OCR/HTR et graphie arabe Les manuscrits arabes à l'heure de la reconnaissance automatique des écritures

Noëmie Lucas

▶ **To cite this version:**

Noëmie Lucas. OCR/HTR et graphie arabe Les manuscrits arabes à l'heure de la reconnaissance automatique des écritures. 2022. hal-03822459

HAL Id: hal-03822459
https://hal.science/hal-03822459v1

Submitted on 30 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Moyen-Orient et Mondes Musulmans
Groupement d'Intérêt Scientifique

ISBN 978-2-493818-01-0

CGIS▸3 4●2022 OCR / HTR et graphie arabe

OCR / HTR et graphie arabe

Les manuscrits arabes à l'heure de la reconnaissance automatique des écritures

Cahier du GIS N°3

Avril 2022

OCR / HTR et graphie arabe

Les manuscrits arabes à l'heure de la reconnaissance automatique des écritures

Cahier du GIS N°3

Avril 2022

OCR / HTR et graphie arabe

OCR / HTR et graphie arabe

Les manuscrits arabes à l'heure de la reconnaissance automatique des écritures

Cahier du GIS N°3

Rapport rédigé par Noëmie Lucas et achevé en 2021

GIS Moyen-Orient et Mondes Musulmans

Avril 2022

MEMBRES DU GROUPEMENT D'INTÉRÊT SCIENTIFIQUE
MOYEN ORIENT ET MONDES MUSULMANS

CNRS
Collège de France
EHESS (L'École des hautes études en sciences sociales)
EPHE (École Pratique des Hautes Études)
ENS Lyon
FMSH (Fondation maison des sciences de l'homme)
IFAO (Institut français d'archéologie orientale)
INALCO
INHA (Institut national d'histoire de l'art)
IRD (Institut de recherche pour le développement)
Aix-Marseille Université
Université Paris I Panthéon-Sorbonne
Université Sorbonne Nouvelle
Sorbonne université
Université de Paris
Université Paris-Nanterre
Unversité Lumière Lyon 2
Université Lyon III Jean Moulin
Université Paris Dauphine — PSL
Sciences po Aix
Sciences Po Grenoble
Sciences Po Lyon
Sciences Po Paris
Université de Tours
Université de Strasbourg

LE GROUPEMENT D'INTÉRÊT SCIENTIFIQUE
MOYEN ORIENT ET MONDES MUSULMANS

Créé par le CNRS au 1er janvier 2013, le Groupement d'intérêt scientifique (GIS) Moyen-Orient et mondes musulmans a vocation à fédérer l'ensemble des équipes de recherche françaises qui se consacrent à ce champ, dans toutes les disciplines concernées (histoire, géographie, sciences politiques, sociologie, anthropologie, histoire de l'art, littérature, islamologie, philosophie, archéologie). Au 1er avril 2022, le GIS rassemble 40 équipes de recherche (UMR et équipes d'accueil) et 10 UMIFRE / ÉFÉ (Écoles françaises à l'Étranger) ; ces unités relèvent en tout de 25 tutelles, universités, grandes écoles et organismes de recherche

Le champ couvert par le GIS s'intéresse en priorité au Moyen-Orient et au Maghreb, en raison d'une longue tradition française d'étude des sociétés qui les composent. Il n'entend pas pour autant ignorer les autres « mondes musulmans » : cette expression se réfère à une histoire et à une culture communes, développées au-delà des mondes arabe, persan et turc dans de vastes espaces qui s'étendent à l'Inde et à l'Asie du Sud-Est, en passant par l'Asie centrale et l'Afrique subsaharienne, et elle englobe toutes les sociétés qui sont aujourd'hui concernés par le fait islamique en Europe et en Amérique. Le champ d'intervention du GIS couvre donc de très vastes régions de la planète et possède par définition une délimitation mouvante. Seul un nombre limité des équipes du GIS s'insère entièrement dans le champ thématique du groupement. Dans la plupart des cas, il s'agit d'équipes généralistes ou thématiques, au sein desquelles œuvrent quelques chercheurs et enseignants-chercheurs spécialistes du Moyen-Orient et des mondes musulmans.

Notons enfin que le champ du GIS MOMM est par son objet d'étude naturellement enclin aux collaborations internationales, ou du moins à une ouverture internationale, que ce soit avec les régions étudiées ou avec les mondes académiques européens et anglo-saxons. C'est dans cette perspective internationale qu'il convient donc de replacer l'ensemble des constats et des réflexions qui seront faits dans ce rapport.

012—013

INTRODUCTION

En 2020, le Groupement d'intérêt scientifique (GIS) Moyen-Orient et mondes musulmans présentait son livre blanc *Vers la science ouverte ? La transition numérique et la recherche sur le Moyen-Orient et les mondes musulmans en France*. Celui-ci offrait un état des lieux factuel, sans prétention d'exhaustivité, de la transition numérique dans le champ de la recherche concernée en cherchant à rendre compte des tendances et des évolutions passées et récentes. Les éléments de diagnostic posés à partir des données présentées ont permis de proposer des initiatives visant à accompagner le renforcement et le développement de la transition numérique de la recherche française sur le Moyen-Orient et les mondes musulmans.
Parmi les priorités assorties d'actions proposées dans ce livre blanc, le développement de la fouille et de l'analyse des textes figurait en bonne place et était décrit comme suit :

> Il apparaît crucial de rattraper le retard accumulé par la recherche française en matière de fouille et d'analyse de textes en arabe, turc et persan. Leur graphie non-latine soulève une difficulté particulière de lecture automatique que l'intelligence artificielle, par le biais de l'apprentissage-machine, est aujourd'hui en mesure de lever sans trop de peine, pour peu que l'on s'en donne les moyens. L'enjeu est d'ordre à la fois technologique (mise à niveau internationale), scientifique (nouveaux fronts de la connaissance) et économique (maîtrise des coûts). Une fois le processus d'OCéRisation opérationnel, il est impératif de pouvoir se familiariser avec les outils et les méthodes d'identification et d'extraction d'information, de classification des textes et d'analyse sémantique et génétique[1].

Cette priorité doit être mise en relation avec un contexte documentaire et patrimonial spécifique. Les bibliothèques françaises, au premier titre desquelles, la Bibliothèque nationale de France (BnF), la Bibliothèque universitaire des langues et des civilisations (BULAC) ou la Bibliothèque nationale universitaire de Strasbourg (BNUS), conservent des corpus de manuscrits orientaux importants. Au sein de ces derniers, les manuscrits arabes, ainsi que persans et turcs, figurent en bonne place. Faute de financement et de priorisation, ces fonds ne sont pas tous complètement signalés et catalogués de manière numérique et il reste encore un travail conséquent pour que ces collections soient numérisées dans leur entièreté.

Ces corpus nombreux de manuscrits arabes en France posent, d'une part, des questions patrimoniales qui dépassent les ambitions de ce rapport et, d'autre part, invitent à interroger la manière dont ces corpus sont exploités par les chercheurs.

L'étude de ces manuscrits est notamment l'œuvre des philologues. La philologie renvoie aux sciences du texte, à savoir l'édition de texte, la critique textuelle, l'histoire du livre et l'histoire de la transmission. Cette étude des langues et de leur littérature à partir de documents a connu et connaît de nombreux renouvellements avec l'application de méthodes de calcul et

[1] M. Volait, N. Lucas, *Vers la science ouverte ? La transition numérique et la recherche sur le Moyen-Orient et les mondes musulmans en France : État des lieux et perspectives*, GIS Moyen-Orient et mondes musulmans, 2020, p. 99. https://halshs.archives-ouvertes.fr/halshs-02937983

l'utilisation de l'ordinateur. Aussi appelle-t-on philologie numérique ou philologie computationnelle une philologie qui se place à l'intersection des sciences classiques du texte et des méthodes computationnelles (à l'image de la modélisation, de l'apprentissage machine ou de la statistique). On parle d'une philologie tournée vers les données qui cherche une répartition optimale entre l'intelligence humaine et l'intelligence artificielle[2].
La philologie computationnelle peut être appliquée à tous les champs de la philologie[3] :

- *production de données – édition* : acquisition du texte (HTR), structuration (XML (encodage)), enrichissements et indexation (lemmatisation) ;
- *analyse de données – critique* : style et question d'attribution, langues/variation des dialectes, analyses de tradition textuelle, modélisation de la transmission, etc…

La question de la reconnaissance automatique des écritures, en l'occurrence manuscrites dans le cas des manuscrits, se pose au moment de l'acquisition du texte. Pour produire une édition d'un texte, il est nécessaire de récupérer ces textes et les contenus qui sont l'objet de l'étude. Cette phase d'acquisition du texte est par ailleurs une étape préalable à toute entreprise d'analyse ou de fouille de texte. Cette récupération peut prendre plusieurs formes :

- la transcription manuelle selon des critères que l'on se fixe ;
- le téléchargement depuis des entrepôts de données textuelles déjà existants dans des formats immédiatement adaptés ;
- la transcription assistée par ordinateur à l'aide d'un algorithme permettant la reconnaissance optique des caractères imprimés (OCR) ou la reconnaissance optique des écritures manuscrites (HTR).

Ce rapport porte sur cette dernière modalité d'acquisition du texte et sur ses enjeux pour la graphie arabe. Il se veut un état de la question de l'OCR et de l'HTR pour les documents historiques rédigés en arabe avec un accent particulier sur les écritures manuscrites. Après avoir expliqué le fonctionnement théorique des OCR/HTR et de leurs infrastructures, nous préciserons les enjeux posés par les écritures non-latines et en priorité l'arabe. Le deuxième temps de ce rapport entend dresser un état des lieux des moyens disponibles (jeux de données, logiciels, interfaces, projets) en 2021 pour s'engager dans un travail d'extraction de textes en graphie arabe via la transcription assistée ou automatisée. Ce panorama non exhaustif permettra de dresser un bilan des acquis et des besoins. Le dernier temps de ce rapport propose un retour sur un projet collaboratif et collectif de développement d'un modèle de reconnaissance des caractères manuscrits arabes, réalisé au cours de l'année 2020/2021 et coordonné par l'auteure de ce rapport.

[2] T. Andrews, « The Third Way: Philology and Critical Edition in the Digital Age », *Variants*, 10, 2013, p. 61-76.

[3] J.-B. Camps, « Introduction à la philologie computationnelle. Science des données et science des textes : De l'acquisition du texte à l'analyse », conférence donnée le 7 décembre 2020 dans le cadre de la formation en ligne « Étudier et publier les textes arabes avec le numérique » : https://www.youtube.com/watch?v=DK7oxn-v0YU&t=1s

016
PRINCIPALES CONCLUSIONS : RECOMMANDATIONS POUR ACCOMPAGNER LES CHERCHEURS FRANÇAIS DANS L'UTILISATION DES SYSTÈMES OCR/HTR POUR L'EXPLOITATION DES DOCUMENTS HISTORIQUES

Mettre l'accent sur la « recherche fondamentale » de l'HTR de l'arabe, dans les documents historiques, dont les manuscrits, en

- encourageant des projets techniques avec des ambitions patrimoniales qui favorisent les collaborations entre chercheurs, ingénieurs et conservateurs. Les premiers relais sont donc les bibliothèques conservant des collections de manuscrits, au premier rang desquelles, en France, la BnF, la BULAC et la BNUS.
- garantissant que les projets impliquant le recours à des systèmes HTR mettent à disposition leurs jeux de données dans des formats ré-exploitables par d'autres et assurent ainsi le transfert de compétences et l'innovation future.

Stimuler la demande scientifique en France pour l'utilisation de ces technologies en

- INFORMANT : Face à la dispersion de l'information et à l'éclatement des initiatives, ainsi qu'aux impératifs de communication institutionnelle, il importe de privilégier un espace numérique d'agrégation des contenus qui n'a pas vocation à créer du contenu nouveau mais vise plutôt à éditorialiser et à organiser les informations éparpillées.
- FORMANT :
  - En développant une formation complète sur l'étude des manuscrits arabes/orientaux à l'heure du numérique : il s'agit de mettre en commun les compétences existantes en France pour développer un diplôme universitaire à destination des étudiants, ingénieurs et/ou chercheurs disposant des compétences linguistiques suffisantes pour exploiter ces matériaux.
  - En organisant des formations ponctuelles : soit sous la forme de projets à durée de vie limitée comme des *hackathons* ; soit sous la forme d'un accompagnement individuel d'un chercheur via une résidence numérique.

Consolider le consortium Huma-Num DISTAM, créé en 2022 et adossé à l'unité Études aréales du Campus Condorcet, par le recrutement de personnels dédiés, afin de lui permettre d'apporter, sur les questions singulières posées par les études aréales traitant des langues non-latines :

- un accompagnement des chercheurs et de leurs projets, en particulier dans la définition de leurs objectifs et l'établissement des besoins et des possibilités numériques ;
- des services d'information et de formation mais aussi une assistance à la mise en relation des équipes ingénieurs/chercheurs ou la mise en place des partenariats public/privé ;
- un soutien technique et financier.

En sus de sa mission d'initiation aux humanités numériques aréales, DISTAM pourrait ainsi développer et soutenir une pépinière de projets aréaux sur les textes en graphies non-latines. Les communautés concernées disposeraient par ce biais d'une interface d'échange à plusieurs échelles : entre les équipes, entre les équipes et les entreprises, entre la France et le reste du monde sur ces questions.

REMERCIEMENTS

Le présent rapport a été réalisé dans le cadre d'un contrat postdoctoral CNRS du GIS Moyen-Orient et mondes musulmans dédié à la philologie numérique des textes en alphabet arabe financé dans le cadre du plan SHS 2020 par le ministère français de l'Enseignement supérieur, de la Recherche et de l'Innovation et dans le cadre d'une résidence à la BULAC entre septembre 2020 et août 2021. Les informations présentées ont été collectées par des lectures et des entretiens individuels, au cours des différentes formations et évènements suivis, ou organisés, par l'auteure de ce rapport.
L'auteure de ce rapport remercie chaleureusement toutes les personnes qui ont contribué à la réalisation de ce rapport, en acceptant notamment de répondre à ses questions : Daniel Stökl Ben Ezra, Matthew Miller et Chahan Vidal-Gorène notamment. Ses remerciements sont par ailleurs adressés à toute l'équipe de la BULAC. Que les participants au *hackathon* soient remerciés pour avoir permis la réalisation du projet RASAM et avoir contribué au développement d'un outil précieux pour les chercheurs. Merci à Mercedes Volait et Éric Vallet qui ont accompagné la réalisation de ce rapport et ont participé à sa structuration. Merci pour leur relecture attentive, ainsi qu'à celle de Benjamin Guichard et Chahan Vidal-Gorène.

AVERTISSEMENT

Ce rapport n'a pas été rédigé par une ingénieure ou une *data scientist*, mais par une chercheuse en sciences humaines en formation dans le domaine des humanités numériques. Il s'adresse en priorité à des collègues non spécialistes des humanités numériques mais qui s'interrogent et souhaitent disposer d'un support d'accompagnement dans leurs démarches d'élaboration de projets numériques impliquant des entreprises d'extraction de textes.

[I]
LA RECONNAISSANCE AUTOMATIQUE DES ÉCRITURES IMPRIMÉES ET MANUSCRITES : MÉTHODES ET ENJEUX

022 [I] A FONCTIONNEMENT D'UN OCR/HTR

Cette première partie du rapport s'intéresse au fonctionnement des OCR/HTR. Après avoir rappelé la différence entre les deux, nous présenterons la logique de fonctionnement d'un OCR ou d'un HTR.

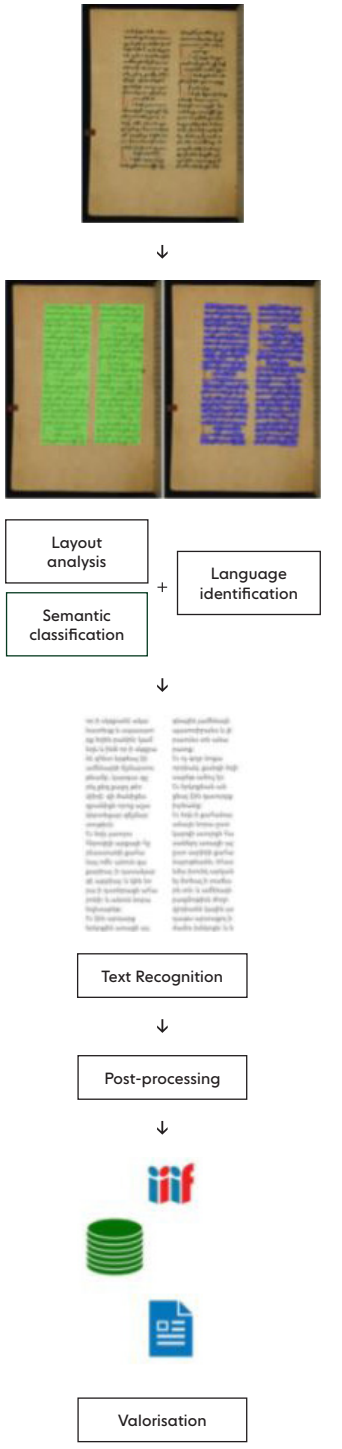
Définitions et remarques préliminaires

L'OCR – *Optical character recognition* - consiste en la conversion d'une image en un texte, ou plus exactement, il s'agit de la conversion d'un texte imprimé, écrit ou inscrit, en un texte codé par une machine. Dans les études informatiques, l'OCR est l'une des principales tâches de l'Analyse de documents images (ou *DIA – Document Image Analysis*). On désigne généralement par OCR ou HTR un logiciel de reconnaissance automatique de texte qui analyse une image numérisée pour en extraire le texte.

On distingue OCR – *Optical character recognition* et HTR - *Handwritten text recognition*. Le premier s'applique aux caractères imprimés, le second aux écritures manuscrites. La séparation entre les deux tient notamment à des enjeux techniques, concernant en particulier la mise en page, et donc à des résultats, différents.

Tandis que l'OCR est aujourd'hui considéré comme un problème résolu, pour les langues latines du moins (taux d'erreur par caractère (CER*) de moins de 2 %) avec des logiciels libres ou propriétaires, l'HTR est longtemps resté très peu fonctionnel car les difficultés sont plus nombreuses. Les données sont en effet plus complexes et les variables sont en plus grand nombre. Les premiers tests en matière d'HTR datent des années 1980, mais c'est véritablement grâce au développement de l'intelligence artificielle (et des progrès en matière de puissance de calcul) et des réseaux de neurones depuis une dizaine d'années que les choses ont pu évoluer. Même s'il n'existe pas de modèle généraliste, y compris pour les langues latines, il s'agit d'un champ de recherche très actif. La spécialisation progressive (ou *fine-tuning**) des modèles sur une écriture donnée, un type de main ou un type de manuscrit, peut facilement atteindre un CER* d'environ 5 %, voire moins en cas de mise en page simple.

Fig.1
Chaîne de traitement HTR/OCR – *blog calfa*



En quelques mots, la méthode de l'apprentissage profond* (appelé en anglais *deep learning*) consiste à soumettre à un réseau de neurones des bases de données proposant ce que l'on souhaite faire reconnaître sous de multiples facettes. Le réseau extrait les informations de ces bases de données et apprend à généraliser l'objet ou le texte et à le reconnaître quel que soit le contexte. L'entraînement est ainsi réalisé par la fréquence et l'habitude. L'OCR ou l'HTR ne consiste pas à abandonner à une intelligence artificielle le travail de transcription d'un texte. Il s'agit d'une démarche interactive entre l'homme et la machine. Pour que la machine puisse prédire le contenu d'un texte, il faudra d'abord compter sur une phase de transcription par l'homme suivie d'une phase de calcul par la machine. L'étape suivante consistera en une phase de correction par l'homme puis d'une phase de réentraînement jusqu'à parvenir à un modèle satisfaisant.

La chaîne de traitement d'un OCR/HTR

La reconnaissance des écritures [imprimées et manuscrites] implique donc une collaboration entre l'homme et la machine que la présentation de la chaîne de traitement d'un OCR/HTR entend expliciter. Dans cette partie, nous revenons sur les différentes étapes de cette chaîne de traitement.

Le schéma ci-dessus [FIG.1] permet de visualiser les différentes étapes d'une chaîne de traitement OCR/HTR. Chaque étape de cette chaîne va être explicitée dans les développements qui vont suivre. Pour réaliser cette description, nous nous sommes notamment appuyée sur les conférences données par Jean-Baptiste Camps et Chahan Vidal-Gorène au cours de différents évènements organisés par le GIS Moyen-Orient et mondes musulmans en lien avec les humanités numériques, ainsi que sur la thèse de Benjamin Kiessling[4].

Étape 1 : Acquérir des données
La première étape consiste à récupérer les images dont on souhaite extraire les données textuelles. Cette récupération peut se faire de plusieurs manières, par exemple via une bibliothèque numérique en ligne qui utilise le protocole IIIF* ou auprès de banques d'images. Une des conditions requises est de disposer d'images qui ont été numérisées.

Il faut noter que les différences quant à la qualité de la numérisation peuvent avoir des impacts lors de l'élaboration des modèles OCR/HTR. Les formats des images utilisables peuvent varier en fonction des plateformes d'annotation qui sont utilisées. Il est donc conseillé de toujours vérifier les formats pris en compte (par exemple : .jpeg, .tiff, .pdf).

Étape 2 : Traitement des images/ Preprocessing
Une fois les images récupérées, et selon le type de projet, il peut s'avérer nécessaire de prétraiter les données, ici les images. Il importe de mettre l'accent sur la dimension « optionnelle » de cette étape. En effet, elle dépend du modèle utilisé. Si ce modèle est entraîné avec des données qui ne sont pas prétraitées, alors le prétraitement des images ne sera pas nécessaire.

[4] B. Kiessling, *Advances in Optical Character Recognition for Historical Arabic Documents*, thèse de doctorat en informatique, mathématique et applications sous la direction de Marc Bui, soutenue le 13 avril 2021, Université PSL, École Pratique des Hautes Études. Nous remercions Benjamin Kiessling de nous avoir communiqué sa thèse de doctorat.

Certaines plateformes d'annotation intègrent cette fonctionnalité (ORC4all, eScriptorium), tandis que d'autres non (Transkribus, Calfa Vision).

Le type de prétraitement dépend de l'outil utilisé. Il peut être nécessaire de binariser l'image, ou de la retourner par exemple. Les deux principales opérations réalisées sont : la normalisation et la binarisation des images.

Normalisation : plusieurs tâches spécifiques sont englobées par ce terme :
- la réduction du « bruit » : cela consiste à corriger par exemple une faible luminosité, c'est-à-dire éliminer des « bruits » provenant de facteurs que l'on pourrait qualifier d'extérieurs. Cette réduction passe par l'application de filtres sur l'image ;
- la correction du désalignement (rotation, retournement) ;
- la déformation (qui comprend la correction de la perspective et des distorsions de l'objectif).

Ces deux dernières tâches ont pour but de corriger des déformations introduites pendant la numérisation de l'image.

Binarisation : La binarisation consiste à transformer une image en couleur ou en niveaux de gris en image en noir et blanc en classant chaque pixel d'une image dans deux classes : avant-plan et arrière-plan. La binarisation permet notamment de supprimer du bruit.

Dans tous les cas, l'objectif général de cette étape consiste en l'amélioration de la qualité de l'image. Après cette étape, les images sont prêtes à être traitées.

Étape 3 : Analyse de la mise en page/ Layout Analysis
Cette étape consiste à détecter les zones d'intérêts dans une image : c'est-à-dire les zones de textes, les lignes, les colonnes ; éventuellement différencier les marges, les réclames, etc. Cette étape inclut l'analyse de la mise en page, de la langue et des centres d'intérêts. Ces derniers peuvent être spécialisés en fonction des besoins. Autrement dit, on peut spécialiser une intelligence artificielle pour qu'elle ne récupère que certains types d'informations sur un document. Une fois encore, la détection des zones d'intérêts implique de déterminer au départ quelles zones sont considérées comme des zones d'intérêts (prises en compte ou pas des marges).

L'analyse de la mise en page constitue une étape cruciale de tout système OCR (voir *infra* I.B). L'efficacité de la reconnaissance des caractères et des mots à proprement parler dépend directement de ce travail réalisé sur la segmentation de la page. En d'autres termes, même si le modèle de reconnaissance des caractères est performant, une segmentation incorrecte ou incomplète empêchera tout résultat satisfaisant. On peut considérer plusieurs approches de la segmentation des pages :

Text region/Les « régions de texte »
Il s'agit de détecter les types de texte présents dans l'image : texte principal, marge, titre, réclame, etc., mais aussi tout autre ensemble graphique que l'on souhaite discriminer comme des tableaux, des figures, des lettrines, etc.

Baseline/Les bases d'écriture

L'état de l'art pour la segmentation des lignes de texte dans des documents historiques considérés comme difficiles consiste en ce qu'on appelle « les bases d'écriture » ou *baselines*. La ligne de base est une ligne imaginaire sur laquelle les lettres reposent, bien que les lettres aient fréquemment des parties (appelées descendantes) qui plongent en dessous. Il s'agit donc de la ligne virtuelle joignant le bas des lettres dépourvues de queue[5].

Il existe d'autres approches comme l'approche par *center line* ou *topline*. Il s'agit, de la même manière, de détecter une ligne d'écriture, une ligne fictive pour localiser le texte dans une région de texte. On notera cependant que certaines appellations sont privilégiées pour certaines langues, comme *topline* pour l'hébreu car les caractères sont alignés par le haut des lettres et non par le bas. Ceux qui n'utilisent pas l'approche par ligne préfèrent l'approche dite par *bounding boxes*. Dans ce cas, une sorte de boîte encadre la ligne de part et d'autre. On peut aussi citer une approche par polygones qui encadrent le mot et ses caractères.

Polygones

Souvent les *baselines* ne sont pas suffisantes pour extraire complètement la ligne de texte. Les *baselines* sont donc souvent associés à des polygones.
Ce n'est qu'une fois cette étape essentielle de la segmentation réalisée que la récupération du texte peut se faire. De la reconnaissance de la mise en page dépend l'extraction du texte dans un second temps.

Il faut préciser que l'état de l'art en matière d'analyse de la mise en page est directement tributaire du paradigme de détection des lignes de base défini par les *datasets* cBAD développés dans le cadre de la compétition éponyme Competition on Baseline Detection organisée en 2017 et 2019. Pour les résultats et *datasets* de la compétition de l'ICDAR 2019 : https://zenodo.org/record/2567398#.YbCq4L3P02w

Étape 4 : Récupération du texte – transcription

Cette étape repose sur des algorithmes de prédiction du texte. Le réseau de neurones transforme une information qui est de l'ordre de l'image en information qui est de l'ordre du texte. Dans le cadre de ce rapport, nous n'entrons pas dans les détails techniques expliquant le fonctionnement des différents systèmes d'OCR et de leurs approches car cela nécessiterait d'apporter des précisions mathématiques et informatiques qui dépassent les compétences de l'auteure de ce rapport.

À noter Pour un état résumé de la question, nous invitons les lecteurs à consulter l'article suivant : Francesco Lombardi, Simone Marinai, « Deep learning for Historical Document Analysis and Recognition – A survey », *Journal of Imagining*, 6/110, 2020[6].

En l'absence de modèles généralistes pour la langue et/ou la graphie considérée, il faut passer par une étape d'entraînement. Cette étape suppose de transcrire « à la main » un certain nombre de pages pour obtenir une vérité terrain*. Ces pages servent à entraîner le réseau de neurones. Cela revient à lui apprendre à lire la graphie. On va ensuite tester les résultats obtenus. Si

[5] Définition proposée par *Codicologia* : http://codicologia.irht.cnrs.fr/
[6] https://doi.org/10.3390/jimaging6100110. https://www.mdpi.com/2313-433X/6/10/110#cite

ces résultats ne sont pas complètement satisfaisants mais rendent tout de même compte de performances encourageantes, on pourra se servir de ce premier modèle pour prédire le texte d'une trentaine de pages supplémentaires que l'on va corriger pour accroître le réservoir de vérité terrain jusqu'à parvenir à un modèle satisfaisant.

On peut préciser qu'en fonction des architectures développées, l'approche en matière de reconnaissance de texte pourra privilégier une approche au caractère ou une approche au mot. Il est également possible de fournir au logiciel un modèle de langues pour enrichir son lexique même si cette approche doit être considérée avec précaution dans le cadre de l'HTR en raison notamment des risques de sur-correction. La pertinence de recourir à un modèle de langues pourra donc dépendre du besoin et du projet : s'agit-il de réaliser une édition diplomatique ? Une édition normalisée ?

Il importe d'ajouter par ailleurs que les résultats qui seront obtenus une fois la vérité terrain constituée dépendront bien entendu de l'architecture HTR/OCR utilisée, car ce sont cette architecture et les algorithmes qui la sous-tendent qui privilégient une approche ou une autre, approche qui doit être adaptée à la graphie considérée. Dans le cas de modèles existants pour la graphie des documents étudiés, il est également nécessaire de transcrire quelques pages du corpus considéré pour affiner le modèle utilisé sur le corpus étudié. C'est en particulier nécessaire car, dans le cas des écritures manuscrites, les variations d'un document à l'autre, pour une même graphie, sont nombreuses et il faut donc transcrire un nombre de pages qui dépendra du niveau de polyvalence du modèle HTR existant pour la graphie considérée. Plus le modèle est polyvalent, moins le nombre de pages à transcrire au départ sera élevé. En outre, dans le cas d'un document à la graphie très régulière, particulièrement propre et nette, un modèle qui ne serait pas nécessairement très générique pourra se spécialiser plus rapidement que sur des documents présentant beaucoup de variations.

À noter Les étapes 3 et 4 fonctionnent avec des logiciels généralement fondés sur de l'apprentissage profond* mais pas toujours. Depuis vingt ans, le leader sur le marché est ABBYY, un logiciel propriétaire qui a résolu tous ces problèmes, avant l'explosion de l'intelligence artificielle, en mettant l'accent sur le traitement de l'image.

Étape 5 : Post-traitement pour améliorer les résultats

Cette étape vise à enrichir ou à nettoyer les résultats obtenus. Ce post-traitement peut prendre plusieurs formes. Par exemple, il peut s'agir d'appliquer un modèle de langues et de vérifier les résultats en fonction de ce modèle. Dans le cas où l'OCR ou HTR donnerait le mot « chameaau » qui n'est pas un mot qui existe dans le modèle de langue, il sera corrigé en « chameau ». Le post-traitement peut prendre d'autres formes comme un travail sur la séparation des mots ou une explicitation des abréviations.

Valorisation et exploitation des OCR/HTR

Étape 6 : Valorisation - Et après ? À quoi ça sert un OCR ou un HTR ?

Comme cela a été précisé en introduction, les outils d'OCR et d'HTR permettent d'extraire un texte initialement disponible au format image, pour le transformer en texte qui offre alors des possibilités de fouille. La reconnaissance des écritures imprimées ou manuscrites nous permet d'obtenir un texte que l'on pourrait qualifier de brut. En d'autres termes, le texte obtenu présente un format qui permet différentes sortes de nouvelles actions.

À ce stade, il importe de retenir que le texte acquis est exploitable par le chercheur mais qu'il nécessite des enrichissements supplémentaires afin d'être exploité à des fins de publication ou d'analyse computationnelle. Le texte brut pourra être par exemple structuré pour en arriver à un texte qui reflète les grandes subdivisions matérielles et intellectuelles de la source. On pourra par ailleurs lui appliquer une série de normalisations (lemmatisation par exemple). Cette normalisation permettra d'aboutir à un texte enrichi. Le travail entrepris sur le texte résultat de l'OCR ou de l'HTR dépend du projet de départ et des besoins de chacun. Le travail réalisé sur le texte ne sera pas le même si l'objectif est :

- le développement d'un moteur de recherche pour faire de la recherche plein texte et la mise en place de possibilités de recherche dans les manuscrits via une interface de visualisation des manuscrits ;
- la collation de manuscrits ou l'édition critique de celui-ci ou de ceux-ci.

L'enrichissement du texte extrait peut aussi consister en un encodage de plusieurs types d'entités nommées et en l'alignement entre ces entités nommées et des thesaurus ou des bases de données prosopographiques externes.

Il ne s'agit que de quelques exemples mais ce qu'il faut retenir est que la récupération du texte via l'OCR/HTR autorise de nombreuses opérations qui sont possibles et/ou facilitées par ce mode d'acquisition du texte ; que l'on pense par exemple à des analyses lexicographiques ou à l'élaboration d'une édition critique. Cette valorisation n'est pas l'objet de ce rapport mais nous souhaitons insister sur deux dimensions :

- L'OCR/HTR est l'une des premières étapes du travail sur un texte lorsque cette méthode d'acquisition du texte est privilégiée. C'est aussi pour cette raison qu'il est très important de définir les objectifs scientifiques d'une telle démarche en amont car ceux-ci peuvent avoir un impact sur la nature du texte qui sera acquis.
- La conversion du texte accessible au format image en format texte de manière plus ou moins automatique facilite grandement les étapes suivantes visant à le valoriser, en particulier quand ce texte n'est pas disponible autrement qu'à l'état d'images numérisées ou d'édition papier.

Il existe un certain nombre d'architectures disponibles sur le web qui permettent de faire de l'OCéRisation et qui atteignent des CER* de 5 %. Il y a beaucoup d'approches différentes et nous ne sommes pas entrée dans leurs détails pour privilégier une approche globale et générale de la logique des étapes de traitement. Nous reviendrons sur les outils dans le second temps de ce rapport.

029 [I] B
ENJEUX DE L'HTR ET DE L'OCR POUR LES LANGUES NON LATINES EN 2021, AVEC UN ACCENT PARTICULIER SUR L'ARABE

Les langues non-latines à l'heure de l'OCR/HTR

Il faut dire d'emblée que la plupart des écritures et des langues ne disposent pas de système d'OCR efficient. Par ailleurs, la plupart de la littérature mondiale n'a pas été encore numérisée. En matière de documents arabes anciens, au premier titre les manuscrits, nous disposons d'importants corpus numérisés mais ceux-ci ne représentent encore qu'une part restreinte de l'ensemble des fonds existants. Pour ne donner qu'un exemple avec lequel nous sommes familiers : alors que la BULAC détient plus de 2 500 manuscrits arabes, 273 (en date du 1er novembre 2021) sont actuellement numérisés et consultables sur le site de la BINA (la bibliothèque numérique de la BULAC), soit environ 10 %.

Concernant l'écriture arabe, il importe de préciser que la graphie arabe n'est pas uniquement utilisée pour écrire la langue arabe. Les caractères arabes ont servi, moyennant quelques adaptations, pour d'autres langues que l'arabe comme le turc ottoman ou osmanli (langue officielle de l'Empire ottoman jusqu'en 1923) ; le persan et l'ourdou (langue parlée dans le nord de l'Inde et au Pakistan), deux langues indo-européennes qui s'écrivent en alphabet arabo-persan.

Il faut par ailleurs noter qu'il n'existe pas un seul système d'écriture arabe. Les formes et les styles de graphies sont multiples et leur utilisation varie en fonction des périodes historiques, des types de document concernés et/ou de la région considérée. Les styles les plus connus sont le coufique, le *naskhī*, le *nastaʿlīq*. En matière d'écriture arabe imprimée, il existe également différentes polices de caractères dont l'identification et le classement peuvent être déterminants pour l'élaboration de modèles OCR adaptés à l'arabe imprimé.

L'un des enjeux pour les langues non-latines en matière d'OCR et d'HTR tient à ce que la plupart des modèles génériques, que ce soit pour les écritures ou les mises en page, sont conçus pour les graphies latines et ne sont souvent pas adaptés aux spécificités de ces écritures. Aussi, les documents historiques en écritures non-latines exigent généralement de créer des jeux de données à partir de zéro.

Enjeux techniques et pratiques

Transcription et compétences

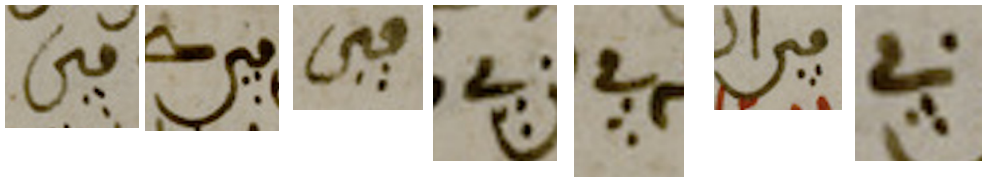
La transcription des écritures non latines manuscrites, telles qu'on peut les lire dans les manuscrits par exemple, exigent des compétences paléographiques importantes. Ces compétences de lecture des manuscrits sont essentielles, en particulier pour participer à élaborer la vérité terrain* et les modèles HTR correspondant aux écritures considérées. Or pour les graphies non-latines, le nombre des spécialistes n'est pas forcément suffisant au regard des besoins, et ce nombre diminue encore lorsque l'on cherche à identifier les spécialistes disponibles pour réaliser ces tâches. L'existence d'un vivier de philologues, formés, compétents et disponibles varie bien entendu en fonction des langues considérées. Chahan Vidal-Gorène, fondateur et président de Calfa, faisait par exemple remarquer que le nombre de spécialistes pour l'arménien était très limité par rapport aux besoins[7]. Même si, pour les manuscrits arabes, on peut supposer que le nombre de personnes compétentes et disponibles pour travailler sur ces textes est plus important au regard du dynamisme du champ des études arabes et assimilées, il n'en demeure pas moins que ce constat est partagé, en raison notamment d'un déficit des études en philologie arabe et du nombre de personnes disponibles disposant effectivement de ces compétences. Matthew Miller, Assistant Professor of Persian Literature and Digital Humanities à l'Institut Roshan pour les études persanes de l'Université du Maryland, faisait un constat similaire pour l'arabe aux États-Unis. Cette nécessaire main d'œuvre de spécialistes plus ou moins manquante participe à expliquer aussi que certains collègues cherchent à développer des moyens pour créer des modèles en évitant autant que possible la tâche de transcription (*voir infra* II.E concernant les projets d'OpenITI).

Caractéristiques des écritures cursives arabes

Les écritures cursives, en comparaison de l'imprimé, se caractérisent en général par une très grande variété. Une même main ne produit pas une écriture constante tout au long d'un manuscrit, ni même d'une seule page. On en donne pour preuve, dans un seul manuscrit arabe d'une même main issu des fonds patrimoniaux de la BULAC, la réalisation de la préposition في qui peut varier de la sorte :

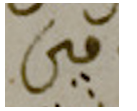
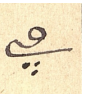
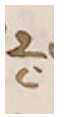
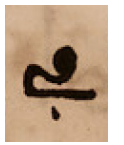
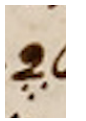
[7] Chahan Vidal-Gorène, « HTR et langues peu dotées, exemple de l'arménien », intervention dans le cadre de l'ANF Digital Areal, le 9 juin 2020 : https://www.youtube.com/watch?v=v91-BWIARNA

Tableau n° 1 : Réalisations de في dans le MS.ARA.609



Dans 9 manuscrits différents mais tous copiés dans une même famille d'écriture, dite écriture maghrébine :

Tableau n° 2 : Réalisations de في dans neuf manuscrits différents

MS.ARA.609	MS.ARA.1977	MS.ARA.417
		
MS.ARA.1926	**MS.ARA.1957**	**MS.ARA.1922**
		
MS.ARA.1925	**MS.ARA.1929**	**MS.ARA.1944**
		

À cette grande variété d'écritures et aux variations dans la forme des lettres au sein même d'un seul type d'écriture ou de main, il faut ajouter d'autres difficultés potentielles pour l'HTR en matière de détection des langues non-latines :

- ligatures et glyphes spéciaux ;
- *scriptio continua* (la *scriptio continua* désigne l'absence de ponctuation, d'espaces et de séparation entre les mots ou entre les phrases) ;
- grande variation dans l'orthographe, morphologie, syntaxe ;
- coquilles du copiste ;
- césures, abréviations, lettres suscrites et souscrites ;
- titres stylisés ou calligraphiés.

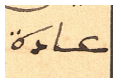
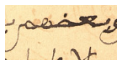
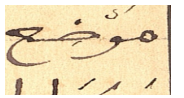
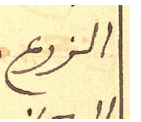
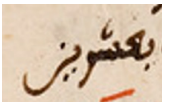
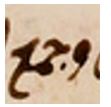
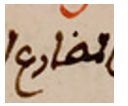
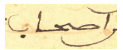
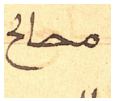
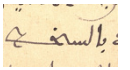
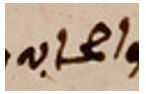
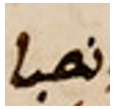
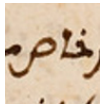
Toutes ces difficultés potentielles se retrouvent dans les manuscrits arabes (voir III) et y sont parfois exacerbées à l'image notamment des lettres et des

mots suscrits ou souscrits ainsi que des chevauchements de ligne à ligne dus notamment aux courbes de certaines des lettres. L'absence de césure entre les mots est un autre des enjeux particulièrement représentés dans le cas des manuscrits arabes. Une autre des difficultés rencontrées en matière de détection des graphies arabes est liée à l'utilisation des signes diacritiques, permettant de distinguer les lettres. De nombreuses ambiguïtés dans la prédiction des graphies arabes manuscrites résultent de points diacritiques erronés ou aléatoirement disposés dans le mot[8].

Ajoutons par ailleurs deux caractéristiques des graphies arabes :

- Ce sont des graphies attachées et le nombres de ligatures est donc non négligeable et connaît des variations importantes.
- La réalisation des lettres en arabe varie en fonction de la position de la lettre dans le mot. Une même lettre aura donc en moyenne trois réalisations différentes, sans compter les ligatures [TABLEAU N°3].

Tableau nº 3 : Exemple des lettres ص et ع dans deux manuscrits arabes issus des collections patrimoniales de la BULAC

	Initiale	Médiane	Finale	Isolé/ligature
ع				
MS.ARA.417				
MS.ARA.1929				
ص				
MS.ARA.417				
MS.ARA.417				

Benjamin Kiessling en rend également compte dans sa thèse. Nous lui empruntons le tableau reproduit ci-après[9].

[8] Des constats similaires sont faits dans B. Kiessling, D. Stökl Ben Ezra, M. T. Miller, « BADAM: A Public Dataset for Baseline Detection in Arabic-script Manuscripts », 2019. https://arxiv.org/ftp/arxiv/papers/1907/1907.04041.pdf ; B. Kiessling, *Advances in Optical Character Recognition for Historical Arabic Documents, op. cit.*, p. 46 et ss ; M. Ibn Khedher, H. Jmila, M. A. El-Yacoubi, « Automatic processing of Historical Arabic Documents: A comprehensive Survey », *Pattern Recognition*, 100, 2020, p. 1-17. https://www.sciencedirect.com/science/article/pii/S0031320319304455

[9] B. Kiessling, *Advances in Optical Character Recognition for Historical Arabic Documents, op. cit.*, p. 47.

Tableau nº 4 : Les 28 lettres de l'alphabet arabe (tableau repris à Benjamin Kiessling)

isolated	initial	medial	final	name	transliteration
ا	ا	ـا	ـا	ʾalif	ā
ب	بـ	ـبـ	ـب	bāʾ	b
ت	تـ	ـتـ	ـت	tāʾ	t
ث	ثـ	ـثـ	ـث	thāʾ	th
ج	جـ	ـجـ	ـج	jīm	j
ح	حـ	ـحـ	ـح	ḥāʾ	ḥ
خ	خـ	ـخـ	ـخ	khāʾ	kh
د	د	ـد	ـد	dāl	d
ذ	ذ	ـذ	ـذ	dhāl	dh
ر	ر	ـر	ـر	rāʾ	r
ز	ز	ـز	ـز	zayn	z
س	سـ	ـسـ	ـس	sīn	s
ش	شـ	ـشـ	ـش	shīn	sh
ص	صـ	ـصـ	ـص	ṣād	ṣ
ض	ضـ	ـضـ	ـض	ḍād	ḍ
ط	طـ	ـطـ	ـط	ṭāʾ	ṭ
ظ	ظـ	ـظـ	ـظ	ẓāʾ	ẓ
ع	عـ	ـعـ	ـع	ʿayn	ʿ
غ	غـ	ـغـ	ـغ	ghayn	gh
ف	فـ	ـفـ	ـف	fāʾ	f
ق	قـ	ـقـ	ـق	qāf	q
ك	كـ	ـكـ	ـك	kāf	k
ل	لـ	ـلـ	ـل	lām	l
م	مـ	ـمـ	ـم	mīm	m
ن	نـ	ـنـ	ـن	nūn	n
ه	هـ	ـهـ	ـه	hāʾ	h
و	و	ـو	ـو	wāwʾ	w/ū
ي	يـ	ـيـ	ـي	yāʾ	y/ī

Le fait que l'on ait affaire à une graphie attachée dans laquelle la forme du caractère varie en fonction de sa position dans le mot explique notamment que les approches OCR/HTR aux caractères ne sont pas les plus adaptées pour obtenir des résultats de reconnaissance des écritures exploitables. L'approche au mot semble donc mieux pensée pour traiter les graphies arabes[10]. La question de la meilleure approche linguistique pour les graphies arabes doit, dans tous les cas, être sérieusement posée.

[10] Sur ce point, voir la troisième partie de ce rapport et les résultats obtenus dans le cadre du projet RASAM.

C'est par exemple l'objet de l'article publié par Thomas Milo et Alicia Gonzalez Martinez en 2020 dans la revue *Égypte/Monde arabe*, « A New Strategy for Arabic OCR based on Script Analysis and Synthesis[11] ».

Dans cet article, les auteurs présentent une stratégie visant à améliorer l'OCR en arabe qui consiste à éliminer temporairement les points de désambiguïsation afin de réduire les classes de graphèmes (plus petite unité d'un système d'écriture) qui partagent le même élément de base à des archigraphèmes uniques. Par ailleurs, le comportement contextuel des archigraphèmes arabes est redéfini comme une fusion en blocs de lettres selon un système basé sur des règles appelées « grammaire de script ». Ce bloc de lettres est ainsi défini comme l'unité minimale de la formation de l'écriture arabe. Autrement dit, l'unité minimale n'est plus le caractère mais le bloc de lettres. Les auteurs donnent l'exemple suivant dans le résumé de leur article : « Par exemple, le mot **بحوث** se compose de deux blocs de lettres, des groupes d'allographes fusionnés entourés d'un espace graphique, **بحو** et **ب** (BGW B) ». À partir d'un corpus arabe d'environ 85 millions de mots, ils ont constitué une liste d'environ 47 000 blocs de lettres archigraphémiques uniques. La seconde étape consiste à synthétiser toutes ces formes théoriques pour chaque bloc de lettres à partir de modèles informatiques de styles d'écriture arabes spécifiques (*ruqʿa*, *naskhī*, *nastaʿlīq*). La désambiguïsation intervient dans un dernier temps en utilisant des informations linguistiques, recueillies en partie dans le corpus de mots considéré. Dans cet article, l'approche consiste donc à partir d'un répertoire de formes théoriques. Il ne s'agit donc pas d'apprentissage profond* et d'intelligence artificielle.

L'équipe de Calfa, comme cela sera montré dans la dernière partie de ce rapport (voir III), propose une architecture HTR pour les écritures manuscrites dites maghrébines qui privilégie une approche aux mots et aux groupes de mots plutôt qu'aux caractères, en utilisant l'intelligence artificielle. Cette approche s'est avérée pertinente au vu des résultats obtenus. Dans ce cas, l'intelligence artificielle se constitue son propre lexique de mots et de mots en contexte.

Un enjeu de taille : l'analyse de la mise en page

Les spécialistes de l'HTR et de l'OCR expliquent que les difficultés les plus importantes ne concernent pas la reconnaissance de caractères en eux-mêmes. Il est aujourd'hui possible de développer des modèles avec une quantité de données limitées. L'enjeu de taille réside dans l'analyse de la mise en page (voir étape 3 de la chaîne de traitement), autrement dit la détection des zones de textes puis l'extraction des lignes de texte. La première difficulté est donc celle de la segmentation, la capacité à distinguer ce qui est « texte » et ce qui n'est pas du texte.

Prenons deux pages d'un même manuscrit issu des collections patrimoniales de la BULAC, le MS. ARA. 1947 [FIG.2]. Il s'agit des *Maqāmāt d'al-Ḥarīrī*, un ouvrage de la littérature arabe classique très connu. La page de gauche (97) présente une mise en page relativement simple avec 14 lignes par page, une réclame en bas à gauche et une note en haut à droite. À

[11] A. González Martínez, T. Milo, « A New Strategy for Arabic OCR based on Script Analysis and Synthesis », *Égypte/Monde arabe*, Troisième série, 22, 2020, p. 21-30. https://journals.openedition.org/ema/13146?lang=en

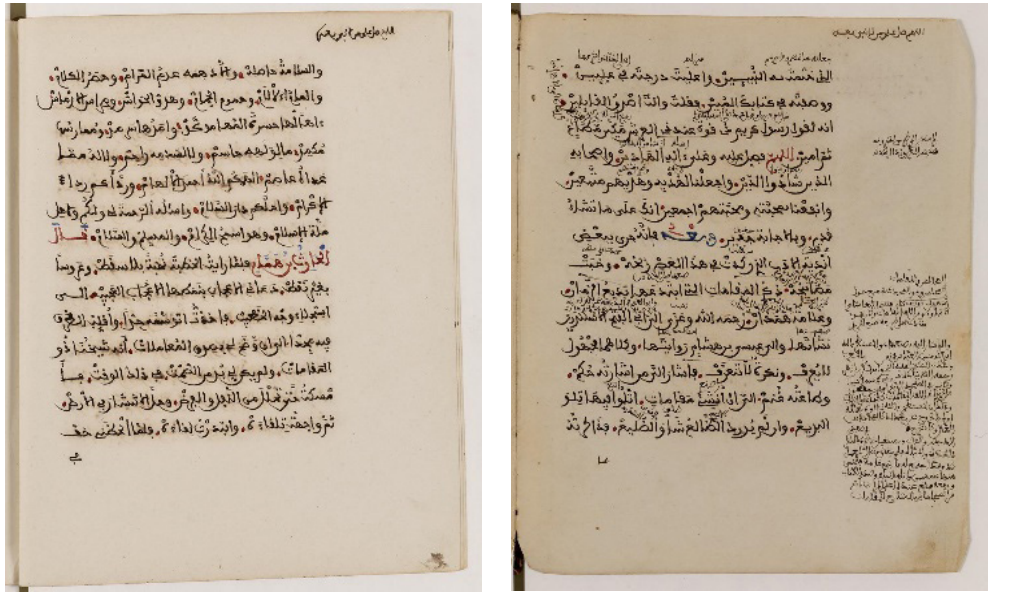
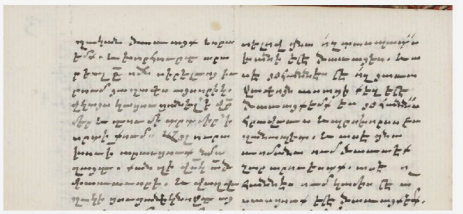


Fig. 2
Deux folios du MS. ARA.1947 : Al-Ḥarīrī (1054-1122), *al-Maqāmāt al-Ḥarīrī*, Bulac, MS.ARA.1947, folio 2v (à droite) et folio 97v (à gauche)

l'exception du fait que le texte se lit de droite à gauche, la mise en page ne présente pas de difficultés majeures a priori. Sur la page de droite (4), extraite du même manuscrit, la mise en page est plus complexe en raison, d'une part, de l'ajout de notes marginales, qui ont néanmoins le mérite d'être clairement identifiables, et surtout d'un nombre important de notes interlinéaires dont le texte est dans une moitié des cas rédigé en diagonale. Cette seconde mise en page représente de nombreux enjeux pour l'étape d'analyse de la mise en page car ces notes interlinéaires compliquent la reconnaissance, sinon de la zone principale du texte, du moins des lignes. Cette difficulté est renforcée dans le cas où le chercheur souhaite singulariser ces notes au moment de l'analyse de la mise en page et doit donc apprendre à l'intelligence artificielle à les singulariser.

Il n'est pas inutile de rappeler qu'un algorithme de pure reconnaissance de caractères fait de la reconnaissance de caractères que l'on pourrait qualifier de « pure et brute ».

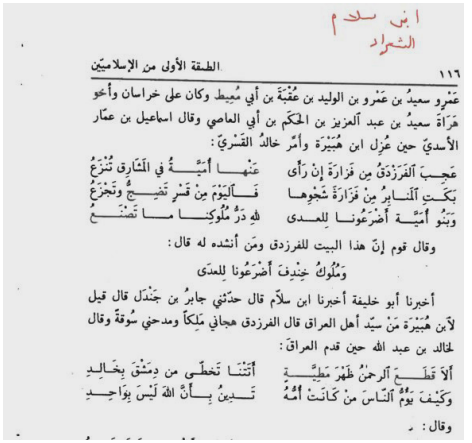
Dans l'exemple ci-après, emprunté à Chahan Vidal-Gorène [FIG.3], on peut observer une page d'un manuscrit arménien qui contient deux colonnes. L'algorithme d'analyse de la mise en page n'a pas fonctionné et n'a pas pris en compte cette mise en page en colonne. Le résultat n'est donc pas satisfaisant. Pourtant le taux d'erreur par caractère (CER*) est de 0 %. Il n'y a donc aucune erreur de reconnaissance au niveau des caractères mais le texte se présente de manière compactée. Cette analyse de la mise en page fautive conduit par ailleurs à la création de mots qui ne peuvent pas être analysés car ce ne sont pas des mots.



ms M5, f. 3b (Matenadaran, Erevan)

ղականծառայքնորասելովգնաոչտասպատս
եմք.ևխորհուրդարախանիեթէծառայես.ևա
րեալըոմնսիրելույիւեյօհաննէսթէոչյառա
րումյաղագայսորիկ.ջագոյնասացիքեզեթէ
վկայսկացուցանել'իվրծառայքեմքեսյօհաննս
մերևդումիտրտմիր'իհրավառսևապրոխորսսբա
սրտիքում..Այլուրաղանապետ.ևասէգնա
խուիարասցուքնմառոմանաումծառաէք
զայդ.քանզիվսկածչարարուեստք.ասէո
միառաւորի.ևվաղվահաննէսումկամիսթէա
ղակիցուցանէկնոջդայսասցուքեթէծառայքեմք.

Fig. 3
Exemple présenté par Chahan Vidal-Gorène au cours de sa conférence : « HTR/OCR pour graphies non latines : approches et bonnes pratiques »[12]



ابن سلام
الشعراء

الطبقة الأولى من الإسلاميّين ١١٦

عَمْرو سعيدُ بن عَمْرو بن الوليد بن عُقْبَةَ بن أبي مُعيط وكان على خراسان وأخو هَرَاةَ سعيدُ بن عبد العزيز بن الحَكَم بن أبي العاصي وقال اسماعيل بن عمّار الأسديّ حين عَزَل ابن هُبَيْرَة وأمّر خالدَ القَسْريّ:

عَجِبَ ٱلفَرَزْدَقُ مِن فَزارَةَ إِنْ رَأَى ... عَنْهـا أُمَيَّـةُ فِي المَشَارِقِ تُنْزَعُ
بَكَـتِ ٱلمَنَـابِرُ مِنْ فَزارَةَ شَجْوِهـا ... فَـٱلْيَوْمَ مِنْ قَسْرٍ تَضِـجُّ وتَجْزَعُ
وبَنُو أُمَيَّـةَ أَضْرَعُونَـا لِلعِـدى ... لِلهِ دَرُّ مُلُوكِنَـا مـا تَصْنَـعُ

وقال قوم إنّ هذا البيت للفرزدق ومَن أنشده له قال:

وَمُلُوكُ خِنْدِفَ أَضْرَعُونا لِلعِدَى

أخبرنا أبو خليفة أخبرنا ابن سلّام قال حدّثني جابرُ بن جَنْدَل قال قيل لابن هُبَيْرَة مَنْ سيّد أهل العراق قال الفرزدق هجاني مَلِكاً ومدحني سُوقةً وقال لخالد بن عبد الله حين قدم العراق:

أَلَا قَطَـعَ ٱلرحمنُ ظَهْرَ مَطِيَّـةٍ ... أَتَتْنَـا تَخطّـى من دِمَشْقَ بِخَـالِدِ
وكَيْفَ يَؤُمُّ النّاسَ مَنْ كَانَتْ أُمُّهُ ... تَـدِينُ بِـأَنَّ اللهَ لَيْسَ بِوَاحِـدِ

وقال: ر

١١٩

ابن سلام

الشرار

الطبقة الأولى من الإسلاميين عمرو سعيد بن عمرو بن الوليد بن عقبة بن أبي مميط وكان على خراسان وأخو مراة سعيد بن عبد العزيز بن الحكم بن أبي العاصي وقال اسماعيل بن عمار الأسدي حين عزل ابن هبيرة وأمر خالد القري: عجب الفرزدق من فزارة إن رأى عنها أمني في الشارق تنزع

بكت المنابر من قرار شجو ما اليوم من قشر تضج وتجزع

وبنو أية أضرعونا للعدي له دز مونا ميات وقال قوم إن هذا البيت للفرزدق ومن أنشده له قال : وممو خندف أضرعونا للعدى أخبرنا أبو خليفة أخبرنا ابن . **سلام قال حدثني جابر بن جندل قال قيل** لابن هبيرة من سيد أهل العراق قال الفرزدق هجاني ملكة ومدحني شوقة وقال لخالد بن عبد الله حين قدم العراق : . ألا قطع الرحمن ظهر مطية أنا تخطى من تنشق بخالد وكيف يؤم الناس من كانت أنه تبين بأن الله ليس بواجي وقال : ه العميرى ليث كان بجيلة

Fig. 4
Google Drive et l'OCR de l'arabe : exemple de résultat (original à gauche et OCR à droite)

[12] Conférence donnée le 8 décembre 2020 dans le cadre de la formation en ligne « Étudier et publier les textes arabes avec le numérique » : https://youtu.be/iHCTx4yJwAs

Dans cet autre exemple, nous avons utilisé la fonction OCR de Google Drive, qui utilise la technologie de Tesseract (cela consiste à charger un document PDF sur le Drive proposé par Google et d'ouvrir le document en Word. Cette manipulation est permise par l'utilisation d'algorithme de reconnaissance optique des caractères) [FIG.4]. Il ne s'agit pas de commenter ici la performance de l'algorithme en matière de reconnaissance automatique des caractères mais plutôt de montrer que la mise en page de l'édition n'a pas été correctement détectée. Vous observez à gauche la page imprimée initialement au format PDF et à droite le résultat de la conversion du document image en texte. La mise en page n'est pas respectée, en particulier celle des vers de poésie présentés en colonnes dans la page de gauche.

En d'autres termes, même dans le cas de l'imprimé, les langues non latines posent des problèmes spécifiques et l'analyse de la mise en page est l'un d'entre eux. L'importance de cet enjeu de la segmentation pour les manuscrits et les documents imprimés pour les langues non latines est reconnue par les collègues du champ investis dans les travaux de recherche sur l'OCR et l'HTR. C'est d'ailleurs la raison pour laquelle certains jeux de données se concentrent principalement sur l'analyse de la mise en page à l'image de BADAM en 2019[13]. Dans cet article, les auteurs présentent un jeu de données (voir II) dédié à l'analyse de la mise en page des manuscrits arabes.

Enjeux liés à la nature des documents historiques

En plus des dégradations possibles qui peuvent advenir au moment de la numérisation et qui nécessitent alors d'être corrigées au moment du prétraitement des images, d'autres types de dégradations peuvent concerner les manuscrits numérisés ou à numériser à des fins d'HTR :

- des dégradations chimiques provoquées par les variations de température, l'humidité, la pollution ou la lumière. Elles entraînent généralement une modification des couleurs voire une décoloration et peuvent avoir des impacts sur les encres et les pigments.
- des dégradations biologiques causées par des animaux ;
- des dégradations provoquées par les hommes à l'image des annotations ajoutés aux documents, les rayures[14].

[13] Voir B. Kiessling, D. Stökl Ben Ezra, M. T. Miller, « BADAM: A Public Dataset for Baseline Detection in Arabic-script Manuscripts », art. cit.

[14] M. Ibn Khedher, H. Jmila, M. A. El-Yacoubi, « Automatic processing of Historical Arabic Documents: A comprehensive Survey », art. cit., p. 3.

[II]
ÉTAT DES LIEUX SUR L'OCR/HTR DE L'ARABE

Cette partie du rapport entend offrir un aperçu de l'état de l'art sur la question de l'OCR/HTR de l'arabe dans les documents historiques. Le sujet peut être traité à partir de plusieurs perspectives qui, dans une certaine mesure, s'entrecroisent. La structuration de cette partie du rapport a donc été pensée pour rendre compte au mieux de la manière dont les recherches en matière d'HTR/OCR pour les langues non-latines, en particulier l'arabe, s'organisent.

Par certains égards, la typologie proposée invite à la redondance car certains projets scientifiques produisent ou ont produit des jeux de données ou s'engagent dans le développement d'interfaces d'annotation permettant la création de modèles HTR. On pourra donc trouver un même projet mentionné dans la catégorie « *Dataset* » et dans la catégorie « Projets scientifiques ». Cependant, il nous a semblé important de privilégier cette approche car ces différentes catégories reflètent la multiplicité des entrées dans la question de l'HTR et, dans le même temps, comment le tout fait ou ne fait pas système.

Notons qu'à quelques exceptions près, nous mettons l'accent sur les outils, *datasets* et interfaces en accès libre, au code source ouvert et gratuits, même si nous n'occulterons pas les outils obéissant à d'autres modèles économiques et juridiques.

L'ambition de cette partie qui propose des « fiches » succinctes sur les différentes entrées de chaque catégorie est de fournir un guide et des jalons pour mieux comprendre le champ de la recherche sur l'HTR/OCR de l'arabe et d'agréger quelques ressources et liens utiles à celui ou celle qui souhaiterait en apprendre plus.

Cette partie s'organise de la manière suivante : le premier point abordé est celui des conférences internationales. Celles-ci rythment à calendrier fixe la recherche dans le domaine et permettent de se faire une idée de l'état et de l'avancement des recherches sur la question. Elles constituent donc de bons indicateurs des recherches en cours. Dans un deuxième temps, nous présenterons les jeux de données disponibles pour l'arabe. Ces jeux de données constituent la condition de possibilité des développements des modèles HTR/OCR, ils sont donc indispensables. Le troisième point s'intéresse aux outils disponibles pour le chercheur qui souhaite OCéRiser ou HTéRiser un document et mettra donc en évidence les logiciels et les plateformes existantes. La dernière catégorie de cette partie abordera quant à elle la manière dont les recherches en HTR/OCR sur l'arabe sont intégrées dans certains projets de recherches européens et internationaux.

041 [II] A
LE POINT SUR LES CONFÉRENCES

Les chercheurs en SHS et en particulier en histoire se sont saisis récemment des enjeux de la reconnaissance automatique des caractères pour leur travail de recherche comme en témoigne la naissance en 2014 de la conférence DATeCH : Digital Access to Textual Cultural Heritage, là où les deux autres conférences (ICDAR et ICFHR), plus anciennes, réunissent en priorité des spécialistes des technologies.

Il importe de préciser que, dans les conférences plus généralistes ICDAR et ICFHR, les recherches sur l'OCR et HTR de l'arabe ne portent qu'en minorité sur l'étude de documents historiques. Il faut distinguer ce qui relève de recherche de vérité terrain* ou d'amélioration de points précis de segmentation, de *post-processing*, etc. et ce qui touche à des réponses à des besoins d'OCR et d'HTR « contemporains » comme l'identification de panneaux de signalisation, que l'auteure de ce rapport n'est pas en mesure de commenter. Aussi, nous mettrons en particulier l'accent sur les recherches qui s'intéressent aux documents historiques, et en particulier aux manuscrits.

Il faut noter néanmoins que nous pouvons souvent rattacher les articles, posters et recherches présentés lors de ces conférences, en particulier lorsqu'ils touchent à l'étude des documents historiques, à des projets de recherche ou d'infrastructures.

ICDAR et ICFHR

Deux conférences internationales traitent directement des questions d'OCR et d'HTR. Elles permettent par ailleurs de prendre la mesure des avancées sur ces questions et des enjeux et questions traitées. Il s'agit de l'International Conference on Document Analysis and Recognition (ICDAR) et de l'International Conference on Frontiers in Handwriting Recognition (ICFHR).

Précisons d'emblée que ces conférences ne sont pas seulement l'occasion de présenter des projets en cours et les résultats de leurs recherches. Elles sont aussi l'occasion de faire effectivement avancer la recherche lors des différentes compétitions qui ont lieu.

Ces deux conférences s'insèrent dans le contexte particulier de l'IAPR: International Association for Pattern Recognition[15]. Officiellement fondée en 1978, l'association internationale pour la reconnaissance des formes/modèles est une association internationale composée d'organisations scientifiques et professionnelles à but non lucratif (de portée nationale, multinationale ou internationale) s'intéressant à la reconnaissance des *patterns*, à la vision par ordinateur et au traitement des images au sens large.

L'association est organisée en comités techniques/*technical committees* qui s'intéressent chacun à un des aspects du très large domaine de *pattern recognition*. Au fur et à mesure, les comités se sont d'ailleurs multipliés en relation avec l'intérêt grandissant pour le domaine. En 2008, l'IAPR compte 18 comités techniques dont le TC11 – *Reading systems*[16].

IAPR-TC11 concerne les théories et les applications des systèmes de lecture. Ce comité technique s'intéresse aux travaux sur le traitement des documents et à l'analyse et à la reconnaissance des informations contenues dans les documents (OCR et HTR). Ces sujets concernent à la fois la reconnaissance de textes imprimés et la reconnaissance de documents manuscrits, à la fois hors ligne (à partir d'images) et en ligne (à partir de dispositifs informatiques basés sur le stylo et le toucher). Tous les niveaux du traitement des documents sont couverts (images, coordonnées planes, extraction et sélection des caractéristiques, algorithmes pour segmenter, analyser le contenu des documents).

Les conférences de l'ICDAR et l'ICFHR font partie des activités menées par ce comité technique[17]. Elles se tiennent tous les deux ans dans des villes différentes.

ICDAR ou International Conférence on Document Analysis and Recognition

La première conférence s'est tenue en 1991 en France. En 2021, la 16e conférence a eu lieu à Lausanne. Évènement international de référence, la conférence réunit les scientifiques et les praticiens impliqués dans l'analyse et la reconnaissance de documents[18].

ICFHR ou International Conference on Frontiers in Handwritten Recognition

L'ICFHR était auparavant un atelier (*workshop*) intitulé IWFHR International Workshop on Frontiers of Handwriting Recognition, qui s'est tenu de 1990 à 2006. Il s'agit d'une conférence depuis 2008. Contrairement à l'ICDAR qui traite des questions de reconnaissance et d'analyse des documents en général, l'ICFHR est spécialement intéressée par le domaine de la reconnaissance des écritures manuscrites.

[15] https://iapr.org/aboutus/
[16] http://www.iapr-tc11.org/mediawiki/index.php?title=IAPR-TC11:Reading_Systems
[17] http://www.iapr-tc11.org/mediawiki/index.php/Conferences
[18] https://www.icdar.org

Quelques remarques sur l'OCR/HTR de la graphie arabe au cours de ces conférences

Afin de prendre la mesure des recherches menées en matière d'OCR/HTR pour les graphies arabes, nous avons consulté les sommaires des actes des conférences de l'ICDAR et de l'ICFHR de ces dix dernières années. Le début des années 2010, 2012 notamment, marque un tournant sur les questions d'OCR et d'HTR car il correspond à la popularisation des approches en apprentissage profond* et le recours à l'intelligence artificielle.

Tableau nº 5 : Recherches présentées sur l'OCR/HTR des graphies arabes lors des conférence ICDAR et ICFHR depuis 2011

Conférence	Nombre d'articles en lien avec l'OCR/HTR de l'arabe
ICDAR 2011	7 articles et 4 compétitions
ICFHR 2012	7 articles
ICDAR 2013	13 articles et 1 compétition
ICFHR 2014	11 articles
ICDAR 2015	16 articles
ICDAR 2017	4 articles et 2 compétitions
ICFHR 2018	1 article (RASM)
ICDAR 2019 (1re participation du *workshop* ASAR)	2 articles
ICFHR 2020	1 article
ICDAR 2021	6 articles et 3 compétitions (tous dans le cadre d'ASAR)

Le tableau ci-dessous résume le dépouillement effectué :

Ce rapide recensement semble rendre compte d'un pic des recherches en 2015 puis d'une diminution, du moins en matière de représentation, des contributions à ces conférences. L'hypothèse que nous formulons est que la popularisation des approches en intelligence artificielle et apprentissage profond s'est accompagnée, dans un premier temps, d'une accélération des recherches sur l'OCR/HTR de l'arabe mais qu'elle semble connaître depuis quelques années une forme de stagnation.

Il est à noter que la plupart des contributions ne concernent pas les documents historiques et qu'elles sont plutôt consacrés, soit à des recherches de solution sur des points très précis, soit à des questionnements contemporains (identification des écritures manuscrites contemporaines, lecture de panneaux de signalisation, etc.). En d'autres termes, les recherches sur l'OCR et l'HTR de l'arabe constituent un champ de recherche actif du côté des ingénieurs en informatique. Cependant, la consultation des sommaires des actes de ces conférences rend compte d'un champ assez peu actif du côté des chercheurs en sciences humaines. Autrement dit, les chercheurs en humanités « arabes » ne se sont investis que de manière limitée sur la question de l'OCR et de l'HTR des graphies arabes pour leurs documents historiques. Cette observation est confirmée lorsque l'on se penche sur des conférences plus spécialisées comme DATeCH.

DATeCH : Digital Access to Textual Cultural Heritage

Plus récemment, en 2014, la conférence DATeCH a vu le jour. Cette conférence rassemble des chercheurs et des praticiens à la recherche d'approches innovantes pour la création, la transformation et l'exploitation de documents historiques sous forme numérique. Cette conférence interdisciplinaire, se déroule à l'intersection de l'informatique, des humanités (numériques) et des études sur le patrimoine culturel.

Trois conférences ont eu lieu : à Madrid en 2014, à Göttingen en 2017 et à Bruxelles en 2019. La dernière conférence en date était organisée conjointement par le centre de compétences IMPACT, l'Instituut voor de Nederlandse Taal, DARIAH-BE (Digital Research Infrastructure for the Arts and Humanities, initiative européenne pour soutenir la recherche en sciences humaines et sociales sur les objets numériques) et CLARIN-Flandres (Common Language Resources and Technology Infrastructure, infrastructure européenne créée pour le partage d'outils autour du langage). La 4e conférence était prévue en 2021 à Bologne mais elle ne semble pas avoir eu lieu.

Communications/papers en lien avec l'arabe dans cette conférence

Lors de la première conférence DATeCH, aucune des recherches présentées ne concernait des documents en arabe. On remarque cependant qu'en 2017, une recherche concernait l'arabe[19] et qu'en 2019, on en comptait deux[20].

ASAR : International Workshop on Arabic and Derived Script Analysis and Recognition

L'Atelier sur la reconnaissance et l'analyse des écritures arabes et dérivées a organisé cette année sa quatrième édition, dans le cadre de la conférence internationale ICDAR, à Lausanne (septembre 2021).

La première édition date de 2017 et s'est tenue à Nancy dans le cadre d'une collaboration entre le laboratoire LORIA de l'Université de Lorraine et le REGIM-LAB de l'Université de Sfax. Cet atelier était alors soutenu par l'Association professionnelle des ingénieurs électriciens et électroniciens (IEEE), notamment les sections françaises et tunisiennes et par IAPR (mentionné plus haut). La deuxième édition a eu lieu à Londres en 2018 à l'Institut Alan Turing.

Comme le nom de l'atelier l'indique, celui-ci s'intéresse aux méthodes de reconnaissance et d'analyse des documents en graphies arabes ou assimilées. On notera que l'atelier est principalement suivi et animé par des ingénieurs en informatique et que la plupart des communications et des articles ne concernent pas les documents historiques. Ainsi dans les actes de l'atelier de la première édition paru sur IEEE Xplore[21], sur les

[19] A. Gonzalez Martinez, T. Feige, T. Eich, « Clear-cut methodology for Arabic OCR and post-correction with low technical skilled annotators », *DATeCH2017: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage Göttingen 2017,* New York, ACM, 2017, p. 67-70.

[20] *DATeCH2019: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage,* New York, ACM, 2019. Dans la section « Applications » : E. Mohamed, Z. A. Sayyed, « Arabic-SOS: Segmentation, Stemming and Orthography Standardization for Classical and pre-Modern Standard Arabic », p. 27-32.
Dans la section « Digitisation of historical languages » : A. Gonzalez Martinez, T. Milo, « A New Strategy for Arabic ICR: Archigraphemes, Letter Blocks, Script Grammar and shape synthesis », p. 93-96. Ce dernier article fait directement écho à celui qui a été publié en 2020 par la revue *Égypte/Monde arabe* dans le dossier intitulé « L'archivage numérique dans le monde arabe » : « A New Strategy for Arabic OCR based on Script Analysis and Synthesis », art. cit.

[21] https://ieeexplore.ieee.org/xpl/conhome/8054539/proceeding?pageNumber=2

34 contributions, 5 portaient sur des documents historiques. Et dans les actes de la deuxième édition[22], 2 contributions sur les 32 semblent considérer des documents historiques.

Depuis la troisième édition, l'atelier ASAR est organisé dans le cadre de la conférence ICDAR : à Sydney en septembre 2019 (15th International Conference on Document Analysis and Recognition) et cette année à Lausanne. Dans le cadre du travail réalisé sur le *dataset* RASAM (voir III), nous avons participé, avec Chahan Vidal-Gorène, Clément Salah, Aliénor Decours et Boris Dupin, à cet atelier qui s'est tenu le 6 septembre 2021 à Lausanne.

Il apparaît que l'atelier est plutôt fréquenté par des ingénieurs qui présentent des recherches sur des points très précis d'OCR ou d'HTR de l'arabe. Sur les six présentations et trois compétitions, deux contributions (dont RASAM) concernaient des documents historiques.

Précisons pour clore cette section que cet atelier ainsi que les autres ateliers organisés dans le cadre de ces conférences accordent une place importante aux compétitions. Celles-ci consistent à mettre à disposition des données puis à comparer les résultats de différentes équipes internationales[23].

[22] https://ieeexplore.ieee.org/xpl/conhome/8465566/proceeding

[23] Pour un exemple, voir le *dataset* RASM ci-dessous ou la présentation des projets HTR menés par la British Library.

046 [II] B
BASES DE DONNÉES ET *DATASETS*

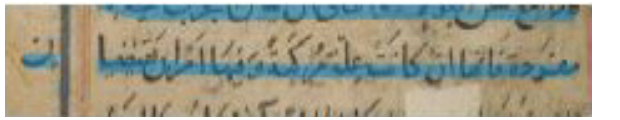
Les bases de données et les *datasets* (ou jeux de données) sont essentiels dans le processus de création des modèles OCR et HTR. Couramment utilisés en *machine learning*, ils regroupent un ensemble de données cohérentes qui peuvent se présenter de différentes manières (textes, chiffres, images). Ces *datasets* ou jeux de données constituent une mécanique essentielle en intelligence artificielle car ils permettent l'expression des algorithmes. Ils constituent en somme la matière première et c'est la raison pour laquelle le flux de travail de création d'un modèle OCR/HTR à partir de zéro exige dans un premier temps la création de données dites de terrain.

Il faut signaler dans ce contexte l'initiative HTR United menée par Thibault Clérice (EnC-PSL) et Alix Chagué (Inria)[24]. Il s'agit d'une organisation Github qui permet la mise en commun des *datasets* et des modèles pour l'HTR/OCR. HTR-United propose ainsi un catalogue de jeux de données et de modèles de reconnaissance automatique des écritures. Si l'immense majorité du catalogue porte sur les langues latines, notamment le français, le projet est ouvert à toutes les langues et styles.

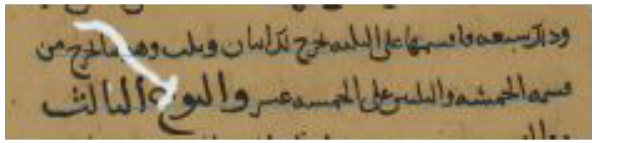
La liste non exhaustive des *datasets* proposée ci-dessous a notamment été élaborée à partir du dépouillement des actes des deux principales conférences internationales rattachées à ce champ (ICDAR et ICFHR). Notons qu'un certain nombre de ces *datasets* sont directement rattachés à des

[24] Site : https://htr-united.github.io/. Pour en savoir plus : « HTR-United : Mutualisons la vérité de terrain ! », communication au colloque « Publier, partager, réutiliser les données de la recherche : les *data papers* et leurs enjeux », Lille, MESH, novembre 2021 : https://hal.archives-ouvertes.fr/hal-03398740/document

Fig. 5
Illustration de l'article consacré à BADAM. Titre de la figure « Examples of annotation guideline application (baseline indicated with opaque blue polyline) ». Reproduite avec l'aimable autorisation de Benjamin Kiessling

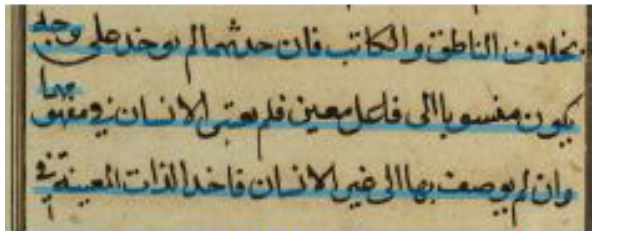


(a) Annotation of dislocated fragments in margin



(b) Holes in writing surface

(c) Per-word baseline annotation through imaginary baseline



(d) Separate annotation of heaped elements with complete overlap vs single baseline for partial overlap

(e) Joint annotation of half-verses as a single baseline

(f) Separated annotation of slanted half-verses

projets scientifiques ou des infrastructures plus larges et sont donc mentionnés dans la présentation correspondante. Nous nous bornons ici à donner une description succincte de ces *datasets*. Pour plus d'informations nous renvoyons le lecteur aux articles correspondants.

Pour l'arabe, le nombre des *datasets* disponibles reste relativement limité. Par ailleurs, il apparaît que peu de *datasets* fournissent des données suffisantes pour pouvoir véritablement exploiter les jeux de données mis à disposition en *open source* sur des GIT par exemple. Nous incluons seulement les *datasets* existant pour l'arabe ces dernières années (2014-2021). Les *datasets* plus anciens ne sont pas pris en compte dans cette liste. Notez enfin qu'il faut désormais ajouter à cette liste RASAM [2021] qui sera présenté en détails dans la dernière partie de ce rapport. Ce *dataset* propose 300 pages de manuscrits arabes en graphie dite maghrébine.

BADAM [2019]

Le *dataset* comprend 400 images annotées provenant de différentes périodes et dont les caractéristiques sont la diversité en matière de sujets, de styles et de complexité de la mise en page [FIG. 5]. Ces manuscrits sont issus de la culture manuscrite persane et arabe. Le *dataset* porte en priorité sur la détection de la mise en page et en particulier sur la ligne de texte (*baseline*).

Les images proviennent de 42 manuscrits issus de quatre collections numériques de manuscrits en langues arabe et persane : la Qatar Digital Library, la bibliothèque numérique du Walters Art Museum, la Beinecke Rare Book and Manuscript Library de Yale et la collection des manuscrits de bibliothèques de l'Université de Pennsylvanie.

La majorité du corpus est copié dans le style *naskh*, le reste comprend du *thuluth*, *nasta'līq*, et coufique.

Dans l'article accompagnant la présentation de ce *dataset*, les auteurs décrivent un système de *baseline* pour l'extraction des lignes. Leurs conclusions portent notamment sur le fait que même les méthodes les plus avancées ont des difficultés à segmenter l'écriture arabe avec autant de précision que pour les manuscrits latins.

- Le *dataset* est disponible ici : https://zenodo.org/record/3274428

À propos Voir Benjamin Kiessling, Daniel Stökl Ben Ezra, Matthew T. Miller, « BADAM: A Public Dataset for Baseline Detection in Arabic-script Manuscripts », 2019. https://arxiv.org/ftp/arxiv/papers/1907/1907.04041.pdf

RASM [2018 et 2019]

Le *dataset* a été élaboré afin de fournir la vérité terrain* nécessaire pour une compétition organisée dans le cadre de l'ICFHR 2018 qui visait à tester l'analyse de la mise en page, la détection des lignes et la reconnaissance de texte. Le *dataset* se compose de manuscrits scientifiques de la bibliothèque numérique du Qatar. Il comprend 100 images qui ont été annotées à différents niveaux : région, polygones, lignes et textes.

En 2019, il a été augmenté dans le cadre de la deuxième compétition RASM 2019, organisée lors de l'ICDAR 2019. Il comprend désormais 120 images numérisées (TIFF) provenant d'une sélection de manuscrits scientifiques arabes copiés entre le X^e^ et le XIX^e^ siècles. Le *dataset* contient également des transcriptions de vérité terrain (XML) pour chaque page qui peuvent être utilisées pour l'entraînement de logiciels de reconnaissance optique de caractères (OCR) ou de reconnaissance de textes manuscrits (HTR) sur des textes manuscrits arabes historiques. Le dossier contient donc les

images et la vérité terrain utilisées dans le cadre de ces deux compétitions.

- Le *dataset* est disponible ici : https://bl.iro.bl.uk/concern/datasets/f866aefa-b025-4675-b37d-44647649ba71?locale=en

À propos Christian Clausner, Apostolos Antonacopoulos, Nora Mcgregor, Daniel Wilson-Nunn, « ICFHR 2018 Competition on Recognition of Historical Arabic Scientific Manuscripts - RASM2018 » dans *ICFHR 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), Niagara Falls, NY, USA, 2018*, Piscataway, IEEE, 2018, p. 471-476. https://doi.org/10.1109/ICFHR-2018.2018.00088

KERTAS [2018]

Ce *dataset* a été développé par une équipe de l'Université du Qatar à Doha. Leur *dataset* est composé de manuscrits arabes et il a été développé pour tester les algorithmes les plus récents de détection de la date et de l'auteur de ces manuscrits. KERTAS est un jeu de données de documents historiques qui est pensé pour aider à dater automatiquement les manuscrits arabes avec plus de précision et d'efficacité. Il comprend 2 000 images issues de 95 manuscrits répartis sur une période de quatorze siècles. Plus de la moitié (57) provient de la collection de manuscrits de la Bibliothèque nationale du Qatar. Les manuscrits du *dataset* sont catégorisés selon le siècle de leur publication/rédaction. Cela a été fait à partir des informations indiquées dans les notices documentaires de ces manuscrits. Les informations ont ensuite été vérifiées à partir de différentes sources avant que le manuscrit ne soit ajouté à la base de données. Le *dataset* prend la forme d'une base de données qui comprend les images des manuscrits et un dosser XML avec les métadonnées.

- Dataset : non disponible en ligne à notre connaissance

À propos Kalthoum Adam, Asim Baig, Somaya Al-Maadeed, *et al.*, « KERTAS: Dataset for Automatic Dating of Ancient Arabic Manuscripts », *IJDAR*, 21, 2018, p.283–290. https://doi.org/10.1007/s10032-018-0312-3

WAHD [2017?]

Ce *dataset* a été composé dans le cadre d'un travail sur les méthodes permettant d'identifier les copistes dans les documents historiques arabes mené par une équipe de l'Université Ben Gourion. Il s'agit d'une base de données portant sur 353 manuscrits provenant du *Islamic Heritage Project* (333) et de la *National Library in Jerusalem* (20). Ces 353 manuscrits ont été copiés par 302 personnes différentes dont seules 23 sont connues (11 scribes ont copié 42 mss et les 12 autres ont copié un ms chacun). Les copistes de 302 manuscrits restants sont inconnus. Ce que contient effectivement ce *dataset* n'est pas très clairement indiqué.

- Le *dataset* WAHD (appelé aussi WAHAD) est disponible ici : https://www.cs.bgu.ac.il/~vml/wahad.html

À propos Alaa Abdelhaleem, Ahmed Droby, Abdelkader Asi, *et al.*, « WAHD: a database for writer identification of Arabic historical documents », dans *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, Piscataway, IEEE, 2017, p. 64–68. *L'article se concentre sur les méthodes d'identification des copistes. Il présente par ailleurs le dataset KHATT* (voir *infra*)

VML-HD [2017]

La base de données est composée de cinq livres écrits par différents auteurs dans les années 1088-1451. Elle comprend 668 pages entièrement annotées au niveau des « sous-mots ». Pour chaque page, des boîtes de délimitation aux différents sous-mots ont été appliquées et la séquence de caractères a été annotée.

La base de données comprend 159 149 apparitions de sous-mots composés de 326 289 caractères sur un vocabulaire de 5 509 formes de

sous-mots. La base de données est conçue pour former et tester les systèmes de reconnaissance des « sous-mots » arabes manuscrits.

- Le *dataset* est disponible ici : https://majeek.github.io/tutorials/vmlHD/

À propos Majeed Kassis, Alaa Abdalhaleem, Ahmad Droby, *et al.*, « VML-HD: the Historical Arabic Documents Dataset for Recognition Systems », dans *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, Piscataway, IEEE, 2017, p. 11–14.

Compléments sur les datasets développés par le Visual Media Lab (VML) de l'Université Ben-Gurion du Negev

Les deux *datasets* mentionnés précédemment ont été développés par la même équipe, dirigé par Jihad el-Sana. D'autres *datasets* peuvent être consultés sur le site de Berat Kurar Barakat, une doctorante en sciences informatiques : https://www.cs.bgu.ac.il/~berat/ ; ils concernent l'arabe mais aussi l'hébreu et portent sur des documents historiques (manuscrits). Le site propose par ailleurs une liste d'articles publiés par l'équipe qui permet ainsi de se faire une idée des travaux qu'elle mène et de ses axes de recherche. Elle travaille actuellement sur de nouveaux champs de la recherche en OCR/HTR, en particulier l'analyse de la mise en page sans supervision. Voir la contribution « Unsupervised learning of text line segmentation by differentiating coarse patterns[25] » que l'équipe a présentée lors de la conférence ICDAR de 2021.
Deux remarques :

- Bien qu'elle travaille sur l'analyse et la reconnaissance de documents historiques et donc sur l'HTR, il semble qu'il s'agisse d'une approche d'abord informatique et technique au sens où ces travaux ne semblent pas adossés à des projets de recherche en sciences humaines et sociales sur ces documents.
- Les *datasets* mis à disposition par cette équipe fournissent des données qui ne sont pas vraiment interopérables, ce qui est en partie lié au fait que la plupart de ces *datasets* ont cinq ans ou plus.

KHATT [2014]

Il s'agit d'un *dataset* d'arabe moderne manuscrit collecté auprès de 1000 scribes de différents pays. Chacun d'entre eux a écrit six paragraphes. Ce jeu de données a été pensé pour la recherche en matière d'identification du copiste/écrivant, vérification de l'auteur, segmentation des lignes et *pattern recognition*.

- Pour accéder au *dataset* : http://khatt.ideas2serve.net/

À propos Sabri A. Mahmoud, Irfan Ahmad, Wasfi G. Al-Khatib, *et al.*, « KHATT: an open Arabic offline handwritten text database », *Pattern Recognition*, 47/3, 2014, p. 1096-1112. https://doi.org/10.1016/j.patcog.2013.08.009. Sabri A. Mahmoud, Irfan Ahmad, Mohammed Alshayeb, *et al.*, « KHATT: Arabic offline handwritten text database », *13th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, Picataway, IEEE, 2012, p. 447–452, 2012. [Best Poster Award Winner] https://doi.org/10.1109/ICFHR.2012.224

HADARA80P [2014]

Ce *dataset* a été développé dans le cadre du projet HADARA (voir l'article « HADARA – A Software System for Semi-Automatic Processing of Historical Handwritten Arabic Documents[26] ») développé à la Technische Universität

[25] https://arxiv.org/abs/2105.09405
[26] http://www.imaging.org/site/PDFS/Reporter/Articles/Rep28_3_ARCH2013_PANTKE_PG161.pdf

Braunschweig vers les années 2013-2014 par une équipe de l'université en collaboration avec des chercheurs en informatique israéliens. Ce projet, qui réunissait des ingénieurs informatiques mais aussi des linguistes et des historiens avait plusieurs objectifs : la numérisation de livres difficilement accessibles et le développement d'outils pour traiter et archiver des manuscrits numérisés, en se concentrant sur l'analyse de documents historiques en arabe manuscrit.

Ce *dataset* a été créé en particulier dans le cadre d'une recherche sur la reconnaissance des mots (*word spotting*) dans les documents historiques avec une approche non segmentée en partant du constat que les quelques *datasets* existants concernaient tous les graphies latines et non l'arabe. Ainsi le *dataset* HADARA80P comprend 80 pages provenant d'un manuscrit arabe du XV^e^ siècle composé de 250 pages. Il a été créé pour le développement et l'évaluation des systèmes de reconnaissance de mots non segmentés. Il comprend des images en haute résolution avec des polygones, des informations complémentaires avec des balises au niveau du mot et un ensemble de 25 mots-clefs pré-définis.
Dataset : non trouvé en ligne.

À propos Le *dataset* a été présenté lors de la conférence ICFHR 2014 : Werner Pantke, Martin Dennhardt, Daniel Fecker, *et al.*, « An Historical Handwritten Arabic Dataset for Segmentation-Free Word Spotting - HADARA80P », dans *2014 14th International Conference on Frontiers in Handwriting Recognition*, Piscataway, IEEE, 2014, p. 15–20.Power-point de la présentation : http://www.icfhr2014.org/wp-content/uploads/2015/02/ICFHR2014-Pantke.pdf

Dans leur état des lieux intitulé « Automatic processing of Historical Arabic Documents : A Comprehensive Survey », les auteurs classent les *datasets* disponibles comme suit[27] [TABLEAU N.6] :

À propos Pour accéder à des *dataset* plus anciens, on peut consulter le site du TC11 : http://www.iapr-tc11.g/mediawiki/index.php/IBN_SINA:_A_database_for_research_on_processing_and_understanding_of_Arabic_manuscripts_images

[27] M. Ibn Khedher, H. Jmila, M. A. El-Yacoubi, « Automatic processing of Historical Arabic Documents: A comprehensive Survey », art. cit.

Tableau nº 6 : Résumés des travaux de recherche/*datasets*

Application	Category	Publication	Dataset	Year
WRITER CLASSIFICATION				
Writer Classification	Model-free	[41]	IHP	2014
		[19]	WAHD	2015
		[12]	IHP/KHATT	2017
	Mode-based	[40]	IHP	2014
DATA RETRIEVAL				
Word spotting	Model-free	[11]	Private (23 pages)	2011
		[87]	Private (20 pages)	2013
		[105]	Private (40 pages)	2013
		[10]	Private (23 pages)	2014
		[53]	Private (20 pages)	2014
		[81]	Private (12 pages)	2014
		[55]	VML	2016
		[39]	HADARA80P	2017
	Mode-based	[84]	Private (20 pages)	2008
		[36]	Ibn-Sina	2015
		[58]	Ibn-Sina	2016
		[22]	VML	2018
TEXT ANALYSIS				
Text Line Detection	Top-down	[17]	Private (217 pages)	2011
		[83]	Private (315 pages)	2014
	Hybrid	[31]	Private (836 pages)	2014
Layout Analysis	Model-free	[14]	Private (38 pages)	2014
		[54]	Private (38 pages)	2016
	Model-based	[29]	Private (38 pages)	2012
		[23]	Private (38 pages)	2018
Text Preprocessing	Text skew correction	[35]	Private	2010
	Image binarization	[8]	Private	2014
		[37]	Private	2015
TEXT CLASSIFICATION				
Text Recognition	Model-free	[15]	40,000 sub-words	2012
	Model-free	[5]	22,218 sub-words	2017
		[4]	4,124 sub-words	2018
Text Alignment	Model-free	[32]	(6 pages)	2015
		[80]	Private (6 pages)	2013
		[16]	Not mentioned	2011
	Model-based	[56]	Public (72 pages)	2017
DATASETS				
Writer classification	Not degraded	[3]	WAHD Dataset: IHP subset	2017
		[3]	WAHD Dataset: NLJ subset	2017
Word spotting	Not degraded	[71]	IBN-SINA	2010
		[74]	HADARA80P	2014
		[52]	VML-HD	2017
Quality Assessment	Degraded	[91]	MHDID	2018
		[92]	VDIQA	2010
SOFTWARES				
Data Retrieval	Word spotting	[70]	–	2011
	Annotation	[57]	Private (>10 pages)	2014
		[24]	–	2013
Text Analysis	Text Preprocessing	[25]	National library of Tunisia	2007
	Simplifying the reading	[13]	Public (25 pages)	2015
Text Recognition	Text Classification	[95]	HADARA80P *and others*	2010

053 [II] C
LOGICIELS ET INTERFACES

Les chercheurs qui se lancent dans des projets impliquant des systèmes OCR ou HTR pour extraire leurs textes le font généralement par la médiation d'interfaces ou de plateformes qui sont les interfaces graphiques supportant des logiciels.

Ces logiciels ou outils, qui peuvent être en *open source* comme OCRopus, Tesseract, Calamari ou Kraken, permettent théoriquement de réaliser toutes les étapes nécessaires à la transcription automatique. Même s'ils sont tout à fait fonctionnels et puissants, les auteurs de l'article consacré à la présentation d'OCR4all rappellent justement que leur utilisation peut être assez compliquée pour l'utilisateur car :

- ils ne disposent pas dans la plupart des cas d'une interface graphique confortable ;
- ils s'appuient sur des formats d'entrée et de sorties qui peuvent être différents ;
- les procédures d'installation et de configuration peuvent être complexes et sujettes aux erreurs par l'utilisateur ;
- pour des utilisateurs non familiers de la technique, la courbe d'apprentissage peut être abrupte[28].

Ces aspects posent particulièrement problème pour les utilisateurs inexpérimentés, qui englobent en fait la majorité des chercheurs en sciences humaines et sociales, et c'est la raison pour laquelle les interfaces et plateformes que nous allons évoquer, sont, elles, pensées pour être plus adéquates et offrir la chaîne complète d'opérations pour l'OCR/HTR.

[28] C. Reul, D. Christ, A. Hartelt, *et al.*, « OCR4all –An Open-Source Tool Providing a (semi-) Automatic OCR Workflow for Historical Printings », *Applied Sciences*, 9/22, 2019, p. 1-54, p. 3.

S'il existe donc de nombreuses plateformes pour annoter des images, la question se pose de savoir dans quelle mesure elles sont ergonomiques pour un public de non-spécialistes des humanités numériques et si leur prise en main peut être rapide.

Dans cette partie, nous présenterons des logiciels et des interfaces/plateformes. Nous ne décrivons pas toutes les plateformes existantes mais certaines d'entre elles qui considèrent notamment les langues non latines et leurs graphies. Il s'agit d'une sélection et il nous faut d'ores et déjà préciser qu'aucune de ces interfaces n'a été développée précisément pour l'arabe, même si eScriptorium et, surtout Calfa, ont mis les langues orientales non latines à l'honneur. Le logiciel Kraken a par ailleurs été développé dans le cadre d'une recherche sur l'OCR de l'arabe. Nous verrons également que certaines interfaces, qui ne sont pas spécialisées dans les langues non latines, les prennent tout de même en considération

Il faut préciser qu'une plateforme d'annotation efficace doit comprendre des tâches automatisées, des options simples et limitées pour favoriser le travail collaboratif et prendre en compte plusieurs types de langues. L'interopérabilité* des formats est par ailleurs importante à considérer.

Dans le cas de l'arabe, langue pour laquelle nous ne disposons ni de modèles généralistes HTR par défaut ni de beaucoup de modèles spécialisés, un chercheur qui souhaiterait utiliser un système HTR pour extraire le texte de son ou de ses manuscrits, devra constituer une vérité terrain* pour entraîner un modèle. La question se pose donc des possibilités offertes par ces outils pour faire une base de données rapidement et être rapidement utilisables. Une des approches que l'on peut envisager est le *crowdsourcing*, ou production participative. Dans ce cas, on fait appel à des gens que l'on forme pour les besoins du projet afin de construire une base de données. Cette dimension n'est pas négligeable dans le processus d'OCR/HTR et une interface d'annotation efficace doit également prendre en compte cette dimension et pouvoir s'adapter au travail collaboratif.

Nous mentionnerons ici à la marge les logiciels propriétaires pour l'OCR, notamment le leader des logiciels propriétaires ABBYY avec son logiciel FineReader. Nous souhaitons par ailleurs indiquer qu'aucune des technologies présentées dans ce rapport ne fera l'objet d'une évaluation par l'auteur de ce rapport. Il ne s'agit pas non plus d'indiquer au lecteur quel outil il doit privilégier en fonction de ses besoins. Nous mettons ici l'accent sur ce qui existe en l'état actuel de nos connaissances pour qui souhaite se lancer dans la réalisation d'un projet OCR ou HTR sur des documents historiques en arabe.

En septembre 2021, le magazine en ligne *The Digital Orientalist* a publié une double contribution d'Ishida Yuri, de l'Université d'Okayama, et Shinoda Tomoaki de l'Université de Tokyo, dans lesquels ils proposaient une évaluation de la précision des systèmes d'OCR faciles d'utilisation et peu coûteux pour l'arabe[29]. Comme ces deux chercheurs le signalent, les évaluations de la précision des OCR, notamment pour l'arabe, sont rares. Aussi,

[29] https://digitalorientalist.com/2021/09/17/a-study-on-the-accuracy-of-low-cost-user-friendly-ocr-systems-for-arabic-part-1/
https://digitalorientalist.com/2021/09/24/a-study-on-the-accuracy-of-low-cost-user-friendly-ocr-systems-for-arabic-part-2/

nous avons pensé qu'un résumé des principales conclusions de cette double contribution serait un bon complément à l'introduction de l'état des lieux des logiciels et des outils que nous allons présenter, puisque nous n'évaluerons pas l'efficacité de ces outils pour l'arabe, sauf lorsque ces informations sont renseignées ailleurs.

L'objectif des tests réalisés par ces deux chercheurs est de guider les chercheurs dans leur choix de plateformes et d'outils lorsqu'ils disposent d'images et souhaitent les convertir en texte. C'est la raison pour laquelle ils ont mis l'accent sur les interfaces graphiques *user-friendly*. Notons que leur étude ne porte que sur l'arabe imprimé (pour les écritures manuscrites, les auteurs renvoient en note à Transkribus) et qu'ils ont volontairement mené leurs tests à partir d'exemples qui ne présentaient pas d'enjeux de mise en page.

Les résultats obtenus sont les suivants : des 17 systèmes OCR testés, seuls 11 pouvaient effectivement traiter l'arabe[30]. Les tests ont été réalisés à partir de 4 textes et l'évaluation a été faite avec un logiciel appelé OCR Evaluation qui calcule la précision en comparant le fichier texte généré par l'OCR à un fichier test «vérité terrain*». Les résultats sont affichés sous forme de statistiques, avec les taux d'erreur au caractère et au mot. Il ressort de cette évaluation que les quatre systèmes les plus robustes sont : i2OCR, OCR Space, Google Drive et Fine Reader (ABBYY). Précisons que le CER moyen et le WER (*Word Error Rates*) moyen les plus satisfaisants sont ceux de l'OCR Fine Reader (ABBYY), qui est actuellement le leader sur le marché des logiciels propriétaires. Les trois autres architectures présentées comme robustes sont, elles, des solutions gratuites mais on ne peut manquer de préciser que leurs résultats, pour le CER en tout cas, sont en dessous de celui de ABBYY puisqu'ils sont, au mieux, à trois points de CER de Fine Reader (ABBYY).

Les auteurs précisent que la correction manuelle d'erreurs reste une étape essentielle et que, par ailleurs, les résultats peuvent grandement varier d'une page à l'autre. Ils concluent leur billet en conseillant de choisir l'outil OCR qui fonctionne le mieux en fonction des documents utilisés sans se concentrer sur les taux d'erreur car certains systèmes peuvent être plus performants que d'autres en fonction des types de documents.

Les recherches menées par ces deux chercheurs, synthétisées dans leur contribution à *The Digital Orientalist*, avaient été présentées en mars 2021, en japonais, au cours d'un atelier organisé par the Islamic Trust Studies consacré aux humanités numériques au Japon. Il importe de préciser que cet exemple de contribution rappelle que des travaux et recherches sont en cours au Japon sur les questions d'OCR et d'HTR de l'arabe. Cependant, l'accès aux recherches en cours est souvent difficile en raison de la non connaissance de la langue japonaise.

[30] Les 11 systèmes sont : Convertio (https://convertio.co/); ABBYY Fine Reader PDF (https://pdf.abbyy.com/pricing/), Foxit Phantom PDF (https://www.foxit.com/shopping/) ; Free Online OCR (https://www.newocr.com/), Gold/Sakhr (http://www.sakhr.com/index.php/en/solutions/ocr); i2 OCR (https://www.i2ocr.com/free-online-arabic-ocr), OCR Convert (https://www.ocr-convert.com/arabic-ocr), OCR Space (https://ocr.space/), Online Convert Free (https://onlineconvertfree.com/ocr/arabic/), Sotoor (https://rdi-eg.ai/optical-character-recognition/), Google Drive.

Kraken

Kraken est un logiciel OCR *open source* qui a été développé par Benjamin Kiessling, d'abord en Allemagne à l'Université de Leipzig (Alexander von Humboldt Chair for Digital Humanities) en 2017. Benjamin Kiessling a depuis rejoint l'École pratique des hautes études (EPHE) et le projet Scripta. Au départ, Kraken est une amélioration d'OCRopus avec l'ajout de réseaux de neurones LSTM (couche contextuelle). Il prend en charge les graphies de droite à gauche et inclut une interface de transcription rudimentaire pour une utilisation *offline*. Le logiciel a été développé pour l'arabe au départ mais il traite aujourd'hui d'autres graphies dont un grand nombre de graphies latines. Benjamin Kiessling a publié un certain nombre d'articles en lien avec Kraken que nous avons mentionnés tout au long de ce rapport.

Kraken est un outil d'analyse de mise en page et d'HTR fondé sur de l'apprentissage profond*. eScriptorium, que nous présentons plus loin, a été développée pour interagir avec Kraken et son moteur pour l'OCR/HTR. Kraken est écrit en python et a été conçu pour fonctionner avec un très large spectre d'écritures différentes. Il est pensé pour être modulaire et flexible et permet au chercheur de modifier le modèle en fonction de ses besoins, à condition bien sûr que celui-ci dispose d'une bonne compréhension des logiciels et des processus OCR/HTR au départ et d'une aisance avec les logiques du code. Kraken n'est donc pas approprié pour la majorité des utilisateurs des sciences humaines et c'est notamment la raison pour laquelle eScriptorium a été développé.

Kraken est utilisé par eScriptorium, qui en est donc, en quelque sorte, l'interface graphique. Le logiciel reste disponible pour ceux qui voudraient l'utiliser sans eScriptorium. Il a aussi été utilisé par l'équipe du projet OCR4all qui a développé et amélioré, à partir de Kraken, Calamari.

À propos Benjamin Kiessling, « Kraken – A Universal Text Recognizer for the Humanities » : https://dataverse.nl/dataset.xhtml?persistentId=doi:10.34894/Z9G2EX. « Using Kraken to Train your own OCR Models » : https://digitalorientalist.com/2019/11/05/using-kraken-to-train-your-own-ocr-models/. Site de Kraken : http://kraken.re/master/index.html. GIT de Kraken : https://github.com/mittagessen/kraken

OCR4all

Contexte

OCR4all a été développé au départ dans le cadre du projet Kallimachos[31]. Kallimachos est un centre de recherche en humanités numériques réunissant des chercheurs en humanités et des informaticiens, financé par le ministère fédéral de l'Éducation et de la Recherche dans le cadre du programme de financement des e-Humanités en Allemagne. Initié par l'Université de Würzburg, il associe également le DFKI (Centre de recherche allemand pour l'intelligence artificielle) de Kaiserslautern (OCR) et l'Université d'Erlangen-Nüremberg (informatique linguistique).

Les objectifs de cette infrastructure sont notamment la supervision et la coordination d'éditions numériques et l'application d'analyse quantitative par le biais de différentes méthodes de fouille de textes comme la modélisation thématique et la reconnaissance des entités nommées. Structure intégrée pour les données de recherche en SHS, Kallimachos comprend également le développement de composants logiciels et l'établissement de flux de travail prototypiques à intégrer dans les infrastructures existantes.

[31] http://kallimachos.de/kallimachos/index.php/Project_description

OCR4all a donc été développé dans ce contexte, au sein du Center for Philology and Digitality de l'Université de Würzburg sous la direction de Christian Reul, avec Frank Puppe (qui détient la chaire d'intelligence artificielle et de sciences informatiques appliquées), Christoph Wick, Uwe Springmann et de nombreux étudiants et assistants.

Interface et utilisation

Outil *open source*, OCR4all vise à encapsuler un flux de travail OCR complet dans une seule application Docker, garantissant une installation facile et une indépendance de plateforme. OCR4all fonctionne aussi bien sur Linux, Mac ou Windows. Il s'agit d'une interface web qui couvre toutes les étapes d'un flux de travail OCR, du prétraitement à l'analyse du document, en incluant l'analyse de l'image (segmentation des régions textuelles et non textuelles d'une page), l'apprentissage du modèle et la reconnaissance des caractères des régions textuelles. Des estimations des taux d'erreur restants sont également fournies.

L'objectif est de mettre les capacités d'outils de pointe comme OCRopus ou Calamari à la disposition de tout utilisateur dans le cadre d'un flux de travail semi-automatique compréhensible et applicable. Cet objectif est atteint en fournissant aux utilisateurs une interface graphique qui se veut confortable et facile à utiliser et une approche modulaire, permettant un processus de correction efficace entre les différentes étapes du flux de travail afin de minimiser les effets négatifs des erreurs consécutives.

Calamari est un moteur OCR basé sur OCRopy et Kraken utilisant Python3. Il est conçu pour être à la fois facile à utiliser à partir de la ligne de commande mais aussi modulaire pour être intégré et personnalisé à partir d'autres scripts python. À chaque étape du processus, il est ainsi possible de choisir l'outil avec lequel on souhaite travailler. En d'autres termes, il y a plusieurs modèles d'analyse de la mise en page ou de reconnaissance des caractères qui sont proposés. Par exemple, l'utilisateur peut utiliser le modèle d'analyse de la mise en page de Calfa puis utiliser Tesseract pour l'OCR. L'un des avantages de cette plateforme est donc notamment qu'elle inclut différentes architectures d'OCR.

Et l'arabe ?

L'interface n'a pas été spécifiquement développée pour l'arabe ou pour les langues non latines. Elle a été développée au départ pour les tous premiers documents imprimés, incunables, afin de traiter la complexité des types et des conceptions de mise en page de ces textes. On notera néanmoins que la possibilité d'utiliser différentes architectures d'OCR/HTR et son caractère modulaire en font un outil théoriquement adapté pour l'arabe.

Perspectives

Depuis l'été 2020, OCR4all collabore avec OCR-D[32], qui a également développé une interface. L'objectif de cette collaboration n'est pas seulement le partage d'informations concernant les interfaces ou les implémentations logicielles mais aussi les développements à venir dans le domaine de l'OCR. Cette collaboration vise par ailleurs à réaliser une convergence technique des deux projets.

À propos Calamari : GIT : https://github.com/Calamari-OCR. https://calamari-ocr.readthedocs.io/en/latest/. Site du projet : http://www.ocr4all.org/en/about.html#workflow. Informations et guide d'installation et d'utilisation en anglais et en allemand sont proposés. Dans cette description : OCR-D project : https://ocr-d.de/en/about. C. Reul, D. Christ, A. Hartelt, *et al.*, « OCR4all – An Open-Source Tool Providing a (semi-) Automatic OCR Workflow for Historical Printings », *Applied Sciences*, 9/22, 2019, p. 1-54. https://arxiv.org/abs/1909.04032

Le projet eScriptorium

Contexte

Ce projet s'inscrit dans le cadre de Scripta, une initiative de recherche stratégique et interdisciplinaire (IRIS) sur l'histoire et les pratiques de l'écrit dont l'objectif est de fédérer la recherche dans ce domaine au sein de Paris Sciences & Lettres (PSL). Elle est le fruit de l'association de six établissements de PSL (EPHE, EFEO, ENC, ENS, EHESS, Collège de France). Scripta réunit environ une centaine de chercheurs en humanités, sciences sociales, humanités numériques et sciences de l'informatique.

L'interface s'adresse aux chercheurs en « études religieuses » entendues dans un sens très large comprenant la philologie, la paléographie, l'épigraphie, la papyrologie, la linguistique mais aussi l'anthropologie ou la musicologie. Les documents et les textes sont le principal cœur de cible et les supports sont divers et variés (parchemins, papier, papyrus, pierre, feuilles de palme). L'objectif d'eScriptorium est de combiner des outils informatiques avec des outils numériques manuels pour la transcription et l'annotation de textes et d'images (paléographiques, philologiques, historiques et linguistiques). Le développement de la plateforme a commencé en novembre 2018 même si Kraken (voir *supra*) est en développement depuis 2017.

Interface

eScriptorium est une infrastructure HTR *open source* qui est le résultat d'un travail collaboratif au sein duquel Benjamin Kiessling et Robin Tissot ont joué un rôle déterminant. Elle permet de doter Kraken d'une interface graphique. Le modèle de l'analyse de la mise en page est celui développé par Benjamin Kiessling avec Kraken. Il se caractérise par l'identification d'une zone unique de texte.

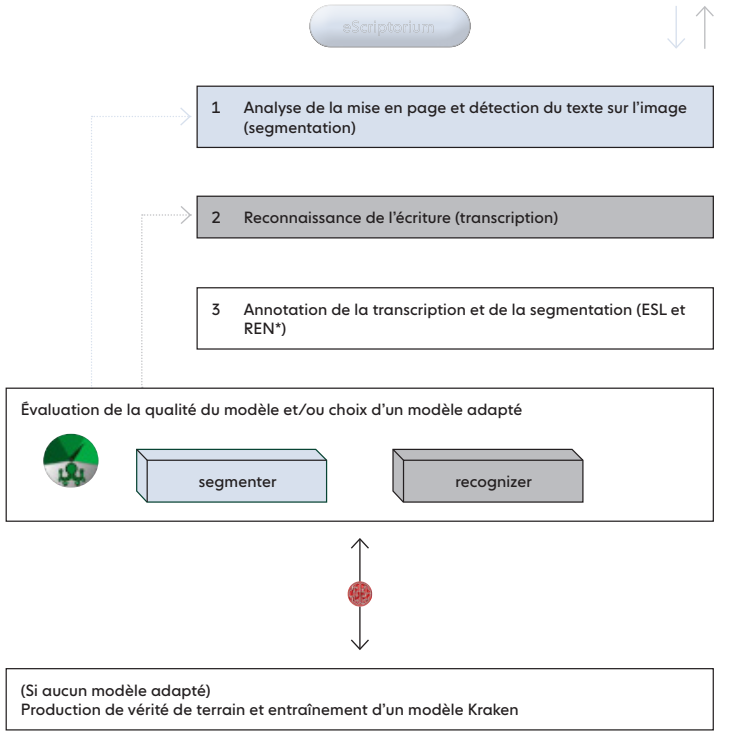
Utilisation

Les utilisateurs interagissent avec la plateforme via une interface web utilisateur graphique. Cette interface web est optimisée pour Chrome et Firefox.

[32] Il s'agit d'un projet allemand financé par le DFG (German Research Foundation) qui porte sur l'OCR des documents historiques allemands entre les XVI[e] et XVIII[e] siècles, qui a développé un écosystème ouvert visant à améliorer l'OCR de ces documents. Lire : https://ocr-d.de/en/ ; voir : https://elag.org/mini-elag-october-20-2020/ocr-d-an-open-ecosystem-for-improving-ocr-on-historical-documents/

Fig. 6
Résumé graphique de l'infrastructure eScriptorium
(graphique réalisé à partir de celui présenté par Alix Chagué dans le cadre d'une présentation du projet Lectaurep)

VERS LA DÉFINITION D'UNE INFRASTRUCTURE DE RECHERCHE ET DE SERVICE POUR LA REM



*REN : Reconnaissance d'Entités Nommées

UN TRAVAIL DE PÉRENNISATION À TOUS LES NIVEAUX

- Mise en *open source* de tous les développements informatiques
- Large ouverture des documentations et composants pédagogiques
- Travail inter-institutionnel sur le renforcement des capacités de calcul et de leur hébergement
- Vers une coordination des établissements concernés dans la gestion des accès
- Perspective d'un service relié à Huma-Num et DARIAH à terme

RÉCAPITULATIF DE LA CHAÎNE DE TRAITEMENTS

1. Téléchargement des images et métadonnées dans eScriptorium depuis un serveur IIIF ;
2. Analyse de la mise en page et détection des segments de texte à l'aide d'un modèle Kraken sélectionné grâce à Kraken-Benchmark ;
3. Transcription du texte à l'aide d'un modèle Kraken sélectionné grâce à Kraken-Benchmark ;
4. Annotation de la structure logique et des entités nommées dans la transcription obtenue ;
5. Injection de toutes les données dans le fichier pivot TEI lié à l'image sur le serveur IIIF et/ou export dans différents formats standardisés et à jour.

LIENS ET RESSOURCES

- Blog LECTAUREP : https://lectaurep.hypotheses.org/
- Serveur Inria pour eScriptorium : http://traces6.paris.inria.fr/
- Code source du fork eScriptorium : https://gitlab.inria.fr/almanach/lectaure/escriptorium/
- Code source de Kraken-Benchmark : https://gitlab.inria.fr/dh-projects/kraken-benchmark
- Code source d'Aspyre : https://gitlab.inria.fr/dh-projects/aspyre-gt

eScriptorium entend gérer l'essentiel de la chaîne de traitement des documents. L'importation des images est prévue pour être en .jpg, .pdf ou .png et l'interface dispose également d'un import avec IIIF*. Il est à noter que eScriptorium permet aussi via l'API* d'exporter des transcriptions déjà existantes et de créer des tableaux paléographiques. Le plus est que l'interface est installable sur son propre ordinateur ou bien, à l'échelle d'une institution comme une université, sur les serveurs de celle-ci pour être mis à la disposition de ses membres. Il importe de préciser cependant que l'installation sur sa propre machine n'est pas sans limites : si l'on peut théoriquement entraîner des modèles, la puissance n'est pas suffisante pour une seule machine pour faire seul des entraînements. Il est par ailleurs possible d'utiliser la plateforme en ligne en se créant un compte.

Et l'arabe ?
Les projets relatifs au développement de modèle HTR pour l'arabe ne sont actuellement pas menés par les équipes de Scripta mais par le projet OpenITI qui est associé à eScriptorium.

Comme cela a été précisé, en amont, Kraken a notamment été développé pour l'arabe imprimé mais, aujourd'hui, l'essentiel du travail est mené en collaboration avec l'équipe d'OpenITI et du projet KITAB[33]. L'essentiel des projets traités par eScriptorium porte sur les langues et graphies latines, même si les langues non-latines peuvent être prises en charge et font l'objet de projets. C'est le cas notamment des projets autour de l'hébreu et des manuscrits hébraïques, dont ceux menés par Daniel Stökl Ben Ezra.

Perspectives du projet pour la suite
L'ambition annoncée du projet eScriptorium est de couvrir toute la chaîne de traitement des manuscrits/documents, jusqu'à l'étape d'édition de ces textes ou de leur comparaison, d'où les perspectives de développement suivantes :

- développement informatique : Le repérage de mots-clefs (*keyword spotting*) ainsi que la classification automatique des manuscrits à des fins de datation et d'archivage et d'établissement de la provenance font partie des perspectives, de même que des outils de linguistique de corpus et de reconnaissance d'entités nommées.
- *deep annotation*/annotation profonde : l'un des développements prévus est de permettre des annotations structurées dites « profondes » des textes et des images. En d'autres termes, il s'agit de pouvoir ajouter des informations spécifiques sur les images et les textes, comme des informations linguistiques ou des entités nommées, voire des détails concernant les écritures, etc. Ces annotations impliquent notamment d'encoder le texte en ce sens en utilisant par exemple les standards de la TEI.
- une plateforme de publication : L'ambition de ce projet est notamment que cette interface fasse aussi office d'espace de publication en ligne, sous la forme d'éditions électroniques avec analyse de l'écriture directement rattachée aux images.

[33] Porté par l'Université Aga Khan, le projet ambitionne de développer des outils numériques pour étudier la circulation des textes arabes prémodernes, en partenariat avec l'équipe d'Open-ITI : https://kitab-project.org/

TEI Publisher[34] est l'outil envisagé pour ce module de développement mais d'autres possibilités de publications pourraient être considérées, via notamment l'utilisation d'API*.

À noter — Pour un chercheur seul, la prise en main peut être difficile si elle n'est pas accompagnée d'un contact avec l'équipe du projet. Par ailleurs, il n'y a pas de modèles fournis par défaut, à l'exception de celui de la mise en page. Il est cependant possible de charger des modèles pour les appliquer aux pages en cours de traitement. Par défaut, aucun modèle d'analyse de texte, à l'exception de celui pour l'analyse de la mise en page, n'est disponible sur l'interface. Il faut donc, soit importer ses propres modèles ou des modèles récupérés ailleurs pour la langue considérée, ou bien entraîner avec Kraken ses propres modèles sur la plateforme. Dans ce dernier cas, l'utilisateur peut annoter et transcrire un certain nombre de pages puis lancer l'entraînement d'un modèle. Les réseaux utilisés sont alors ceux de Kraken.

À propos — The eScriptorium VRE for Manuscripts Cultures: https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/. Le projet et l'interface eScriptorium sont complètement *open source*, c'est-à-dire que les codes UI et AI sont accessibles : UI Code : https://gitlab.inria.fr/scripta/escriptorium ; AI Code : https://github.com/mittagessen/kraken. Pour un tutoriel expliquant « pas à pas » comment utiliser l'interface, voir les deux tutoriels d'Alix Chagué : Prendre en main eScriptorium : https://lectaurep.hypotheses.org/documentation/prendre-en-main-escriptorium

Transkribus et le projet READ-COOP

Le projet
La plateforme de transcription automatique des textes manuscrits Transkribus a été développée depuis 2015 à l'Université d'Innsbruck dans le cadre de deux projets européens : le projet tranScriptorium (2013-2015) et le projet READ, financé par le programme de recherche européen H2020. READ ou *Recognition and Enrichment of Archival Documents* est un projet qui a pris fin en 2019. Le développement de la plateforme Transkribus visant à fournir des services pour la reconnaissance, la transcription et la recherche automatisée de documents historiques était l'objectif principal de ces deux projets. READ s'est ainsi appuyé sur les recherches menées dans le cadre de tranScriptorium pour établir de nouvelles normes en matière de reconnaissance des textes manuscrits mais aussi de repérage des mots, d'analyse de la mise en page ou d'identification automatique des auteurs et des domaines connexes.

Depuis 2017, des conférences des utilisateurs de Transkribus sont organisées, trois ont eu lieu pour le moment (2017, 2018 et 2020). Elles ont pour objectif de partager les expériences des utilisateurs.Depuis 2019, et la fin du projet, le modèle économique de READ a été revu. Une coopérative européenne READ-COOP-SCE a été créée en 2020 et elle propose désormais des services payants pour traiter de grandes quantités de documents ou pour adapter le logiciel à des besoins spécifiques. Les services gratuits sont donc désormais limités en performance et en nombre de pages. La coopérative compte actuellement 80 membres répartis dans une vingtaine de pays

Utilisation
Le moteur de reconnaissance de texte manuscrit de Transkribus, qui est fondé sur l'intelligence artificielle, doit être préparé avec des données d'apprentissage, obtenues par la transcription d'une centaine de pages (au moins 20 000 mots), en établissant la correspondance ligne à ligne entre l'image du texte numérisé et sa transcription.

L'outil est notamment sollicité pour numériser un volume de pages manuscrites important, que l'on souhaite ensuite encoder et traiter

[34] https://teipublisher.com/index.html

informatiquement, et/ou que l'on souhaite éditer. Il faut noter que la plateforme n'a pas été développée pour les langues orientales au départ. L'outil peut être utilisé à plusieurs fins dont notamment :

- la création de données d'apprentissage pour le moteur HTR afin de développer un modèle pour apprendre à déchiffrer une écriture ;
- l'exécution d'un HTR sur des documents afin d'obtenir des transcriptions générées automatiquement ;
- la transcription des documents pour une édition scientifique.

Le site propose un ensemble de ressources utiles dont des tutoriels à lire ou à regarder pour prendre en main la plateforme : https://readcoop.eu/transkribus/resources/

De nombreuses descriptions en ligne expliquent comment utiliser Transkribus, par exemple : http://regis-schlagdenhauffen.eu/wp-content/uploads/2018/01/Comment-utiliser-Transkribus-%E2%80%93-en-10-%C3%A9tapes-ou-moins.pdf

Et l'arabe ?

On pourra noter que l'outil a été utilisé pour les manuscrits arabes ou ottomans dans plusieurs projets. La conservatrice des collections asiatiques et africaines de la British Library (voir *infra*), Adi Keinan-Schoonbaert, a par exemple publié les résultats d'un test effectué à partir des jeux de données RASM : https://blogs.bl.uk/digital-scholarship/2020/01/using-transkribus-for-arabic-handwritten-text-recognition.html

Elle observe que plus le réseau de neurones dispose de données, plus les prédictions sont bonnes. Ainsi, avec un ensemble d'apprentissage de 100 pages, Transkribus a transcrit automatiquement le reste des 20 pages avec un taux de précision de 86,43 %.

À propos Le lien vers le billet suivant offre une fiche utile sur Transkribus avec des liens vers des présentations de la plateforme, la liste de projets utilisant Transkribus, etc. : https://bbm.hypotheses.org/category/transkribus. On notera parmi les projets listés : *le Periodicals of Hakki Tarik Us Collections*, un projet sur les périodiques ottomans http://www.tufs.ac.jp/common/fs/asw/tur/htu/list1.html. (voir NYU Abu Dhabi et les projets HTR pour le turc ottoman, ci-dessous). Emmanuelle Perrin, Philippe Chassignet, « Rendre visible la face cachée de l'iceberg », *ArchéOrient – Le Blog*, 27 mars 2020 : https://archeorient.hypotheses.org/14807. Site de Transkribus : https://readcoop.eu/transkribus/. Plateforme en ligne : https://transkribus.eu/lite/

Tesseract

Tesseract est un logiciel d'OCR (donc pour l'imprimé) disponible sous licence Apache[35], une licence gratuite et *open source*. Le logiciel a été développé au départ par les ingénieurs de Hewlett Packard à la fin des années 1980. Son développement a été repris par Google en 2005 après la publication du logiciel sous licence Apache. Le logiciel en est actuellement à sa quatrième version qui peut être retrouvée ici : https://tesseract-ocr.github.io/

Les langues non latines sont prises en compte par le logiciel depuis la version 3, notamment l'arabe et l'hébreu par exemple. Il est indiqué que la version 4 reconnaît 116 langues et les graphies associées.

Il importe de préciser que Tesseract n'est pas fourni avec une interface graphique (voir l'introduction de cette partie), il peut être utilisé comme

[35] https://www.apache.org/licenses/LICENSE-2.0

backend. Par ailleurs le modèle d'analyse de la mise en page de Tesseract ne repose pas sur de l'apprentissage profond* mais le modèle d'analyse des caractères, quant à lui, en bénéficie.

Tesseract est le logiciel utilisé sur Google Drive, lors de la conversion d'un document PDF en format Word. Plusieurs chercheurs qui ont utilisé cette fonction pour faire des tests en ont rendu compte dans des billets. On peut recommander notamment :

- Pour l'arabe imprimé (avec l'évaluation également de la fonction « traduction ») : « Automatic Arabic Translation Using Google : A Test » : https://digitalorientalist.com/2021/03/09/automatic-arabic-translation-using-google-a-test/
- Pour le syriaque : « Brief Notes on OCR and the Automated Transcription of Syriac Books » : https://digitalorientalist.com/2020/05/17/brief-notes-on-ocr-and-the-automated-transcription-of-syriac-books/
- Pour les textes en japonais : « Google Docs and OCR : Some Experiments Transcribing Japanses Language Texts » : https://digitalorientalist.com/2021/04/09/google-docs-and-ocr-some-experiments-transcribing-japanese-language-texts/

À propos https://github.com/tesseract-ocr/tessdoc. Lire aussi l'article sur Wikipédia qui est sourcé et indique de nombreuses références pour en savoir plus : https://en.wikipedia.org/wiki/Tesseract_(software)

Calfa Vision – Une plateforme d'annotation pour les langues orientales

Le projet Calfa est né en 2014 du constat de l'extrême difficulté de trouver des ressources papier pour l'étude de l'arménien classique. À partir de ce constat, partagé par des membres de la section d'arménien de l'INALCO, un projet de création d'un dictionnaire en ligne d'arménien classique-français a été proposé. L'équipe de Calfa a ensuite étendu l'idée en agrégeant et enrichissant de très nombreux contenus pour faire de leur site calfa.fr la référence la plus complète en lexicographie arménienne classique : traductions en anglais et en italien, exemples tirés de la littérature arménienne, dictionnaire unilingue, synonymes, étymologie, conjugaison, etc.

Dans un second temps, fin 2016, le projet s'est concentré sur le développement d'une technologie utilisant l'intelligence artificielle pour la reconnaissance des caractères manuscrits, focalisé sur l'arménien d'abord puis sur d'autres langues orientales. Calfa est spécialisée dans le développement d'outils sur mesure à destination des professionnels du patrimoine (que ce soit les bibliothèques, les universités ou les centres de manuscrits) qui ont besoin de traiter des données (manuscrits anciens, archives manuscrites, documents imprimés anciens) à des fins de préservation, de valorisation et d'exploitation de leurs documents. En plus de travailler sur l'HTR, Calfa propose également de l'analyse de textes automatique, du *post-tagging* ou de la lemmatisation pour exploiter au mieux les corpus.

Au départ, le projet a été centré sur l'arménien mais Calfa s'intéresse plus globalement aux langues orientales (arménien, syriaque, géorgien, arabe, chinois). La technologie est donc spécialisable en fonction des besoins d'un projet. Depuis 2021, Calfa est engagée dans plusieurs projets de reconnaissance automatique des caractères manuscrits.

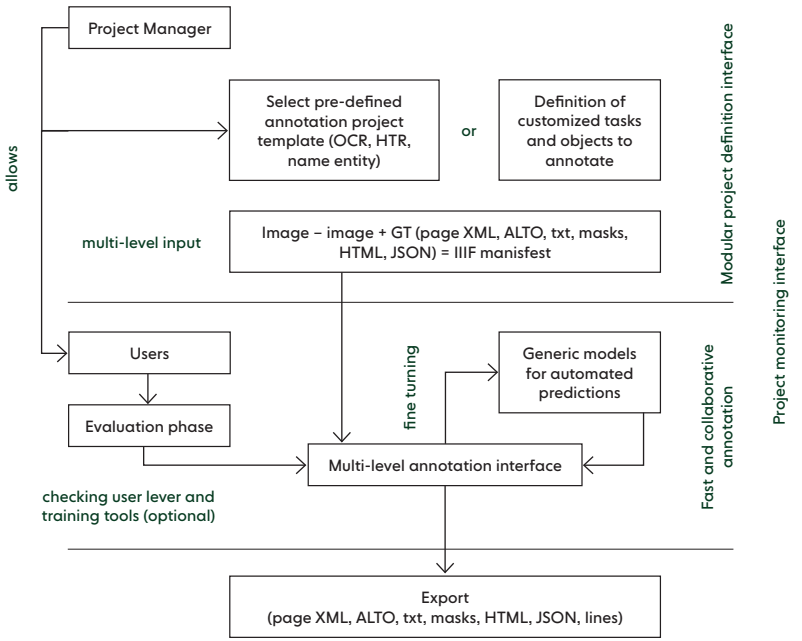


Fig. 7
Plan d'exploitation détaillé de la plateforme Calfa Vision (extrait de l'article : « A Modular and Automated Annotation Platform for Handwritings: Evaluation on Under-resourced Languages », p. 8)

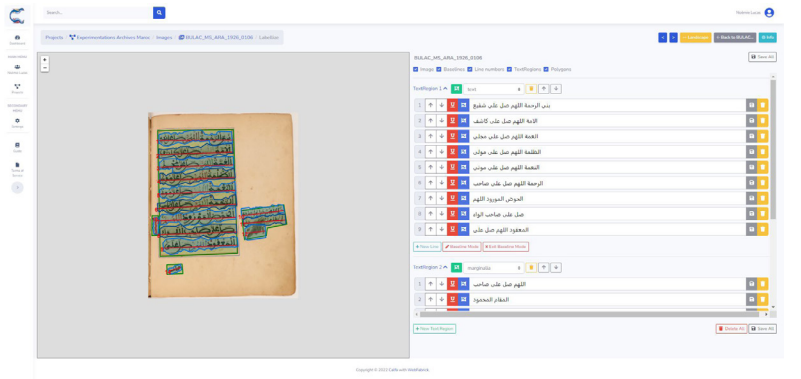


Fig. 8
Transcription de la page 106 du manuscrit MS ARA 1926 avec Calfa Vision (capture écran du site en février 2021)

La plateforme
https://vision.calfa.fr/
Il s'agit d'une plateforme en ligne en libre accès qui permet une annotation à plusieurs niveaux [FIG. 7]. Celle-ci propose une architecture modulaire permettant de créer et de suivre différents types de projets d'annotation. Les concepteurs de la plateforme précisent que l'objectif de la plateforme est d'aider à la création de données personnalisables à plusieurs niveaux de manière rapide et le plus facilement possible. Pour obtenir rapidement des modèles robustes pour les langues orientales considérées, et pour lesquelles les modèles basés sur les langues latines sont moins appropriés, l'accent est mis sur le *fine-tuning**.

Par ailleurs, l'interface proposée à l'utilisateur est pensée pour être ergonomique. L'utilisateur peut gérer son projet via une interface de gestion et dispose des statistiques en temps réel. De plus, la plateforme permet le travail collaboratif avec un nombre illimité d'utilisateurs par projet. L'organisateur du projet peut accéder aux statistiques des membres de son projet. Les données que la plateforme peut prendre en compte sont :

- soit une image ;
- soit une image avec une vérité terrain* préexistante ;
- soit un lien ver un manifeste IIIF* qui permet d'importer les métadonnées de tout le manuscrit.

En sortie, les exports peuvent être de plusieurs natures : image et fichier de vérité terrain (texte, XML, JSON, des masques).

Utilisation
L'interface propose des options similaires aux autres plateformes d'annotation. L'interface fournit volontairement peu d'options par défaut, pour accompagner l'utilisateur dans la familiarisation avec le logiciel. L'affichage est personnalisable selon les préférences des utilisateurs (affichage image-texte : soit gauche-droite, soit l'inverse, ou haut-bas ; ou encore texte sur l'image). La plateforme permet la correction des données prédites ou téléchargées, y compris la correction des *baselines* et des polygones. Plusieurs niveaux d'annotation sont possibles :

- les différentes zones de texte et leur caractérisation : texte, tableau, image, titre, réclame, etc. ;
- la ligne de texte qui peut être définie par : une *baseline*, un polygone environnant et le texte.

La détection automatique est disponible pour les zones de texte et les lignes de texte. Calfa propose plusieurs modèles de détection pour les régions de texte et les lignes de base. Ces modèles ont été entraînés à partir d'une base de données composée de documents en arménien et augmentée de différents documents manuscrits en arabe, hébreu, grec, syriaque et géorgien. Les jeux de données entraînés comprennent également des documents imprimés comme des journaux.

La stratégie privilégiée par l'équipe Calfa repose sur la capacité de ses modèles à être rapidement « affinables » en fonction des projets d'annotation. L'approche par *fine-tuning** est pensée pour faciliter la création de données. Concernant la prédiction des textes, la plateforme propose pour l'heure des modèles HTR pour quatre écritures orientales : l'arménien, le géorgien, le syriaque et l'arabe maghrébin. La mise à disposition de ces modèles est réservée aux partenaires de Calfa pour le moment.

La plateforme d'annotation Calfa Vision est donc une plateforme modulaire qui garantit une adaptabilité en fonction des besoins des projets et des langues considérées. Cette modularité est l'une des spécificités de la plateforme, ainsi que le fait qu'elle a été développée spécifiquement pour les langues orientales. Pensée par ailleurs pour être prise en main par des utilisateurs sans compétences informatiques, la plateforme est conçue pour faciliter le travail collaboratif.

À propos C. Vidal-Gorène, B. Dupin, A. Decours-Perez, T. Riccioli, « A Modular and Automated Annotation Platform for Handwritings: Evaluation on Under-resourced Languages », dans J. Lladós, D. Lopresti, S. Uchida (eds), *Document Analysis and Recognition – ICDAR 2021. ICDAR 2021. Lecture Notes in Computer Science*, vol 12823, Cham, Springer, 2021, p. 507-521. https://doi.org/10.1007/978-3-030-86334-0_33. « How AI can help researchers to transcribe their manuscripts (HTR)? », Calfa Team, 31 août 2021 : https://calfa.fr/blog/28. Calfa OCR : https://calfa.fr/ocr. Un guide d'utilisation est par ailleurs disponible sur la page d'accueil de l'interface : https://vision.calfa.fr/app/guide. Ce guide explique, étape par étape, comment utiliser l'interface et explicite toutes les fonctionnalités disponibles.

QUID DE L'INTEROPÉRABILITÉ ENTRE LES JEUX DE DONNÉES ET LES OUTILS ?

Nous avons listé et décrit succinctement un certain nombre de jeux de données et d'outils (logiciels, plateformes) qui sont utilisés dans le cadre de l'OCR/HTR de l'arabe. Pour certains d'entre eux, nous avons notamment montré les liens qui pouvaient exister mais que peut-on dire plus largement de l'interopérabilité* entre les jeux de données et les interfaces ? Toutes les interfaces présentées peuvent-elles traiter tous les jeux de données ?

Il apparaît que l'interopérabilité n'est nullement garantie et que les logiciels et les interfaces ont été développés pour un certain type de structuration de données. Aussi, lorsque les données récupérées ne suivent pas le protocole, un travail supplémentaire est nécessaire pour pouvoir utiliser ces données. Pour ne donner qu'un exemple, Kraken, et par extension eScriptorium, ne prend en compte que des images dans lesquelles toutes les lignes sont polygonisées. Aussi, dans le cas d'un *dataset* pour lequel le choix a été fait de ne pas polygoniser toutes les régions de texte, Kraken ne pourra pas traiter les images de ce *dataset* à moins de tout repolygoniser, ce qui engendrera un temps incompressible de travail supplémentaire.

Il faut aussi considérer le cas des *datasets* qui ont été créés il y a cinq ans ou plus et qui ne sont parfois plus utilisables car ils utilisent des formats qui ne sont plus standards ou plus privilégiés. Pour garantir l'interopérabilité dans le temps de ces *datasets*, il est nécessaire de concevoir des API*. Cette API doit être prévue en amont car, une fois le projet terminé, la question des moyens et des disponibilités pour prendre en charge ce développement est plus difficilement envisageable.

068 [II] E
OCR ET HTR DANS LES PROJETS DE RECHERCHE

Dans cette section, sont considérées les recherches menées sur l'OCR/HTR des graphies arabes dans le cadre de projets qui ne sont pas technologiques au départ, au sens où ils n'ont pas été conçus ou élaborés pour développer des outils, mais au sein desquels l'HTR en tant que méthode d'acquisition de texte vient répondre à des questions scientifiques et/ou patrimoniales. Nous ne proposons ici que quelques exemples et souhaitons d'ores et déjà pointer l'absence de projet français dans la liste. Nous nous concentrons par ailleurs sur l'HTR, même si certains projets mentionnés concernent plutôt, ou aussi, les recherches en OCR.

Le projet HTR pour les manuscrits scientifiques arabes de la British Library

La British Library conserve une importante collection de manuscrits arabes qui comprend environ 15 000 œuvres religieuses, historiques, littéraires et scientifiques conservées dans 14 000 volumes[36]. Depuis 2012, en partenariat avec la Bibliothèque nationale du Qatar, 1 921 252 images dont 76 092 de la collection de la British Library ont été numérisées sur la bibliothèque numérique du Qatar[37].

Depuis environ quatre années, la conservatrice des collections africaines et asiatiques, Adi Keinan-Schoonbaert, et la *Digital Curator*, Nora McGregor, portent un projet autour de l'HTR et des manuscrits arabes intitulé *Automatic Transcription of Historical Handwritten Arabic Texts*. Leur

[36] Sur la collection : https://www.bl.uk/collection-guides/arabic-manuscripts

[37] Pour consulter les manuscrits numérisés : Qatar Digital Library. 80 manuscrits arabes sont disponibles. Pour plus d'informations sur les manuscrits numérisés et les hyperliens, deux billets peuvent être consultés:
- Arabic scientific manuscripts go live in Qatar Digital Library
Forty more Arabic scientific manuscripts go live in Qatar Digital Library
- Concernant les autres manuscrits numérisés par la British Library: Une petite collection de manuscrits arabes comprenant des textes théologiques druzes a également été numérisée et est disponible ici Digitised Manuscripts. Consulter également Images Online.

objectif est d'entraîner un logiciel à lire les manuscrits historiques arabes afin de rendre consultables les manuscrits numérisés en rendant ces derniers disponibles à la recherche en texte intégral.

Leur projet est de poursuivre la recherche en matière de reconnaissance des textes arabes manuscrits. Il vise à soutenir la recherche dans ce domaine en fournissant un ensemble d'images ouvertes et de données de vérité terrain* de textes arabes manuscrits historiques, afin que les collections arabes historiques continuent de bénéficier des développements de pointe en matière de reconnaissance de textes manuscrits (HTR). Le projet a connu plusieurs phases directement en lien avec les deux grandes conférences ICFHR et ICDAR (voir *supra*).

La démarche choisie a été celle de la transcription collaborative. Pour ce faire, les chercheurs ont utilisé la plateforme *open source From the Page* (équivalente à l'outil français de transcription collaborative Transcrire) pour laquelle un développement a été réalisé par le département des études numériques de la British Library afin de permettre la transcription de droite à gauche. Les dix premières pages ont été transcrites par une équipe de la British Library puis 36 personnes ont transcrit 85 pages tirées de 9 manuscrits en dix jours. Cette première étape de *crowdsourcing* a été suivie de l'organisation d'un atelier de transcription des manuscrits arabes scientifiques. La suite du projet a pris la forme de deux compétitions organisées dans le cadre des deux grandes conférences internationales ICFHR et ICDAR dont sont issus un article et plusieurs billets de blog.

ICFHR 2018 – Competition on Recognition of Historical Arabic Scientific Manuscripts

Cette compétition a été organisée en partenariat avec l'Institut Alan Turing et le PRImA Research Lab[38], partenaires de la British Library au mois d'août 2018 à Niagara Falls (USA). En partant de la vérité terrain créée (*crowdsourcing* et *workshop*), la compétition a mis l'accent sur la recherche d'une solution optimale pour transcrire précisément et automatiquement les manuscrits arabes scientifiques.

Les participants disposaient des images numérisées et de leurs fichiers XML attachés. Les images provenaient de collections de manuscrits numérisés. Les fichiers indiquaient les différentes régions de textes et lignes ainsi que leur transcription. Cette vérité terrain était utilisée pour entraîner les systèmes de reconnaissance de texte des participants à automatiquement identifier la graphie arabe sur 85 autres images, fournies par l'équipe. Cette collection représentait un échantillon pertinent pour traiter des différentes difficultés concernant l'analyse de la mise en page et l'OCR, parmi lesquelles la présence de lignes de texte non droites, les transparences, l'encre délavée, les décorations, la présence de régions de forme non rectangulaire, la variation de la largeur des colonnes de texte, la variation de la taille des caractères et divers problèmes liés au vieillissement et à la numérisation.

Trois défis distincts étaient proposés et portaient sur : la segmentation de la page (en régions), la détection des lignes de textes et la

[38] Un laboratoire de recherche sur « Pattern Recognition and Image Analysis ».

reconnaissance du texte. Trois méthodes proposées par les compétiteurs ont été évaluées : Google Cloud Vision API, KFCN, proposée par Berat Kurar de l'Université Ben-Gurion du Negev et RDI, proposée par Hany Ahmed de la RDI Company (Université du Caire).

Il faut noter que RDI-Corporation dispose de son propre système d'OCR pour l'arabe manuscrit historique, qui a été élaboré à partir de différents manuscrits historiques. En guise de comparaison, Tesseract OCR 3.04, 4.0 (beta) et ABBYY FineReader Engine 11 ont été également utilisés, lesquels sont plutôt optimisés pour les textes imprimés.

Résultats

Les résultats pour *Page Layout Analysis*/Analyse de la mise en page - *Text Line Segmentation*/Segmentation des lignes de texte - Précision de l'OCR/*OCR Accuracy* peuvent être résumés comme suit :

- KFCN obtient 87,9 % de bonne segmentation pour la segmentation de la page, RDI 81,6 % pour la reconnaissance des lignes. En matière d'OCR, RDI obtient les meilleurs résultats, 78,1 % sur les textes orignaux et 85,4 % sur les textes normalisés.

Notons que l'article[39] se termine sur le paragraphe suivant :

> *« De bons résultats ont été obtenus dans les trois défis. Les domaines à améliorer sont notamment la séparation des régions. Les notes marginales proches du texte principal ont souvent perturbé les méthodes de numérisation. Des dictionnaires historiques spécialisés pourraient encore améliorer les résultats de l'OCR. »*

ICDAR 2019 – Competition on Recognition of Historical Arabic Scientific Manuscripts

La deuxième compétition a été organisée à Sydney en septembre 2019 dans le cadre de l'ICDAR. Les mêmes objectifs étaient poursuivis partant d'une vérité terrain* d'une centaine de pages, considérant en plus les marges. 120 images et leurs fichiers XML associés, 20 pages pour entraîner les modèles et 100 pages pour évaluer l'entraînement étaient fournies aux participants.

Seule RDI Company de l'Université du Caire a candidaté pour cette compétition en proposant trois modèles différents pour la segmentation des lignes de texte et l'OCR. Au cours de l'évaluation des résultats, le laboratoire de recherche PRImA a effectué une comparaison avec les systèmes utilisés dans l'industrie et le monde académique : Tesseract 4.0, ABBYY FineReader Engine 12 et Google Cloud Vision API.

Résultats

- *Page Layout Analysis*/Analyse de la mise en page
  Les tests ont été réalisés avec Tesseract4, FRE12 et Google et il ressort que Google Cloud Vision obtient les meilleurs résultats,69,3 % de bonne reconnaissance, marges inclues [Tesseract4 et FRE12 n'atteignent pas 43 %].

[39] C. Clausner, A. Antonacopoulos, N. Mcgregor, D. Wilson-Nunn, « ICFHR 2018 Competition on Recognition of Historical Arabic Scientific Manuscripts - RASM2018 », art. cit.

- *Text Line Segmentation*/Segmentation des lignes de texte
  Les trois méthodes proposées par RDI Company obtiennent de bien meilleurs résultats que Tesseract4 et FRE12 mais n'atteignent pas 80 % (maximum 77,6 %) et la détection des marges reste limitée.
- Précision de l'OCR/*OCR Accuracy*
  Les tests réalisés (en intégrant les marges et en n'évaluant que sur le texte principal) montrent que des trois Tesseract4, FRE12 et Google, Google obtient les meilleurs résultats avec plus de 60 % de bonne reconnaissance, alors que le moteur n'était pas spécifiquement entraîné et optimisé pour la compétition. Quant aux modèles de RDI, ils ne dépassent pas 78,72 % pour le plus performants des trois (soit un moins bon résultat que pour RASM 2018 – 85,44 %)

Les vérités terrain créées sont disponibles en *open access* ici : https://bl.iro.bl.uk/concern/datasets/f866aefa-b025-4675-b37d-44647649ba71?locale=en

En septembre 2019, les porteurs de ce projet annonçaient que la suite consistait à tester leurs matériels avec Transkribus et possiblement Kraken. Un billet posté en janvier 2020 décrit le test réalisé avec Transkribus. Il faut noter que la British Library est l'un des membres fondateurs de READ-COOP (voir *supra*). Pour le test, 4 modèles différents ont été créés afin de voir comme les algorithmes de reconnaissance de Transkribus géraient un ensemble d'apprentissage croissant.

Il en ressort que la précision de l'HTR s'est améliorée à chaque itération de l'entraînement. Plus les réseaux de neurones du moteur de Transkribus disposent de données d'entraînement, meilleurs sont les résultats. Avec un ensemble d'apprentissage de 100 pages, Transkribus a réussi à transcrire automatiquement le reste des 20 pages avec un taux de précision de 86,43 %.

À propos Pour les résultats de la compétition ICDAR 2019 : https://blogs.bl.uk/digital-scholarship/2019/09/rasm2019-results.html?_ga=2.80085330.438617550.1629730734-427714573.1629470599. Pour le travail entrepris par la British Library en matière d'OCR/HTR, spécifiquement d'HTR : Adi Keinan-Schoonbaert, « Using Transkribus for Arabic Handwritten Text Recognition » : https://blogs.bl.uk/digital-scholarship/2020/01/using-transkribus-for-arabic-handwritten-text-recognition.html. Présentation du travail d'Adi Keinan-Schoonbaert https://www.bl.uk/people/experts/adi-keinan-schoonbaert?_ga=2.31276855.965525789.1629470599-427714573.1629470599#

COBHUNI Project

Depuis 2015, le projet COBHUNI, financé par l'ERC et piloté par le Professeur Thomas Eich à l'Institut Afrique-Asie de l'Université de Hambourg, comprend une dimension OCR. Le projet porte sur les idées concernant la vie prénatale dans l'histoire islamique, sur la manière dont ces idées ont évolué au cours des 1 400 ans d'histoire de l'Islam. L'objectif est de montrer quels facteurs ont influencé ces idées en fonction des époques.

L'analyse de la littérature exégétique sur le Coran et les Hadiths constitue une importante part du travail réalisé par l'équipe de ce projet. Leur objectif est d'identifier les liens et les chevauchements entre les textes. L'étude de ce vaste matériel est réalisée à partir de textes arabes numérisés par l'application d'outils d'analyse linguistique computationnelle. Le recours à ces méthodes a pour objectif d'identifier des modèles de citation. Pour ce faire, le corpus comprend des textes déjà numérisés et accessibles sur le web en mode

texte, des textes numérisés dans le cadre du projet et des textes saisis à partir de manuscrits numérisés ou d'éditions de texte de mauvaise qualité.

Parmi les approches méthodologiques privilégiées par l'équipe de ce projet afin de disposer du plus de données possible, le recours à des logiciels OCR figurait en bonne place. L'objectif était d'utiliser un OCR pour extraire les textes des documents numérisés dans le cadre du projet. Les membres du projet ont d'abord utilisé Tesseract avec, pour la post-correction le recours à l'extension de relecture de Mediawiki qui supporte l'Unicode et les écritures sinistroverses.

Dans l'article qui a résulté de la présentation faite à DATECH, ils annonçaient se tourner vers d'autres outils comme Kraken, qui était alors développé à l'Université de Leipzig. Depuis, les recherches présentées en lien avec l'OCR et l'arabe ne semblent pas directement en lien avec ce projet. Il s'agit notamment des recherches réalisées par Alicia Gonzalez Martinez, la spécialiste en linguistique computationnelle du projet, avec Thomas Milo. Ces deux chercheurs ont présenté leurs recherches sur l'OCR à partir d'une reconnaissance optique des caractères arabes basée sur l'analyse et la synthèse de l'écriture, en 2019 lors de la conférence DATeCH, et plus récemment en 2020 dans un article paru dans la revue *Égypte/Monde arabe*.

À propos
Alicia González Martínez, Tillmann Feige, and Thomas Eich. « Clear-cut methodology for Arabic OCR and post-correction with low technical skilled annotators », *DATeCH2017: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, New York, ACM, 2017, p. 67-70. https://doi.org/http://dx.doi.org/10.1145/3078081.3078103
Pour se renseigner sur le projet : https://www.cobhuni.uni-hamburg.de
Alicia González Martínez, Thomas Milo, « A New Strategy for Arabic ICR: Archigraphemes, Letter Blocks, Script Grammar and shape synthesis », *DATeCH2019: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage Brussels 2019*, New York, ACM, 2019, p. 93-96. https://dl.acm.org/doi/10.1145/3322905.3322928
« A New Strategy for Arabic OCR based on Script Analysis and Synthesis », *Égypte/Monde arabe*, Troisième série, 22, 2020, p. 21-30. https://journals.openedition.org/ema/13146?lang=en

NYU Abu Dhabi et l'HTR du turc ottoman

À la New York University à Abu Dhabi, David Wrisley (Associate Professor of Digital Humanities) et Süphan Kirmizialtin (Visiting Assistant Professor of Middle Eastern History) sont investis dans un projet d'OCR sur les documents en turc ottoman, à partir notamment d'un corpus de périodiques provenant du fonds *Hakki Tarik Us*, comprenant 1 500 périodiques pour un nombre de pages d'environ 400 000. Le nombre limité d'archives ottomanes numérisées, et plus généralement de corpus de documents historiques en turc ottoman disponibles, constitue l'un des enjeux pour la recherche sur ces questions. L'objectif de ce projet est de créer un corpus de documents en turc ottoman qui soit fouillable et de rendre disponible ce corpus, imprimé dans une langue qui n'est plus parlée et dont l'accès est rendu difficile par le nombre réduit de personnes la maîtrisant. Pour ce faire, la plateforme d'annotation utilisée est Transkribus (voir *supra*).

À propos
« Automated Transcription of Non-Latin Script Periodicals: A Case Study in the Ottoman Turkish Print Archive », paru en 2020 : https://arxiv.org/abs/2011.01139

The Ottoman Text Recognition Network (OTRN)

Ce réseau de chercheurs travaillant sur l'Empire ottoman a été fondé récemment, probablement en 2020 (leurs premières activités datent de cette année-là) et est fédéré par Yavuz Köse, titulaire de la chaire en études ottomanes et turques à l'Université de Vienne. Il réunit des chercheurs internationaux ;

parmi les fondateurs, on compte en particulier des chercheurs issus des universités allemandes et turques. Ces derniers partagent le constat suivant : l'étude de l'Empire ottoman dépend d'archives qui sont dans leur immense majorité des documents manuscrits et qui présentent un certain nombre de difficultés. En plus des spécificités des écritures arabes (voir I.B), leur corpus compte par ailleurs un grand nombre de textes turcs rédigés avec des graphies différentes de l'arabe comme l'arménien, l'hébreu ou le grec. C'est donc pour essayer de pallier ces difficultés que ce réseau s'intéresse à l'HTR en postulant que ces méthodes peuvent les accompagner que ce soit dans la transcription des documents ou dans la conversion de ces documents dans des formats fouillables.

Parmi les initiatives qu'ils ont menées depuis 2020, on compte une série de conférences organisée entre octobre 2020 et février 2021, qui n'était néanmoins pas uniquement consacrée à la reconnaissance automatique des caractères. D'après les titres, deux conférences étaient spécifiquement dédiées à cette question, celle de Süphan Kirmizialtin (voir *supra*) et celle de Aysu Akcan et Yavue Köse de Vienne.

Ce réseau compte organiser un *workshop* portant sur ces questions de reconnaissance automatique tous les ans. La première édition s'est tenue en février 2021. L'objectif de cet atelier d'une journée était notamment de présenter les premiers résultats du travail mené à partir de la plateforme Transkribus. La deuxième édition est prévue en février 2022. Il s'agira d'un évènement de trois jours qui dépassera la question de l'HTR ; il est organisé en partenariat avec le département de sciences informatiques de l'Université de Hambourg et en collaboration avec le Centre autrichien pour les humanités numériques et l'héritage culturel (ACDH-CH The Austrian Centre for Digital Humanities and Cultural Heritage).

À propos
https://otrn.univie.ac.at/.
Sur les évènements passés et à venir: https://dh-ottoman.univie.ac.at/

ERC Mamlukisation of the Mamluk State II (MSS II) – IHODP et le projet Corpus

Dans le cadre du projet ERC Mamlukisation of the Mamluk State II, hébergé par l'Université de Gand et dirigé par Jo Van Steenbergen de 2017 à 2021, une plateforme intitulée IHODP pour « Islamic History Open Data Platform » a été élaborée. IHODP se présente comme une plateforme *open source* pour les chercheurs et elle entend connecter les trois projets qu'elle héberge à savoir, le Mamluk Proposopography project (MPP), la Bibliography of 15th Century Arabic Historiography (BAH) et le projet CORPUS : Texts from Late Medieval Egypt and Syria.

CORPUS consiste en la mise à disposition de textes permettant la recherche plein texte. 82 textes sont ainsi disponibles au format électronique. Ils proviennent d'une part du corpus d'OpenITI (voir *infra*) et d'autre part, il s'agit de textes édités qui ont été OCéRisés par l'équipe du projet, notamment par Manhal Makhoul.

Afin de réaliser cette extraction de textes imprimés, plusieurs plateformes ont été utilisées. L'équipe a d'abord eu recours à Tesseract, dont les résultats étaient surtout satisfaisants dans le cas de mise en page simples et de ligatures nettes. Les erreurs restaient importantes dans le cas des textes vocalisés notamment. Puis, la plateforme Transkribus a été employée et plusieurs modèles ont été entraînés par l'équipe. Les résultats, quoi que

satisfaisants, restaient limités dans le cas des passages vocalisés (les signes vocaliques étaient transcrits sur une ligne à part) et de mises en page plus complexes. Après que l'utilisation de la plateforme est devenue payante, l'équipe du projet s'est tourné vers Calfa et la plateforme Calfa Vision. Dans ce cadre, 30 000 pages imprimées en arabe ont été traitées avec un résultat de 99 % en moyenne.

À propos Concernant le projet : https://www.mms.ugent.be/
Pour accéder à IHODP : https://www.mms.ugent.be/IHODP/
Projet Corpus : https://ihodp.ugent.be/corpus/

OpenITI – Open Islamicate Text Initiative et la reconnaissance automatique des écritures imprimées et manuscrites

L'Open Islamicate Text Initiative (OpenITI) est un projet mené par des chercheurs de l'Aga Khan Université (AKU), l'Université de Vienne/Université de Leipzig (LU) et l'Institut Roshan pour les études persanes à l'Université du Maryland (College Park) depuis 2016. Les principaux porteurs sont Sarah Savant (AKU), Maxim Romanov (Université de Vienne[40]) et Matthew Miller. Ils sont accompagnés par un conseil consultatif interdisciplinaire composé de spécialistes en études islamiques, persanes et arabes, et en humanités numériques.

Le constat de départ formulé par ces chercheurs était que le patrimoine textuel du monde islamique est immense et que les fonds de manuscrits arabes, persans, ottomans, etc. sont nombreux, ce qui les rend particulièrement adaptés pour l'application de méthodes d'analyse computationnelles comme celles appliquées aux textes dans d'autres traditions linguistiques. Le problème des fonds numérisés ou numériques existant pour les textes arabes et persans est de trois ordres : (1) les standards internationaux de données ne sont pas nécessairement respectés, (2) les métadonnées scientifiques sont manquantes, (3) toutes les traditions historiques ne sont pas également représentées.

L'objectif du projet est de construire le premier corpus scientifique de textes islamiques prémodernes pouvant être exploité par une machine. Pour le moment, OpenITI n'est pas une librairie numérique disposant d'une interface d'utilisation, d'un environnement de lecture et de fonctions de recherche mais son objectif est de fournir « une interface textuelle […] pour de nouvelles formes d'analyse macro-textuelle et de recherche numérique » pour les textes écrits par des musulmans et/ou dans des contextes où les musulmans étaient socialement et culturellement dominants. Cela comprend donc des textes écrits par des musulmans et des non-musulmans, sur tous les sujets imaginables, dans un nombre varié de langues, dont l'arabe, le turc, le syriaque, le moyen-persan et le néo-persan.

La dernière version du corpus comprend 10 202 fichiers texte. Chaque fichier contient le texte intégral d'une œuvre. Certaines œuvres disposent de plusieurs versions (numérisations d'éditions papier distinctes, numérisations distinctes de la même édition ou numérisations identiques de la même édition). Les 10 202 fichiers textes représentent 6 236 œuvres distinctes rédigées par 2 582 auteurs. Les textes d'OpenITI proviennent de diverses sources. La majorité (8 431 textes) ont été extraits de douze

[40] Maxime Romanov est désormais à l'Université de Hambourg en qualité de « Emmy Noether Junior Research Group Leader » pour le projet « The Evolution of Islamic Societies: Algorithmic Analysis into Social History ».

collections de textes numériques (dont al-Maktaba al-Shāmila, al-Jāmiʿ al-Kabīr et ShiaOnlineLibrary) ; un nombre limité mais en augmentation a été numérisé par l'équipe du projet KITAB par transcription manuelle ou en utilisant l'OCR (26 textes) ou fourni par les utilisateurs d'OpenITI (19 textes)[41].

Le projet OpenITI comporte en effet un volet de recherche sur les possibilités de reconnaissance optique des caractères. Les membres de l'équipe ont d'abord concentrés leurs efforts sur l'OCR pour les imprimés en caractères arabes et depuis peu, ils travaillent sur le développement de modèles d'HTR pour les manuscrits.

Ce travail a débuté dans le cadre d'une collaboration avec Benjamin Kiessling commencée en 2016. Benjamin Kiessling est l'ingénieur qui a développé Kraken (voir *supra*). En 2017, les membres du projet ont présenté les premiers résultats de leurs recherches entreprises en matière d'OCR pour les imprimés arabes avec Kraken dans un article intitulé « Important new development in Arabographic optical character recognition (OCR) »[42].

Avec cet article, un *dataset* de 7 000 lignes de textes en arabe imprimé a été mis à la disposition de tous sur un GIT à des fins d'entraînement de modèles OCR de l'arabe ou de test[43]. À partir de l'édition du *Kitāb al-Buldān* d'Ibn al-Faqīh, les chercheurs ont développé un modèle d'entraînement (avec 1 000 lignes transcrites). Ayant obtenu une bonne reconnaissance de 97,56 % (99,68 % si on ne prenait en compte que l'arabe, donc sans ponctuation et espace), ils ont ensuite testé ce modèle sur six autres éditions de manuscrits arabes qui différaient en matière de typographies, de graphies, de conventions orthographiques et aussi de qualité de l'image. Quatre des éditions utilisées étaient de bonne qualité et les deux autres, d'une qualité moindre. Or les meilleurs résultats en matière de reconnaissance des caractères ont été obtenus sur les deux éditions présentant une qualité moindre. Sur ces éditions, la reconnaissance a été d'environ 97 % pour l'arabe uniquement et d'environ 93 % si on prend en compte la ponctuation et les espaces. Pour les quatre autres, les résultats obtenus de reconnaissance n'ont pas dépassé 90 % pour l'arabe seul, se situant en moyenne à 75 % avec la ponctuation et les espaces. Les auteurs de l'article expliquent ces résultats en mettant en avant deux éléments :

- Les polices de caractères des éditions pour lesquelles les résultats ont été les meilleurs étaient plus proches de l'édition initiale ayant servi à développer le modèle.
- La qualité de l'image n'a pas un impact sur le taux de reconnaissance aussi important qu'ils l'avaient imaginé au départ.

Enfin, l'article précise par ailleurs que pour entraîner de nouveaux modèles à partir du modèle initial, il fallait transcrire 800 lignes de texte.
Depuis les auteurs travaillent à identifier toutes les polices de caractères

[41] Toutes les informations et chiffres ont été tirés de : https://openiti.github.io/documentation/. Nous invitons le lecteur à consulter cette page très fournie et précise sur le projet et le GIT.

[42] Matthew Miller, Benjamin Kiessling, Maxime Romanov, Sarah Savant, « Important New Development in Arabographic Optical Character Recognition (OCR) », *al-ʿUsūr al-Wusṭā*, 25, 2017, p. 1-13.

[43] GitHub - OpenArabic/OCR_GS_Data: Double-checked Gold Standard Data for Training and Testing OCR Engines

arabes utilisées dans les éditions de textes pour créer une bibliothèque de modèles OCR en fonction des polices de caractères et des styles éditoriaux également.

Ce travail est effectué dans le cadre de OpenITI AOCP (*Open Islamicate Text Initiative Arabic-script OCR Catalyst Project*), qui est financé par The Andrew W. Mellon Foundation depuis juin 2019[44]. Au départ, le projet visait à transformer CorpusBuilder 1.0, une chaîne de traitement OCR et une interface de post-correction, en un *pipeline* de production de texte numérique ergonomique avec un OCR amélioré et une fonctionnalité d'export. Il comprenait notamment :

- la création d'un OCR pour l'arabe imprimé avec un travail sur les polices de caractères de l'imprimé. Ce travail inclut l'établissement d'une typologie distinguant un peu plus d'une douzaine de types de police et l'entraînement de modèles sur l'ensemble de celles-ci afin d'avoir un modèle généraliste pour toutes les polices de caractères.
- l'amélioration de la reconnaissance des régions de texte dans la page. En effet, le modèle utilisé jusqu'alors était un modèle d'analyse de la mise en page identifiant une seule zone de texte.

Depuis 2020, l'équipe d'OpenITI collabore avec eScriptorium et utilise leur interface (voir *supra*). Par ailleurs, le projet OpenITI inclut désormais non seulement des recherches sur l'OCR mais également sur l'HTR. En effet, depuis 2020, les travaux se portent sur la reconnaissance automatique des caractères manuscrits. Dans cette perspective, Matthew Thomas Miller a reçu un financement de la NEH (National Endowment for the Humanities, agence fédérale américaine destinée à soutenir la recherche) pour un projet intitulé « Automatic Collation for Diversifying Corpora: Improving Handwritten Text Recognition (HTR) for Arabic-script Manuscripts » (ACDC) qui est co-piloté par David Smith (Northeastern University – Khoury College of Computer Sciences).

ACDC

L'objectif de ce projet est d'améliorer les résultats des modèles HTR pour les graphies arabes manuscrites en développant un outil de collation qui crée automatiquement de grandes quantités de données d'entraînement à partir des textes numériques et des images de manuscrits, en évitant l'étape d'annotation par l'humain de chaque manuscrit qui exige un temps de travail important. Le projet entend mener cette tâche en exploitant et prolongeant les capacités d'alignement des textes qu'offrent l'outil Passim (https://github.com/dasmiq/passim) et Kraken afin d'aligner des transcriptions de différents exemples de manuscrits, provenant d'HTR et de qualité limitée, avec des textes numériques existants pour produire des données d'entraînement avec une supervision minimale. À terme, avec cette approche qui vise à accélérer le processus de production de données d'entraînement, les chercheurs de ce projet entendent créer des modèles HTR généralisés pour l'arabe et le persan.

[44] https://medium.com/@openiti/openiti-aocp-9802865a6586 ; https://www.endpointdev.com/blog/2019/09/openiti-arabic-ocr-catalyst-project/

Ce projet vise ainsi à pallier l'absence de données d'entraînement, autrement dit ces jeux de données présentés précédemment, et à augmenter le plus rapidement possible ces données, qui sont essentielles pour le développement des modèles. C'est d'ailleurs la raison pour laquelle Matthew T. Miller précise, dans sa description du projet, que l'équipe produira et mettra à disposition des données d'entraînement en libre accès et des tutoriels pour utiliser cet outil qui sera disponible sur la plateforme eScriptorium.

À propos Article sur le projet OpenITI : Mathew T. Miller, Maxim G. Romanov, Sarah Bowen Savant « Digitizing the Textual Heritage of the Premodern Islamicate World: Principles and Plans », *IJMES*, 50/1, 2018, p. 103-109. https://openiti.github.io/documentation/. Voir aussi: https://matthewthomasmiller.github.io/2020/NEH-Awards-HTR-Grant-Roshan-UMD/

078 [II] F OBSERVATIONS ET CONCLUSIONS PROVISOIRES

À l'issue de cet état des lieux des recherches, des jeux de données, des outils et des projets en lien avec l'HTR et les écritures arabes, plusieurs observations peuvent être dégagées. Nous structurons celles-ci en quatre thèmes qui ressortent, selon nous, de la présentation proposée

HTR, intelligence artificielle et graphies arabes : le point sur l'état des recherches et des possibilités

À partir de 2010, à la faveur des avancées techniques autorisées par les développements de l'intelligence artificielle et de l'apprentissage profond*, la reconnaissance automatique des écritures manuscrites a connu d'importants progrès offrant aux philologues des possibilités nouvelles. Les recherches sur les écritures arabes ne sont pas restées à l'écart de ces avancées. Les conférences internationales dédiées à ces technologies en témoignent ainsi que la création en 2017 d'un atelier qui a entièrement porté sur les écritures arabes, l'International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR), initié par deux équipes française et tunisienne (voir II.B).

Les architectures OCR/HTR s'intéressent de plus en plus à l'élaboration de modèles spécialisés sur les graphies non latines, en particulier l'arabe et tentent de traiter les différents enjeux présentés dans la première partie de ce rapport, en particulier ceux propres à l'analyse de la mise en page et aux spécificités des graphies arabes. Il faut signaler ici le travail pionner de Benjamin Kiessling avec le développement de Kraken.

On remarque néanmoins que la plupart des recherches et des articles publiés sur ces questions concernent peu les documents historiques en arabe et sont plutôt consacrés à des recherches sur les reconnaissances des écritures cursives dans le cadre de questionnements pratiques contemporains. Cette observation est confirmée au vu des éditions de la récente

conférence DATeCH, lancée en 2014 et entièrement consacrée à l'analyse des documents historiques avec le numérique. Les contributions sur l'arabe sont presque inexistantes. Cela nous semble être le signe d'une part d'un nombre limité de projets de recherche en cours sur des documents historiques en arabe qui utilisent l'HTR dans leur processus d'acquisition des textes, et plus largement qui font appel aux méthodes computationnelles dans l'analyse de ceux-ci, et d'autre part, d'une faible collaboration entre les équipes d'ingénieurs qui travaillent sur ces questions et les chercheurs en sciences humaines.

Les collaborations entre ingénieurs et chercheurs en sciences humaines

Les verrous empêchant un dynamisme des recherches sur l'HTR de l'arabe sur des documents historiques (presse, manuscrits, etc.) sont moins de nature technique, même si l'on ne peut nier l'existence de ces enjeux propres à la nature des documents considérés et aux graphies, que de nature relationnelle ou humaine. Ce qui manque est une collaboration réelle entre les équipes d'ingénieurs et les chercheurs en sciences humaines. Ce déficit de collaboration s'explique selon nous par un déficit de demande de la part des potentiels utilisateurs de ces technologies qui sont les chercheurs étudiant ces textes.

Pourtant, il faut rappeler ici que la collaboration est essentielle et indispensable pour le développement des jeux de données que ce soit à l'étape de la constitution de la vérité de terrain ou pour la réalisation des tests. Ainsi apparaît-il par exemple que certaines équipes d'ingénieurs s'intéressent précisément aux documents historiques en arabe, à l'image du Visual Media Lab (VML) de l'Université Ben-Gurion du Negev, mais que leurs recherches et les *datasets* créés ne semblent pas reliés à des projets scientifiques sur ces corpus de manuscrits.

À l'inverse, on peut signaler deux exemples de collaboration entre chercheurs, ingénieurs, mais aussi conservateurs qui rendent compte des avantages de cette mise en commun des intérêts et des compétences :

- tout d'abord, les projets entrepris par la British Library en partenariat avec la bibliothèque numérique du Qatar d'une part et avec l'Institut Alan Turing et le PRImA Research Lab d'autre part ;
- d'autre part, le projet OpenITI, qui vient de chercheurs dans le domaine des études arabes et persanes et qui dépasse la seule question de l'HTR mais qui, pour répondre à des besoins scientifiques, participe activement aux recherches sur l'OCR (développement de Kraken par Benjamin Kiessling dans le cadre de ce projet) et l'HTR (récent partenariat avec eScriptorium et utilisation de l'interface et travail aux États-Unis avec le Khoury College of Computer Sciences à la Northeastern University).

Dans les deux cas, les partenariats entre institutions et projets et les collaborations entre compétences sont la clef du dynamisme de ces recherches et de leur visibilité.

Il est intéressant de noter que dans le cas de collaborations, sinon réussies, du moins démontrant un réel dynamisme, les recherches réalisées sont présentes dans toutes les catégories susmentionnées dans cet état des lieux, à savoir : participation aux conférences internationales, création de

jeux de données, projets de recherche utilisant des interfaces et outils, voire développant certains outils

Des outils, quelques leaders et peu d'architectures pensées pour l'arabe

Aucune des principales plateformes utilisées aujourd'hui (en particulier Transkribus qui semble être utilisée dans de nombreux projets, mais c'est le cas également d'OCR4all qui est plus récente, et également d'eScriptorium) n'ont été pensées précisément pour les langues orientales et les langues non-latines. La seule plateforme qui se distingue en cela est la plateforme développée par Calfa, Calfa Vision, qui propose des services HTR et OCR spécifiquement dédiés aux langues orientales.

Évaluer ces outils et ces plateformes n'est pas l'objet de ce rapport et la question n'est pas pertinente, à moins de se poser à l'échelle d'un projet spécifique ou d'un corpus de documents. En d'autres termes, à la question de l'efficacité respective de ces différents outils pour les textes historiques en arabe, il n'est pas possible de répondre en général. Tout dépend :

- de la nature du projet, de ses objectifs et de ses besoins ;
- du budget et des financements alloués à la partie OCR/HTR du projet ;
- de la nature du corpus et des documents (voir II.C) ;
- des compétences en informatique des chercheurs participant au projet en informatique et de leur familiarité avec les humanités numériques et l'HTR.

Un champ des études françaises sur le Moyen-Orient, les mondes musulmans et le Maghreb en marge

Le dernier point sur lequel nous souhaitons mettre l'accent dans ces observations est l'absence des équipes et des chercheurs français travaillant dans le champ des études sur le Moyen-Orient, les mondes musulmans et le Maghreb. Cela ne veut pas dire que des chercheurs, à titre individuel, ne font pas usage de ces technologies pour acquérir leurs textes, mais nous n'avons pas trouvé trace de communication à ce sujet. Il y a manifestement un manque de demande, alors même que des outils ont été développés en France : nous pensons en particulier à eScriptorium qui est un projet de Scripta-PSL ou à Calfa.

Ce constat est cohérent avec l'état des lieux et les conclusions rédigés dans le livre blanc *Vers la science ouverte ?*, publié en 2020. On notera néanmoins qu'aux échelles européenne et internationale, le nombre de projets initiés par des chercheurs travaillant sur des manuscrits arabes et ayant l'intention de recourir aux technologies de l'HTR est également limité. Un dynamisme semble repérable du côté des études sur les textes ottomans, mais pour ce qui est des manuscrits arabes et/ou persans, sur lesquels nous avons mis l'accent, le seul projet que nous avons identifié est celui d'OpenITI dont l'ambition est généraliste et n'est donc pas directement attachée à des projets précis sur des corpus limités.

Cela n'exclut pas, une fois encore, que des équipes et des projets travaillent sur ces questions et fassent usage des outils existants. Nous en donnons pour preuve les quelques échos que nous avons eu de collègues japonais que nous savons engagés dans des projets. Du côté des pays du Maghreb et du Proche et Moyen-Orient, l'impression qui se dégage est que les recherches et les développements en informatique sur ces questions portent en priorité sur des questionnements extérieurs aux documents

historiques, ce qui fait également écho à la situation des sciences humaines et sociales et des humanités en général dans un certain nombre de pays de la région.

En guise d'ouverture : entretien avec Chahan Vidal-Gorène

Chahan Vidal-Gorène est le fondateur de Calfa (voir *supra*), qui a remporté en 2019 le prix Télécoms Innovations. Dans le cadre de son activité avec Calfa, il mène de nombreux projets, par exemple avec la congrégation des pères mekhitaristes de Venise ou avec la BULAC. Il prépare en parallèle à l'École nationale des Chartes-PSL une thèse sous la direction de Marc Smith et Aram Mardirossian en paléographie intitulée *Questions de paléographie arménienne : l'évolution de l'écriture à travers l'étude des fragments*.

Quelles sont selon vous les prochains enjeux techniques à dépasser en matière d'HTR des langues non latines ? Quels seront les prochains terrains de recherche en matière d'HTR ?

Au-delà du perfectionnement progressif des architectures OCR/HTR, rappelons au préalable trois problématiques qu'il convient d'interroger :

- le résultat : la reconnaissance du texte doit être distinguée de la reconnaissance du caractère proprement dit ;
- les données : la pertinence du paradigme de la nécessité de disposer d'une quantité de données importantes peut être interrogée ;
- le besoin : la nécessité de définir les besoins dans un *process* OCR/HTR est centrale.

Au niveau des OCR, la reconnaissance des caractères imprimés est classiquement considérée comme un problème réglé. On obtient des CER* inférieurs à 3 % même avec des systèmes qui n'utilisent pas l'intelligence artificielle (IA). L'HTR, lui, est un nouveau champ de recherche qui a décollé grâce à l'utilisation de l'apprentissage profond* *(deep learning)*. Sur ce plan, on commence à avoir des résultats très bons avec les écritures des manuscrits anciens. Quelle que soit la graphie, et sous réserve d'avoir des données en quantité suffisante, on peut obtenir des CER inférieurs à 8 %, voire 5 % pour certaines d'entre elles, notamment les écritures latines. En vérité, il n'est pas difficile d'obtenir un bon taux de reconnaissance au niveau des caractères car il s'agit d'un problème classique de vision par ordinateur, consistant à reconnaître une forme simple dans une image. Les copistes des manuscrits étaient suffisamment soigneux pour que la variabilité des formes ne soit pas aussi importante que nous pourrions l'imaginer, en tout cas bien moins grande que dans d'autres projets de vision par ordinateur. Avec des algorithmes de *machine learning* et des données, on obtient donc des résultats. Cependant une reconnaissance du caractère qui fonctionne bien ne signifie pas pour autant que la reconnaissance du texte, elle, sera bonne.

Le chaîne de traitement classique pour les OCR et HTR est d'abord d'analyser la mise en page, d'identifier les zones et les lignes de textes ; puis, une fois cette identification réalisée, on identifie dans quel ordre cela va se lire. Ensuite, la ligne de texte est extraite. Ce n'est que dans un second temps que la reconnaissance de caractères de la ligne est réalisée. Or, quand on tient compte de toute cette chaîne de traitement, les occasions sont nombreuses pour que les résultats s'écroulent. Il reste en effet de nombreux enjeux car les deux premières étapes sont extrêmement compliquées, variables et spécifiques à un ensemble de documents, en particulier l'analyse de la mise en page et le sens de lecture de cette mise en page. Ainsi pourrais-je obtenir un CER* de 0 %, mais si les lignes ne sont pas lues dans le bon ordre, est-ce pour autant utile ? Ou même, avoir un CER de 0 % pour le texte reconnu, mais avoir manqué 3 zones de texte sur 4 ? Cette distinction reconnaissance de texte / reconnaissance de caractères apparaît fondamentale.

C'est pourquoi, il faut noter que, parmi les nombreuses architectures OCR/HTR qui sortent aujourd'hui, la plupart se concentre uniquement sur la reconnaissance de caractères et ne font pas du tout d'analyse de la mise en page. En d'autres termes, elles considèrent uniquement une image de ligne en entrée et une ligne de texte en sortie. La tâche d'identification de cette ligne est laissée à l'utilisateur et à d'autres outils, comme des plateformes HTR dédiées.

Le paradigme classique en vision par ordinateur est de considérer que plus il y a de données, mieux le système fonctionnera. La logique est donc celle de systèmes pensés pour être alimentés par d'énormes bases de données pour progresser. Dans le cadre d'un projet de transcription, cette logique peut aboutir à une fuite en avant dans la création de données pour essayer de palier des taux de reconnaissance inférieurs à ceux attendus. Or créer des données est un processus long et coûteux ; certaines graphies, certaines collections ou certains manuscrits exigent une expertise de la langue ou au moins une solide expérience en lecture de manuscrits. Rien ne permet d'être sûr que cela soit ainsi possible pour toutes les langues et rien ne permet par ailleurs de dire que l'on puisse le faire à une échelle suffisamment grande pour obtenir des modèles généralistes.

Ce paradigme a donc des limites pour les langues notamment non latines, surtout qu'il n'est pas une garantie de résultats. La question qui se pose alors est celle des stratégies qui peuvent être mises en place pour simuler une quantité de données ou compenser au contraire la petite taille de la base de données : miser sur la qualité des bases de données plutôt que sur leur taille, réfléchir à des architectures spécifiques à ces langues, ne pas dépendre du caractère ou d'une seule approche d'annotation, etc.

La reconnaissance des caractères est effectivement faisable et l'on peut parvenir à des résultats exploitables à condition de solutionner la question de l'analyse de la mise en page en amont. Pour autant, la reconnaissance de caractères ou de textes ne veut rien dire si les besoins qui ont conditionné ce travail n'ont pas été définis au préalable. La reconnaissance de caractères n'est jamais qu'une étape dans un processus de recherche. Il importe aussi de pouvoir établir pour quels besoins un certain type de modèles a été créé. Si l'on fait uniquement de la pure reconnaissance de caractères à partir d'images, cela ne sert à rien car ce n'est potentiellement pas exploitable. En effet, pour des manuscrits anciens par exemple, quid de la césure du texte,

quid de l'espacement inter-mots, quid des coquilles du copiste, etc. ? Il faut donc définir des choix éditoriaux et un cahier des charges pour chaque projet, ce qui limite donc, une fois encore, le développement de modèles généralistes.

Dans l'article « Handling Heavily Abbreviated Manuscripts: HTR Engines vs Text Normalisation Approaches[45] », Jean-Baptiste Camps et moi-même nous sommes posés cette question pour un manuscrit latin comportant de très nombreuses abréviations. Certes, il était possible d'en faire la reconnaissance de caractères pour obtenir un texte qui était au caractère près ce qu'il y avait dans l'image. Chaque signe abréviatif était donc transcrit par un caractère spécifique, ce qui aboutissait à une édition diplomatique mais limitait considérablement la recherche plein texte, impliquant un travail postérieur de nettoyage à réaliser. Nous avons donc également construit un modèle détaillant les abréviations immédiatement : à un caractère de l'image pouvait correspondre 5 ou 6 caractères dans le texte. Ce modèle, très efficace, a répondu à un besoin concret qui était celui de pouvoir lire le texte immédiatement. En revanche, le modèle créé est difficilement généralisable : il est intrinsèquement très spécifique à ce manuscrit là et ne peut pas forcément s'intégrer dans d'autres projets car il a été particulièrement entraîné à lire les abréviations de cet ensemble de mains. Cette question de la spécialisation des modèles est totalement similaire pour les questions de mise en page. Entraîner un modèle à reconnaître des actes notariés ou des catalogues de manuscrits donnera des résultats très bons, mais ils ne dépasseront pas facilement le cadre du projet concerné, ce qui n'est pas mal en soi.

Ces trois enjeux correspondent à trois terrains de recherche dans le domaine de l'HTR que l'on retrouve particulièrement dans le cas des langues peu dotées. Ce sont celles qui sont le plus confrontées à ces enjeux-là. Il n'y a pas de difficulté majeure à traiter des projets très spécifiques car le modèle d'analyse de la mise en page et celui de la reconnaissance de caractères seront spécialisés sur le projet. En revanche, il ne s'agira pas d'un modèle généraliste et il aura fallu beaucoup de données pour l'entraîner.
Il faut donc distinguer l'HTR intégré dans des projets de recherche spécifiques, qui va bien marcher, et la question de la reconnaissance de caractères manuscrits au sens large pour les graphies non-latines.

Par ailleurs, j'ajouterai un autre enjeu de recherche lié à la définition même du *deep learning**: on apprend par l'exemple. On fournit des exemples différents à la machine pour qu'elle en construise un concept. Avec cette méthode, il est toujours possible de trouver un exemple qui ne corresponde pas du tout à ce qu'on avait dans l'ensemble d'entraînement et qui ne sera donc pas bien reconnu par le *process*. Et c'est là peut-être l'une des « limites » de l'IA car dans cette définition consistant à apprendre par l'exemple, il faut nécessairement une masse de données importante pour que l'on atteigne quelque chose de pertinent : il faut couvrir un ensemble de cas de figures suffisamment important pour que le modèle soit efficace et généraliste. Et au bout d'un moment, à force de données, l'IA aura-t-elle appris

[45] Dans E. H Barney Smith., U. Pal (eds.), *Document Analysis and Recognition – ICDAR 2021. Workshops. ICDAR 2021, Lecture Notes in Computer Science*, vol. 12917, Cham, Springer, 2021, p. 306-316. https://link.springer.com/chapter/10.1007/978-3-030-86159-9_21

quelque chose ou bien lui aura-t-on fourni tellement d'exemples qu'elle ne fera que reproduire ce qu'elle a vu en apprentissage ?

On peut dans ce sens se demander si d'autres approches pourraient être développées. Pour l'HTR, du fait du grand nombre de variations, l'approche par l'exemple semble pertinente dans des projets spécifiques. Mais, si l'on reprend l'exemple d'ABBYY, leader sur le marché pour l'imprimé, ou de Tesseract, ces logiciels n'utilisent initialement pas de *deep learning** pour l'analyse de la mise en page, et pourtant cela fonctionne très bien. Des systèmes entraînés à reconnaître la mise en page de documents imprimés, comme ceux de Calfa, sont peut-être plus polyvalents mais demandent aussi plus d'investissement par rapport à des solutions reposant sur du traitement de l'image classique. On peut donc légitimement se demander si le *deep learning* est indispensable pour toutes les tâches ?

Dans les cas où on observe que l'apport de données ne suffit pas pour obtenir des modèles efficaces, d'autres stratégies, plus fines, peuvent être considérées. Par exemple, l'utilisation de modèles de langues pourrait permettre de compenser la quantité de données.

Pour l'heure, la mise à disposition de modèles HTR généralisés pour l'arabe ne semble pas imminente. Pensez-vous que la tâche soit possible ? Est-elle souhaitable ?

L'HTR généraliste est peut-être atteignable pour les manuscrits anciens en raison de la régularité dans les écritures. L'HTR généralisé en soi me semble être un objectif très compliqué à atteindre en l'état actuel des approches, car, nous l'avons vu, le cœur du problème consiste à pouvoir exploiter des données qui permettraient d'avoir une polyvalence extrême. Malheureusement, on ne peut pas uniquement miser, humainement parlant, sur un *dataset* totalement polyvalent. Nous pourrions disposer d'un *dataset* constituant une base solide de convergence pour pouvoir dans un deuxième temps le rendre polyvalent et l'ouvrir à de nouveaux besoins. Encore une fois, c'est ce dernier point qui revient au centre de la réflexion. En général les besoins sont spécialisés, par exemple la lecture des abréviations, et ce qui est donc développé n'est pas nécessairement interopérable*. C'est une approche que nous favorisons pour l'arabe maghrébin.

Mais finalement, a-t-on besoin aujourd'hui d'un modèle généralisé ? C'est une question qu'il faut se poser. Est-ce même souhaitable ? Le développement de modèles généralisés pour les manuscrits anciens nécessite des budgets très conséquents – concrètement des ressources humaines pour annoter des documents – pour un public qui n'est pas nécessairement là.

Pour l'arabe manuscrit, on entend souvent que ce qui manque serait notamment les données de terrain, autrement dit les *datasets* disponibles. Quelle est votre opinion à ce sujet ?

Il est vrai qu'il n'existe pas beaucoup de *datasets* pour l'arabe. Il y a des choses très spécifiques, pour l'analyse de la mise en page, un peu pour la transcription, mais extrêmement limitées à un certain type de documents : plutôt de l'imprimé lorsque l'on a du texte et plutôt du manuscrit lorsque l'on a de la mise en page. Certains ont été présentés aux dernières conférences ICDAR et ICFHR, comme le *dataset* RASAM en septembre 2021 à

Lausanne (ICDAR 2021). On fait donc face à un manque de données, en tout cas en *open access*. Cela ne signifie pour autant pas qu'il n'existe pas des *datasets* développés par des équipes, des entreprises privées, des laboratoires de recherche qui leur permettent de faire cela.

Il apparaît néanmoins que la reconnaissance de l'arabe manuscrit est un véritable enjeu, ce qui signifie qu'il n'y a pas une entreprise ou un laboratoire de recherche qui a trouvé la solution. Donc, même s'il y a des données qui existent, la problématique de l'HTR de l'arabe reste largement ouverte.

Est-ce que le manque de données suffit à expliquer que l'HTR de l'arabe est un champ de recherche encore en tâtonnement ? On peut se poser la question. D'un côté, dans le cadre du *hackathon* sur les écritures maghrébines (voir III), l'apport de données qui étaient bien faites et bien ciblées a permis d'obtenir des résultats rapides. Ce sont des résultats qui sont assez inédits dans le monde de l'HTR de l'arabe. Cela signifie que la question de la quantité de données n'est pas un verrou suffisant pour expliquer l'absence de modèles HTR pour l'arabe. On touche ici à une des limites de la polyvalence des architectures qui existent, qu'elles soient *open* ou pas. Ces dernières utilisent ce qui existe pour les alphabets latins puis elles l'appliquent à l'arabe. Si cela avait bien marché, on aurait déjà des architectures disponibles et des modèles efficaces pour l'arabe manuscrit.

À moyen terme, la question de la reconnaissance de l'écriture arabe des manuscrits anciens est quelque chose qui va sûrement se solutionner rapidement. À noter tout de même qu'aux dernières ICDAR, les recherches en cours sur l'HTR de l'arabe incluent autant de solutions avec que sans intelligence artificielle.

L'état des lieux non exhaustif rend compte d'un réel dynamisme pour ces questions d'OCR et d'HTR pour les documents historiques. On observe notamment le développement de plusieurs plateformes (eScriptorium, OCR4all ou Transkribus) ou d'ailleurs la plateforme Calfa Vision à l'origine de laquelle vous êtes. Quel est selon vous l'avenir de ces différentes plateformes ?

Ces plateformes répondent à un besoin précis dont on parlait précédemment. S'il n'est pas nécessaire de faire des modèles généralistes, créer des modèles spécialisés pour traiter une collection, un manuscrit privé, etc., est une approche totalement pertinente et un véritable besoin existe. Ces plateformes permettent précisément cela : partir d'un modèle un peu polyvalent que l'on spécialise à force d'annotations sur quelque chose de nouveau pour obtenir des résultats convaincants dessus. Toutes ces plateformes répondent à ce besoin et elles le font très bien, chacune à leur manière.

L'avenir de ces plateformes dépendra de leur modularité avec l'avancement de la recherche. Si l'on prend eScriptorium, Transkribus ou OCR4all, ces trois plateformes sont identiques sur les annotations produites, c'est-à-dire que l'on a un niveau de région de textes, puis de lignes de textes, extraites avec des polygones, etc. Donc on retrouve à chaque fois cette hiérarchie : région – ligne – texte. Cela correspond aux architectures de l'état de l'art. Cette structure de document est néanmoins potentiellement difficile à mettre en œuvre car, pour chaque page, il faut travailler sur trois

niveaux d'information et faire en sorte que ces trois niveaux d'information soient correctement réalisés, et soient raccords avec les autres modèles existants. Cela implique beaucoup de temps et exige un travail triple qui, pour une langue peu dotée, limite d'autant plus le nombre de personnes pouvant y consacrer du temps. Ces plateformes sont dépendantes de ce choix-là, qui est défendable aujourd'hui mais qui ne le sera peut-être pas demain. Cela signifie qu'à l'avenir, un gros investissement sera nécessaire pour les mettre à jour, ou alors elles resteront sur ce format-là de création de données dont l'efficacité sera peut-être plus limitée à l'avenir, ou peut-être pas. Ainsi, si l'enjeu est de créer des systèmes généralistes ou des systèmes qui permettent de limiter l'investissement humain pour avoir beaucoup de données, alors ces plateformes ne répondront plus forcément à ce besoin.

Une remarque finale ?

Il importe d'insister sur la différence entre les possibilités offertes par l'intelligence artificielle pour la reconnaissance proprement dite des caractères et ses conséquences et ce que cela veut dire au niveau de la reconnaissance du texte plus généralement. Il faut s'interroger sur ce que cela veut dire un « texte ». Le texte répond à un besoin et il faut identifier ces besoins. Dans ce contexte, partir dans un projet de reconnaissance de caractères et dire que l'on va créer des modèles de reconnaissance des écritures qui soient généralistes, cela reste très évasif au fond. Il faut que cela réponde aux besoins d'un projet pour que cela soit efficace ; comme aux besoins d'une édition par exemple. Il faut avant tout se demander : qu'est-ce que je veux (résultat), qu'est-ce que j'ai (données) et qu'est-ce je voudrais (besoin) ? C'est la raison pour laquelle le *hackathon* a été une réussite sur ce point : cohérence du *dataset* et objectif de transcription très clair. Cela doit être au cœur de la réflexion, de même qu'il ne faut jamais oublier que l'OCR et l'HTR ne sont que des briques dans un *process* bien plus grand.

[III] EXPÉRIMENTATION : RETOUR SUR UN PROJET DE *HACKATHON* AUTOUR DE LA GRAPHIE ARABE MAGHRÉBINE

Dans cette partie du rapport, nous présentons le projet réalisé au cours de l'année 2021 autour du développement d'un modèle HTR pour les manuscrits arabes en graphie maghrébine. Ce projet est le résultat d'un partenariat entre le GIS MOMM, la BULAC et Calfa et il s'inscrit directement dans le cadre des missions formulées par le GIS MOMM pour la philologie numérique des manuscrits arabes.

L'idée a été d'associer le développement de la reconnaissance automatique de l'écriture manuscrite arabe à un projet collaboratif de formation. Le projet a consisté, sous la forme d'un *hackathon,* à entraîner un moteur de reconnaissance automatique de l'écriture en vue d'élaborer un modèle spécifique pour la graphie arabe dite maghrébine.

Cette technologie pourra bénéficier à toute la communauté des chercheurs et à des étudiants travaillant sur des manuscrits maghrébins dans le cadre de projet d'édition ou tout autre projet d'analyse et de fouille de textes.

091 [III] A CONTEXTE ET OBJECTIFS

Ce projet a vu le jour dans le cadre d'une résidence de chercheur à la BULAC. La BULAC a en effet hébergé l'auteure de ce rapport, dont les activités étaient en lien avec ses collections. Par ailleurs, le partenariat établi entre la BULAC et le CNRS comprenait l'organisation d'activités conjointes s'inscrivant dans le cadre des missions de ce post-doctorat et des activités de la BULAC.

En 2018, la BULAC avait établi un partenariat, avec Calfa autour de la valorisation de ses collections arméniennes, déjà via le perfectionnement de modèles HTR, facilitant ainsi notre rencontre avec l'équipe de Calfa. Ce projet de *hackathon* répondait à plusieurs objectifs distincts qui ont pu entrer en synergie :

[1] Le développement d'un modèle HTR pour l'arabe manuscrit avec un CER* similaire aux autres HTR spécialisés pour les graphies latines et la création d'un *dataset* ouvert pour les écritures maghrébines qui soit représentatif et fonctionnel.

[2] La valorisation de la collection des manuscrits maghrébins de la BULAC.

[3] La mise en place de modalités de formation aux humanités numériques adaptées et pertinentes.

1 Le développement d'un modèle HTR pour l'arabe manuscrit avec un CER similaire aux autres HTR spécialisés et la création d'un *dataset* ouvert pour les écritures maghrébines qui soit représentatif et fonctionnel

L'exploration des possibilités concrètes d'analyse et de fouille de textes ouvertes par l'OCéRisation constitue l'un des objectifs de l'axe prioritaire portant sur la philologie numérique des textes en alphabet arabe. Pour pouvoir explorer ces possibilités, il importe d'avoir, dans un premier temps, déverrouillé, plus précisément, la question de l'OCéRisation, en particulier pour l'arabe cursif, autrement dit de développer des modèles HTR. Ce n'est qu'une fois cette modalité d'extraction des textes opérationnelle et efficiente que les possibilités d'analyse et de fouille de textes pourront être explorées.

Comme nous l'avons montré dans les parties précédentes, il n'existe pas encore, pour la graphie arabe manuscrite, de modèle de reconnaissance automatique des caractères « généralistes », même si des projets témoignent de cette ambition. En matière de modèles spécialisés, même si les efforts sont en cours, le champ doit encore faire l'objet de développements importants (voir II).

C'est la raison pour laquelle, l'accent a été mis, au cours de ce projet, sur le développement d'un modèle HTR pour les manuscrits arabes à partir de la question suivante : peut-on entraîner rapidement un modèle HTR pour l'arabe manuscrit qui atteigne des taux d'erreur de caractères similaires aux modèles d'ores et déjà entraînés pour d'autres langues non-latines ? Cet objectif technique et expérimental partait par ailleurs du principe qu'il était préférable de tirer parti de technologies déjà existantes pour tester leur adaptabilité plutôt que de tout construire à partir de rien. Cela a été possible grâce à la collaboration étroite avec Chahan Vidal-Gorène et toute l'équipe de Calfa. L'idée était donc d'évaluer l'utilisation de la plateforme Calfa Vision pour le traitement d'une langue d'écriture non-latine, qui était nouvelle pour cette plateforme, en privilégiant cependant une interface qui avait été développée pour les langues orientales dès l'origine.

En résumé, le premier objectif prenait la forme d'une preuve de concept, puisqu'il s'agissait, à partir d'une réalisation expérimentale concrète et préliminaire illustrant une méthode de développement, d'en démontrer la faisabilité. Le deuxième objectif dépendait de la réussite du premier puisque de la preuve de la faisabilité découlent les possibilités scientifiques permises par la mise à disposition d'un modèle HTR pour l'arabe maghrébin.

2 La valorisation de la collection des manuscrits maghrébins de la BULAC

D'emblée, le travail expérimental réalisé dans le cadre de ce *hackathon* avait été circonscrit à la collection des manuscrits arabes maghrébins de la BULAC. En effet, à cette priorité « philologie numérique » était adossée un autre des axes d'intérêt du GIS MOMM, à savoir les études sur le Maghreb. Les priorités formulées dans le livre blanc *Vers la science ouverte ?* incluent d'ailleurs ces études sur le Maghreb à travers la question de la structuration de manière pérenne et visible du signalement, et de l'archivage digital et de la constitution de ressources numériques sur le Maghreb.

Partenaire de ce projet, la BULAC conserve le deuxième fonds de manuscrits arabes en France avec 2 458 unités documentaires identifiées. Le fonds comprend une part importante de manuscrits provenant du Maghreb qui ont été copiés en écriture dite maghrébine. Ces manuscrits issus des fonds patrimoniaux sont l'objet d'une politique de valorisation qui prend notamment la forme d'une numérisation et d'une mise en ligne sur le site de la BINA (Bibliothèque Numérique Aréale), la bibliothèque numérique de la BULAC inaugurée en octobre 2019[46].

Le développement d'un modèle de reconnaissance de l'écriture maghrébine s'inscrit ainsi dans un engagement de la communauté scientifique française pour les études sur le Maghreb et pour une mise en valeur des fonds d'archives et de manuscrits maghrébins conservés dans les collections françaises[47]. L'entraînement de ce modèle HTR pour les manuscrits en graphie maghrébine avait ainsi pour objectif de contribuer à mettre en lumière ces collections tout en offrant des possibilités uniques d'étude de ceux-ci que ce soit dans le cadre d'un projet d'édition ou d'analyse des données et de fouille de texte. Ces considérations ont ainsi déterminé le choix des manuscrits maghrébins, et en particulier d'une graphie spécifique, à savoir la graphie dite maghrébine (*ḫaṭṭ maġribī*).

Quelques remarques sur les « écritures maghrébines »[48]

Les écritures dites « maghrébines » sont également appelées « écritures occidentales » ou « écritures arrondies ». On désigne par ce terme une variété de styles qui présentent des caractéristiques communes. Ces écritures, qui trouvent leurs origines au X^e^ siècle, ont largement été utilisées en Occident islamique – al-Andalous et Afrique du Nord – ainsi qu'en Afrique subsaharienne jusqu'au XX^e^ siècle. L'histoire et les origines de ces graphies ont fait

[46] Bibliothèque numérique inaugurée en janvier 2019 : https://bina.bulac.fr . Les manuscrits numérisés de la BULAC sont par ailleurs mis en ligne conjointement sur la BINA (CMS Omeka) et sur Internet Archive : https://archive.org/details/bulac

[47] Voir le travail mené par le GIS Moyen-Orient et mondes musulmans (http://majlis-remomm.fr), notamment le projet Digi#Magh (https://digimagh.hypotheses.org).

[48] Les paragraphes ci-dessous sont tirés de C. Vidal-Gorène, N. Lucas, C. Salah, A. Decours-Perez, B. Dupin, « RASAM – A Dataset for the Recognition and Analysis of Scripts in Arabic Maghrebi », dans E. H Barney Smith., U. Pal (eds.), *Document Analysis and Recognition – ICDAR 2021. Workshops. ICDAR 2021, Lecture Notes in Computer Science*, vol. 12916, Cham, Springer, 2021, p. 265–281. https://doi.org/10.1007/978-3-030-86198-8_19

l'objet d'un important débat scientifique[49]. O. Houdas est l'un des premiers, à la fin du XIXe siècle, à s'être penché sur cette écriture faisant de la ville de Kairouan, au cœur de l'actuelle Tunisie, le lieu de sa naissance[50]. Cette hypothèse, longtemps admise, a d'abord été battue en brèche par F. Déroche qui, à l'extrême fin du XXe siècle, fit l'hypothèse que l'écriture des papyrus égyptiens serait à l'origine de deux types d'écriture : les écritures dite maghrébines d'une part, les écritures livresques abbassides d'autre part[51]. Cependant, la théorie de F. Déroche a, à son tour, rapidement été abandonnée, au profit de l'origine andalouse de ces écritures. Les travaux les plus récents, en particulier ceux d'U. Bongianino, ont ainsi permis de mettre en avant les différents itinéraires (des livres aux Corans, de l'Andalousie au Maghreb) que suivirent ces écritures entre le Xe et XIIIe siècle[52].

Les caractéristiques arrondies de ces graphies s'expliqueraient par l'instrument utilisé pour écrire, à savoir des calames fabriqués à partir de larges roseaux coupés en deux dans le sens de la longueur avec une extrémité coupée en pointe et non pas en biseau comme en Orient[53]. En conséquence, les écritures maghrébines forment une famille d'écritures rondes se caractérisant par un certain nombre de formes communes, au premier rang desquelles les courbes très accentuées[54].

[49] Pour une vue d'ensemble de ce débat, voir : U. Bongianino, *The Origins and Developments of Maghribī Rounds Scripts*, thèse de doctorat, sous la direction de Jeremy Johns, Université d'Oxford, 2017, I, p. 9-14 ; N. Ben Azzouna, « Les corans de l'Occident musulman médiéval : état des recherches et nouvelles perspectives », *Perspective*, 2, 2017, p. 112-114.

[50] O. Houdas, « Essai sur les écritures maghrébines », *Nouveaux mélanges orientaux*, 1886, p. 85-112.

[51] Fr. Déroche, « Traditions et innovations dans la pratique de l'écriture du Maghreb pendant les IVe/Xe et Ve/VIe siècle », dans : S. Lancel (dir.), *Afrique du Nord antique et médiévale : numismatique, langues, écritures et arts du livre, spécificité des arts figurés*, actes du VIIe colloque international réunis dans le cadre du 121e congrès des Sociétés historiques et scientifiques (Nice, 1996), Paris, S. Lancel, 1999, p. 233-247.

[52] U. Bongianino, *The Origins and Developments of Maghribī Rounds Scripts*, *op. cit.*

[53] N. Ben Azzouna, « Les corans de l'Occident musulman médiéval : état des recherches et nouvelles perspectives », art. cit., p. 112.

[54] Pour une présentation détaillée des spécificités de ces graphies, voir : N. Van Den Boogert, « Some notes on Maghribi Script », *Manuscripts of the Middle East*, 4, 1989, p. 30-43 ; U. Bongianino, *The Origins and Developments of Maghribī Rounds Scripts*, *op. cit.*, p. 14-16 ; repris et synthétisé dans C. Vidal-Gorène, N. Lucas, C. Salah, A. Decours-Perez, B. Dupin, « RASAM – A Dataset for the Recognition and Analysis of Scripts in Arabic Maghrebi », art. cit.

3 La mise en place de modalités de formations aux humanités numériques adaptées et pertinentes

La recherche des formations les plus adaptées pour acquérir des compétences en matière d'humanités numériques a constitué l'un des axes du travail réalisé au cours de cette année 2020/2021, en particulier pour les questions de philologie numérique. Un premier évènement en décembre 2020 intitulé « Étudier et publier les textes arabes avec le numérique » avait proposé une formation de trois jours entièrement en ligne qui visait à sensibiliser la communauté scientifique à la philologie numérique en introduisant aux possibilités offertes par le numérique pour l'étude et la publication des textes arabes. L'ambition était par ailleurs de mener une réflexion autour des enjeux propres aux textes arabes à partir de la présentation de projets en cours. Le caractère dématérialisé de la formation a permis à un nombre important de personnes de participer (104 participants) dans un contexte pandémique ayant des impacts concrets sur la vie scientifique. *A contrario*, cette formation, qui avait été initialement prévue en présentiel, n'a pas pu proposer des ateliers pratiques de prise en main d'outils numériques.

Cette première formation a permis de mobiliser un groupe d'étudiants et de chercheurs, lors d'un *hackathon*. Ce concept, provenant du monde du codage informatique, désigne une façon de rassembler, sur un temps relativement court, une communauté pour travailler sur une question numérique. Dans le cadre de ce projet, nous avons rassemblé un certain nombre de spécialistes pour travailler sur l'annotation de documents manuscrits en écriture arabe maghrébine. Les participants disposaient ainsi des compétences linguistiques et scientifiques (lecture de l'arabe, familiarité avec les manuscrits arabes) et venaient se former au fonctionnement de l'OCR/HTR et en particulier à la plateforme d'annotation développée par Calfa.

L'hypothèse, qui a été vérifiée, consistait à postuler qu'il est plus aisé de se former à un outil et à une chaîne d'opérations spécifiques à partir d'un projet plutôt que théoriquement.

MS.ARA.609, folio 72v | MS.ARA.1977, page 1 | MS.ARA.417, page de titre

Fig. 9
Exemples de folio des trois manuscrits du corpus RASAM

Description du corpus *[Dataset]*

À partir d'un nombre de manuscrits maghrébins issus des collections patrimoniales de manuscrits arabes de la BULAC, nous avons privilégié une approche dite de *crowdsourcing* (ou production participative) pour entraîner un modèle HTR pour une graphie arabe spécifique : le *ḫaṭṭ maġribī*.

La sélection des manuscrits qui composent le *dataset* a été réalisée avec des spécialistes du Maghreb de différentes périodes historiques. À partir des retours sur leurs pratiques philologiques et en particulier sur les types de manuscrits avec lesquels ils étaient le plus souvent amenés à travailler, nous avons pu sélectionner, dans la bibliothèque numérique de la BULAC (BINA), les manuscrits les plus pertinents pour ce *hackathon*.

Nous avons par ailleurs fait le choix de limiter le corpus à un petit nombre de manuscrits dans le but de développer un modèle sur une forme dominante [modèle pré-générique][55].

Trois manuscrits composent donc le *dataset* : deux des manuscrits ont été choisis parmi les 250 manuscrits arabes mis en ligne (le MS.ARA.1977 et le MS.ARA.609) ; le troisième et dernier manuscrit du corpus (le MS.ARA.417) a été numérisé à notre demande et ajouté au *dataset* dans un second temps. (NB : il est désormais disponible sur la bibliothèque numérique BINA)

Les manuscrits devaient répondre à plusieurs critères dont la diversité des thèmes, la représentativité du type de graphie maghrébine ainsi que la pluralité des mises en page. En conséquence, deux des manuscrits appartiennent au genre historique (*ʿilm al-taʾrīḫ*) là où le troisième a trait au droit des successions (*fiqh al-farāʾiḍ*).

[55] Nous explorons actuellement la polyvalence du modèle développé sur d'autres manuscrits à graphie maghrébine.

Modalités pratiques et organisation

En raison des conditions sanitaires d'une part et afin de garantir un travail avec un groupe limité de personnes d'autre part, nous avons décidé de limiter le groupe de participants à une dizaine de personnes, soit une quinzaine avec les organisateurs.

De décembre à début janvier, le groupe s'est aisément constitué dans le prolongement de quelques annonces et avec une communication volontairement limitée. Les critères de sélection des participants étaient : leur compétence de lecture en arabe, une familiarité avec les manuscrits et/ou le champ d'études, un engagement à être présent à toutes les séances.

Du janvier à fin avril, le groupe s'est réuni à six reprises le mardi après-midi en présentiel puis en format hybride. Le *hackathon* s'est tenu à la BULAC dans la salle de formation RJ 24 et dans les salles informatiques RJ21 et RJ22, permettant ainsi de travailler directement sur les ordinateurs mis à disposition. Une séance a par ailleurs eu lieu dans la bibliothèque de l'IRHT qui nous a été, à cette occasion, présentée par Muriel Roiland et Élise Voguet.

La première séance, le 19 janvier 2021, a été pensée comme une séance de présentation du projet et de formation à la plateforme Calfa Vision. Après cette séance introductive, les participants disposaient d'une dizaine de jours pour prendre en main la plateforme en s'entraînant sur une page de manuscrit. Le deuxième temps de ce *hackathon* a consisté à décider collectivement des choix de transcription afin de garantir une transcription uniforme. À cette occasion, de nombreuses discussions ont eu lieu et un certain nombre de points ont été débattus. Pour faciliter ce travail collaboratif de moyenne ampleur, nous avons utilisé un outil de gestion de projet en ligne appelé Trello (Trello.com) qui permettait de suivre en ligne la progression du projet et, pour chacun, constituait un outil d'organisation de son travail. Par ailleurs, des étudiants en stage à Calfa ont aidé les participants, en particulier dans le travail de vérification de la mise en page. Après une première phase de travail dévolue à la vérification des prédictions d'analyse de la mise en page de toutes les pages composant le *dataset*, nous avons entamé le tra-

Tableau nº 7 : Description des manuscrits du corpus RASAM

		MS.ARA.609	MS.ARA.1977	MS.ARA.417
Date		1146/1734	1259/1843	1292/1875 copié sur le manuscrit de la bibliothèque d'Alger n°1061
Description physique		Écrit sur papier (210 × 175 mm), le manuscrit se compose de 202 pages – 100 feuillets rédigés, les pages 1 et 2 sont restées blanches – où l'écriture court sur 25 lignes chacune. L'encre principale de ce manuscrit est le noir ; cependant, le copiste a fait usage, à certaines occasions, d'une encre rouge. Ce manuscrit présente de nombreux chiffres, fractions ou tableaux. Les chiffres suivent la numérotation indo-arabe. Présence de réclames ; ajouts, corrections et gloses peuvent être présents dans les marges latérales, inférieures et supérieures.	249 pages, recueil composé de trois parties. Écrit sur papier (305 × 210 mm) ; l'écriture du manuscrit court sur 31 lignes – à l'exception des trois dernières pages qui en contiennent 27 ainsi que des pages 67-68, 202-204 et 245-246 qui sont laissées blanches L'essentiel du texte est écrit à l'encre noire, avec ça et là des éléments en encres rouges et vertes. Ce manuscrit présente une série d'indications marginales : en plus des réclames en bas de page, des corrections et notes sont présentes tout au long du texte. Les caractéristiques de l'encre et de l'écriture nous amènent à faire l'hypothèse qu'il s'agit là de l'œuvre du copiste.	Il consiste en 48 feuillets de 12 lignes par page avec moins de 10 mots par ligne. L'écriture de ce manuscrit, très soignée, est principalement réalisée en noir bien que le scribe ait pu faire usage de rouge et de bleu de façon sporadique. Dans les marges latérales, une autre main, qui semble identique à celle de la note de catalogage laissée en première page, a ajouté les noms des beys, en arabe, et certaines dates. Cette même main semble être à l'origine des marques de vocalisation laissées sur certains folios.
Description	Auteur	Le juriste malikite ʿAbd al-Raḥmān al-Aḫḍarī (m. 953/1546 ou 983/1575)	La majorité de ce recueil (p. 1-201) est occupé par un traité d'histoire intitulé *al-Ǧumān fī muḫtaṣar aḫbār al-zamān* composé par l'historien andalou Abū ʿAbd Allāh Muḥammad b. ʿAlī b. Muḥammad al-Šuṭaybī (m. 963/1556), disciple du grand juriste malikite Aḥmad Zarrūq (m. 899/1493). Le deuxième texte, long de 38 pages (p. 205-243), traite des coutumes et pratiques se rapportant au prophète Muḥammad ; quant au troisième texte (p. 247-249), il enregistre une série de propos du savant al-Ḥasan b. Masʿūd al-Yūsī (m. 1102/1691), originaire du nord-ouest du Moyen Atlas marocain, au sujet d'une mission confiée par le prophète Muḥammad aux tribus berbères afin de conquérir le Maghreb.	Œuvre de Ḥasan Ḫūǧah, secrétaire du Bey Hassan (1817-1831)
	Œuvre	*Šarḥ al-Durrat al-Bayḍā* ʾ		*Tārīḫ Bāyāt Wahrān*
	Thème	Traité en vers sur l'arithmétique, les successions et les testaments. Le poème et son commentaire concernent la science des successions (*ʿilm al-farāʾiḍ*) et les connaissances arithmétiques qu'elle implique.		Histoire de Beys d'Oran au XIIIe siècle
	Copiste	Abū l-Qāsim b. Muḥammad b. Abū l-Qāsim al-Duraydī	Muḥammad b. Mubārak al-Barāšī	
		Consultable sur le site de la BINA : https://bina.bulac.fr/ARA/MS.ARA.609. Notice dans *Calames* : http://www.calames.abes.fr/pub/#details?id=Calames-20171218101240580l	Consultable sur le site de la BINA : https://bina.bulac.fr/ARA/MS.ARA.1977 ; Notice dans *Calames* : http://www.calames.abes.fr/pub/#details?id=Calames-20194894163481	Consultable sur le site de la BINA : https://bina.bulac.fr/ARA/MS.ARA.417 Notice dans *Calames* http://www.calames.abes.fr/pub/#details?id=Calames-20201231412432681

vail de transcription/relecture des pages.
De manière similaire à tous les projets impliquant une production participative, la participation des personnes investies a été plus ou moins active. Aussi peut-on considérer que la majorité du travail a été effectuée par un groupe de plus ou moins 5 personnes.

Spécificités techniques

Le *dataset* comprend 300 images, 676 régions, 7 540 lignes et 483 725 caractères. Les annotations ont été réalisées automatiquement sur la plateforme Calfa Vision, puis vérifiées manuellement au cours du *hackathon* collaboratif. L'ensemble des éléments constitutifs d'une page de manuscrit a été annoté.

Fig. 10
Capture écran de la plateforme Calfa Vision en février 2021

Pour chaque image, nous avons réalisé :

- une annotation sémantique des régions (en vert dans l'image ci-dessus). L'analyse de la mise en page consiste à identifier toutes les zones de textes (régions). 5 catégories ont été retenues : le texte, les marges, les réclames, les tableaux et les nombres (fractions, chiffres, dates).
- une annotation des lignes de base (en rouge dans l'image ci-dessus). L'annotation par ligne de base (*baseline*) est appropriée pour les écritures arabes en raison de la courbure parfois prononcée de celles-ci, et garantit l'interopérabilité* avec les autres *datasets* (par exemple BADAM). Chaque segment de phrase a sa propre *baseline*.
- un polygone encadrant chaque ligne associée à une ligne de base (en bleu dans l'image ci-dessus). Chaque ligne de texte est ensuite extraite avec un polygone encadrant. Les polygones ont été relus manuellement pour intégrer tous les traits constitutifs d'un caractère donné, y compris les ascendants et les descendants, ainsi que les diacritiques associées. Des chevauchements entre certains polygones des lignes demeurent mais les résultats ont démontré que ceux-ci avaient peu d'impact sur les prédictions. Par contre, la non-prise en compte des diacritiques dans les polygones semble participer à l'introduction d'erreurs supplémentaires dans la reconnaissance des écritures.
- la transcription. Afin de garantir une transcription uniforme par les transcripteurs, un cahier des charges précis a été décidé

collectivement. Nos transcriptions restituent les espaces en arabe, même lorsqu'il n'y a visuellement aucun espace discernable dans le manuscrit. Compte tenu de la grande variété de morphologies de caractères dans les manuscrits maghrébins, nous avons privilégié l'approche par mots plutôt que l'approche par caractères où la séparation des mots est gérée en post-traitement. Cette approche participe à expliquer la qualité de nos résultats (voir *infra*).

Nous avons choisi de privilégier une transcription au plus près du texte. La transcription devait notamment respecter les usages orthographiques du copiste quand bien même elles s'écartaient significativement des usages de l'arabe standard. En conséquence, la fréquente confusion entre le *ḍād* et le *ẓāʿ* et le *ṣād* et le *ṭāʿ*, en particulier dans le MS.ARA.609 a été respectée. Par ailleurs, les démonstratifs *hāḍā*, *hāḍihi* et à certaines occasions *ḍālika* qui s'écrivent en arabe moderne sous une forme défective ou à l'aide d'un *alif* suscrit sont souvent réalisés sous leurs formes archaïques avec un *alif* médian : nous conservons dans nos transcriptions alors la réalisation de cet *alif*. [Tous les choix de transcriptions sont détaillés dans l'article].

En plus de permettre le travail collaboratif en ligne sur une même image en temps réel, la plateforme Calfa comprend des modèles pour la prédiction de la mise en page et de l'HTR. Ces modèles sont automatiquement évalués et affinés, en fonction des corrections apportées par le contributeur au projet, afin d'accélérer la tâche de vérification des images suivantes. Trois tâches principales ont été définies pour annoter chaque image. Les participants travaillaient par paire à :

[1] Annoter et vérifier l'analyse de la mise en page : Il s'agissait de vérifier les prédictions de détection des zones de texte et des lignes de textes ; et de les corriger.

[2] Transcrire du texte. Une fois la mise en page vérifiée, la page a été transcrite manuellement. Une fois les modèles HTR suffisamment précis, certaines pré-transcriptions ont été réalisées puis manuellement corrigées. (Pour l'évaluation du gain de temps, voir *infra*).

[3] Vérifier l'extraction des lignes avec un polygone englobant pour chaque ligne transcrite.

Lorsque les trois tâches sont terminées, une vérification complète est effectuée par une équipe d'administrateurs.

Enjeux et difficultés

Dans le tableau ci-après [TABLEAU N°8], nous avons reproduit les principales caractéristiques des écritures maghrébines : les caractéristiques sont tirées du travail d'U. Bongianino ; les réalisations théoriques sont tirées de l'article de N. Van de Boogert sur lequel s'appuie U. Bongianino[56].

[56] Le tableau est tiré de l'article « RASAM – A Dataset for the Recognition and Analysis of Scripts in Arabic Maghrebi » et a été réalisé par Clément Salah (Université de Lausanne/Sorbonne Université).

Les enjeux et difficultés de ce projet croisent en grande partie les difficultés listées précédemment (voir I). On les retrouve par exemple dans ce folio du MS.ARA.609 :

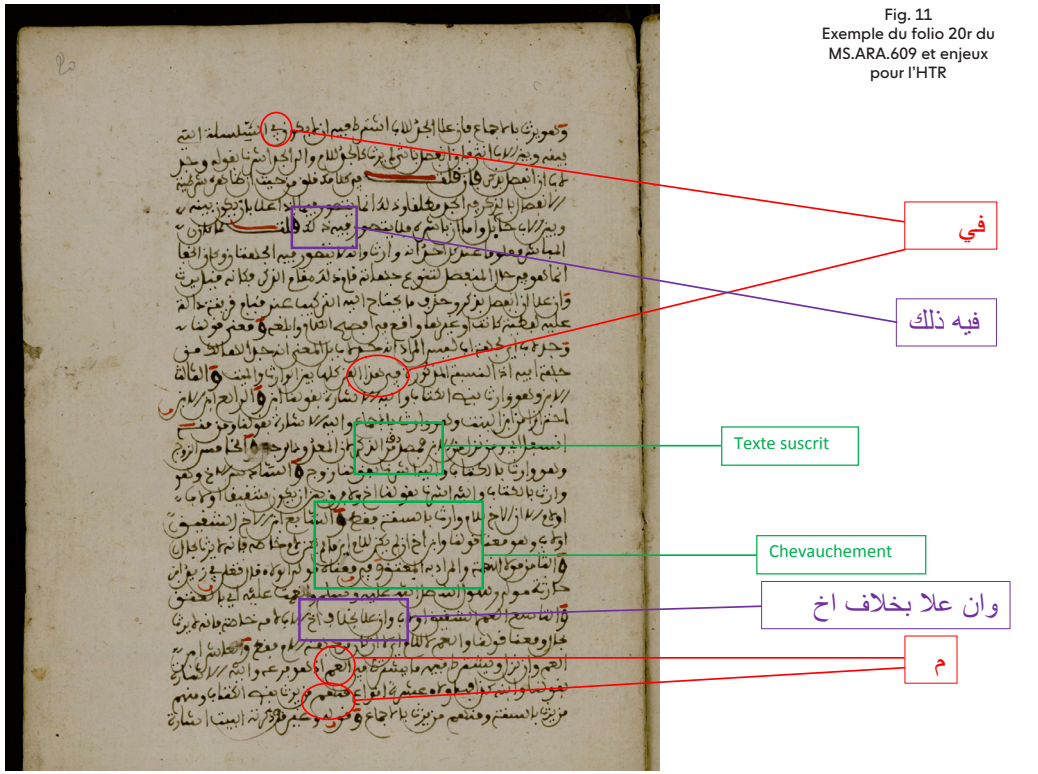


Fig. 11
Exemple du folio 20r du MS.ARA.609 et enjeux pour l'HTR

- variation dans la forme des lettres au sein même d'un seul type d'écriture, d'une même main et d'une même page : dans cette page, le م par exemple même en position identique dans le mot [rappel : en arabe, la réalisation des lettres diffère en fonction de leur position dans le mot (initiale, médiane, finale ou isolée)] ;
- ligatures et glyphes spéciaux (auxquels se couple parfois la question de la variation) comme dans le cas de في dans cette page ;
- lettres suscrites et souscrites ;
- enjeu de la césure entre les mots. Cette question est d'importance dans le cas de la graphie manuscrite arabe. La difficulté de la séparation entre les mots a pu être traitée dans ce travail grâce à l'approche non seulement au caractère mais aussi au mot (dans son sens informatique). Aussi, l'apprentissage par la machine des mots transcrits et la constitution d'un lexique, permet une prédiction plus juste de ceux-ci. (*sur l'image ci-dessus : voir les éléments en violet*)

- le chevauchement d'une ligne à l'autre. En raison du caractère arrondi des écritures dites maghrébines, c'est un enjeu spécifique à cette écriture.

On note par ailleurs que ces manuscrits présentent un texte incurvé dans la marge (voir ligne 12 de la fig. 11 par exemple). L'absence d'espace entre les mots et la multiplicité des formes pour un même caractère rendent la lecture difficile et nécessitent souvent la connaissance préalable du mot pour déchiffrer le caractère. En outre, l'utilisation de signes diacritiques, parfois erronés ou décalés, entraîne de fortes ambiguïtés entre les lettres qui peuvent être facilement confondues sans cela.

Difficultés et enjeux au niveau de la mise en page : Dans les manuscrits sélectionnés pour ce projet, les enjeux liés à l'analyse de la mise en page et les difficultés rencontrées étaient liées à la présence de notes marginales copiées horizontalement comme verticalement et la présence de tableau et de chiffres. Ces difficultés se concentraient en particulier sur le manuscrit MS.ARA.609. Dans le cas des deux autres manuscrits, il n'y avait pas d'enjeux particuliers pour l'identification des zones de texte dans la page. Les deux manuscrits se distinguaient néanmoins par, dans le cas du MS.ARA.417, une mise en page aérée (12 lignes par page et peu de mots par ligne) et, à l'inverse, dans le cas du MS.ARA. 1977, 32 lignes par page et une trentaine de mots par ligne, soit une écriture dense.

Tableau nº 8 : Caractéristiques et réalisations des écritures maghrébines

Letters	Characteristics	Theoretical realization	Examples from mss ARA.1977, ARA.609 and ARA.417
bā' *tā'* *ṭā'* *fā'*	(I) Isolated position: concave form – (II) Final position: closing denticle in the shape of an inverted comma	ب ـب	وركب لصاحب مغلوب وركب لصاحب مغلوب
dāl *ḏāl*	Isolated, median and final positions: concave downstroke and final downward spur (*dāl kāfiyya*)	د ـد	يزيد دنانير الديان يزيد دنانير الديان
dāl *ḏāl*	Final position: marked semicircular descender, resembling the letters *rā' and zā'*	ـد	بعد فقد محمد بعد فقد محمد
sīn *šīn* *ṣād* *ḍād* *qāf* *nūn*	Final position: exaggerated semi-circular descenders, often descrbed as 'swooping' or 'plunging', stretching below the following word	ن ـن	من ماله كان بن عبد من ماله كان بن عبد
ṣād *ḍād* *ṭā'* *ẓā'*	Oval or semi-circular body and lack of denticle	ط ـط	سبط الاصطلاح قسنطينة سبط الاصطلاح قسنطينة
'ayn *ġayn*	Initial position: oversized curl	عـ	عايشه عم عادة عايشه عم عادة
kāf	Initial and median positions: semicircle topped by a diagonal stroke	كـ ـكـ	وكتب كذلك كثير وكتب كذلك كثير
mīm	Final and isolated positions: long curved tail in two variants (concave or convex)	ـم م	ايام تقدم اسلام ايام تقدم اسلام
hā' *tā'* *marbūṭa*	Isolated position: drawn in he shape of a '6', sometimes inverted	ه ـه	يذكره هاذه اياه يذكره هاذه اياه

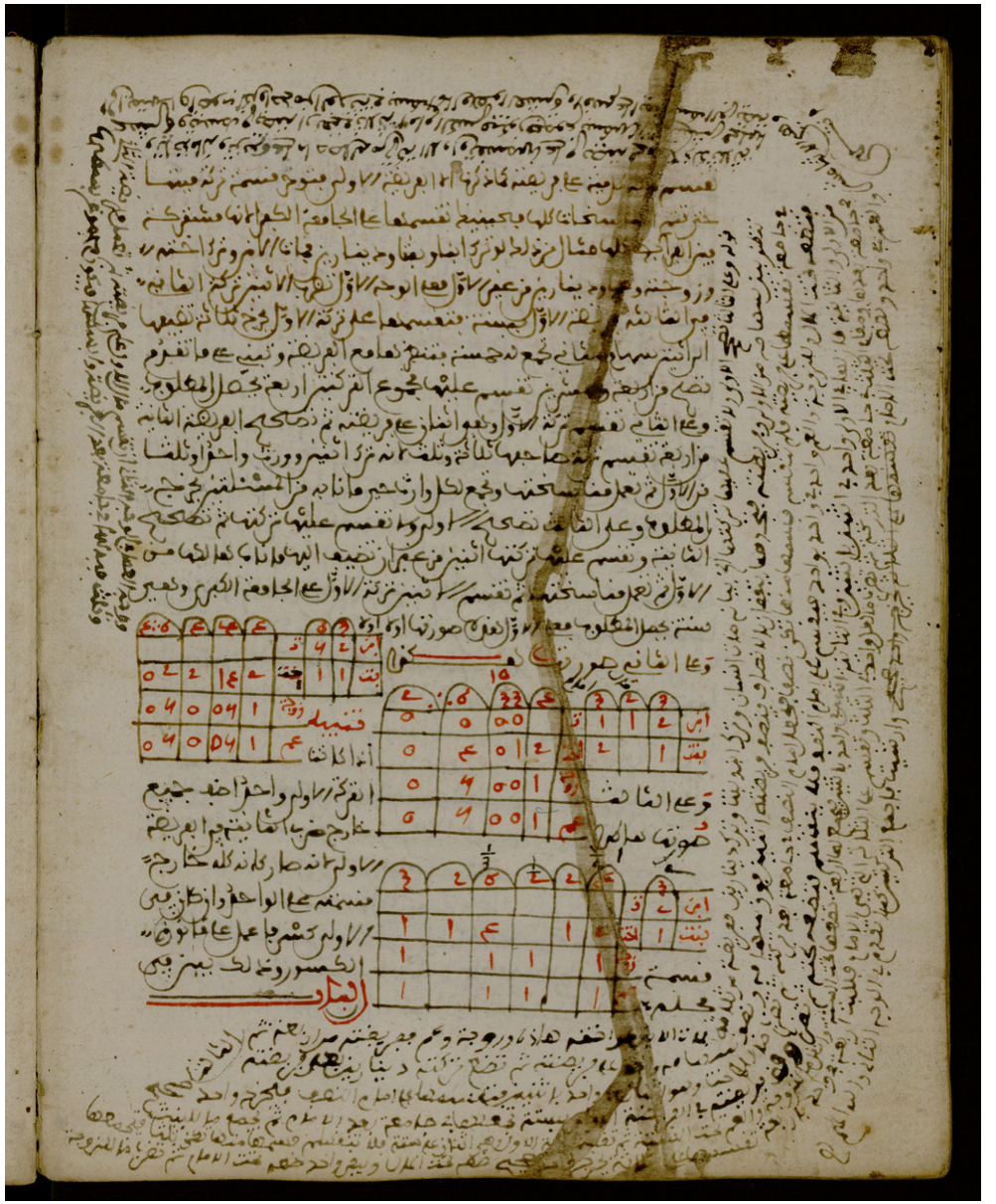


Fig. 12 : Exemple du folio 61v du MS. ARA.609

RÉSULTATS

L'évaluation des résultats porte d'une part sur les modèles d'analyse de la mise en page et des lignes de texte et d'autre part sur l'HTR à proprement parler ; les deux étant nécessairement corrélés pour garantir des résultats satisfaisants (voir I.B).

Cette partie ne fait que résumer les principaux résultats du travail réalisé. Voir l'article pour les détails techniques relatifs aux calculs de ces résultats.

Mise en page

Le modèle d'analyse de la mise en page développé par Calfa a démontré une grande polyvalence et une adaptation rapide aux manuscrits du *dataset* du projet de *hackathon* grâce au *fine-tuning**.

Pour les régions de texte, le modèle atteint une précision de 85,34 % en moyenne (97,80 % pour le texte principal), et de 97,69 % pour les lignes de texte. Les marges et réclames, parfois proches du texte principal, peuvent être difficiles à identifier. Les tableaux et la numérotation, en raison de leur très faible proportion dans le corpus annoté, sont, pour le moment, encore prédits très imparfaitement. Au niveau des polygones, une pertinence de 94 % est mesurée quelle que soit la courbe de la ligne. La principale difficulté rencontrée concerne les points diacritiques qui ne sont pas toujours englobés dans le polygone et nécessitent une correction manuelle.

HTR et CER

Pour des écritures inédites (pour une intelligence artificielle) et avec un nombre limité de données, nous sommes parvenus à un CER* moyen de 4,8 %, autrement dit 4,8 caractères sur 100 sont erronés lors de la prédiction. Ce pourcentage n'a rien à envier aux résultats obtenus pour des HTR spécialisés sur des langues latines puisqu'ils sont en moyenne de 5 %.

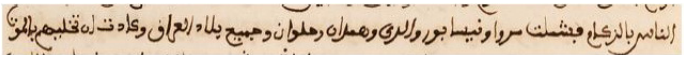
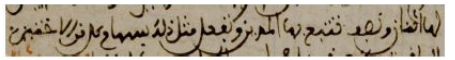
Quelques éléments sur les types d'erreurs de la prédiction

Les principales erreurs de prédiction au niveau du caractère portent sur des lettres situées en position initiale ou en position finale dans le mot. Parmi les erreurs les plus fréquentes, on note que le *nūn* est confondu avec le *rā'* ou le *zāy*, de même que le *dāl* et le *rā'* ou, lorsqu'ils sont en positions initiale ou médiane, le *ḍād* et le *ḥā'*.

Les coupures entre les mots peuvent être mal identifiées, ce qui conduit à une mauvaise prédiction des mots. Nous remarquons par ailleurs que lorsque plusieurs caractères comprenant des diacritiques suscrites ou souscrites se suivent, la prédiction de la séquence est souvent incorrecte et aléatoire. Cela n'est sans doute pas sans lien avec l'irrégularité de la répartition de ces points par le copiste.

Dans le tableau ci-dessous présentant deux lignes extraites de deux manuscrits du *dataset* (ligne supérieure, MS.ARA.1977 et ligne inférieure, MS.ARA.609), nous observons effectivement des erreurs de prédictions pour les caractères en début de mot ou en fin de mot. La confusion entre le *dāl* et le *rā'* est également observable والـدى /والـرى . Il faut noter à cet égard que cette difficulté d'identification entre les deux lettres avait été soulignée par les transcripteurs eux-mêmes.

Tableau n° 9 : Exemples de prédiction pour les MS.ARA.1977 et MS.ARA.609

	
Pred	الناس بالزكام فشملت مروا ونيسابور والدى وهمدان رحلوان وجميع بلاد العراق وكادت ان تخليهم بالوت
GT	الناس بالزكام فشملت مروا ونيسابور والرى وهمدان وحلوان وجميع بلاد العراق وكادت ان تخليهم بالوت
	
Pred	لها اثنان ونصف تتبع بها الدين ويفعل مثل ذلك بهام كل من الاختينة
GT	لها اثنان ونصف تتبع بها الدين ويفعل مثل ذلك بهام كل من الاختين

Plateforme Calfa Vision : gain de temps et évaluation d'autres modèles

L'approche avec Calfa est de travailler à une spécialisation de leurs modèles déjà intégrés dans la plateforme pour d'autres manuscrits et d'autres langues non-latines au premier rang desquelles l'arménien[57]. Certains de leurs modèles comme les modèles d'analyse de la mise en page sont déjà polyvalents [FIG.13].

Sur l'évaluation du gain apporté à l'analyse de la mise en page, Calfa note que l'application du modèle par défaut de Calfa sur les manuscrits arabes connaît une nette amélioration après 50 images corrigées. Après 100 images corrigées, le résultat est encore meilleur avec une bonne distinction entre la marge et le texte principal. Il fallait en moyenne 7 minutes pour corriger l'analyse de la mise en page prédite. Après 50 images, le temps était en moyenne de 3 minutes 50 et après 100 images de moins de 2 minutes. Cette amélioration rapide a été permise par le *fine-tuning** automatique des modèles.

[57] Pour plus de détails, voir C. Vidal-Gorène, B. Dupin, A. Decours-Perez, T. Riccioli, « A Modular and Automated Annotation Platform for Handwritings: Evaluation on Under-Resourced Languages », art. cit.

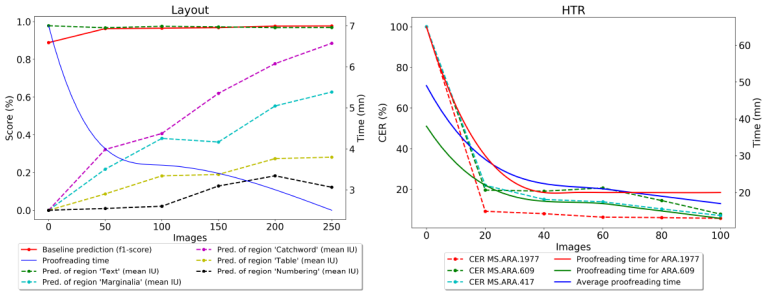


Fig. 13
Évolution du *fine-tuning* et effets sur le temps de relecture pour l'analyse de la mise en page et les transcriptions

Le même type d'évaluation du gain de temps permis par le *fine-tuning* a été réalisé sur la transcription. Pour une page de MS.ARA.1977, il fallait jusqu'à 1h15 pour transcrire une page de zéro. Les évaluations ont montré qu'à partir de 20 pages/images transcrites, le modèle avait déjà suffisamment appris. Même avec un CER* atteignant alors 15 %, le temps de correction était d'environ 30 minutes. À la fin du *hackathon*, moins de 20 minutes étaient nécessaires pour corriger la prédiction automatique de la page.

L'équipe de Calfa a donc observé une convergence des modèles qui vient aider réellement la tâche d'annotation. Ces évaluations ont été réalisées sur une écriture identique. Calfa a par ailleurs évalué ce qui se passait lorsque l'on entraînait un modèle à partir de MS.ARA.1977 et MS.ARA.609 et qu'on l'appliquait à un autre manuscrit, ici le MS.ARA.1977. Sans spécialisation, le CER obtenu était de 5,9 % permettant de conclure que l'apport d'images du manuscrit annoté offrirait un CER encore plus bas.

Interprétation des résultats

Les résultats montrent que l'utilisation d'une plateforme collaborative de transcription automatique comme Calfa Vision – qui intègre des modèles génériques pouvant être spécialisés sur un ensemble de données spécifiques – est une stratégie pertinente et appropriée pour la création de données et pour l'entraînement de systèmes HTR efficaces, en particulier pour les langues non-latines, au premier titre desquelles, dans le contexte ci-présent, l'arabe.

Les différents tests permettent par ailleurs de dire que les données du *hackathon* constituent un socle solide pour élaborer d'autres modèles spécialisés par la suite. Il importe de préciser par ailleurs que le *dataset* mis à disposition du public sur un GIT est un *dataset* complètement réutilisable et exploitable par les autres équipes.

Livrables et diffusion

Mise à disposition de la base de données

Chaque image est associée à un fichier XML généré par la plateforme Calfa Vision et comprend l'intégralité des annotations produites (informations sur la mise en page et la transcription). Le RASAM *dataset* de ce projet est librement disponible sur un GIT : https://github.com/calfa-co/rasam-dataset

Présentation des résultats

- Le 8 juin 2021 dans le cadre des Rendez-vous de la philologie numérique : « Intelligence artificielle et *khaṭṭ maghribī* : Résultats d'un *hackathon* pour la reconnaissance de texte automatique de l'arabe manuscrit ».

- Le 30 juin 2021 au cours d'une table ronde « Accompagner la transition numérique des études aréales : les actions-pilotes et les résidences numériques du consortium DISTAM » organisée dans le cadre du 4e Congrès des études sur le Moyen-Orient et les mondes musulmans.

Les résultats et le *dataset* ont par ailleurs été présentés à la conférence ICDAR 2021, qui s'est tenue à Lausanne, au cours du workshop « Arabic and Derived Script Analysis and Recognition » (ASAR) le 6 septembre 2021 : Chahan Vidal-Gorène, Noëmie Lucas, Clément Salah, Aliénor Decours-Perez, Boris Dupin, « RASAM - A Dataset for the Recognition and Analysis of Scripts in Arabic Maghrebi ». Cette participation a par ailleurs donné lieu à la publication d'un article :

Chahan Vidal-Gorène, Noëmie Lucas, Clément Salah, Aliénor Decours-Perez, Boris Dupin, « RASAM – A Dataset for the Recognition and Analysis of Scripts in Arabic Maghrebi », dans E. H Barney Smith, U. Pal (eds.), *Document Analysis and Recognition – ICDAR 2021. Workshops. ICDAR 2021*, *Lecture Notes in Computer Science*, vol. 12916, Cham, Springer, 2021, p. 265–281. L'article peut être consulté à cette adresse : https://rdcu.be/cxi4a

Communication autour de ce projet et de ses résultats

- Billet sur le Carreau de la BULAC rédigé par Claire Camberlein « Intelligence artificielle et graphie dite maghrébine. Un projet collectif et collaboratif sur l'OCR et les manuscrits maghrébins » : https://bulac.hypotheses.org/32837
- Présentation du partenariat sur le site de la BULAC : https://www.bulac.fr/partenariats-scientifiques-et-appui-aux-projets-de-recherche
- Présentation du projet sur le carnet de recherche Philaranum : https://philaranum.hypotheses.org/219
- Présentation du projet sur le site de Calfa : https://calfa.fr/blog/20
- Présentation des résultats sur le site de Calfa : https://calfa.fr/blog/26

Rétribution des participants

Ce projet a été réalisé dans le cadre d'un partenariat signé par les trois parties : Calfa, BULAC et le CNRS. Ce partenariat établit que les transcripteurs ayant pris part à ce projet pourront utiliser le modèle HTR développé dans le cadre du *hackathon* pour leurs projets de recherche respectifs, dans la limite de 1500 pages par personne et en s'engageant à respecter les conditions générales d'utilisation de la plateforme Calfa Vision.

109 [III] D PERSPECTIVES

Pour la BULAC

La richesse des collections patrimoniales de manuscrits arabes mérite un investissement en matière de valorisation. Pour l'heure, 712 unités documentaires ont été cataloguées, ce qui représente moins de 30 % de l'ensemble. Il serait donc important de trouver les moyens de financer une accélération de cette entreprise de signalement et de catalogage afin, dans un second temps, de poursuivre les campagnes de numérisation de ces manuscrits. Il semble qu'une part non négligeable du fonds de manuscrits arabes soit composée de manuscrits maghrébins copiés dans l'écriture traitée dans le cadre de ce travail expérimental. Un catalogage complet du fonds permettrait de vérifier cette intuition. Pour l'heure, sur les 712 unités documentaires cataloguées, 250 ont été numérisées et sont disponibles sur la BINA et sur Archive.org. Sur ces 250 manuscrits numérisés, 77 sont catalogués en écriture maghrébine, soit environ 30 %.

En attendant de pouvoir compléter le catalogage et la mise en ligne de l'intégralité de la collection, il pourrait être intéressant de mettre l'accent sur des sous-collections spécifiques au sein de ce fonds. Augustin Jomier, maître de conférences à l'INALCO, travaille depuis quelques temps sur la collection du Cheikh El Haddad (1790-1873), chef de l'importante confrérie soufie de la Rahmaniyya ; à partir de 1860, il joua un rôle important dans la mobilisation qui conduisit au soulèvement de la Kabylie en 1871[58].

Par ailleurs, certaines collections ont déjà été traitées en tant que telles, cataloguées et numérisées. C'est le cas de la collection constituée par Paul Geuthner et son épouse Warburga Seidl dans le cadre des activités de

[58] Cette collection en cours d'identification fait actuellement l'objet d'une exposition à la BULAC. Pour en savoir plus, lire ici : https://www.bulac.fr/cheikh-el-haddad-et-linsurrection-de-la-kabylie-des-vestiges-de-papier-1871-2021

la Librairie orientaliste Paul-Geuthner, fondée en 1901. Cette collection a été acquise par la BULAC en novembre 2016 auprès de la galerie Laure Soustiel. Elle comprend 88 manuscrits arabo-berbères dont 38 manuscrits en écriture maghrébine.

Aspects techniques

Le travail d'amélioration du modèle, en particulier de sa polyvalence, peut être poursuivi dans plusieurs directions.

- Tester la polyvalence du modèle sur différents types de mise en page
  Le travail réalisé au cours de ce *hackathon* a montré que les éléments comme les tableaux ou les numérotations étaient plus difficilement prédits par les modèles d'analyse de la mise en page. Des efforts pourraient donc être apportés de ce côté-là en augmentant les données d'entraînement.

  Au cours de nos annotations et transcriptions, nous avons volontairement non discriminé les titres. Un entraînement spécifique du modèle d'analyse de la mise en page pourrait être réalisé en ce sens. Par ailleurs, les manuscrits considérés dans ce *dataset* comprennent de nombreux signes et marques, qui ne sont pas des signes de ponctuation mais plutôt des signes décoratifs ou à d'autres fins de séparation : comme entre deux vers, ou en fin de ligne, ou de section. Ces signes pourraient faire l'objet d'un travail spécifique.
- Tester la polyvalence du modèle sur d'autres types de graphies maghrébines
  L'évaluation de cette polyvalence sur d'autres graphies appartenant à la famille des écritures maghrébines participe de l'évaluation de la possibilité de créer rapidement des modèles dédiés à une main ou à un ensemble de documents.

  Des premiers tests ont été réalisés sur un corpus de 5 manuscrits issus des fonds de la BULAC et appartenant tous à la collection Geuthner. Ces expérimentations montrent que le modèle développé dans le cadre du *hackathon* est polyvalent et que le nombre de pages nécessaires pour spécialiser le modèle est faible. D'autres tests méritent d'être réalisés afin de rendre le modèle pré-générique créé dans le cadre de RASAM le plus polyvalent possible et donc le plus utile possible pour les collègues souhaitant se saisir de cette technologie pour leur exploitation des textes.

  Ces recherches sur la polyvalence du modèle sont actuellement en cours dans le cadre d'une nouvelle édition du *hackathon* organisée par le GIS MOMM, la BULAC et Calfa et animé par Antoine Perrier (CNRS, IREMAM) et Chahan Vidal-Gorène.
- S'intéresser à la diversité des thèmes pour enrichir le lexique de mots arabes du modèle
  L'une des hypothèses que nous formulons pour expliquer les bons résultats obtenus est l'approche choisie par Calfa pour le modèle HTR, une approche au mot (au sens informatique) plutôt qu'au caractère qui permet notamment un respect de la césure satisfaisante entre les mots. Ce faisant, cela implique que l'IA apprend des mots et se constitue un lexique à partir des transcriptions réalisées.

Comme nous l'avions précisé plus haut, les thèmes des manuscrits choisis pour le *hackathon* sont l'histoire et le droit. Aussi le lexique que s'est constitué la machine est-il particulièrement lié à ces thèmes, et pour être encore plus précis, aux thèmes du droit de l'héritage pour le droit. On peut donc formuler l'hypothèse que le modèle HTR obtiendra de meilleurs résultats pour des prédictions sur des manuscrits portant sur des thèmes similaires.

Aussi pourrait-il être intéressant d'entraîner le modèle sur d'autres thèmes en essayant de voir comment fonctionne une spécialisation que l'on pourrait qualifier d'ultra. À cette fin, un domaine comme la grammaire ou le droit, pour lesquels la BULAC dispose de fonds importants, et déjà signalés, pourrait être intéressant.

CONCLUSION GÉNÉRALE

Ce rapport rend ainsi compte du fait que les défis techniques spécifiques aux manuscrits arabes (partie 1) sont dépassables à terme et que des outils existent d'ores et déjà (partie 2). Même si la majorité des outils et des architectures ont été pensés pour les langues latines, les langues orientales, et notamment l'arabe, ont bénéficié ces cinq dernières années d'un intérêt grandissant (développement de Calfa, eScriptorium, projet OpenITI). Nous avons montré que le nombre de jeux de données ouverts existant pour l'arabe manuscrit dans des documents historiques était limité. Bien que la quantité de données ne soit pas un critère aussi déterminant qu'il puisse paraître du point de vue technique, comme le montre notamment l'exemple des travaux menés par l'équipe de Calfa, on ne peut ignorer que les recherches sur l'HTR de l'arabe gagneraient à disposer d'un plus grand nombre de jeux de données d'entraînement ouverts et réutilisables par d'autres. En effet, il apparaît qu'un certain nombre de jeux de données mis en ligne par les équipes sont souvent assez peu exploitables par d'autres, dans la mesure où ils ont été créés pour répondre aux besoins d'un projet spécifique sans penser à leur réemploi.

En d'autres termes, plus de «recherche fondamentale» sur l'HTR de l'arabe serait bienvenue. Par «recherche fondamentale», nous entendons la mise à disposition de jeux de données qui constituent des vérités terrain pour le développement de modèles HTR adaptés aux manuscrits. Cela suppose de travailler sur des projets susceptibles de servir la communauté scientifique des études arabes dans son ensemble. Autrement dit, les objectifs définis au départ, ces mêmes objectifs qui déterminent des besoins techniques, devront être pensés pour être réutilisables et interopérables*. Une autre manière de considérer cette «recherche fondamentale» consiste à dire que lors de l'élaboration de projets scientifiques précis impliquant l'utilisation d'architecture HTR, il faut penser le schéma de données en concordance avec les objectifs du *FAIR**.

À l'image du rapport, cette conclusion n'insistera pas sur les questions techniques propres à l'intelligence artificielle et sur les verrous informatiques qui constitueront les enjeux de demain. Les temporalités de la recherche et du développement sur ces questions sont bien différentes des temps de la recherche en sciences humaines et sociales. Les outils et les plateformes que nous utilisons aujourd'hui sont le résultat de recherches en cours depuis des décennies qui connaissent des résultats exploitables depuis un peu moins d'une dizaine d'années. Les ingénieurs et les *data scientists* concentrent désormais leur effort sur de nouvelles questions et il y a peu de doute que les barrières techniques en lien avec les enjeux spécifiques des manuscrits arabes seront levées dans un avenir proche; à condition bien sûr que les besoins soient présents. On touche ici à la question déterminante de la demande scientifique pour l'utilisation de ces technologies. C'est là un verrou qui nous semble beaucoup plus déterminant, en particulier dans le contexte de la recherche française à l'échelle internationale, qu'il s'agisse de sa représentation ou de son positionnement.

À la question, la France dispose-t-elle des compétences nécessaires pour mener ce travail, il nous semble que le projet autour du *dataset* RASAM (https://github.com/calfa-co/rasam-dataset) atteste l'existence des

compétences dans tous les domaines clefs qui entourent la recherche sur l'HTR/OCR: à savoir les compétences en développement informatique, les compétences en philologie arabe et les compétences en curation de données. Avec un nombre limité de participants et un temps pour réaliser le projet lui-même ramassé (7 mois), il a été possible de parvenir à des résultats, tant du point de vue technique (le modèle spécialisé développé atteint des résultats équivalents à ceux obtenus pour les graphies latines) qu'en matière de formation (les participants savent désormais prendre en main une interface d'annotation et peuvent se lancer dans leur propre projet) et en matière scientifique et de partage de technologie (RASAM est actuellement utilisé par des collègues européens et internationaux et a été récemment ajouté aux modèles proposés par Transkribus). Mais cela a pu être mené à bien car:

- le projet et l'initiative plus large étaient soutenus et pilotés par le Groupement d'intérêt scientifique Moyen-Orient et mondes musulmans (GIS MOMM) avec le soutien financier du ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. Ce soutien a permis la mise à disposition d'une coordinatrice dédiée au projet.
- le projet a bénéficié d'une collaboration (par le biais d'un partenariat) entre trois partenaires: le GIS MOMM d'une part, la BULAC d'autre part et Calfa. Cette collaboration est à considérer à deux niveaux. Le premier niveau est le partenariat entre organismes publics que sont les 25 institutions tutelles du GIS, représentées par le CNRS, et le GIP BULAC. Les compétences du premier s'allient alors aux compétences du deuxième pour valoriser les collections patrimoniales de ce second partenaire. Le second niveau s'articule entre ces deux entités publiques et une troisième entité privée, une start-up spécialisée en reconnaissance et compréhension des graphies orientales qui collabore activement avec des institutions patrimoniales, en plus de son activité économique.
- les objectifs avaient été clairement énoncés et précisés au départ.

Ce projet circonscrit confirme ainsi que les compétences existent en France. La difficulté porte néanmoins sur deux points:

- Les compétences disponibles restent limitées en nombre, ce qui implique des limites également à leur capacité d'investissement.
- La formulation de besoins clairs, adaptés à l'état de l'art demeure lente et difficile, voire inexistante du côté des chercheurs, faute de sensibilisation ou de formation suffisante.

Parce que la communauté des chercheurs travaillant sur les manuscrits arabes n'a pas embrassé les nouvelles technologies computationnelles, ou du moins l'a fait dans une bien moindre mesure, comparativement aux philologues des manuscrits latins, les projets scientifiques incluant de l'HTR ou de l'OCR restent peu nombreux voire inexistants. Pourtant, il ne fait aucun doute, au regard de l'expérimentation menée au cours du printemps 2021, que les conditions nécessaires sont minimes pour parvenir à des résultats du point de vue technique.

Il existe d'ores et déjà des outils pour l'OCéRisation des textes en alphabet arabe (imprimé et cursif) au plan européen et international. On pense par exemple à OCR4all ou eScriptorium (voir II). Sur ce point, on peut donc conclure qu'il n'y a pas véritablement de verrous, si ce n'est, et ces limites ne sont pas négligeables, la capacité pour les chercheur.e.s de prendre en main ces outils et de pouvoir véritablement en bénéficier, car ces outils, comme nous l'avons rappelé, peuvent être limités en capacité de calcul, et/ou gourmands en main d'œuvre pour l'administration et la maintenance et donc dans le suivi des projets individuels ou de groupes restreints qui nécessitent une expertise élevée et des développements techniques spécifiques.

Partant du principe que les enjeux techniques peuvent être dépassés et le seront d'ici cinq à dix ans, il importe de se demander quelle serait la stratégie la plus pertinente pour permettre à la communauté des chercheurs français travaillant à partir de documents historiques arabes de prendre en main cette méthode d'acquisition des textes pour leur propre recherche et de participer aux développements d'outils qui soient adaptés à leurs besoins précis et aux langues et graphies assimilées qu'ils traitent.

Il nous semble que la remarque de Chahan Vidal-Gorène sur la définition des besoins doit ici être rappelée (voir II). Or ces besoins sont à considérer selon plusieurs échelles :

À l'échelle des projets bien sûr : Nous n'avons cessé de mettre l'accent sur l'importance, dans un processus d'HTR, de la définition des besoins. En d'autres termes, pourquoi choisir d'utiliser cette méthode d'extraction de texte et comment ? Afin de définir aux mieux ces besoins et objectifs, l'interaction avec les ingénieurs et, plus largement, le dialogue entre chercheurs en SHS et *data scientists* sont cruciaux et il est indispensable qu'ils aient lieu en amont d'un projet donné et non pas une fois que le projet a été élaboré et a reçu d'éventuels financements. Un projet HTR réussi doit avoir été pensé au départ en prenant en compte tous les aspects dont : les besoins, les objectifs, les outils, les enjeux techniques, les résultats attendus, le type d'analyse sur textes extraits, la publication des résultats.

À l'échelle du champ plus largement :
Besoins d'information : le travail de recensement entrepris dans ce rapport a permis de prendre la mesure de l'éclatement de l'information, d'une part, et de la difficulté à suivre en temps réel les avancées techniques et les nouveaux outils sur le marché mais aussi de prendre, d'autre part, la mesure des travaux entrepris effectivement par les équipes de chercheurs au niveau international. Il en ressort d'après nous qu'un manque de communication et d'information est criant. Ce manque porte à la fois sur :

- la communication entre les équipes de recherche qui mènent parfois des projets parallèles et gagneraient à échanger ensemble, sinon à travailler de concert ;
- l'accès aux actualités concernant les nouveaux outils, les jeux de données ou les projets en cours.

C'est la raison pour laquelle il nous semble essentiel de réfléchir aux possibilités de communication et de veille technologique, offrant dans le même

temps des moyens d'agrégation de contenus. Ce problème d'information n'est pas spécifique à l'HTR, ni au champ des études sur le Moyen-Orient, les mondes musulmans et le Maghreb. Il est intimement lié à la nature de notre société numérique et à la rapidité de la circulation des informations. Le problème de la dispersion des données ne peut pas être résolu facilement. Rappelons que les projets financés aujourd'hui, que ce soit par l'Agence Nationale de la Recherche ou par l'European Research Council par exemple, requièrent de plus en plus, pour être lauréats du financement, une dimension numérique d'une part, et un support de communication numérique, d'autre part, qui prendra bien souvent la forme d'un carnet de recherche sur la plateforme d'OpenEdition Hypothèses. En plus de cela, les porteurs de ces projets doivent montrer leur capacité à se saisir des pratiques de la science ouverte dans le cadre du traitement de leurs données. La dispersion de l'information, dans ce contexte, et la multiplication des supports de communication paraissent des résultats naturels de ces processus. Il serait néanmoins possible de proposer un espace en ligne de moissonnage de tout ce contenu aujourd'hui disponible en ligne, à l'image par exemple de ce que la BULAC propose avec la Croisée de la BULAC (https://veillebulac.hypotheses.org/), à ceci près qu'il serait intéressant que cet espace en ligne ne moissonne pas uniquement les carnets Hypothèses mais tous les sites d'institutions ou de réseaux touchant aux questions d'HTR/OCR. Ce rapport en répertorie un certain nombre sans cependant être exhaustif.

Cette mise à disposition de l'information ne peut être uniquement réalisée par des supports en ligne, elle nécessite la médiation de(s) (in)formateurs et elle doit concerner des publics aussi différents que les étudiants en études arabes (au sens large), les enseignants, les chercheurs, les documentalistes et les ingénieurs de recherche. C'est là que se croisent les besoins d'information et de formation, qui sont d'ailleurs intimement liés.

Besoins de formation : En lien avec la rédaction de ce rapport, nous avons mené un projet d'HTR basé sur la transcription collaborative des trois manuscrits issus des collections de la BULAC (partie 3) et nous avons, par ailleurs, organisé des évènements de formation (« Étudier et publier les textes arabes avec le numérique ») et d'information (« Les rendez-vous de la philologie »). Aussi, ce travail d'état de lieux sur l'OCR/HTR des manuscrits arabes a été réalisé en parallèle avec une réflexion sur la formation à ces nouvelles technologies et aux approches méthodologiques des textes historiques. Il ressort de ces deux travaux menés conjointement que l'un des verrous majeurs à dépasser est celui de la formation.

1 Une offre de formation complète sur l'étude des manuscrits arabes/orientaux à l'heure du numérique en France est manquante et répondrait pourtant à un besoin réel. Il existe des formations en langues arabes nombreuses et réparties dans de nombreuses universités françaises, l'une des plus connues, à l'échelle européenne et internationale, étant l'INALCO. On compte par ailleurs des philologues, paléographes et codicologues pour l'arabe au sein de nombreuses institutions comme le Collège de France, Sorbonne-Université ou l'Institut de Recherche et d'Histoire des Textes (IRHT) du CNRS. Les

compétences sont donc éclatées et il n'existe pas de formation complète sur l'étude des manuscrits arabes, encore moins une formation qui s'inscrirait dans les humanités numériques et mettrait à l'honneur les méthodes computationnelles dans l'étude de ces manuscrits. Cela ne signifie pas bien sûr qu'aucun collègue n'utilise ces méthodes, ni qu'aucune formation ne les mentionne (cf. par exemple le dernier stage d'initiation au manuscrit médiéval organisé par l'IRHT en mars 2021) mais il n'existe pas de formation dédiée. De plus l'arabe est peu, voire pas présent, dans les formations existantes en humanités numériques, à l'image du master « Humanités numériques » de l'École des Chartes au sein duquel les langues latines sont majoritaires.

La formation complète qui pourrait prendre la forme d'un diplôme universitaire serait proposée à celles et ceux qui disposeraient d'un niveau déterminé en arabe garantissant leur capacité à lire, rédiger et comprendre l'arabe et qui travailleraient ou souhaiteraient travailler sur les manuscrits.

2 Des formations ponctuelles aux outils à partir de projets concrets. Nous renouvelons la certitude que les formations par défaut à des outils ou des plateformes ne fonctionnent pas si l'apprenant.e ne prend pas immédiatement en main l'outil sur un projet concret. Aussi, il nous semble que deux approches doivent être considérées :

- une approche similaire au *hackathon autour de l'HTR de la graphie arabe maghrébine*. En d'autres termes, la réunion de collègues autour d'un projet HTR *ad hoc* sur lequel ils travaillent de concert pour obtenir un résultat collectif aux objectifs clairement définis. Le problème de cette approche est la possibilité de fédérer les collègues en poste, à qui s'adresseraient ces formations ponctuelles, autour de projets extérieurs à leur projet individuel ou collectif mené d'autre part ou, et la tâche n'est pas moins difficile, de définir collectivement ces projets HTR pour qu'ils satisfassent des individualités aux besoins et envies très variés. Cette limite des projets collectifs est la raison pour laquelle il nous semble indispensable de considérer la seconde approche.
- un accompagnement individuel et personnel des collègues dans leur projet. Cela fait directement écho aux résidences numériques lancées à l'origine par l'UAR InVisu (https://invisu.cnrs.fr/residences/) et dont une formule adaptée à l'absence de personnels dédiés a été proposée l'an passé par le GIS MOMM pour trois résidences numériques. Ces résidences posent néanmoins la question de la structure d'accompagnement de ces projets, et de leur développement ultérieur après le temps de pré-maturation.

On pourrait en effet penser, comme cela a pu être fait cette année, que pour chaque résidence, la structure d'accompagnement la plus adaptée serait

choisie. Cependant, si on se place dans le contexte de l'HTR/OCR et plus largement de la philologie numérique des textes arabes et de l'étude des manuscrits arabes, voire orientaux, il apparaît plutôt que ce qui manque est une structure capable de proposer, au niveau national, un service d'accompagnement aux projets numériques sur les manuscrits en écriture arabe et de mobiliser des experts spécialisés sur les différentes méthodes et outils appliqués à ce domaine. Cette structure existe dans d'autres pays dans lesquels elle semble servir des objectifs divers au service des projets scientifiques et de la recherche[59], sous la forme d'un centre pour les humanités numériques et les documents historiques et patrimoniaux à l'image de l'Austrian Centre For Digital Humanities and Cultural Heritage (https://www.oeaw.ac.at/acdh/acdh-ch-home) ou d'un centre pour l'étude des manuscrits orientaux à l'image du Center for the Study of Manuscript Culture à l'Université de Hambourg (https://www.csmc.uni-hamburg.de/about.html), au sein desquels se trouvent des spécialistes en humanités digitales appliquées aux études arabes. La France dispose de son côté d'une infrastructure nationale (TGIR Huma-Num), qui offre un accompagnement de qualité, mais généraliste, et d'un institut spécialisé sur les manuscrits (IRHT), comptant en son sein une section arabe, mais qui n'a pas dans ses missions de jouer de rôle d'accompagnement des projets au niveau national. Le consortium Biblissima +, établi sur le Campus Condorcet pour la période 2020-2030 a vocation à devenir un observatoire de toutes les cultures écrites, depuis le cunéiforme jusqu'aux premiers imprimés, et entend développer ce rôle par une série d'appels ouverts, il constitue un environnement favorable, mais accorde jusqu'à présent très peu de place aux manuscrits arabes et orientaux. Ce rapport confirme ainsi le besoin déjà exprimé par les chercheurs lors de l'élaboration du livre blanc *Vers la science ouverte ?* et qui a débouché sur la préparation du consortium DISTAM (Digital Studies. Africa Asia Middle East) au sein de la TGIR Huma-Num, spécifiquement dédié aux humanités digitales aréales, et dont le premier groupe de travail se penchera notamment sur les questions d'acquisition et de fouille de données textuelles en alphabets extra-européens, préfigurant ainsi la création de ce pôle d'accompagnement des communautés de recherche au niveau national dans ce domaine.

[59] Elle existe aussi, dans une certaine mesure, en France pour les humanités numériques en général : https://iscpif.fr/projects/david-chavalarias/

GLOSSAIRE

API
Interface de programmation d'application (*Application Programming Interface*) qui permet à deux applications de communiquer entre elles. (Source : https://www.redhat.com/fr/topics/api/what-are-application-programming-interfaces)

Apprentissage profond
(deep learning)
Il consiste à soumettre à un réseau de neurones des bases de données proposant ce que l'on souhaite faire reconnaître sous de multiples facettes.

CER
Character Error Rate ou taux d'erreur au niveau des caractères. Il mesure le nombre de caractères erronés lors d'une conversion d'une image en texte par OCR ou HTR. Un CER de 2 % signifie que 2 caractères sur 100 sont erronés.

FAIR
Findable, Accessible, Interoperable and Reusable ; la notion de FAIR et plus précisément de FAIR data recouvre, dans le contexte de la science ouverte et du partage et de l'ouverture des données, les manières de construire, stocker et présenter ou publier des données en permettant que la donnée soit « Facile à trouver, Accessible, Interopérable et Réutilisable » (Source: Wikipédia).

Fine-tuning (spécialisation progressive)
Il s'agit, dans le domaine de l'intelligence artificielle, d'adapter un modèle préexistant à un jeu de données spécifique à la tâche visée.

Interopérabilité
Capacité que possède un produit ou un système, dont les interfaces sont intégralement connues, à fonctionner avec d'autres produits ou systèmes existants ou futurs et ce sans restriction d'accès ou de mise en œuvre. Un des quatre éléments des principes du FAIR.

Protocole IIIF
International Image Interoperability Framework. Protocole qui permet l'échange de fichiers de documents numérisés entre différents systèmes. Ce protocole implique l'existence d'un serveur qui centralise les images, et qui permet de les décrire en tant qu'« objet informatique » de façon standardisée pour ensuite pouvoir les intégrer dans des systèmes tiers comme une interface de visualisation (exemple de Mirador) ou des interfaces de labellisation ou d'annotation.
Sur ce protocole : https://beta.iiif.io/

Vérité terrain
« Ensembles de données annotées et corrigées de manière à fournir au modèle des paires composées d'une image ou d'une portion d'image (entrée) et d'autre part de l'annotation attendue (sortie), qui peut être des coordonnées dans le cas de la segmentation ou un ensemble de caractères pour la transcription » (Source : A. Chagué, T. Clérice, L. Romary, « HTR-United : Mutualisons la vérité de terrain ». https://hal.archives-ouvertes.fr/hal-03398740/document).

Crédits photographiques
Fig. 1, 8, 10 © Calfa
Fig. 2, 9, 11, 12 © Collections patrimoniales numérisées de la Bulac

Groupement d'Intérêt Scientifique
Moyen-Orient et mondes musulmans
Campus Condorcet
Bâtiment de recherche Sud
5, Cours des Humanités
93322 Aubervilliers Cedex
www.majlis-remomm.fr

Texte placé sous licence Creative Commons.
Attribution-ShareAlike 4.0
International (CC BY-SA 4.0)
http://creativecomons.org/licenses/by-sa/4.0

Achevé d'imprimer en avril 2022

Conception graphique : Maison Solide

Moyen-Orient et Mondes Musulmans
Groupement d'Intérêt Scientifique

ISBN 978-2-493818-01-0