

Die in der Software BIT-Alpha eingesetzten Technologien sind die Grundlage der computergestützten Analyse, Identifizierung und Transkription von Handschriften sowie deren Interpretation und Bewertung in der paläographischen Forschung. Der vorliegende Aufsatz stellt die Verfahren der Binarisierung, der Zeilenerkennung und der Bildsegmentierung vor. Bei der graphischen Analyse des Schriftbildes, die eine erste Handschriftenerkennung erlaubt, werden zunächst die Wörter identifiziert und anschließend deren einzelnen Zeichen. Die Transkription des Textes stützt sich zusätzlich auf linguistische Methoden. Die zur Analyse des Schriftbildes herangezogenen Kriterien geben auch eine Hilfe zur Identifizierung des Schreibers. Die Erkennung eines mittelalterlichen handschriftlichen Textes wird im Detail an einem Beispiel dargestellt. Fernziel ist es, Messungen von Schriftzügen und eine normenbasierte Unterscheidung von Schrifttypen zu ermöglichen.

Abstract

The technologies used by the software BIT-Alpha are the basis as well for computer-aided analysis, identification, and transcription of writings as for their interpretation

* Nous désirons exprimer nos remerciements à Monsieur Torsten Schaßan de la Herzog August Bibliothek de Wolfenbüttel pour son soutien et conseil pour la réalisation de cet exposé.

Kodikologie und Paläographie im Digitalen Zeitalter – Codicology and Palaeography in the Digital Age. Hrsg. Malte Rehbein, Patrick Sahle und Torsten Schaßan, unter Mitarbeit von Bernhard Assmann, Franz Fischer und Christiane Fritze. Schriften des Instituts für Dokumentologie und Editorik 2. Norderstedt : Books on Demand, 2009. 197–218.

and evaluation for palaeographical research. The present article presents methods for binarisation, line detection and image segmentation. During the graphical analysis of script which is used as initial approach to script recognition words will be recognised first and after them single characters. The transcription of text then has to be assisted by linguistic methods. The criteria which are drawn on for the analysis of script will help to identify individual scribes. The recognition of a medieval written text is presented in detail. Ideas of how differences between writing styles can be measured and normalized will present future prospects.

1 Introduction

La transcription informatisée de l'image numérique d'un texte en un texte utilisable informatiquement est réalisée communément par des logiciels dénommés OCR pour Optical Characters Recognition. S'il existe aujourd'hui des solutions OCR pour certains documents imprimés, les documents manuscrits restent globalement informatiquement inaccessibles, illisibles.

Sur des documents imprimés récents, la transcription informatisée d'un texte donne, avec l'utilisation de logiciels OCR usuels, des résultats utilisables, en particulier pour une recherche contextuelle. La transcription de documents imprimés anciens, avant 1850, constitue une difficulté demandant l'utilisation d'outils informatiques spécialisés, comme le logiciel BIT-Alpha, en mesure de lire tout document imprimé, indépendamment de l'époque et la langue. BIT-Alpha peut apprendre la fonte de caractères et mêmes les mots directement du texte à traiter. Ainsi BIT-Alpha peut transcrire par exemple des incunables, dont la fonte fut la copie de l'écriture manuscrite des livres anciens. Les technologies informatiques mises en œuvre avec BIT-Alpha, peuvent être la base d'un outil informatique permettant l'analyse de l'écriture manuscrite et constituer une aide à la retranscription de l'écriture manuscrite naturelle. Un logiciel OCR spécialisé sur l'écriture manuscrite est aussi appelé ICR pour Intelligent Characters Recognition, car il est supposé reconnaître des caractères générés par l'intelligence et non par une machine.

C'est cette approche informatique du document manuscrit que nous voudrions exposer ici en présentant notre développement, le logiciel BIT-Alpha_ICR, comme outil d'analyse de l'écriture et d'aide à la retranscription, précisément dans la perspective d'une utilisation en paléographie. Dans l'image numérique, il s'agit en premier de capturer le contenu du document, pour en faire ensuite une analyse, puis une interprétation et une valorisation. La lecture informatisée d'un document suppose donc primairement la capture du contenu de ce document.

Au départ, il s'agit de rendre l'image d'archive, souvent très lourde car en couleur et de haute définition, utilisable informatiquement pour un traitement ICR. Cette première transcription, appelée binarisation, crée une image binaire et voudrait éliminer le bruit

de fond du scanner, les taches et colorisation rendant toute lecture impossible, et cela si possible en gardant le détail de l'écriture, ponctuation et point sur les i ! C'est loin d'être une problématique simple et elle est souvent négligée, alors qu'il s'agit de la base de toute analyse et retranscription informatisée.

Ensuite une segmentation de l'image est nécessaire, pour distinguer dans l'image du document, des régions contenant un texte et celles contenant un élément graphique, afin de les découper pour les traiter séparément et différemment. La segmentation va de paire avec le redressement de l'image pour placer les lignes du texte le plus horizontalement possible.

Une des difficultés majeure du traitement de l'écriture manuscrite constitue en la capture des lignes du texte, celles-ci n'étant naturellement pas droites ou alors soulignées d'un trait de ligne, gênant la reconnaissance et de plus souvent les lignes sont imbriquées, les lettres se chevauchant et se touchant. Une fois le texte isolé et les lignes déterminées, suit l'analyse du contenu en vue d'une transcription. Il s'agit pour chaque ligne de texte, d'en capturer les mots, puis de capturer les lettres et de proposer une reconnaissance de l'écriture.

Nous procéderons en reconnaissance ICR par une double approche, partant en premier de la capture et de l'analyse globale du mot, capturé en tant qu'unité graphique à analyser dans son tout, en recherchant une corrélation avec un mot d'un vocable connu dont le graphisme a été appris au préalable, un peu comme la méthode de lecture dite « globale ».

Ensuite, cette identification globale du mot sera confrontée à une analyse des lettres constitutives du même mot. La capture individuelle des lettres suppose la séparation des caractères entre eux, ce qui constitue une vraie difficulté dans le cas de l'écriture liée naturelle. Avec la capture et l'analyse des caractères, pourra être envisagée leur identification en vue de proposer une transcription. Celle-ci est basée sur divers critères corrélés à une analyse de forme. Une fois les éléments constitutifs du texte analysés, il sera procédé à leur interprétation pour valorisation. Pour être pertinente, la transcription du texte devra s'appuyer sur des considérations linguistiques en complément aux informations de source purement graphique.

Les critères d'analyse des lignes du texte, des mots et des caractères permettront aussi une aide à l'identification du scribe. Une idée de mesure et de normalisation de la différentiation d'écritures est envisagée. Enfin les éléments graphiques peuvent être édités et constituer une base de données permettant aussi d'identifier le scribe ou servir une étude spécifique.

En résumé BIT-Alpha_ICR constitue un outil permettant l'analyse de l'écriture manuscrite et sa retranscription avec l'aide d'un computeur. Une quantification normée de la différentiation d'écritures serait d'un grand intérêt pour la paléographie. En perspective, une application vers un traitement ICR automatisé peut être envisagée ainsi que la lecture ICR de textes en écriture liée comme l'arabe ou l'hindi.

2 Capture du contenu

Préalablement, il convient de lever un éventuel malentendu : nous voulons traiter informatiquement l'écriture manuscrite de documents anciens ou dits « existants », c'est-à-dire partant de l'image numérique, généralement scannée, d'un document existant sous forme papier ou équivalent. Il s'agit donc d'un traitement d'image et de l'analyse du tracé d'une écriture « existante ». Cette démarche ne doit pas être confondue avec la reconnaissance de l'écriture manuscrite « vivante » obtenue par capture du mouvement au moment même de l'acte d'écrire et de la génération de l'écriture, par exemple sur un capteur graphique, comme une tablette graphique ou un écran tactile. Dans ce cas, sont aussi disponibles pour l'analyse informatique des informations dynamiques, comme la vitesse du tracé, l'ordre d'exécution des différents traits, la force d'appuis, etc. Autant d'éléments dynamiques permettant une reconnaissance finalement plus aisée et communément utilisée aujourd'hui sur certains appareils portables. Dans ce cas d'analyse d'informations dynamiques, il est usuel de parler de reconnaissance « on-line » en opposition à l'analyse d'image d'un texte scanné, appelée reconnaissance « off-line » et qui constitue ici notre propos.

En partant d'une image numérique d'archive, souvent de haute définition, typiquement de 300 à 600 dpi couleurs, il s'agit en premier d'obtenir une image utilisable pour un traitement de reconnaissance de caractères. Au final, il faut obtenir une image binaire et d'une définition de 300 à 400 dpi, une définition moindre rendant la distinction impossible entre des caractères proches et une image de trop haute définition faisant de chaque caractère une pièce unique.

2.1 La binarisation

Le procédé de binarisation constitue une opération informatique de préparation au traitement de reconnaissance et qui n'est pas toujours intégrée dans un logiciel OCR ou ICR. Elle constitue pourtant une étape majeure dans l'analyse et demande grand soin, car une mauvaise binarisation peut détruire beaucoup d'informations contenues dans le facsimilé scanné de l'original. Par contre la binarisation doit rendre le document utilisable pour une reconnaissance et doit pour cela effectuer des opérations délicates : supprimer le bruit de fond du scanner, tenir compte des variations de contraste et de luminosité sur la page scannée, éliminer les taches gênantes, nettoyer les bords et zone de marges sans texte, obtenir le plus de détails possibles dans la zone texte, c'est-à-dire éliminer la face arrière du document visible par transparence ou parce que l'encre à diffuser dans le papier, mais garder la ponctuation et les points sur les i.

A côté d'autres outils de binarisation plus usuels, comme un seuillage de luminosité ou de couleur, mis aussi en œuvre dans BIT-Alpha mais avec la particularité d'une localisation par région de texte, c'est-à-dire différent en haut et en bas de la même page, le

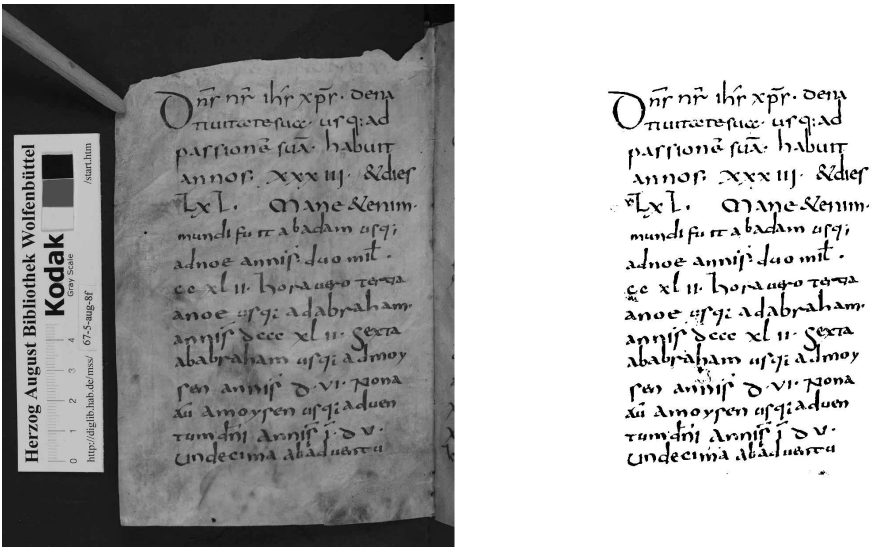


Fig. 1. Un document numérique et sa binarisation par BIT-Alpha. Wolfenbüttel, Herzog August Bibliothek, Cod. Guelf. 67.5 Aug. 8°, fol. 2v

logiciel BIT-Alpha propose aussi une binarisation utilisant l'algorithme de Niblack, modifié par nos soins pour en corriger certaines imperfections, afin de satisfaire au mieux les impératifs contradictoires de la binarisation. Cet algorithme reconnu comme un des mieux adaptés pour ce faire, procède d'une analyse dit « dynamique » de contraste local et cela individuellement autour de chaque caractère. De plus une analyse de la répartition gaussienne des contrastes, permet de distinguer les zones textes, des zones de marge, afin de les traiter avec des seuillages différents. Ce procédé génère une image binaire de bonne qualité et c'est le type de binarisation que nous préconiserons pour le traitement de documents manuscrits. On y ajoutera éventuellement un prétraitement d'image permettant d'éliminer une couleur correspondant à une tache gênante, comme la coloration intempestive en rouge de certaines majuscules, ce qui les rendrait illisibles informatiquement.

La figure 1 montre un exemple de binarisation avec l'algorithme de Niblack modifié, tel que mis en œuvre dans BIT-Alpha.

2.2 La segmentation

Après avoir nettoyé et réduit l'image d'archive en une image binaire plus légère et plus facilement traitable informatiquement, il est nécessaire de distinguer dans cette image les zones contenant un texte de celle contenant une image ou des éléments graphiques. Cette séparation de l'image en différentes régions est appelée communément segmentation.

Dans le cas d'un outil d'aide à la transcription d'un texte manuscrit où l'intervention humaine est prévue pour finaliser la transcription, cette étape peut paraître secondaire dans la mesure où l'opérateur peut lui-même isoler la zone texte l'intéressant. Malgré tout, elle est d'intérêt en permettant d'isoler et par là de capturer, les différents éléments graphiques. Dans le futur, un procédé entièrement automatique ne pourrait se concevoir sans segmentation, telle qu'elle est mise en œuvre aujourd'hui pour des livres anciens ou des journaux par BIT-Alpha.

La segmentation va souvent de paire avec l'orientation de la page, rotation dénommée aussi redressement, afin d'obtenir du mieux possible un texte aux lignes plus ou moins horizontales, ce qui simplifie grandement la capture des lignes et la reconnaissance des caractères. Concernant un texte manuscrit, le redressement sera de toute autre difficulté si nous avons affaire à un texte écrit bien droit, avec éventuellement des lignes guide que suit l'écriture manuscrite, comme c'est souvent le cas, ou un texte aux lignes courbes et écrit « en vrac ».

Pour le redressement d'un texte manuscrit, nous préconiserons un procédé suivant les travaux de Frank Le Bourgeois de l'I.N.S.A. Lyon et partant d'une analyse différentielle de la variance bidirectionnelle dans l'image, telle que décrite dans la publication citée en bibliographie.

2.3 La capture des lignes

La capture des lignes d'un texte manuscrit est la première difficulté majeure. En effet, même si les lignes de texte suivent un trait du document, les caractères sont souvent très imbriqués et se touchent d'une ligne à l'autre. Parfois un document manuscrit présente des lignes courbes, ce qui complique le traitement informatique. Un procédé usuel donnant de bons résultats sur des textes imprimés aux lignes clairement séparées et rectilignes, ne conviendra pas aux variations de l'écriture manuscrite. BIT-Alpha met en œuvre pour la capture des lignes les travaux de recherche du Laboratoire CNRS L.I.R.I.S. de I.N.S.A. Lyon autour du Professeur Hubert Emptoz (Le Bourgeois et al. 2004) et concernant aussi des applications en paléographie (Emptoz).

Pour la capture des lignes du texte, il est procédé à une analyse différentielle de la variation unidimensionnelle dans l'image et permettant d'obtenir les zones constitutives des lignes du texte, comme visible dans l'image figure 3. Cette méthode mathématique

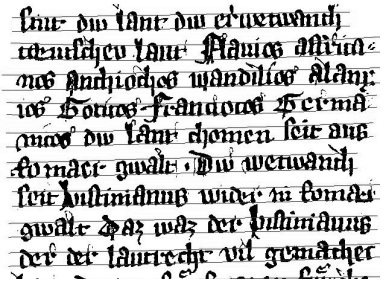


Fig. 2. Capture de lignes dans un texte manuscrit. Wolfenbüttel, Herzog August Bibliothek, Cod. Guelf. 15.2 Aug. 2°, fol. 1v

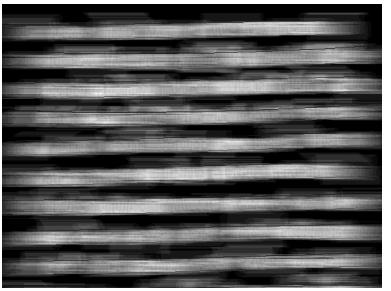


Fig. 3. Analyse différentielle de l'image du texte manuscrit de la figure 2.

est la base du procédé de reconnaissance de lignes de BIT-Alpha, laquelle a été complétée et adaptée pour tenir compte de différentes difficultés inhérentes à la méthode, comme la capture des grandes lettres d'un titre ou de symboles isolés. Cette méthode a le gros avantage de bien reproduire la courbure des lignes de texte et fonctionne indépendamment de l'identification de caractères. Après avoir déterminé la zone constitutive d'une ligne, encore faut il en déduire, à partir d'une suite de points, une fonction mathématique donnant la ligne elle-même. Ce procédé que nous jugeons très performant constitue la quatrième génération de procédés de capture de lignes mis en œuvre dans BIT-Alpha et se montre supérieur aux autres méthodes précédemment utilisées dans BIT-Alpha et partant soit des caractères eux-mêmes, soit d'une analyse de périodicité ou de projection d'ombres. Ces derniers procédés ne pouvant convenir à l'écriture manuscrite qui n'a en général ni caractères séparés et homogènes ni périodicité établie.

Ce procédé d'analyse différentielle a l'avantage, non seulement de permettre de capturer les lignes, mais aussi d'obtenir la base-ligne, et la top-ligne comme visible dans l'image figure 2. Cette information, de première importance pour la reconnaissance individuelle des caractères, donne aussi accès, avec la distance moyenne entre ces lignes, à une grandeur de référence pour la taille de l'écriture, premier élément pouvant caractériser une écriture manuscrite déterminée.

3 L'analyse du contenu

Une fois le texte isolé et les lignes du texte déterminées, il faut procéder à l'analyse des mots et des caractères pour obtenir une identification permettant l'accès au texte, au

contenu. Cette identification servira de base à une première proposition de transcription du texte, basée uniquement sur l'analyse de l'image.

3.1 Analyse des mots

Une fois les lignes du texte déterminées, nous capturerons les mots de chaque ligne pour une première analyse d'identification. Alors que les OCR procèdent généralement pour reconstituer les mots d'une approche partant de la capture individuelle des lettres constitutives, séparées entre elles et réputées identiques dans les imprimés, un ICR est confronté à des lettres souvent liées entre elles et présentant une forte variance graphique. De ce fait, une approche globale considérant en premier le mot entier dans son ensemble s'impose au prime abord.

La capture des mots sera facilitée dans un texte à l'écriture liée ou chaque mot constitue naturellement un ensemble graphique séparé les uns des autres. La chose sera plus délicate dans un texte ou les lettres manuscrites ne seront pas toutes liées et avec une distance entre elles changeante. Une analyse de la distance entre les caractères devrait permettre une première capture des mots. Ici BIT-Alpha dispose d'outil d'analyse assez fin mis au point pour lire des textes imprimés de la Renaissance ou des journaux dans lesquels la distance entre lettres et mots est souvent très fantaisiste et n'est pas plus homogène que dans un texte manuscrit. Cette analyse de la distance entre lettres a lieu individuellement pour chaque ligne.

La figure 4 montre un exemple de capture et d'identification après l'apprentissage d'un mot entier dans BIT-Alpha. Il s'agit du mot « dem » qui est en premier appris en tant qu'un tout. Dans la figure 5, un autre mot « dem » est ainsi reconnu dans une autre partie du texte, affichant un taux de reconaissance moindre.

C'est une particularité de BIT-Alpha de pouvoir, dans une première phase d'apprentissage, enregistrer l'identification donnée par l'opérateur et apprendre, directement du document à traiter, le graphisme des éléments à reconnaître. BIT-Alpha est un OCR du type « self-learning » ou « adaptatif » procédant par une phase d'apprentissage pour reconnaître ensuite les symboles appris au préalable. Ceci est valable pour les mots dans leur globalité graphique et pour les symboles individuellement.

3.2 Analyse des caractères

Pour obtenir une reconnaissance fiable, l'analyse des mots dans leur tout ne suffit pas, il faut compléter par une analyse des caractères constitutifs du mot afin de corréler ces deux informations. Cette analyse suppose de pouvoir capturer les caractères individuellement, c'est-à-dire de pouvoir les considérer isolément. Ceci suppose qu'ils

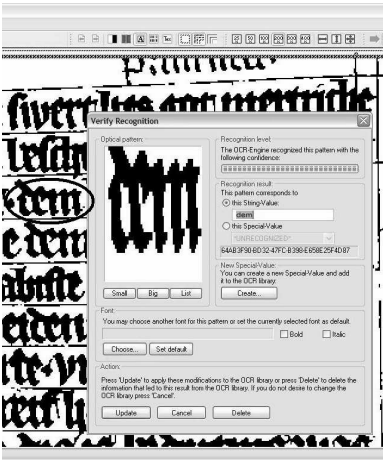


Fig. 4. Capture et apprentissage du mot « dem ».

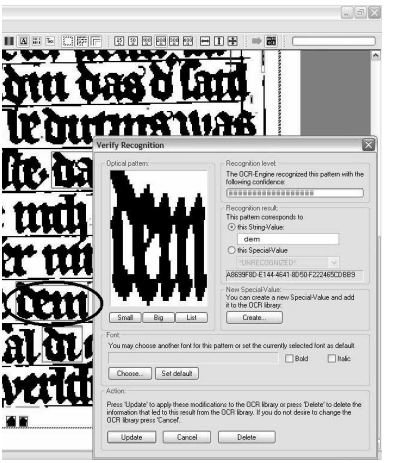


Fig. 5. Reconnaissance d'un autre mot « dem » dans une autre partie du texte.

soient séparés les uns des autres ou que nous disposions d'un procédé pour séparer les caractères collés. Ce procédé que nous avons nommé « disjonction » est intégré dans BIT-Alpha.

Ici aussi, après plusieurs premières générations d'algorithme ou la disjonction avait lieu suivant des critères géométriques, le procédé de disjonction actuellement utilisé dans BIT-Alpha part d'une approche « d'intelligence artificielle » en séparant les caractères lorsque le logiciel ICR reconnaît les composants à séparer. Cette démarche, plus complexe à mettre en œuvre que son simple énoncé, donne dans la pratique d'excellents résultats, confirmés en production de masse avec BIT-Alpha. Nous sommes de l'avis que ce soit le procédé unique ouvrant l'accès à la reconnaissance ICR devant traiter l'écriture naturelle liée.

Dans la reconnaissance ICR d'un fragment de texte, comme un code postal et sa ville ou un nombre en toutes lettres, il est d'usage de procédé d'une démarche atomisant le mot en une suite de traits verticaux pour reconstituer ensuite les « m », « u », « n » et « a » en quelque sorte en comptant le nombre de pieds. Cette approche que nous avons aussi testée, nous semble inférieure à une reconnaissance du caractère dans son intégrité, basée sur la disjonction après reconnaissance ICR du caractère à isoler.

Notons au passage que si un ICR fonctionne sur un fragment de texte assez court et dont les éléments à reconnaître sont « attendus », c'est-à-dire connus et présents dans une liste de villes en référence ou des chiffres déjà lus, la difficulté est toute autre lorsqu'on se trouve confronté à la lecture d'une page d'écriture manuscrite dont le contenu

est inconnu. Concernant les documents manuscrits historiques, la problématique est donc sensiblement différente, dans la mesure où il s'agit de reconnaître plusieurs pages d'un texte au contenu inconnu, mais écrit par une seule et même personne, appelée ici scribe, et dont l'écriture est apprise préalablement et non, à la différence, d'un texte limité et dont le contenu est connu ou du moins attendu, mais par contre écrit par des scribes inconnus et variés, comme c'est le cas pour la reconnaissance ICR par exemple d'adresses postales ou de nombres sur des chèques.

Profitons de cette parenthèse pour souligner que nous n'avons pas la prétention d'avoir une solution automatique pour la lecture ICR de page manuscrite inconnue, un domaine aux difficultés toujours non résolues, mais nous voulons simplement ici présenter notre approche et l'état de notre technologie. Le logiciel BIT-Alpha_ICR pouvant constituer un outil d'aide à la transcription ou d'aide à l'identification du scribe d'un texte. BIT-Alpha offre la possibilité, dans une première phase d'apprentissage, par un simple clic, d'apprendre un élément graphique isolé, tel qu'un caractère.

Les images figures 6 et 7 montrent la capture d'une lettre « a » dans BIT-Alpha, l'apprentissage de cette lettre et sa reconnaissance dans plusieurs mots et après disjonction dans trois autres mots du texte. Dans la figure 6 la lettre « a » est apprise, c'est-à-dire que l'opérateur en a donné la transcription, et elle est donc reconnue à 100 %, ce qui est indiqué par la couleur bleue (voir aussi figure 12 pour les couleurs). Par exemple, le « a » de « ra » qui n'est pas lié, est alors aussi reconnu, sans avoir été appris, avec une haute fiabilité, ce qu'indique la couleur verte. La disjonction n'est pas activée. Les « a » de « ad » et « am » qui sont liés ne sont pas reconnus. La figure 7 montre, après activation de la disjonction, la séparation de la lettre « a » de « ad » et « am » et la reconnaissance individuelle de la lettre « a » séparée du reste du tracé graphique.

Ainsi, il sera possible dans un texte manuscrit où les caractères sont collés entre eux, d'obtenir par disjonction une capture individuelle des caractères, et leur identification afin de reconstituer le mot. Cette méthode de disjonction suppose donc la reconnaissance des symboles à séparer, donc aussi un apprentissage préalable des caractères constitutifs des mots. Dans une écriture où tous les caractères seraient liés, la mise en œuvre de la disjonction suppose alors que lors de la phase d'apprentissage, l'opérateur puisse définir à quel endroit du graphisme du mot il convient de séparer les caractères afin de pouvoir apprendre le symbole individuellement. Cette fonction de disjonction manuelle de lettres collées, spécifique à un ICR, sera implémentée dans BIT-Alpha_ICR.

D'une façon plus générale, nous devons ici insister sur la spécificité du logiciel « self-learning » BIT-Alpha qui part en premier d'une démarche d'apprentissage des éléments à reconnaître : BIT-Alpha ne va reconnaître que le type d'écriture qu'il a auparavant appris. Il est évident que son taux de fiabilité de lecture est alors supérieur a celui d'un OCR « omni-fonte » supposé connaître toutes les types d'écriture et risquant de les confondre toutes. Les différentes bibliothèques de bases de reconnaissance OCR contenant les « empreintes digitales » des différents symboles et graphismes appris par

Fig. 6. Lettre « a » apprise (au centre) avec la fenêtre de dialogue d'apprentissage ouverte, d'autres « a » non liées sont aussi reconnues.

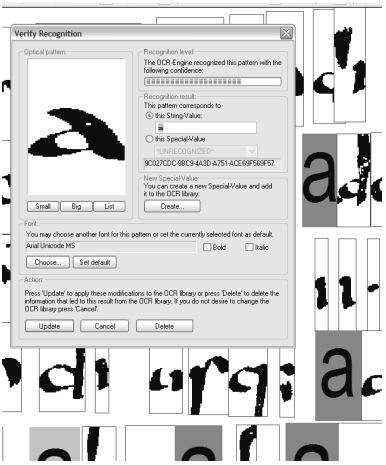


Fig. 7. Autres lettres « a » liées, reconnues une fois la disjonction activée.

BIT-Alpha, sont individuellement enregistrées. Cela permet de les inter-changer pour contrôle individuel ou de les mixer pour constituer une base plus générale. Il est ainsi possible d'identifier une écriture particulière en n'utilisant qu'une base OCR spécifique ne « connaissant » que l'écriture particulière d'un scribe connu afin de vérifier si le texte est bien de ce scribe. Ou bien, au contraire de rassembler en une seule base plus large toutes les écritures apprises afin de transcrire au mieux un texte inconnu.

BIT-Alpha peut apprendre et reconnaître aussi des ligatures comme le « ch » de l'exemple figure 8. Mais BIT-Alpha peut aussi apprendre et donc reconnaître toute suite des caractères convenant à l'analyse du texte.

BIT-Alpha dispose aussi de la possibilité, à l'inverse, de recoller des symboles disloqués, comme le « h » en bas de la figure 6. Exposer ici le détail de cette procédure de « séquenceur » serait fastidieux. Disons qu'elle part simplement du principe de séquences de symboles, par exemple qu'un « l » suivi d'une partie droite de « petit h » donne un « h ». Ces séquences sont individuellement programmables dans BIT-Alpha suivant les nécessités du texte et sont intégrées dans les bibliothèques ICR interchangeables.

Bien plus, la reconstitution des symboles à partir des éléments graphiques constitutifs pour établir une capture pertinente, est dans BIT-Alpha individuellement programmable. Il est ainsi possible avec BIT-Alpha d'obtenir par exemple dans un texte Fraktur

Fig. 8. Exemple d'apprentissage de ligature, ici « ch ».

imprimé ou pour des coursives de la Renaissance, la séparation de « e » ou « o » sous un « s long » tout en recollant des « n » ou « u » disloqués en petits traits verticaux. Les mêmes possibilités existent pour l'écriture manuscrite et le lecteur attentif remarquera que dans la figure 6 le « h », en bas à droite de page, et capturé disloqué alors que la même « h » est capturé intégralement dans la figure 11.

3.3 Analyse du texte

L'analyse d'image nous donne donc pour chaque mot deux identifications :

1. Suite à l'analyse globale des mots :
   La proposition d'un mot, en comparaison aux graphismes de mots entiers, auparavant appris en tant qu'entités géométriques et en mémoire dans une bibliothèque de bases ICR « d'empreintes numériques » de mots.
2. Suite à l'analyse des caractères constitutifs des mots :
   La proposition d'un mot reconstitué par la juxtaposition dans le bon ordre de ses caractères identifiés et reconnus par rapport à une bibliothèque de bases ICR « d'empreintes numériques » de caractères, aussi auparavant appris.

Dans la mesure où BIT-Alpha_ICR pourra de proposer le premier et le deuxième ou troisième choix, une corrélation de ces deux informations devrait permettre une première proposition de transcription du mot à lire informatiquement. A ce stade, tout ce que l'on a pu obtenir comme information pour la transcription du texte vient essentiellement de l'analyse du graphisme, du tracé du texte.

4 Valorisation du contenu

Pour valider la transcription, il faut maintenant faire intervenir des considérations basées sur le langage.

Une fois les lignes, les mots et les lettres capturés et analysés pour obtenir une proposition de mots pour la transcription du texte, celle-ci doit être affinée en corrélation au langage utilisé par l'auteur du texte. Cela suppose d'en connaître effectivement la langue et de disposer d'un lexique de mots ou d'un texte correspondant. Les abréviations pourront être transcrites avec le symbole Unicode correspondant au mieux au symbole original ou bien en les remplaçant par leur forme développée. Enfin BIT-Alpha propose plusieurs possibilités pour l'identification de la personne ayant écrit le texte, désignée ici par scribe.

4.1 Transcription du texte

En combinaison l'analyse du mot dans son ensemble et celle des lettres constitutives de ce mot, le logiciel propose un choix de mots pour la retranscription du texte.

Cette analyse du graphisme propose donc un mot comme résultat d'une analyse ICR purement basée sur le tracé de l'écriture. Ces considérations géométriques seront complétées par des aspects linguistiques de corrélation en référence à une liste de mots connus de la langue du texte à déchiffrer, afin de permettre de restreindre le choix et de valider la transcription proposée. Il sera aussi possible de recoller ainsi des lettres isolées en début ou en fin de mot.

Le premier résultat de lecture ICR va donc devoir être confronté à des considérations linguistiques pour le valider et permettre une identification et proposer une transcription du texte la plus correcte possible. La langue du texte analysée étant supposée connue, le mot proposé par l'ICR, ou les premier, second et troisième choix de mots proposés, vont être corrélés à un lexique de mots de la langue utilisée afin d'en valider le bon choix pour le mot transcrit. Dans BIT-Alpha_ICR nous mettons en œuvre l'évaluation de la distance de Levenshtein (Ringlstetter). Cette distance est une grandeur donnant la mesure du coût pour modifier un mot en un autre mot. BIT-Alpha_ICR est capable de gérer des bases de données considérables, comme par exemple un lexique de 500 000 mots latins.

Mais nous sommes d'avis que ces seules considérations de correction de textes ne suffisent pas dans le cas de la validation d'un mot lu par ICR. En effet il ne s'agit pas de corriger des fautes de frappe par exemple, ni des fautes d'orthographe, mais bien plus, des fautes de lecture ICR. Il nous semble donc nécessaire d'ajouter aux considérations de corrélation basique, une pondération basée sur la probabilité d'erreur de lecture ICR, c'est-à-dire la probabilité de confusion entre deux lettres proches. Par exemple en reconnaissance de l'écriture Fraktur ou en Roman de la Renaissance, le «ſ» long sera souvent confondu avec un «f». Concrètement par exemple un mot d'un texte français lu «ſut» devra plutôt être corrigé « fut » que « sur ». Cette pondération de la correction linguistique nous paraît essentielle en lecture ICR et nous l'avons implémentée dans BIT-Alpha en considérant aussi les travaux de Frank Le Bourgeois et Hubert Emptoz du laboratoire CNRS I.R.I.S. de l'INSA Lyon (Emptoz et al.). Ceci permet à l'opérateur de définir pour une bibliothèque de mots donnée, une matrice de coefficients de corrélation entre les lettres sujettes à confusion. Comme pour tout dans BIT-Alpha l'opérateur maîtrise tous les paramètres et décide lui-même de ce que doit effectuer ou pas le logiciel.

Notons ici pour mémoire que le « séquenceur » déjà expliqué au chapitre 2.3 permet d'implémenter des séquences typiques de la langue, donnons à titre d'exemple et simplement pour la compréhension du principe : « nnn » donne « mm ». C'est un instrument très puissant de correction d'erreurs de lecture et peut ainsi permettre une meilleure transcription du texte.

Dans la mesure où la langue du texte est connue et où un échantillon de texte (transcrit en « texte mode » utilisable informatiquement) de cette langue est disponible, une analyse de la codification du texte permettrait aussi de restreindre le choix des lettres convenant. Il s'agit ici d'une méthode utilisée communément en déchiffrage et prenant en considération la fréquence d'une lettre, ses appareillages avec d'autres lettres (syllabes usuelles) et aussi la statistique de répartition de sa position dans les mots, début, fin, milieu. Théoriquement ces seules considérations permettraient en partant de l'analyse de codification d'une seule page de texte de la même langue (en texte mode), d'identifier automatiquement les lettres d'un texte inconnu mais de cette même langue et donc d'en faire un apprentissage entièrement automatisé, sans intervention humaine.

Tous ces aspects peuvent être l'assise d'une correction, basée sur la langue, de la première lecture ICR basée sur l'analyse du graphisme, et aboutir aussi à une première proposition de texte dans le cadre d'une aide informatisée à la transcription.

Dans les textes manuscrits patrimoniaux, les auteurs utilisent très souvent des abréviations afin d'écrire plus vite. La retranscription du texte sous une forme lisible par un non spécialiste demande la transcription de ces abréviations. Pour pouvoir transcrire une abréviation, encore faut-il pouvoir la capturer dans le texte et l'identifier comme telle. Pour ce faire BIT-Alpha propose l'utilisation des caractères du code Unicode, puisse que BIT-Alpha édite le texte codé en Unicode, c'est-à-dire avec 20 bit par

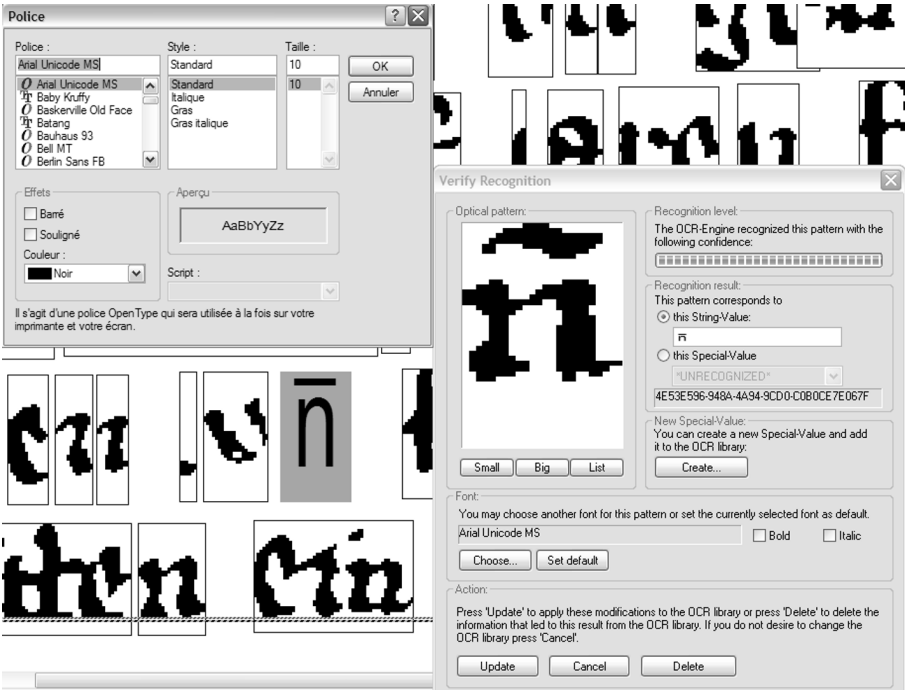


Fig. 9. Capture, apprentissage et reconnaissance d'une abréviation « n̄ ».

caractères. La figure 9 montre la capture d'une abréviation, retranscrite en symbole Unicode, (« n » avec barre supérieure d'abréviation).

Ici, nous ferions référence à MUFI « Medieval Unicode Font Initiative ». Ce groupement de chercheurs propose l'utilisation et la normalisation des symboles Unicode pour transcrire les abréviations de textes manuscrits anciens. BIT-Alpha est en mesure, soit d'utiliser ces symboles Unicode pour les abréviations conformément à l'original du texte, soit d'en donner la transcription développée afin de rendre le texte directement lisible. Concernant le développement des abréviations, on pourra se référer aux travaux du Centre d'Etudes Supérieures de la Renaissance de Tours autour de Madame Marie-Luce Demonet et aussi à la publication sur « Les abréviatures » de la Renaissance (Andrieux-Reix et al.). Bien sûr, avec BIT-Alpha, tout opérateur reste entièrement libre de ses choix concernant la transcription et l'interprétation des abréviations.

Enfin, si BIT-Alpha peut faire la transcription, la corrélation, entre l'image scannée d'un texte et son édition en un fichier « texte-mode » utilisable informatiquement, il est

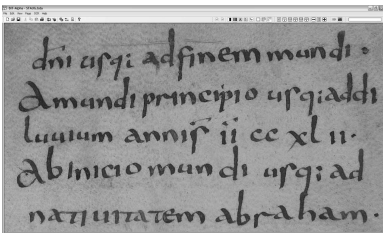


Fig. 10. Le document pris en charge dans BIT-Alpha, image 300 dpi couleur. Wolfenbüttel, Herzog August Bibliothek, Cod. Guelf. 67.5 Aug. 8°, fol. 3r

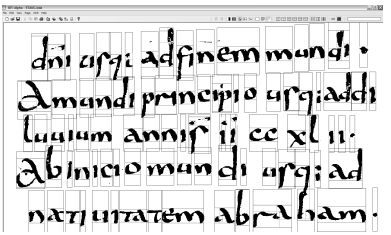


Fig. 11. L'image binarisée (Niblack modifié) et la capture des symboles.

aussi possible de faire la démarche inverse, c'est-à-dire d'apprendre, sur la base d'un texte connu, les caractères d'une écriture particulière, à partir du document scanné correspondant au texte connu. Cette démarche permet de valoriser des textes déjà transcrits par lecture humaine et saisis manuellement, en créant une bibliothèque de base de symboles ICR permettant de retranscrire ensuite d'autres textes scannés écrits de la même main. Si ce procédé d'apprentissage automatique de caractères est relativement facile à effectuer dans le cas d'une écriture où les caractères sont isolés ou bien formés, elle demandera l'intervention de l'opérateur dans le cas de caractères liés et imbriqués afin de bien définir l'endroit du tracé où la disjonction des symboles doit avoir lieu. Malgré tout, cela constitue une aide considérable à l'apprentissage ICR avec BIT-Alpha d'une écriture inconnue et par là de sa transcription.

4.2 Edition du texte

Pour l'édition du texte, BIT-Alpha propose plusieurs possibilités.

Le texte pourra être édité soit en fichier au format XML contenant les mots et leur position, et aussi conformément à la norme ALTO ou TEI, soit sous forme d'un fichier au format PDF présentant le facsimilé du document comme image de fond avec le texte sous-jacent et invisible pour permettre un affichage, appelé « highlighting », en surbrillance dans l'original du texte du mot trouvé lors d'une recherche contextuelle. Un document manuscrit entièrement traité avec BIT-Alpha va maintenant être donné en exemple avec les figures 10 à 17.

4.3 Identification du scribe

En paléographie, l'identification de l'auteur de l'écriture, respectivement de la personne l'ayant écrite de sa propre main, appelée ici scribe, est un sujet de grand intérêt.

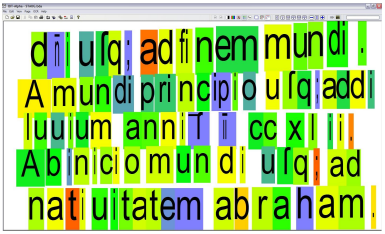


Fig. 12. La reconnaissance ICR des caractères, code Unicode (la couleur indique la fiabilité de lecture, du bleu « parfait » au rouge « douteux »).

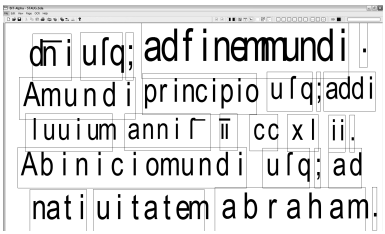


Fig. 13. Le texte transcrit (code Unicode) avec la formation des mots.

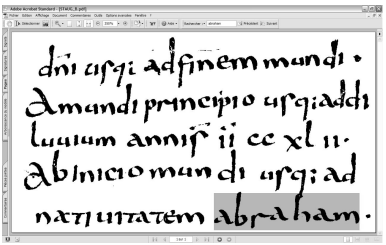


Fig. 14. Image d'un document PDF avec l'image binarisée en premier plan et le texte sous-jacent : mot recherché « abraham », affiché avec highlighting.

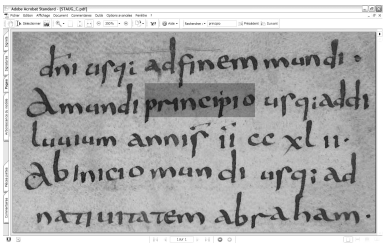


Fig. 15. Image d'un document PDF avec l'image couleur en premier plan et le texte sous-jacent : mot recherché « principio », affiché avec « highlighting ».

En partant de l'analyse détaillée des aspects géométriques du texte, BIT-Alpha peut être un outil d'aide à l'identification du scribe. Pour ce procédé plusieurs grandeurs analysées par BIT-Alpha sont disponibles :

En premier, des grandeurs liées aux lignes :

- Epaisseur de ligne « e »: comme nous l'avons vue au chapitre 2, nous disposons de la distance entre base-ligne et top-ligne (voir figure 2).
- Hauteur de ligne : BIT-Alpha détermine en plus de la base-ligne et de la top-ligne, la ligne du haut des caractères appelée high-line et la ligne du bas appelée bottom-line. La distance entre bottom-line et high-line est aussi une caractéristique de l'écriture. Elle peut être exprimée relative à l'épaisseur de ligne « e ».
- Distance relative entre ligne : cette distance entre les lignes médianes à top- et base-ligne de lignes consécutives, sera exprimée relative par rapport à l'épaisseur

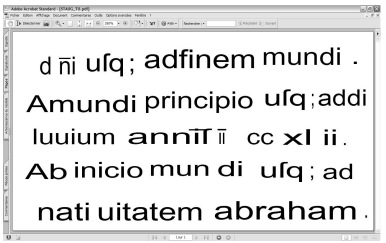


Fig. 16. Image d'un document PDF avec le texte transcrit seul (Unicode).

```
    </a:TextLine>
- <a:TextLine ID="ID_BIT_ea206734-46af-4214-866d-17e00225b7a5" HEIGHT="116.38" WIDTH="874.13"
      HPOS="25.48" VPOS="110.43">
  <a:String HEIGHT="101.94" WIDTH="284.58" VPOS="124.88" HPOS="25.48" CONTENT="Amundi" />
  <a:SP WIDTH="301.57" VPOS="120.2" HPOS="25.48" />
  <a:String HEIGHT="101.09" WIDTH="281.18" VPOS="115.53" HPOS="327.06" CONTENT="principio" />
  <a:SP WIDTH="305.82" VPOS="112.98" HPOS="327.06" />
  <a:String HEIGHT="98.54" WIDTH="112.13" VPOS="110.43" HPOS="632.88" CONTENT="uſq" />
  <a:SP WIDTH="121.48" VPOS="110.43" HPOS="632.88" />
  <a:String HEIGHT="98.54" WIDTH="11.04" VPOS="110.43" HPOS="754.35" CONTENT=";" />
  <a:SP WIDTH="17.84" VPOS="110.43" HPOS="754.35" />
  <a:String HEIGHT="98.54" WIDTH="127.42" VPOS="110.43" HPOS="772.19" CONTENT="addi" />
```

Fig. 17. Image d'un extrait du fichier XML au format ALTO, donnant la position de chaque ligne et de chaque mot.

moyenne « e » des lignes. Cette grandeur exprime la place occupée par le corps de l'écriture par rapport à l'interligne.

- Imbrication du texte : la distance entre la bottom-line et la high-line de la ligne en dessous donne une idée de l'imbrication du texte, cette distance peut être négative et aussi mesurée relativement à la hauteur de ligne « e ».
- Espacement des mots : une distance moyenne entre mots, relativement à l'épaisseur « e » du texte, peut être mesurée par BIT-Alpha.
- Longueur des mots : à partir de la forme des mots identifiés, il est possible de donner une répartition statistique (répartition de Gauss) de la forme des mots, de leur longueur, par rapport d'une part à l'épaisseur « e » du texte et d'autre part par rapport à la hauteur des lignes et d'avoir aussi deux grandeurs aussi caractéristiques de l'écriture.

- Hauteur relative des mots : il est possible de donner la répartition statistique de la hauteur totale relative des mots, par rapport à l'épaisseur « e » du corps du texte. Ceci donne une grandeur significative des allongements de l'écriture, comme boucle de « l » et pied de « p ».
- Courbure des lignes : BIT-Alpha est en mesure de suivre la courbure d'une ligne de texte et ainsi de donner une valeur caractéristique de sa courbure relative. Ceci bien sûr n'a d'intérêt majeur que pour un texte écrit de main libre, sans trait « guide » sur le document.
- Variation de l'inclinaison : il est possible d'obtenir une valeur caractérisant la variation d'inclinaison moyenne d'une ligne à l'autre. Ceci n'a de sens que pour un texte écrit de main libre, sans trait guide sur le document.

La figure 18 montre en plus de la top-line et de la base-line, la bottom- line et la high-line pour chaque ligne du texte manuscrit. Par ailleurs les éléments de mesure de chaque élément graphique constitutif du texte sont visibles.

Nous avons en second les valeurs liées aux caractères eux-mêmes :

Nous pouvons extraire entre autres les paramètres suivants :

- Taille moyenne des caractères du texte.
- Taille moyenne des majuscules, des minuscules.
- Variance entre la plus grande et la plus petite lettre d'une ligne.
- Taille des allongements, relatifs à l'épaisseur du texte « e », comme des barres de « d », « l », « b » etc.
- Inclinaison moyenne des traits verticaux.
- Variance de l'inclinaison des traits verticaux.
- Taille moyenne des accents et points sur « i ».
- Distance moyenne entre les points et accents et la base-line.

Cette énumération n'est donnée qu'à titre d'exemple et il est certainement possible d'imaginer d'autres paramètres géométriques.

La détermination de ces éléments géométriques partant du texte dans son ensemble, des mots globalement et des caractères individuels, permettent de concevoir une matrice de paramètres caractérisant le texte et l'écriture de son scribe. En vue d'une comparaison de plusieurs textes pour en déterminer le scribe, ces paramètres peuvent aussi se voir attribuer une pondération donnant plus d'importance à une considération qu'à une autre. Il pourrait être ainsi possible de définir une « distance » entre écritures, qu'il serait possible de normer et définir ainsi une valeur « analytique » impartiale d'identification d'écriture, respectivement du scribe.[1]

Ainsi BIT-Alpha pourrait être un outil permettant de donner à l'identification du scribe, nous voulons dire de la main ayant écrit le texte, une approche « quantifiée »

[1] Cfr. la contribution de Aussems, Brink en cette volume.

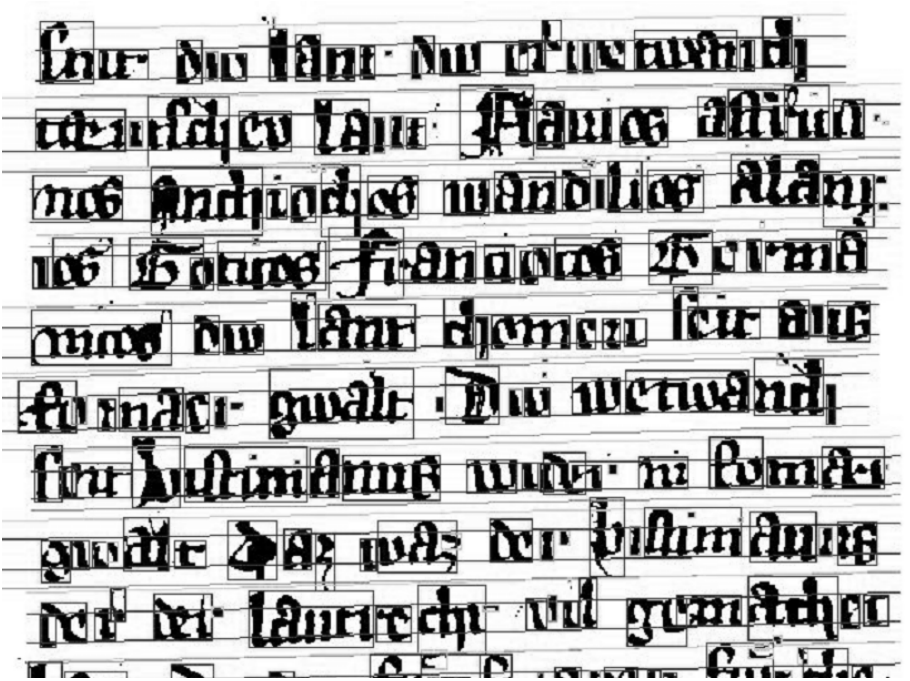


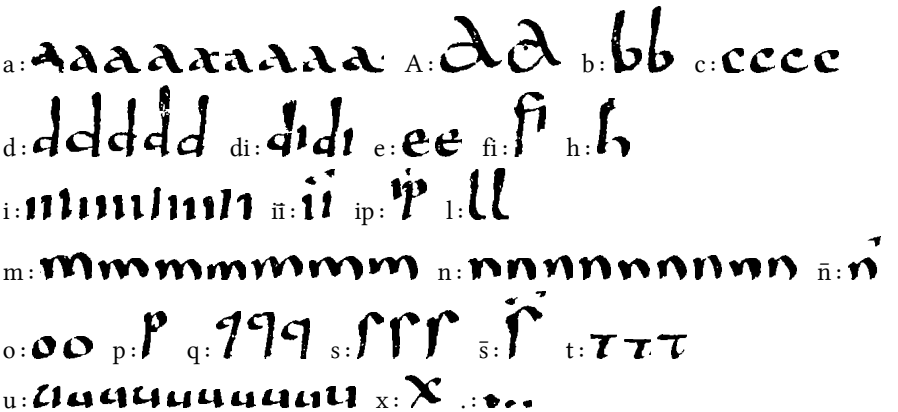
Fig. 18. Capture des éléments graphiques constitutifs d'un texte et des lignes bottom-line, base-line, top-line, high-line.

et « mesurée » pouvant aider le génie intuitif d'un paléographe. Ou permettre à une personne moins expérimentée de vérifier ses présomptions. Les idées, les possibilités et les considérations présentées ici devront être validées et affinées avec des chercheurs et paléographes. La validation de la détermination du scribe présumé d'un texte avec le logiciel BIT-Alpha_ICR passe aussi essentiellement pour la validation de la reconnaissance ICR des caractères du texte en utilisant une bibliothèque base de données ICR comportant les « empreintes numériques » des caractères du scribe considérés et préalablement appris.

Ainsi BIT-Alpha permet la reconnaissance de caractères par rapport à une bibliothèque base de données de symboles ICR pouvant être constituée séparément pour chaque scribe et appris directement d'un document traité. Chaque bibliothèque base de symboles ICR peut être prise en charge pour la reconnaissance au choix de l'opérateur. Ainsi en vérifiant la reconnaissance des symboles avec une base ICR d'un auteur connu, tout doute quant à la validation de l'auteur devrait pouvoir être levé. Le taux de fia-

bilité de lecture est directement affiché dans BIT-Alpha par la couleur des caractères proposés pour la transcription. La base ICR d'un scribe ne correspondant pas au tracé de l'écriture analysée donnera beaucoup de rouge dans la reconnaissance ICR.

Par ailleurs, BIT-Alpha permet de capturer et d'éditer en un clic tous les symboles graphiques d'une page traitée, en classant les lettres dans des dossiers distincts, un dossier pour chaque symbole reconnu. Les symboles capturés sous édités individuellement sous forme d'une image au format Bitmap et correspondant exactement au caractère capturé depuis le graphisme du texte. Nous les avons rassemblés en figure 18, chaque symbole étant une image individuelle utilisable à toute fin voulue.



Une étude de ces symboles devrait permettre au paléographe d'obtenir de précieux renseignements sur l'écriture du scribe et de bien pouvoir finaliser son identification. Il serait ainsi possible de générer une fonte de caractères pour texte permettant informatiquement d'écrire en reprenant les caractères du document.

5 Conclusion

L'approche informatisée du document manuscrit permet de prendre conscience des difficultés du traitement et de la transcription de ce type de document. Seule une approche prenant en compte l'ensemble des considérations, d'une part graphiques venant du tracé de l'écriture et d'autre part linguistiques venant de la langue utilisée, permettent un résultat utilisable.

Nous avons ici présenté nos développements et le logiciel BIT-Alpha, avec ses possibilités. Ce logiciel ICR peut constituer un outil d'aide à la transcription d'un texte manuscrit ou un outil d'aide à l'identification du scribe. Dans l'état de la technologie actuelle, nous considérons que l'intervention du spécialiste pour valider les résultats proposés reste essentielle. Cet exposé permet aussi de percevoir les difficultés de la transcription informatisée « pleine page » d'un texte manuscrit inconnu. Il reste encore

un long chemin de recherche et développement avant une transcription automatique et nous ne voulons ici qu'apporter une contribution dans cette voie.

Mais malgré les difficultés, l'approche informatisée du document manuscrit est très prometteuse et la lecture ICR automatisée de textes manuscrits anciens ou en écritures liées, comme l'arabe ou l'hindi, permettrait l'accès de l'humanité à des trésors encore inaccessibles.

En paléographie, la définition d'une grandeur quantifiée mesurant impartialement la « distance » entre une écriture connue et une écriture analysée, apporterait une contribution d'intérêt aux débats concernant l'identification de certains écrits.

Bibliographie

ALTO : Analyzed Layout and Text Object. <http://www.ccs-gmbh.com/alto/>.

Andrieux-Reix, Nelly, Sonia Branca-Rosoff, and Christian Puech. « Les abreviatures à la Renaissance : enjeux et usages. » *Écritures abrégées (notes, notules, messages, codes).* Bibliothèque de Faits de Langues, org. Paris : Ophrys, 2004.

Emptoz, Hubert, Frank Le Bourgeois, Véronique Eglin, Stéphane Bres, Yann Leydier, Ikram Moalla, and Fadoua Drira. *Computer Assistance for Digital Libraries : Contributions to Middle-ages and Authors' Manuscripts Exploitation and Enrichment*, Second International Conference on Document Image Analysis for Libraries (DIAL 2006). Los Alamitos (CA) [et al.] : IEEE Computer Society, 2006. 265–80.

Le Bourgeois, Frank. *Automatic Metadata retrieval from Ancient Manuscripts*, Documents Analysis Systems (DAS 2004). Lecture Notes in Computer Science n° 3163. Berlin : Springer, 2004. 75–89.

Le Bourgeois, Frank, E. Trinh, Bénédicte Allier, Véronique Eglin, and Hubert Emptoz, *Document Image Analysis solutions for Digital libraries*, International Conference on Document Image Analysis for Libraries (DIAL 2004). Los Alamitos (CA) [et al.] : IEEE Computer Society, 2004. 2–24.

Le Bourgeois, Frank, Hubert Emptoz, Ikram Moalla, and Adel M. Alimi. *Contribution to the discrimination of the medieval manuscript texts : Application in the palaeography*, Document Analysis Systems (DAS 2006). Lecture Notes in Computer Science n° 3872. Berlin : Springer, 2006. 25–37.

Medieval Unicode Font Initiative. <http://www.mufi.info/>

MUFI character recommendation v. 2.0. Ed. Odd Einar Haugen, Bergen, 2006. <http://hdl.handle.net/1956/2003>.

Niblack, Wayne. *An Introduction to Digital image processing.* Upper Saddle River (NJ) : Prentice Hall, 1986. 115–16.

Ringlstetter, Christoph. *OCR-Korrektur und Bestimmung von Levenshtein-Gewichten.* Magisterarbeit Computerlingustik, Centrum für Information und Sprache, Ludwig-Maximilians-Universität-München, Prof. Klaus Uwe-Schulz, 2003.