HAL open science

Préparer un projet incluant l'extraction des textes en graphies non-latines par transcription et HTR

Anaïs Wion, Chahan Vidal-Gorène

▶ **To cite this version:**

Anaïs Wion, Chahan Vidal-Gorène. Préparer un projet incluant l'extraction des textes en graphies non-latines par transcription et HTR : Quelques conseils pratiques. 2023. halshs-04161903

HAL Id: halshs-04161903

https://shs.hal.science/halshs-04161903

Submitted on 13 Jul 2023

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.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Huma-Num IR*

distan
Digital Studies
Africa
Asia
Middle East

Préparer un projet incluant l'extraction des textes en graphies non-latines par transcription et HTR : quelques conseils pratiques

Livrable 2022 - GT1.1 – Février 2023

Rédaction :
Anaïs Wion, Chahan Vidal-Gorène, avec les relectures de Jean-Christophe Peyssard

Sommaire

1. Qu'est-ce qu'extraire un texte depuis une image d'un manuscrit et quelles sont les différentes manières d'y parvenir ? La transcription et l'HTR 3

2. Pour quels usages extraire un texte ? 4

3a. Préparer l'HTR : connaître son corpus 5

3b. Préparer l'HTR : connaître les outils – questions des modèles économiques 6

4. Les différentes phases de l'HTR : segmentation, alignement 7

5. Les différentes phases de l'HTR : entraîner un modèle de reconnaissance de caractères 8

6. Préparer l'HTR : connaître les projets similaires et le champ d'étude 8

7. La phase de post-traitement : corriger et typer les données ; exporter 9

8. Partager et ouvrir ses données : mise à disposition des jeux de données et des modèles 10

Lectures complémentaires 10

En complément de l'approche extrêmement fouillée proposée par le guide « OCR / HTR et graphie arabe, Les manuscrits arabes à l'heure de la reconnaissance automatique des écritures », *Cahier du GIS MOMM* n° 3, rédigé par Noémie Lucas et publié en 2022, qui peut être difficile à prendre en main pour des porteurs de projet, il a semblé utile de proposer des fiches pédagogiques concises. Elles visent à accompagner la mise en place d'un travail nécessitant l'acquisition de textes manuscrits en graphies non-latines. Il s'agit de rendre visible les étapes cruciales de tels projets, en insistant sur les choix qui peuvent être faits à chaque étape et qui déterminent l'avancement du travail. L'objectif de ces fiches synthétiques est de permettre de comprendre ce qui motivent ces choix, de documenter les processus de décision et d'aider à préparer ce travail à trois niveaux différents : en amont, afin de pouvoir anticiper les étapes ; lors du choix d'un outil en fonction des objectifs et du type de corpus ; une fois le texte acquis afin que la phase de post-traitement soit la plus optimisée possible.

1. QU'EST-CE QU'EXTRAIRE UN TEXTE DEPUIS UNE IMAGE D'UN MANUSCRIT ET QUELLES SONT LES DIFFÉRENTES MANIÈRES D'Y PARVENIR ? LA TRANSCRIPTION ET L'HTR

Un document manuscrit numérisé propose une image du texte : il faut donc transformer cette image en texte électronique. Deux méthodes complémentaires sont possibles :

1. La transcription manuelle, il s'agit alors de saisir le texte.
2. La transcription automatique, en utilisant des outils automatisant cette tâche, en particulier l'HTR : Handwritten Text Recognition (reconnaissance d'écriture manuscrite).

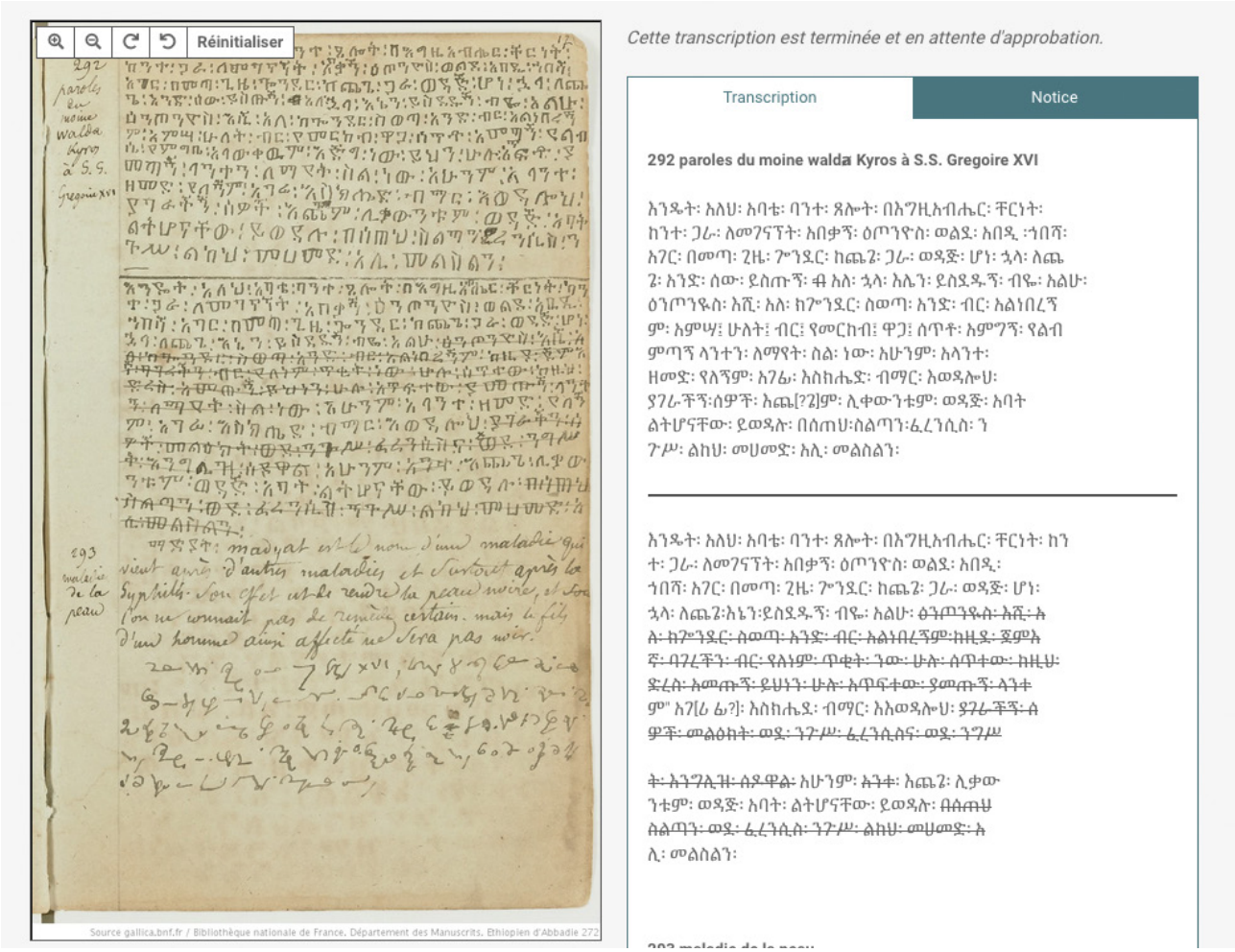

Fig. 1 : Transcription d'un feuillet du carnet BnF Ethiopien 272 d'Antoine d'Abbadie sur l'outil Transcrire. Les choix de transcription visibles ici concernent : le traitement des indexations rédigées dans les marges ; le rendu des lignes barrées ; le rendu des deux a par ɑ̸ et a (cf premier titre) ; le choix de laisser de côté la sténographie dans un premier temps.

Transcrire peut se faire de différentes façons. Au plus simple, on juxtapose une image du texte manuscrit avec un éditeur de texte et on saisit le texte. Mais alors rien ne lie l'image au texte obtenu. Il est aussi possible de transcrire vocalement en lisant le texte qui est reconnu par un logiciel de dictée. On peut également et de façon complémentaire aux techniques précédentes utiliser des outils dédiés à la transcription, tels Transcrire ou From the Page, (tous deux utilisant le plugin Scripto de Omeka-S). Cela permet de travailler de façon collaborative, de gérer l'état d'avancement du travail grâce à des outils de suivi et de validation, et par ailleurs d'établir une relation entre images et textes.

Le travail de transcription est une étape nécessaire pour apprendre à connaître les spécificités d'un corpus en termes de paléographie, mise en page, structuration, contenus. Cela permet de faire les choix scientifiques les mieux adaptés pour établir au mieux la transcription.

L'**HTR** permet d'automatiser ce travail d'acquisition du texte grâce à des outils reposant généralement sur de l'Intelligence Artificielle, en particulier le deep-learning. L'IA est entraînée par l'exemple à la reconnaissance d'écritures manuscrites. Néanmoins, l'utilisation de ces outils demande aussi un temps de travail conséquent qui comprend notamment souvent une phase préalable de transcription manuelle pour l'entraînement de l'IA, qui peut varier entre 10 et 50 pages selon la difficulté et l'homogénéité du corpus.

2. POUR QUELS USAGES EXTRAIRE UN TEXTE ?

Il est nécessaire d'avoir une vision la plus précise possible du projet de recherche et de ses objectifs et problématiques. En effet, le travail d'acquisition du texte pourra différer en fonction de ces derniers. Une édition diplomatique, c'est-à-dire restituant le texte au plus près de l'original, nécessite de préserver chaque particularité du texte (orthographique, paléographique, grammaticale). Mais il n'est pas toujours possible, par exemple, de conserver des glyphes anciens qui ne sont pas supportés par les caractères en unicode et des équivalences devront peut-être être introduites. Si l'édition est normalisée, il faut rapidement décider de l'ampleur de cette normalisation : jusqu'où rétablir une langue moderne afin de rendre accessible le texte et qui en sont les lecteurs potentiels ? Le cas des abréviations exemplifie très bien l'enjeu de ces choix : si les abréviations ont pour but d'être développées, alors la transcription manuelle comme l'HTR peuvent dès le début prendre en charge cette transformation. Mais si au contraire il s'agit de travailler sur les abréviations, alors il faut bien entendu les conserver et en rendre compte au mieux.

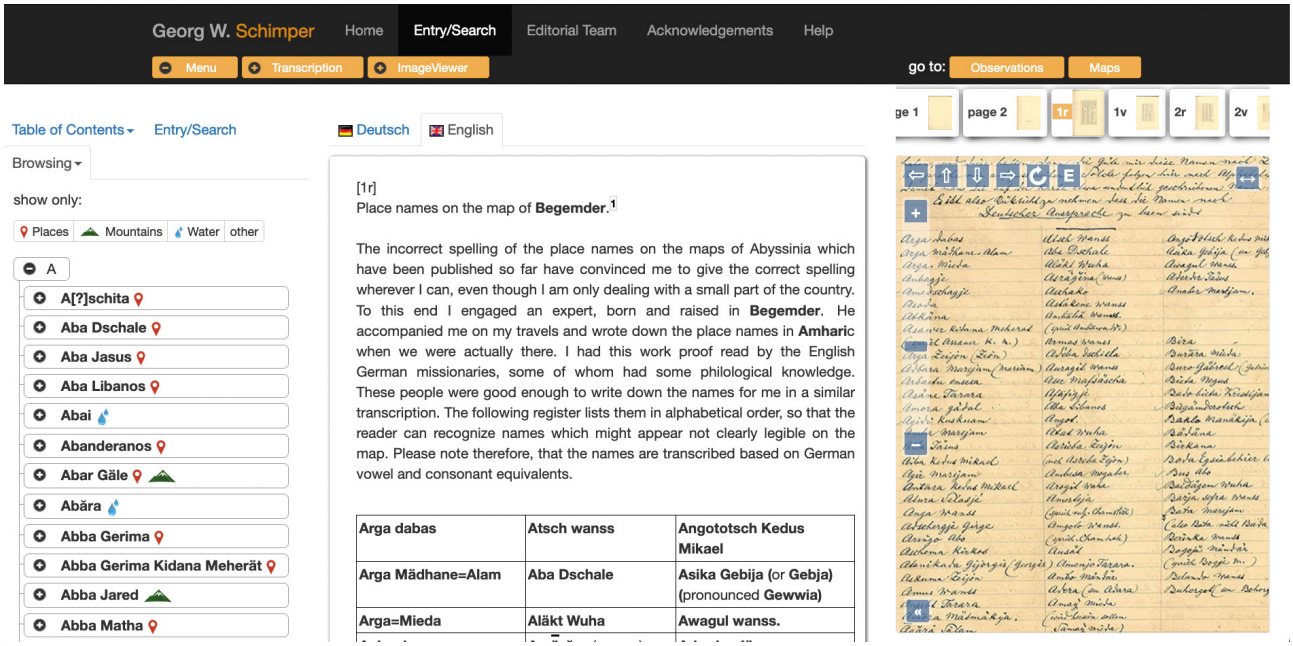

Fig. 2 : L'édition des carnets de notes et des cartes du naturaliste allemand G.W. Schimper. Un exemple d'édition bilingue particulièrement complexe

La rédaction d'un guide à l'attention des transcripteurs/trices permet de mettre à plat ces questions et de décider des réponses à y apporter.

Bien entendu, beaucoup d'éditions donnent à lire à la fois le texte original et sa version normalisée, ce qui inclue donc un travail parfois conséquent de traitement scientifique et éditorial.

Si l'édition est prévue pour être électronique et, probablement, enrichie, c'est-à-dire avec un marquage de certains types de données, il faut prévoir quelles seront les questions de recherche précises ou ouvertes que l'édition permettra d'éclairer. Si le travail scientifique porte sur des vocabulaires spécialisés et anciens, il ne faut pas normaliser à outrance et penser l'usage de la reconnaissance par mots en conséquence.

De la même façon, la prise en compte de la mise en page et de la structuration du texte dans l'édition finale doit être aussi anticipée, autant que possible.

3A. PRÉPARER L'HTR : CONNAÎTRE SON CORPUS

La numérisation de corpus parfois importants dans l'objectif d'en acquérir tout ou partie du texte permet véritablement de les découvrir.

Il est nécessaire de passer un certain temps à transcrire afin de se familiariser avec le corpus étudié.

Il est recommandé de relever les particularités paléographiques et de faire le point sur la disponibilité ou non de caractères unicode pour l'ensemble des graphèmes utilisés.

Il faut aussi explorer la variété des mises en page et des écritures à l'intérieur du corpus, en tâchant de transcrire des exemples simples et complexes de chaque type de feuillets.

Cette familiarisation avec la matière écrite est indispensable pour bien préparer les données qui serviront à l'entraînement de l'HTR.

Il faut en effet avoir transcrit un certain nombre de pages afin de générer des jeux de données validées qui vont permettre aux outils d'HTR d'établir des analogies entre les graphèmes sous forme d'images des manuscrits numérisés et le texte électronique qui en sera extrait. Mais ces jeux de données ne prennent pas en compte que les graphèmes, ils doivent aussi permettre d'établir la reconnaissance des mises en page. Ainsi il faut avoir transcrit des textes dans les différentes mises en page possibles d'un corpus pour anticiper sur la façon de prendre en compte ces mises en page. Un titre, une note marginale, un ajout interlinéaire, un numéro de page ou de paragraphe... sont autant d'éléments qu'il faut distinguer ou non, mais c'est un choix qu'il faut signifier à la machine (cf. point 4).

Durant ce processus, il est important d'assurer la cohérence de l'annotation tout au long de l'annotation.

3B. PRÉPARER L'HTR : CONNAÎTRE LES OUTILS – QUESTIONS DES MODÈLES ÉCONOMIQUES

De nombreux outils existent et, le domaine étant en pleine expansion, ils vont évoluer et se multiplier rapidement dans les années à venir.

Afin de choisir l'outil qui convient le mieux, plusieurs critères peuvent être passés en revue.

L'outil dispose-t-il d'une interface utilisateur ou doit-il être implémenté ? Spontanément, la plupart des chercheurs en SHS choisiront un outil doté d'une interface qui leur évitera d'avoir à développer celle-ci. Ainsi, l'outil Kraken peut être utilisé *per se*, mais il est préférable de l'utiliser une interface telle, par exemple, celle du projet eScriptorium.

L'outil est-il libre, open-source ou propriétaire ? Au-delà d'une question éventuelle de principe, cette question n'a d'intérêt que si votre projet dispose d'un développeur qui peut contribuer à faire évoluer un outil open-source, et si ce dernier dispose d'une communauté active avec laquelle échanger.

Son accès est-il payant ou gratuit ? Là encore, il faut se demander qu'est-ce qu'on paye lorsqu'on paye un service. Et qu'est-ce qui est gratuit quand on pense ne pas payer et quelles institutions payent les différents rendus. Il s'agit là d'une question qui va bien au-delà des outils d'HTR. Ainsi Transkribus est gratuit pour un nombre fixe de "crédits" qui permet de tester l'outil. La version payante permet de traiter un plus grand nombre de données. L'accès à eScriptorium est gratuit mais conditionné à acceptation, son développement et son installation sur des serveurs

fidel	graphie AdA majoritaire (ou des débuts)	graphie AdA alternative (ou de la fin)	transcription simple	**transcription experte**
አ			a	ⱥ (U+2C65)
ኧ	ö	e, o, ĭ, i̯	ö, e, o, i	**ö**, e, o, ĭ, i̯ (U+1E2D)
ዕ	ọ̈		o,	ö
ሐ	ḥ	h'	h	**ḥ (U+1E25), h'**
ኀ			H	ḧ (U+1E27)
ከ	k'		k'	**k'**
ኸ	k̲		k	**k̲ (U+1E35)**
ፕ et/ou ጸ			t	**Ť (U+0164), t'** (U+0165)
ጠ			t	**Ŧ (U+0166), ŧ** (U+0167)
ጨ		ç	t	**ʧ U+02A7**
ኘ	ñ		ñ	**ñ**
ቀ	k’		k’	**k’**
gémination				ّ (l'utilisation du signe de gémination arabe pose des pbs de bidirectionalité, donc mieux vaut suppléer par

Fig. 3 : Guide de transcription pour les carnets d'Antoine d'Abbadie sur Transcrire. On distingue un niveau « simple » et un niveau « expert » pour la transcription des diacritiques

étant dépendants de ceux des établissements de recherche. Dans ces deux cas, la force de travail est celle du projet. L'entreprise Calfa au contraire offre un service payant mais le traitement des données est à sa charge.

Le tableau ci-contre précise ces différents aspects pour chacun des outils cités ci-dessus.

Le budget pour l'utilisation d'un outil doit donc être pensé en amont car, en fonction du coût et des ressources qui seront mises en œuvre, le budget dédié ne pourra pas être imputé sur le même type de dépenses (prestation de service ou recrutement de contractuel·le).

Une dernière question qui doit être examinée avant de choisir un outil est celle de la réutilisation des données. Les modèles de reconnaissance (de la mise en page et des caractères) seront-ils mis à disposition du projet ? Plus important, car les modèles évoluent extrêmement vite, qu'en est-il du partage des données et de la documentation qui les accompagnent, qui permettent de ré-entrainer des modèles, et donc de s'affranchir de l'obsolescence des outils. Il faut alors se demander s'il importe pour le projet d'être en mesure de réutiliser les modèles entraînés, et de les partager, et pourquoi. (voir la fiche “Partage des données”).

Après ces questions portant sur le modèle économique, venons-en aux questions plus techniques qui sont aussi cruciales.

	Compétences et ressources informatiques	Coût	Travail d'annotation	Construction de modèles	Accès	Travail collaboratif
Transkribus expert	ø	Freemium pour test / payant après 500 crédits	chercheur	Transkribus	Sur demande	oui
Kraken	Développeur + serveur	gratuit	chercheur	chercheur	open	oui
eScriptorium	Serveur dans certains cas	gratuit	chercheur	chercheur	Sur demande	oui
Calfa-Vision	ø	gratuit	chercheur	Sur demande	Sur demande	oui
Calfa sur mesure	ø	Payant sur devis	Contractuel (cahier des charges)	Contractuel (cahier des charges)	Sur demande	non

4. LES DIFFÉRENTES PHASES DE L'HTR : SEGMENTATION, ALIGNEMENT

À ce jour, L'HTR se décompose en deux étapes : il va d'abord procéder en reconnaissant d'une part la mise en page (*layout*), puis d'autre part les caractères. Ces deux actions sont complémentaires et distinctes.

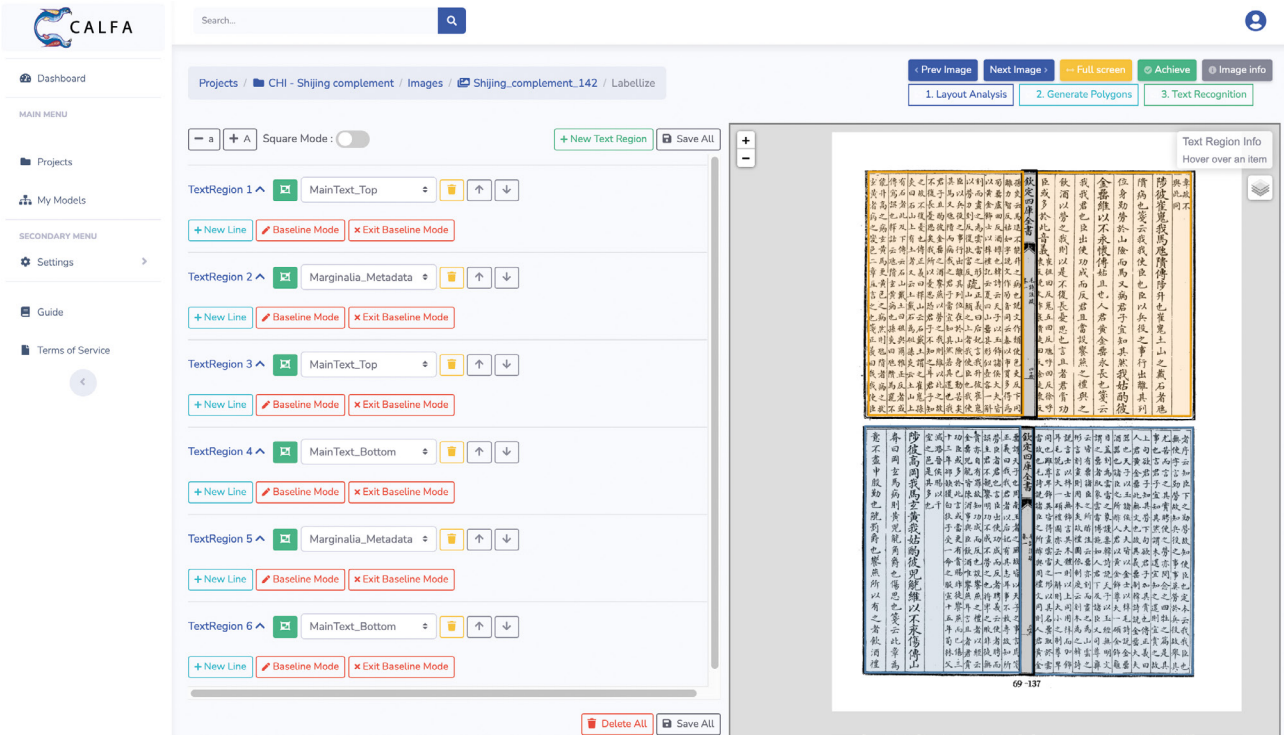

Fig. 4 : Reconnaissance des zones de texte dans un corpus d'imprimé ancien chinois

La mise en page est, dans l'état actuel des processus, définie par : des zones de texte (*text zone*), des zones comprenant les lettres (des polygones) et éventuellement -mais qui devrait à terme disparaître- des lignes de mots (*baseline*). Si les documents imprimés offrent généralement peu de surprise pour définir ces trois phénomènes, il en va tout autrement des documents manuscrits. Les ajouts interlinéaires ou marginaux défient la lecture des zones de textes. Les écritures en *scripto continua* ou excessivement serrées ne permettent pas aisément de différencier

les mots. Les exemples de complexité sont nombreux. Dans les cas complexes, il faut donc annoter manuellement ces zones qui segmentent le texte afin d'entraîner l'HTR à reconnaître ensuite de façon automatique ces éléments.

Cette étape est longue, et de plus elle doit être réalisée sur autant d'exemples caractéristiques qu'en comprend le corpus. D'où la nécessité décrite plus avant de bien connaître les différentes facettes de la mise en page d'un corpus donné.

Comme exemple, vous pouvez consulter un excellent billet de Jonathan Robker décrivant la segmentation effectuée sur des manuscrits hébreux :

Fig. 5 : URL : https://digitalorientalist.com/2022/03/01/introduction-to-escriptorium-htr-for-hebrew-manuscripts-part-1/

Note : l'approche par baseline, plus lourde, reste néanmoins particulièrement adaptée aux mises en pages complexes et aux corpus peu dotés.

5. LES DIFFÉRENTES PHASES DE L'HTR : ENTRAÎNER UN MODÈLE DE RECONNAISSANCE DE CARACTÈRES

Les caractères vont être reconnus soit au niveau du graphème, soit au niveau du mot. Selon le type d'écriture et les graphies, l'une ou l'autre techniques peuvent être privilégiées. Afin de comprendre ce qui est préférable en fonction du corpus, il est souhaitable de prendre le temps de discuter avec des ingénieurs compétents, avec des équipes pouvant faire des retours d'expérience, et de se renseigner sur les projets similaires.

6. PRÉPARER L'HTR : CONNAÎTRE LES PROJETS SIMILAIRES ET LE CHAMP D'ÉTUDE

Il est utile de se renseigner sur les initiatives prises au niveau international afin de s'inspirer de projets éventuellement comparables et d'apprendre de leurs succès ou de leurs erreurs, si celles-ci sont documentées.

Des conférences internationales réunissent régulièrement les projets d'HTR :

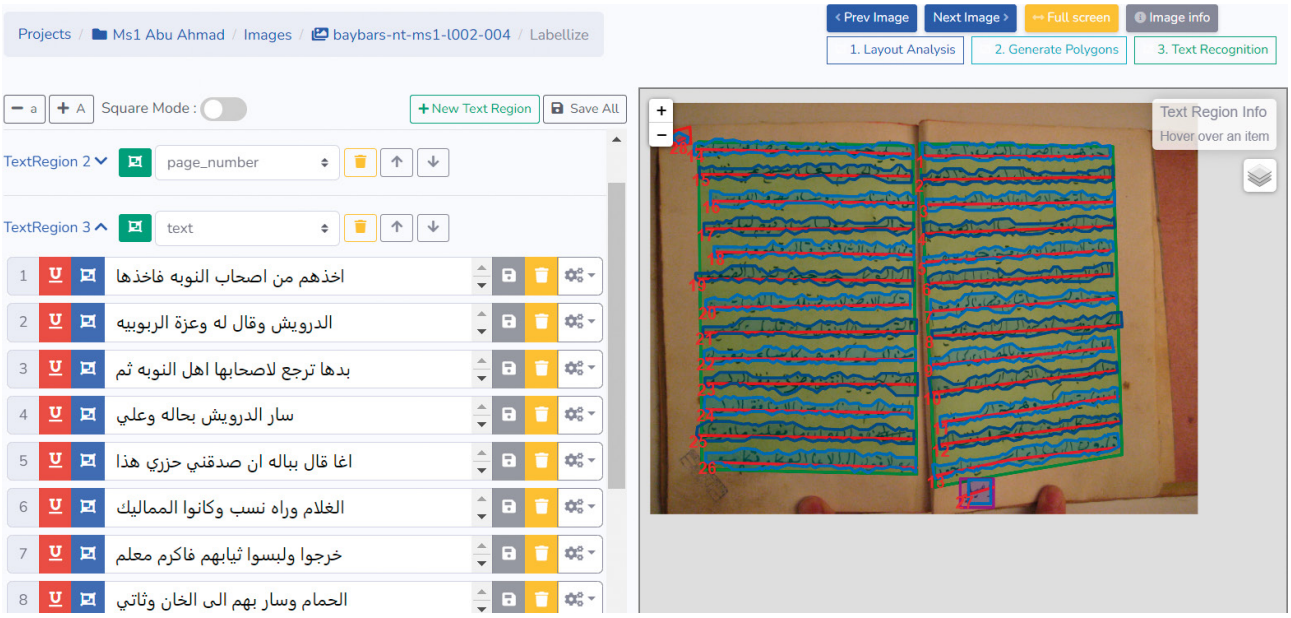

Fig. 6 : Entraînement du modèle HTR à partir de Calfa Vision sur une page du manuscrit dit d'Abu Ahmad de la sīra *de Baybars*

Mentionnons DATeCH : Digital Access to Textual Cultural Heritage, qui est assez généraliste mais traite aussi des questions d'HTR.

Plus spécialisées sont les conférences ICDAR : International Conférence on Document Analysis and Recognition, et ICFHR : International Conference on Frontiers in Handwritten Recognition, qui se réunissaient alternativement tous les deux ans jusqu'en 2023 (ICDAR tous les ans désormais). Leurs programmations, très riches, permettent de connaître les initiatives prises dans le champ des graphies non-latines.

Pour ce qui est de la graphie arabe, il existe aussi l'ASAR : International Workshop on Arabic and Derived Script Analysis and Recognition.

7. LA PHASE DE POST-TRAITEMENT : CORRIGER ET TYPER LES DONNÉES ; EXPORTER

Une fois le traitement par HTR réalisé, il est nécessaire de traiter les données.
Il faut d'abord vérifier la qualité du texte[1], la prise en compte de la mise en page, et probablement faire des corrections. Cette étape peut être faite de manière collaborative ou réalisé par des expert·es, en fonction de la structuration des équipes.

La plupart des outils permettent aussi de qualifier certaines parties du texte : il est ainsi possible de définir s'il s'agit d'un titre, d'un paragraphe, etc... Les outils évoluent vite et certaines de ces tâches de typage peuvent être faites en même temps que l'HTR.

Puis il faut exporter les données produites et validées, pour les utiliser dans l'interface d'édition prévue ou tout simplement en tant que texte brut.

Il est bon d'avoir une compréhension des formats d'exportation possibles. Le format XML-Alto permet de rendre compte de la structuration physique du texte en lien avec les coordonnées sur l'image de référence.

1 La qualité d'un OCR et HTR est classiquement évaluée avec le Character Error Rate (CER), qui donne le taux d'erreur au niveau du caractère : un CER de 3% signifiera que 3 caractères sur 100 sont erronés. Il existe aussi le Word Error Rate (WER), appliqué au niveau du mot : un WER élevé signifiera par exemple que les erreurs sont très dispersées dans le texte puisque de nombreux mots sont concernés, et non pas localisées dans une zone particulièrement difficile.

8. PARTAGER ET OUVRIR SES DONNÉES : MISE À DISPOSITION DES JEUX DE DONNÉES ET DES MODÈLES

Les jeux de données sont classiquement diffusés sur un Git (p. ex. Github, permettant le versionnement des fichiers) ou des sites d'archivage (p. ex. Zenodo), et se composent du fichier image (ou d'un accès à celui-ci dans une bibliothèque numérique) et du fichier contenant la vérité terrain.

Le projet HTR-United permet une mutualisation des jeux de données. L'objectif est de mettre à la disposition de la communauté des jeux de données permettant d'entraîner des modèles et ainsi de rendre plus performantes des analyses qui sinon ne bénéficieraient pas d'expérience antérieure. Il faut pour cela bien entendu choisir des jeux de données comparables aux caractéristiques des corpus qu'on souhaite analyser.

LECTURES COMPLÉMENTAIRES

Chahan Vidal-Gorène, « La reconnaissance automatique d'écriture à l'épreuve des langues peu dotées », *Programming Historian* en français 5 (2023), https://doi.org/10.46430/phfr0023.