La reconnaissance de l'écriture manuscrite hors ligne

Applicabilité à la transcription et l'indexation d'un fonds notarial des Archives cantonales jurassiennes

Micha d'Ans

Ce travail répond à un mandat des Archives cantonales jurassiennes (ArCJ). Dans le cadre de sa stratégie de numérisation, cette institution souhaitait la création d'un rapport analysant les différentes options de transcription et d'indexation de leur fonds de répertoires de notaires pour leur ajouter une composante de recherche plein-texte. Il s'agit d'archives manuscrites importantes qui ont la particularité d'être à la fois des documents d'archives et des instruments de recherche, les notaires s'en servant à l'origine comme index, ils contiennent des références à tous les actes passés dans les études jurassiennes.

Une méthode de traitement automatique, la reconnaissance de l'écriture manuscrite hors ligne, est évaluée et mise en confrontation avec les performances d'options manuelles.

Problématique et état de l'art

Ces dernières années, des campagnes massives de numérisation de documents historiques manuscrits ont été menées par de nombreuses institutions, générant des millions d'images de textes. Les données pertinentes à chaque requête se perdent dans cette masse gigantesque, contraignant les lecteurs à feuilleter sur un écran une quantité de pages rédhibitoire avant de trouver celles qui contiennent les informations qu'ils cherchent. Une vaste partie des documents numérisés attend ainsi d'être transcrite dans un format textuel électronique adapté qui offrirait aux utilisateurs de nouveaux moyens d'indexation, de consultation et de recherche.[1] De nombreux facteurs peuvent s'additionner pour rendre les textes particulièrement difficiles à déchiffrer, tant sur le plan du support matériel (dégradations, vieillissement, modifications) et des pratiques de l'écriture en vigueur à l'époque de leur mise sur papier, que sur celui de la production du scripteur elle-même.

1 KURSHID, Khurram : Analysis and Retrieval of Historical Document Images. Paris. 2009. JUAN, Alfons ; ROMERO, Veronica ; SANCHEZ, Joan Andreu ; SERRANO, Nicolas ; TOSELLI, Alejandro H. ; VIDAL Enrique : « Handwritten Text Recognition for Ancient Documents ». In : JMLR: Workshop and Conference Proceedings 11. 2010.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 2 — http://dx.doi.org/10.18755/iw.2018.28

La transcription et l'indexation de qualité d'une grande quantité d'images textuelles sont difficiles à réaliser manuellement et gagneraient à recevoir l'appui d'une meilleure automatisation de l'analyse d'images. Le développement des systèmes de reconnaissance de l'écriture manuscrite représente une piste intéressante dans l'optique de faciliter l'accès aux collections manuscrites historiques.

Un système de reconnaissance automatique réagit différemment selon les types d'écritures (imprimée ou manuscrite) rencontrés. Une distinction peut être faite à ce stade entre la reconnaissance de l'écriture manuscrite et la reconnaissance optique de caractères. Cette dernière, plus connue sous son acronyme anglais OCR (pour *Optical Character Recognition*), consiste en une traduction mécanique d'images d'éléments textuels manuscrits (lettres, nombres ou symboles), dactylographiés ou imprimés, en des symboles modifiables informatiquement.[2] Les logiciels OCR sont devenus des outils courants présentant d'excellents taux de réussite, au contraire des logiciels de reconnaissance de l'écriture manuscrite qui ne se déclinent pas encore dans des versions diffusables au grand public. En effet, les caractères imprimés, avec leur régularité, autorisent l'utilisation de techniques de reconnaissance rapides et fiables. De son côté, la reconnaissance de l'écriture manuscrite est unanimement jugée plus problématique, car la production de tracés unique à chaque scripteur le rend difficile.

La reconnaissance de l'écriture manuscrite (REM) peut être définie comme l'aptitude d'un ordinateur à transformer une donnée manuscrite représentée par la forme spatiale d'une marque graphique en son équivalente représentation symbolique, sous forme de texte ASCII. De nombreux auteurs[3] distinguent deux secteurs d'application en reconnaissance de l'écriture manuscrite. Cette dernière peut se faire « en ligne », lorsque le texte est saisi directement sur une surface sensible, par exemple avec un stylet. L'écriture est ainsi capturée en temps réel et peut être représentée en une suite de points ordonnés chronologiquement qui donnent de précieuses indications sur la dynamique, la vélocité et la morphologie de l'écriture. Il faut très peu d'entraînement aux programmes pour qu'ils donnent de bons résultats. L'écriture « hors ligne » est obtenue à travers une image par la numérisation d'un texte existant, à l'aide d'un scanneur ou d'un appareil photo. Dans ce cas de figure, les informations temporelles ne sont pas disponibles. De ce fait, les résultats de la

2 ROMERO, Veronica ; TOSELLI, Alejandro H. ; VIDAL, Enrique : Multimodal Interactive Handwritten Text Transcription, World Scientific Publishing. 2012.

3 Ibid. A ce propos voir aussi BERTOLAMI, Roman : Ensemble methods for offline handwritten text line recognition. Berne. 2008. VINCIARELLI, Alessandro : Offline Cursive Handwriting: From Word to Text Recognition. Berne. 2003. PLÖTZ, Thomas ; FINK, Gernot A. : « Markov Models for offline handwriting recognition: a survey ». In: International Journal on Document Analisys and Recognition (IJDAR), vol. 12, n°4. 2009. FISCHER, Andreas : Handwriting Recognition in Historical Documents. Berne. 2012.

reconnaissance automatique de l'écriture manuscrite en ligne sont meilleurs que ceux de la reconnaissance hors ligne.

Diverses tâches peuvent être attribuées à la reconnaissance de l'écriture manuscrite suivant le résultat que l'on souhaite obtenir : reconnaissance de caractères, de mots individuels ou d'une suite de mots.[4] Si la première obtient de très bons résultats, les deux dernières restent des sujets de recherche. La combinaison de multiples éléments génère naturellement plus de complexité. Avec une écriture cursive, les lettres d'un mot sont liées et se suivent en séquence. En conséquence, les méthodes de reconnaissance de caractères traditionnelles ne sont pas applicables aux textes manuscrits.

Trois approches ont été utilisées pour traiter le problème de la reconnaissance du mot :

— *L'approche holistique*, qui surmonte la difficulté de la segmentation en classifiant directement les données des images comme des mots du lexique.
— *L'approche basée sur la segmentation* qui subdivise les mots en de plus petites entités, comme des caractères ou des graphèmes, les classifie à l'aide d'un système OCR, puis les recombine pour construire le mot reconnu.
— *L'approche sans segmentation* qui reproduit le mode de lecture humain, où le cerveau traite les données sur plusieurs niveaux simultanément.

Les systèmes de REM hors ligne classiques suivent en général une architecture composée de trois modules : prétraitement, extraction de primitives et classification.

Le module prétraitement acquiert les données et les conditionne. Ces opérations sont appliquées directement à l'image numérisée. Le bruit est filtré, les tracés manuscrits sont retrouvés et normalisés. La variabilité des styles de texte est réduite. Le module d'extraction des primitives traite l'image afin de représenter les informations qui y sont contenues avec un niveau d'abstraction plus élevé et extrait l'information la plus pertinente pour une classification donnée. Le module de classification détermine l'appartenance d'une forme à une classe à partir des vecteurs de primitives extraits de cette forme.

Différentes approches et logiciels peuvent être utilisés à cet effet :

— *Les machines à vecteurs de supports* : ensemble de techniques d'apprentissage supervisé dont le principe est de maximiser la marge entre les classes. Il s'agit d'outils assez lents, tant en phase d'apprentissage que de reconnaissance. Leur champ d'application est la reconnaissance de caractères manuscrits isolés.

4 BUNKE, Horst ; VARGA, Tamas (2007) : « Off-Line Roman cursive handwriting recognition ». In: Advances in Pattern Recognition. 2007.

— *Les réseaux de neurones récurrents* : systèmes dynamiques capables de pondérer des classes de vecteurs. Ce sont des réseaux de neurones artificiels basés sur un modèle simplifié de neurones biologiques qui permettent certaines fonctions du cerveau (mémorisation associative, apprentissage par l'exemple ou travail en parallèle). Cette technique est bien adaptée à la reconnaissance de formes globales.
— *Les modèles de Markov cachés* : outils statistiques permettant de calculer la probabilité d'appartenance d'une forme à une classe, la forme étant vue comme un ensemble d'observations émises par des états cachés. Les modèles de Markov cachés ont pu être appliqués aux trois niveaux du processus de reconnaissance (caractère, mot et phrase).
— *Les modèles de langage : n-grammes* : modèles utilisés pour quantifier les régularités dans le langage naturel et prédire le mot suivant dans une séquence de mots. Cela est rendu possible par le fait que la position et l'enchaînement des mots dans le langage naturel sont sujets à des règles syntaxiques précises. Les modèles de langage statistiques tentent d'établir des prédictions d'apparitions de mots à partir de grands corpus de textes.

Il existe également des systèmes hybrides combinant modèles de Markov cachés et réseaux de neurones artificiels donnant des résultats prometteurs.

Pratiques institutionnelles

Dans la perspective de comparer différentes stratégies et pratiques de transcription et d'indexation de registres semblables à ceux compris dans le fonds de répertoires de notaires des Archives cantonales jurassiennes, dix-sept institutions ont été approchées par courriel ou téléphone. Il s'agit principalement d'archives d'Etat d'autres cantons, car elles étaient les plus enclines à avoir déjà fait des recherches sur des sujets similaires à ceux abordés dans ce travail. Un questionnaire a été élaboré et leur a été soumis. Les questions suivantes leur ont été posées :

— Votre institution a-t-elle déjà transcrit ou indexé des registres manuscrits du même type que des répertoires de notaires ? Si oui, a) comment avez-vous procédé ? b) Quelles autres méthodes avez-vous évaluées ?
— Votre institution a-t-elle déjà eu recours à la reconnaissance de l'écriture manuscrite hors ligne pour transcrire ou indexer des documents ? Si oui, a) avec quelle efficacité ? b) Comment avez-vous ensuite utilisé les résultats de la reconnaissance ?

Les réponses reçues ont été synthétisées dans un tableau divisé en deux grandes parties en fonction des documents abordés, à savoir les répertoires de notaires ou

des documents similaires. Trois sous-catégories précisent pour chaque partie s'il est question de numérisation, de transcription ou d'indexation.

Force est de constater que la reconnaissance de l'écriture manuscrite n'est utilisée par aucune des institutions qui ont répondu au questionnaire. Celles qui se sont penchées sur la question ont jugé la transcription manuelle plus adéquate à leurs projets. La plupart des archives contactées se servent des répertoires de notaires directement comme index et n'extraient pas de données supplémentaires. De ce fait, en adoptant une méthode de transcription et d'indexation automatique, les Archives cantonales jurassiennes se différencieraient des tendances institutionnelles du paysage archivistique suisse.

Le fonds des répertoires de notaires des Archives cantonales jurassiennes

Les Archives cantonales jurassiennes ont leur siège à l'Hôtel des Halles à Porrentruy et conservent des fonds datant essentiellement des XIXe et XXe siècles, sur environ 4 km de rayons. Elles réunissent les archives des anciens services administratifs du Canton de Berne (avant 1979), des services administratifs de la République et Canton du Jura (depuis 1979 à aujourd'hui), ainsi que les registres paroissiaux et de l'état civil antérieurs à 1875. En plus des archives produites par les instances officielles, les ArCJ conservent les archives privées de personnalités, d'entreprises, de sociétés et d'associations.

L'une de leurs missions fondamentales étant la mise à disposition de l'information, les Archives cantonales jurassiennes ont mis en place une stratégie de numérisation à long terme, appelée projet SIGMA. Ses objectifs principaux sont de rendre accessibles, sans limitations de temps ni de lieu, des copies numériques de haute qualité d'un portefeuille de dix-sept archives analogiques, dont font partie les répertoires de notaires, et d'améliorer la recherche dans leur contenu par une recherche plein-texte.

Le fonds des répertoires de notaires des Archives cantonales jurassiennes comprend les registres établis entre 1815 et 1978 par 180 notaires, répartis dans les districts dans lesquels ils ont exercé, et classés par ordre alphabétique. Les volumes de ce fonds peuvent être de différents types :

— Les *répertoires* de *minutes*, de *testaments* et de *brevets* sont des registres établis par le notaire, relevant dans l'ordre chronologique tous les actes passés dans son étude et indiquant essentiellement le numéro de l'acte, la date à laquelle il a été établi, sa nature et les noms des principales parties concernées.

— Les *répertoires alphabétiques annuels* offrent les mêmes informations, mais recensent les actes sur l'année en cours en fonction des clients classés par ordre alphabétique.
— Les *grands livres* sont la transcription sur un document unique de la totalité des mouvements de comptabilité et sont classés par clients.

Une étude de l'ensemble des répertoires n'étant pas envisageable par manque de temps, la constitution d'un échantillon de 26 registres (un peu moins de 5 % du volume initial) a été menée sur la base d'une sélection aléatoire dans les trois districts jurassiens. A une exception près, tous les répertoires de l'échantillon proviennent de notaires différents. Ils couvrent l'ensemble de la période de création des documents du fonds et offrent ainsi une bonne vue d'ensemble des pratiques notariales pratiquées alors. La grande majorité des registres de l'échantillon sont en bon état et tous ont été désacidifiés en 2014. Ils comprennent un total de pages à numériser allant de 16 à 516. En extrapolant, on peut estimer que les répertoires du fonds contiennent en moyenne 228 pages, pour un total de 121'752 pages. Les notaires n'ayant pas rempli entièrement tous les registres, il est possible d'observer qu'en moyenne seuls 2/3 de leurs pages comprennent des écrits. De ce fait, le nombre total de pages à transcrire et indexer s'élève à un peu plus de 81'000.

Certains répertoires contiennent des titres de colonnes en caractères imprimés, mais l'ensemble du fonds est manuscrit. La qualité et les styles d'écriture varient grandement selon les scripteurs qui peuvent être plusieurs par registre. La structure interne des registres est semblable pour les répertoires de minutes, de testaments et de brevets. Il s'agit de documents présentant les informations sous forme de tableau, sur une ou deux pages.

Même si en général les registres sont tenus avec soin, des variations intrinsèques à l'écriture, liées à l'exécution des tracés, peuvent en complexifier le déchiffrement. Les notaires ont utilisé différents des moyens tels que le soulignement, ou les variations de style, de taille et d'instrument d'écriture pour mettre en évidence certaines données qui leur étaient particulièrement utiles. Pour gagner du temps et se simplifier la vie, ils ont également pris parfois quelques libertés avec la structure rigide des colonnes, se sont servis d'abréviation ou ont apporté des modifications physiques au registre lui-même. Dû à la composition de certaines encres ou à la qualité du papier, le recto est visible en transparence dans certains répertoires.

Tous ces faits contribuent à l'hétérogénéité du fonds et influent à leur manière sur le choix des méthodes de transcription ou d'indexation qui leur seront appliquées. Si les mises en évidence peuvent faciliter le repérage des mots et favoriser un traitement automatique pour une transcription partielle, les variations de style et d'instrument, les débordements de colonnes, les abréviations, de même que les altérations du support peuvent s'avérer être un frein à la reconnaissance de

l'écriture manuscrite. Dans l'optique d'un traitement automatique, il faut également ajouter à la liste des difficultés les différentes formes d'agencement des éléments formant les dates (jour, mois, année) et les informations plus ou moins exhaustives renseignant les parties (nom, prénom, filiation, profession, origine, domicile) selon les notaires. Ces variations empêchent une extraction de données uniformes et imposent des ajustements préalables.

Evaluation et recommandations

Transcrire et indexer des documents manuscrits sont des opérations chronophages qui nécessitent de grands efforts. Il n'est donc pas étonnant que la recherche s'intéresse à des solutions partiellement ou complètement automatiques pour diminuer l'apport humain dans le traitement des tâches.

Des approches différentes peuvent être envisagées en fonction des objectifs que l'on se fixe en ce qui concerne la transcription. Il est possible de distinguer deux sortes de qualité pour cette dernière, selon qu'elle soit complètement correcte ou qu'elle contienne une certaine quantité contrôlée d'erreurs. Les transcriptions du premier type correspondent aux conventionnelles transcriptions paléographiques de manuscrits, alors que celles du second type peuvent être utilisées comme métadonnées pour l'indexation, la consultation et l'interrogation de documents.

L'indexation contribue à la description du contenu des documents, à la localisation et au repérage de l'information, aux rapprochements et aux regroupements. Elle se décline traditionnellement en indexation manuelle et automatique. Cette dernière utilise des méthodes algorithmiques pour procéder à la création d'un index, sous la forme d'une liste de descripteurs, à chacun desquels est associée une liste des documents ou parties de documents auxquels ce descripteur renvoie. Lorsqu'un utilisateur effectue une recherche d'informations, le système rapprochera la requête de l'index pour établir une liste de réponses. Dans le cadre de la reconnaissance et récupération automatiques les systèmes transcrivent puis indexent l'intégralité des images de documents textuels. Une transcription complète est ainsi réalisée, dans laquelle une recherche plein-texte est possible. Il s'agit d'une technique de recherche dans un document électronique ou une base de données textuelle, qui consiste, pour le moteur de recherche, à examiner tous les mots de chaque document enregistré et à essayer de les faire correspondre à ceux fournis par l'utilisateur. Si l'on souhaite avoir une composante de recherche plein-texte, il est nécessaire d'avoir une transcription parfaite et donc d'effectuer un travail de post-édition sur la transcription obtenue automatiquement avant d'utiliser le système d'indexation. Même les systèmes de reconnaissance manuscrite hors ligne de pointe présentent

des taux d'erreurs dépassant les 20 %,[5] ce qui rend les travaux de relecture et de correction obligatoires.

Les seules branches de la recherche qui obtiennent des résultats satisfaisants sont celles qui travaillent avec un vocabulaire très restreint, comme le tri postal d'adresses ou la lecture automatique de montant de chèques ou de formulaires. En effet, le fait d'avoir à choisir entre un nombre moins élevé d'alternatives génère moins d'erreurs de reconnaissance.

Une autre fonction de recherche avancée permet à l'utilisateur de sélectionner et d'associer les champs indexés auxquels il désire limiter sa recherche, pour ensuite utiliser les mots-clés requis à l'intérieur de ces champs, il s'agit du word spotting. Ce système de reconnaissance repère toutes les occurrences d'un même mot dans une collection de documents, permettant de réaliser une transcription partielle des images de documents manuscrits dans laquelle ne figurent que les mots que l'on choisit d'indexer. On peut ensuite interroger le document au moyen d'une recherche par champs. Le word spotting permet d'obtenir une transcription partielle avec une bonne exactitude et peu d'interventions humaines et donne la possibilité d'indexer automatiquement les termes identifiés. Ces deux approches fonctionnent en vue d'obtenir deux résultats différents : respectivement une transcription intégrale ou une transcription partielle.

Etant donnée les taux d'erreur impliqués, les résultats des systèmes de reconnaissance de l'écriture manuscrite doivent être systématiquement relus et corrigés par un expert humain. Une telle solution de post-édition n'est pas très efficace en termes de coût et de temps. En outre, elle n'est que peu appréciée des experts en transcription, qui ne se sentent pas aux commandes du processus. Les prototypes de recherches actuels obtiennent des résultats qui ne dépassent pas une précision de 80 % au niveau du mot et ne peuvent en aucun cas se substituer complètement aux experts humains dans cette tâche. Dans ce cas, la tentation de se passer de l'assistance informatique et transcrire manuellement les documents historiques manuscrits peut être grande.

Il est en effet toujours de miser uniquement sur l'effort humain en faisant appel à des transcripteurs chevronnés. La transcription de documents historiques manuscrits est en général menée par des experts en paléographie qui sont spécialisés dans la lecture d'écritures anciennes. Des connaissances du contexte historique sont en effet souvent nécessaires pour déchiffrer complètement des textes manuscrits peu faciles à aborder. La durée de transcription d'une page de ce type de documents dépend des difficultés qu'elle contient en même temps que des compé-

5 CHEVALIER, Sylvain : Reconnaissance d'écriture manuscrite par des techniques markoviennes : une approche bidimensionnelle et générique. Paris. 2004.

tences des experts. D'après une évaluation des Archives d'Etat du canton de Zurich, une personne habituée transcrit jusqu'à trois ou quatre pages par heure.[6]

Pratiquement, le transcripteur peut, selon ses préférences, saisir le texte soit directement, soit le dicter à un logiciel de reconnaissance automatique de la parole. Cette technique informatique permet d'analyser la voix humaine pour la transcrire sous la forme d'un texte. Il s'agit d'un outil fiable et extrêmement pratique pour ceux que la dactylographie rebute. Qu'elle soit générée par un logiciel de reconnaissance automatique de la parole ou par l'expert lui-même, la transcription doit passer par le contrôle d'un autre expert pour être validée après correction des erreurs qui pourraient s'y trouver.

Le rendement de la transcription manuelle peut être optimisé en augmentant le nombre de transcripteurs. Une grande équipe de transcripteurs doit être supervisée et les tâches de chacun clairement réparties. S'adjoindre les services d'une armée d'experts en transcription n'est cependant pas à la portée de tous les budgets, même si cela représente un gain de temps précieux. Il est également possible de faire appel à des bénévoles. Certaines institutions utilisent ce système de production participative (crowdsourcing) en mettant à disposition, via des interfaces, des images pour lesquelles chacun peut proposer sa transcription ou proposent des espaces numériques de travail en ligne offrant la possibilité aux chercheurs de télécharger leurs documents numérisés, de transcrire et d'annoter les textes, puis de mettre gratuitement les résultats à la disposition des autres utilisateurs.

D'autres approches essaient de combiner méthodes automatiques et manuelles. On peut les qualifier de semi-automatiques. Certains auteurs parlent de CATTI (Computer Assisted Transcription of Texte Images).[7] Leur système prévoit une collaboration entre un transcripteur humain et un système automatique de reconnaissance de l'écriture manuscrite. Le but est de combiner l'exactitude du premier avec la rapidité de traitement du deuxième. Les deux entités doivent interagir par étapes pour générer la transcription finale des images de texte. L'architecture du système est adaptée pour traiter les feedbacks de l'utilisateur. A chaque étape d'interaction, le système propose sa meilleure transcription pour une ligne de texte, et le transcripteur humain qui supervise le processus la valide ou lui apporte des corrections. Le système prend en compte les nouvelles informations pour améliorer

6 Beschluss des Kantonsrates über die Bewilligung eines Beitrages aus dem Lotteriefonds zu Gunsten des Staatsarchivs des Kantons Zürich zur Transkription und Digitalisierung von Kantonsratsprotokollen sowie Regierungsratsbeschlüssen, p. 3.

7 A ce sujet voir CLAWSON, Robert ; BARRETT, Bill : « Intelligent Indexing: A semi-Automated Trainable System for Field Labeling ». In: Proc. SPIE 9402, Document Recognition and Retrieval XXII. 2014. GORDO, Albert ; LLORENS, David ; MARZAL, Andrés ; PRAT, Frederico ; VILAR Juan M : « State: A Multimodal Assisted Text-Transcription System for Ancient Documents ». In: The Eighth IAPR International Workshop on Document Analysis Systems. 2008.

et mettre à jour sa transcription. Cette approche lui permet non seulement de corriger l'erreur en cours, mais aussi toutes les erreurs apparentées plus loin dans le texte. Cela permet de diminuer le travail de correction du transcripteur humain par rapport aux systèmes classiques de reconnaissances de l'écriture manuscrite. Une approche[8] propose un écran tactile sur lequel il est possible d'effectuer des corrections avec un stylet. Le système combine la REM hors ligne pour la transcription et la REM en ligne pour les feedbacks de l'utilisateur. Selon les auteurs, ce moyen s'avère plus naturel que les autres pour produire un texte correct, car c'est le transcripteur qui est dynamiquement aux commandes du système, et non l'inverse.

Les systèmes de ce genre sont cependant des prototypes faits sur mesure, et il n'est pas possible d'en trouver des versions tout public sur le marché.

Applicabilité des méthodes au fonds

Le fonds des répertoires de notaires des Archives cantonales jurassiennes présente un plus haut degré de difficulté en ce qui concerne les approches automatiques. La période de conservation de ce fonds, s'étendant de 1815 à 1978, montre l'évolution des habitudes de productions d'écriture sur plus d'un siècle et demi : des styles d'écriture très différents s'y côtoient et certains registres contiennent également du texte imprimé. Les recours à des abréviations contrastent avec des pratiques plus modernes. Même si la présence de tableaux et de lignes préimprimés structurent et guident le travail des scripteurs, l'utilisation d'instruments d'écriture différents influe sur la morphologie des graphies. Le contenu des colonnes s'articule également de manière plus ou moins libre en fonction de la personne tenant le registre. La taille du vocabulaire est considérable.

La multitude des scripteurs représente la difficulté majeure, car c'est elle qui est la cause des grandes variations dans les styles d'écritures, parfois à l'intérieur d'un même répertoire. En revanche, le bon état matériel des registres est un avantage qui faciliterait la phase de prétraitement. De plus, la structure en tableau permet une localisation du texte et une extraction des lignes plus aisées

Pour traiter ce fonds de manière automatique, il serait nécessaire d'avoir un système de reconnaissance de l'écriture manuscrite hors ligne fonctionnant indépendamment du scripteur, c'est-à-dire être capable de s'adapter sans entraînement ni validation avec des échantillons d'écriture. Il s'agit de la plus haute généralisation possible en ce qui concerne les styles et les instruments d'écriture. Les systèmes de reconnaissances optimisés pour des scripteurs uniques ou des multi-

8 TOSELLI, Alejandro H. ; VIDAL, Enrique ; CASACUBERTA Francisco : Multimodal Interactive Pattern Recognition and Applications. Londres. 2011.

scripteurs fonctionnent en général très mal lorsqu'on leur présente une écriture inconnue.

Une approche de transcription manuelle soulève également bon nombre de difficultés. Comme il a été mentionné plus haut, un transcripteur a une production de trois à quatre pages à l'heure, ce qui représente une vitesse de transcription de vingt minutes par page pour les registres concernés par cette étude. Appliquée au nombre total de pages à transcrire, qui s'élève à 81'168, il est possible d'estimer une durée totale de 27'056 heures de travail pour la l'intégralité de la tâche de transcription manuelle. Cela représente plus de 14 années de travail pour une personne seule transcrivant 8 heures par jour. A ce total, il faudrait ajouter les heures nécessaires à la relecture et à la correction de la transcription. Si plusieurs transcripteurs travaillent sur le projet, la durée totale sera divisée par le nombre de personnes engagées. Pour limiter les erreurs, les transcripteurs doivent travailler en parallèle, et une arbitration doit être faite pour corriger les discordances qui pourraient apparaître entre eux.[9]

L'indexation contribue à la description du contenu des documents, à la localisation et au repérage de l'information, aux rapprochements et aux regroupements d'informations. De nombreuses institutions utilisent les répertoires de notaires tels quels pour retrouver les minutes, les brevets et les testaments, car ces registres sont déjà des index en eux-mêmes. Les Archives cantonales jurassiennes prennent cependant le contre-pied de cette vision des choses en souhaitant offrir à leurs utilisateurs de meilleurs accès à leur fonds en multipliant les possibilités de recherches dans les répertoires. Actuellement, les métadonnées suivantes sont déjà répertoriées dans scopeArchiv : la cote, le district, le notaire, le type de registre et les dates extrêmes de la tenue du répertoire. Les options d'interrogation du document sont ainsi limitées. Pour retrouver une minute particulière, il faut préalablement savoir dans quelle étude de notaire elle a été faite, feuilleter le répertoire et déchiffrer le texte jusqu'à ce que l'information pertinente soit repérée. Proposer de nouvelles métadonnées aux chercheurs faciliterait et améliorerait leur travail.

Les nouvelles données pertinentes qui pourraient être saisies et introduites dans la base de données sont les suivantes : le numéro, la date et la nature des actes, ainsi que les parties concernées (nom, prénom, profession, origine, domicile). Appliquée au fonds des répertoires de notaires, une indexation de ces données permettrait d'effectuer des recherches de tous les actes notariaux d'un comparant, quel que soit le notaire consulté, ou de combiner les informations d'un répertoire de minutes avec celles d'un répertoire alphabétique annuel.

9 SIBADE, Cédric ; RETORNAZ, Thomas ; NION, Thibault, LERALLUT, Romain ; KERMORVANT, Christopher : « Automatic indexing of french handwritten census registers for probate geneaology ». In: Proceedings of the 2011 Workshop on Historical Document Imaging and Processing. 2011.

Une lecture exhaustive de tous les registres de l'échantillon a montré que la nature des actes représentait un lexique de 62 substantifs et syntagmes nominaux. Une telle liste permettrait d'alimenter une base de données en cas d'indexation automatique. Un système de reconnaissance de l'écriture manuscrite fonctionne en effet avec plus d'efficacité lorsqu'il sait d'avance ce qu'il cherche et a pu être entraîné dans ce sens. De la même manière, un dictionnaire des patronymes jurassiens pourrait être implémenté et utilisé pour l'entraînement et la programmation d'un système de word spotting. Si ces domaines contraints sont correctement exploités, des taux de reconnaissance plus que corrects devraient pouvoir être atteints, ce qui permettrait de réaliser une transcription partielle et une indexation automatique du fonds.

Les systèmes de reconnaissance de l'écriture manuscrite hors ligne sont pour l'instant dans une phase de leur développement qui n'est pas encore optimale par rapport aux questions que soulève le mandat des ArCJ.

L'évaluation des registres du fonds notarial a montré que le degré de difficulté qu'ils présentaient pour l'application de techniques complètement automatiques était relativement haut. Le critère le plus problématique réside sans conteste dans la multitude de scripteurs ayant rédigé les textes contenus dans le fonds. Pour entraîner convenablement un système de reconnaissance, la transcription manuelle d'un certain nombre de pages est requise. Or bon nombre de répertoires du fonds n'en contiennent pas assez. Pour être transcrit automatiquement, le fonds nécessite donc un système de reconnaissance indépendant du scripteur. Comme il n'existe pas de tels logiciels sur le marché, la collaboration avec une équipe de chercheur ou une entreprise spécialisée dans le domaine est indispensable si cette option qui est retenue.

Dans l'optique de trouver les méthodes de transcription et d'indexation les plus rentables pour mener à bien leur projet, les Archives cantonales jurassiennes devront scrupuleusement étudier par le prisme du budget et du temps à leur disposition les différentes réponses que recevront leurs appels d'offres. Un critère déterminant sera la solution de recherche qu'elles souhaiteront offrir à leurs utilisateurs : une interface de recherche plein-texte nécessite une transcription complète et parfaite, alors que pour une recherche par mots-clés, une transcription partielle peut s'avérer suffisante.

A partir de ces constats, les recommandations suivantes peuvent être proposées aux ArCJ pour les répertoires de minutes, les répertoires de brevets et les répertoires de testament :

— *En cas de transcription partielle* : l'option idéale semble une approche automatique de type *word spotting*. La réalisation de la tâche pourrait être externalisée sans occasionner des coûts trop importants. En lui implémentant

la liste de la nature des actes et un dictionnaire des patronymes jurassiens, un système de reconnaissance de l'écriture manuscrite devrait être en mesure de trouver toutes les occurrences des mots sélectionnés. Cela permettrait de multiplier les possibilités des recherches déjà offertes par *scopeArchive*.

— *En cas de transcription totale* : le recours à une approche complètement automatique n'est pas recommandé. Au vu des performances actuelles des systèmes de reconnaissance de l'écriture manuscrite, le travail de post-édition est beaucoup trop important et quasiment équivalent à celui d'une transcription manuelle. L'option idéale consisterait plutôt en une approche semi-automatique où l'utilisateur interagit avec une machine. L'idée nécessite cependant de collaborer avec une équipe de chercheurs qui développerait un prototype spécialement adapté à la tâche qui intéresse les ArCJ. Une alternative potentiellement intéressante serait de transcrire manuellement le fonds et de faire simultanément un appel de *crowdsourcing*, en mettant à disposition en ligne les images des documents numérisés pour que des bénévoles participent au processus.

En ce qui concerne les *répertoires alphabétiques annuels* et les *grands livres*, une simple transcription manuelle est recommandée. Les registres de l'échantillon étudié contiennent moins d'information que les autres types de répertoires du fonds. Les patronymes sont clairement mis en évidence, facilitant au maximum une saisie rapide des données.

Informationswissenschaft: Theorie, Methode und Praxis, Bd. 5 (2018), Nr. 2 — http://dx.doi.org/10.18755/iw.2018.28