

RASAM - A Dataset for the Recognition and Analysis of Scripts in Arabic Maghrebi

Chahan Vidal-Gorène, Noëmie Lucas, Clément Salah, Aliénor Decours-Perez, Boris Dupin

▶ **To cite this version:**

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. Document Analysis and Recognition – ICDAR 2021 Workshops, 12916, Springer International Publishing, pp.265-281, 2021, Lecture Notes in Computer Science, 10.1007/978-3-030-86198-8_19 . halshs-03430697

HAL Id: halshs-03430697
https://shs.hal.science/halshs-03430697

Submitted on 16 Nov 2021

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.

RASAM – A Dataset for the Recognition and Analysis of Scripts in Arabic Maghrebi*

Chahan Vidal-Gorène[1,4][0000−0003−1567−6508], Noëmie Lucas[2][0000−0003−2236−6778], Clément Salah[3,5], Aliénor Decours-Perez[4], and Boris Dupin[4]

[1] École Nationale des Chartes – Université Paris Sciences & Lettres, 65 rue Richelieu, 75003 Paris, France
chahan.vidal-gorene@chartes.psl.eu
[2] GIS Moyen-Orient et mondes musulmans – UMS 2000 (CNRS/EHESS), 96 boulevard Raspail, 75006 Paris, France
noemie.lucas@ehess.fr
[3] Sorbonne-Université, Faculté des Lettres, 21 rue de l'école de médecine, 75006 Paris, France
salah.clement@gmail.com
[4] Calfa, MIE Bastille, 50 rue des Tournelles, 75003 Paris, France
lastname.firstname@calfa.fr
[5] Institut d'Histoire et Anthropologie des Religions, Faculté de Théologie et Sciences des Religions, Université de Lausanne, CH-1015, Lausanne, Suisse

Abstract. The Arabic scripts raise numerous issues in text recognition and layout analysis. To overcome these, several datasets and methods have been proposed in recent years. Although the latter are focused on common scripts and layout, many Arabic writings and written traditions remain under-resourced. We therefore propose a new dataset comprising 300 images representative of the handwritten production of the Arabic Maghrebi scripts. This dataset is the achievement of a collaborative work undertaken in the first quarter of 2021, and it offers several levels of annotation and transcription. The article intends to shed light on the specificities of these writing and manuscripts, as well as highlight the challenges of the recognition. The collaborative tools used for the creation of the dataset are assessed and the dataset itself is evaluated with state of the art methods in layout analysis. The word-based text recognition method used and experimented on for these writings achieves CER of 4.8% on average. The pipeline described constitutes an experience feedback for the quick creation of data and the training of effective HTR systems for Arabic scripts and non-Latin scripts in general.

Keywords: Arabic Maghrebi scripts · Dataset · Manuscripts · Layout Analysis · HTR · Crowdsourcing

* This work was carried out with the financial support of the French Ministry of Higher Education, Research and Innovation. It is in line with the scientific focus on digital humanities defined by the Research Consortium Middle-East and Muslim Worlds (GIS MOMM). We would also like to thank all the transcribers and people who took part in the hackathon and ensured its successful completion.

1 Introduction

The automatic analysis of handwritten documents has become a classic preliminary step for numerous digital humanities projects that benefit from the mass digitization policy of heritage institutions. Following the competitions organized in recent years, at ICFHR and ICDAR notably, several robust architectures for layout analysis of historical documents have been developed [8], whose application to non-Latin script documents provide equivalent results [10,14]. The HTR architectures specialized on a type of document or on a hand also achieve a very high recognition score, even though the literature is mostly Latin script based, as well as the proven pipelines composed of character-level HTR and post-processing [7]. The non-Latin, cursive and RTL writings, like the Arabic scripts, remain an open problem in digital humanities with a wide variety of approaches [11]. Although specialized databases have emerged in recent years (see *infra* 2.1), they are often focused on the layout [10,6] and on common documents and writings, leaving out numerous under-resourced written traditions.

We are presenting a new dataset for the analysis and the recognition of handwritten Arabic documents, the first dataset focused on the writings called "Maghrebi scripts", also known as "Western scripts", or "round scripts". This term encompasses a variety of styles that have common characteristics and are poorly represented in digital humanities. These scripts dating back to the 10th century have been widely used however in the Islamic West – al-Andalus and North-Africa –, as well as sub-Saharan Africa until the 20th century[6]. They have numerous specificities that differentiate them from the classical problems met for Arabic handwritten character recognition. The rounded shape of these scripts may be explained by the writing tool used, that is qalams made from large reed straws cut in half lengthwise, with a pointed nib and not a biseled one as in the Islamic East [3]. Therefore, the Maghrebi scripts constitute a family of rounded scripts that share a number of characteristics, first of all very rounded loops, that can be seen in the manuscripts in the present dataset (see *infra* 2.3). The main characteristics[7] of the scripts are displayed in table 1.

The dataset, resulting from a collaborative hackathon held from January to April 2021, intends to cover a large spectrum of the handwritten production in Maghrebi scripts. The choice for a multilevel annotation (semantic and baseline annotation, word-level and character-level transcription) aims to offer to the scientific community a comprehensive dataset dedicated to the creation and evaluation of complete HTR pipelines, from layout analysis to text recognition for this written tradition. After a short presentation of the related work for the datasets, we propose a complete description of the manuscripts, the annotations and the editorial choices made for the transcription. The creation of the dataset

[6] The history and the origins of these scripts have been an important scientific open debate [4,3]. The most recent works, in particular those of U. Bongianino, have foregrounded the different itineraries (from books to qurans, from al-Andalus to the Maghreb) followed by these writings between the 10th and the 13th century [4].

[7] Characteristics are taken from U. Bongianino [4]; theoretical realizations are taken from the article of N. Van de Boogert upon which U. Bongianino draws [13].

has also benefited from a collaborative and semi-automatic work, with architectures dedicated to non-Latin scripts, that we assess with the view to replicate for other under-resourced languages.

Table 1. Some characteristics and realizations of Maghrebi scripts

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

2 Dataset and Arabic Maghrebi manuscripts of the BULAC

2.1 Existing resources for Arabic scripts

Following the competitions organized in recent years, at ICFHR and ICDAR, several datasets for Arabic written documents analysis have emerged. Such is the case of RASM2018 [6], focused on the Arabic scientific manuscripts on Qatar Digital Library, which is used to evaluate the layout analysis of documents in Arabic scripts. The 100 images of the dataset are annotated at different levels: region, polygons, lines and text. The Arabic scripts, however, present unique challenges for text-region detection and baseline detection. Therefore, in order to include these specificities, RASM2018 has been considerably expanded by BADAM [10] with a dataset focused on the Arabic scripts, comprising 400 images annotated at the text-region and baseline level.

These datasets cover a wide-ranging production of texts in non-Maghrebi Arabic scripts to enable the training of dedicated models [10]. More generally, there are other smaller or more specialized datasets, like HADARA80P [12] and VMH-HD [9], annotated at the region and word level, like KERTAS [2], dedicated to manuscripts datation, or like WAHD [1], dedicated to writer identification. Aside from the targeted tasks and the languages concerned, these datasets shed light on the variety of existing perspectives for handwriting, either inspired by the Latin languages, or by word-based approach. A FCN followed by a post-processing for baseline extraction gives robust results even on the most complex layouts [10,14]. Manuscripts in Arabic Maghrebi scripts are largely excluded from the datasets.

2.2 Dataset composition

RASAM is available under an open license[8]. It comprises 300 annotated images extracted from three manuscripts selected among the collections of the Bibliothèque Universitaires des LAngues et Civilisations (BULAC)[9]. The images of the dataset are in JPEG format and have varying resolutions from 96 DPI to 400 DPI. Experiments are carried out on the hackathon results (v1.0, 297 images). Dataset has been expanded in June 2021 (v1.1, 300 images, includes minor corrections).

Thus, two manuscripts of the dataset have been chosen among the 150 Arabic manuscripts available online (MS.ARA.1977 and MS.ARA.609); the third and last manuscript (MS.ARA.417) of the corpus has been digitalized at our request. The variety of topics, the representative type of the Maghrebi script, as well as the diversity of layouts has informed our choice of manuscripts to annotate and

[8] https://github.com/calfa-co/rasam-dataset

[9] The BULAC holds the second biggest fund of Arabic manuscripts in France (2.458 identified documentary units). BULAC collections contains a substantial proportion of the manuscripts copied in Maghrebi script. 150 Arabic manuscripts are available online on the website of the BINA Digital library.

transcript for this dataset. The aim is to obtain polyvalent analysis models for this written tradition. Therefore, two manuscripts belong to the historical genre (*'ilm al-tārīḫ*), whereas the third has to do with inheritance law (*fiqh al-farā'iḍ*). The small number of manuscripts is justified by the necessity to quickly achieve HTR models. The pages are not sequential, and the pages have been randomly selected to cover all the variations of a same copyist and the different layouts within a single manuscript.

2.3 Selected manuscripts

MS. ARA. 1977 : The manuscript MS.ARA.1977[10] consists in a compilation of 249 pages: the most part of which (p. 1-201) is a historical treatise entitled *al-Ǧumān fī muḫtaṣar aḫbār al-zamān* written by the Andalusian historian Abū 'Abd Allāh Muḥammad b. 'Alī b. Muḥammad al-Šuṭaybī (d. 963/1556), disciple of the great Maliki jurist Aḥmad Zarrūq (d. 899/1493). The second 38-page long text (p. 205-243) deals with the customs and practices relating to the prophet Muḥammad; as for the third text (p. 247-249), it is a recollection of the words of a scholar al-Ḥasan b. Mas'ūd al-Yūsī (d. 1102/1691), native from the North-West of the Moroccan Middle Atlas. It deals with the mission that the prophet Muḥammad would have entrusted to the Berber tribes to conquer the Maghreb. The annotation and transcription have been achieved on the first and main part of the compilation, that has been copied by Muḥammad b. Mubārak al-Barāšī around 1259/1843[11]. Compiled on paper (305 x 210 mm.), the manuscript pages contain 31 lines, with the exception of the last three which contain 27 each, and the pages 67-68, 202-204 and 245-246 that are left blank. While the main text is written in black ink, here and there the copyist has used red (e.g. limited to section titles) and green inks (e.g. to indicate poetry verses). This manuscript features a series of marginalia: besides the catchwords at the bottom of the page, numerous corrections and notes are displayed along the text. The characteristics of the ink and the handwriting lead us to assume that they are made by the copyist himself. The same can be assumed for the manuscript MS.ARA.609 (see *infra*).

MS. ARA. 609 : The manuscript MS.ARA.609[12] consists in a treatise in verse on arithmetic, on inheritances and wills. Written by the Maliki jurist 'Abd al-Raḥmān al-Aḫḍarī (d. 953/1513 or 983-1575) around 946/1540, the poem and its commentary are about the science of successions (*'ilm al-farā'iḍ*) and the arithmetic knowledge it required. Abū Zayd 'Abd al-Raḥmān b. Muḥammad al-Aḫḍarī, one of the great names of the Maliki school of the 10th/16th century, was born in 919/1513 near Biskra. He is the author of numerous didactic poems,

[10] مجموع – MS.ARA.1977, Collections patrimoniales numérisées de la BULAC.

[11] Muḥammad b. Mubārak al-Barāšī is also the copyist of the second text. There is no mention for the third text: the paleographical characteristics of the pages lead us to assume that it is the work of another hand.

[12] شرح الدرة البيضاء – MS.ARA.609, Collections patrimoniales numérisées de la BULAC.

often along with their commentaries, in several scholarly fields of study (logic, arithmetic, rhetoric, law). Compiled on paper (210 x 175 mm.), the copy of the manuscript was completed in 1146/1734 by Abū l-Qāsim b. Muḥammad b. Abū l-Qāsim al-Duraydī. The manuscript has 202 pages – 100 written folios, the folios 1 and 2 have been left blank – each page contains 25 lines. The main text is written in black ink, however the copyist has used red ink on several occasions (e.g. tables, numbers or poetry verses). This manuscript holds numerous numbers, fractions and tables throughout the text. The numbers are written in Indo-Arabic numerals. Like in the MS.ARA.1977, catchwords, additions, corrections and glosses are displayed on the lateral, top and bottom margins.

MS. ARA. 417 : The manuscript[13], dated from 1292/1875, was copied on the manuscript n°1061 of the National Library of Algeria. It narrates the history of Beys of Oran in the 13th century: the *Tārīḫ Bāyāt Wahrān*, written by Ḥasān Ḫūğah, secretary of Ḥasān Bey (1817-1831). It distinguishes itself from the two previous manuscripts by its length and its layout: it consists of 48 folios with pages of 12 lines, each of them with less than 10 words per line. The manuscript, very well written, is in black ink, even though the copyist uses red ink sporadically (e.g. chapters headings or some separators). In the lateral margins, another hand, which looks identical to the note of cataloguing in the first page, has added the names of the beys in Arabic and some dates. The same hand seems to have added some vocalizations marks written on some folios and some corrections with a blue ink.



Fig. 1. MS.ARA.1977 (p. 42), MS.ARA.609 (p. 124) and MS.ARA.417 (f. 12v)

3 Ground-Truth Content and Creation

Each image is associated with a pageXML format file describing the entire ground truth and the associated metadata. The annotations have been realized automatically on Calfa Vision platform, then manually checked over the course of a collaborative hackathon.

[13] See bibliographic record on CALAMES.

The overall set of constitutive parts of a manuscript page is annotated. We offer for each image: (i) a semantic annotation of the regions, (ii) an annotation of baselines (polylines), (iii) a polygons framing every line associated to a baseline and (iv) the transcription. In figures, the dataset is comprised of 300 images, 676 text-regions, 7,540 lines and 483,725 characters (v1.1).

3.1 Structure description

Text-region : The layout analysis consisted in identifying all the text-regions. We have defined 5 classes: text (300), marginalia (171), catchword (102), table (53) and numbering (50). The classes are not uniformly distributed but constitute clearly identifiable items. To prevent an overwhelming variety and ambiguity of classes, the 5 defined classes can incorporate some regions that would be otherwise traditionally separated. Thus, the titles, often written in color, are not separated but included in the text class. Likewise, all content in the margins (everything outside from the main text-region) is encompassed in the class marginalia, with the exception of the catchwords. Table and numbering respectively refer to the tables located inside and outside the main text-region and to all fractions within the text (see figure 2).

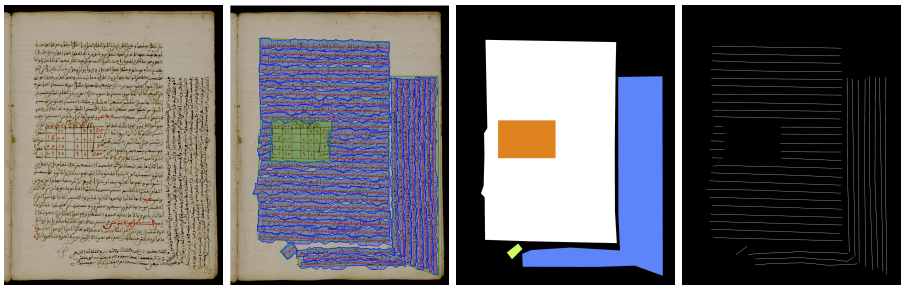


Fig. 2. Semantic classification of text-regions and baselines

Baseline : We adopted the annotation by baseline, suited for the Arabic scripts and interoperable with other datasets of the state of the art. Each segment of a sentence has its own baseline. In the event of overlap, as displayed by figure 2 for the marginal note, there is no continuity in the baseline. Besides, the outline of the latter follows the text line actually present in the manuscript and not a theoretical line that would link two segments of a single sentence (for instance, in the case of a line break for versification or due to a table – see figure 2). The reading order can be managed afterwards in post-processing. The manuscripts present numerous curved lines (see figure 2), in particular at the end of sentences and in the marginal notes. In this case, the baseline follows the same scheme that for BADAM [10], following a theoretical rotation point to match with the line curvature. Markers of verses or other signs for aesthetic purposes have not been taken into consideration (see figure 3, no 3). Lastly, in the event of characters

composed of strokes expanding beyond the body of the character (e.g. the letter *nūn* in figure 3, no 1) and located at the end of a sentence, the baseline has been extended to cover the entire shape of the character, even in the absence of the theoretic writing line.

Polygons : Each line of text is extracted with a surrounding polygon, drawn with an adaptive seamcarve implemented on the annotation platform [14]. Polygons have been manually proofread to integrate all the constitutive strokes of a given character, including ascenders and descenders, as well as the associated diacritics (see figure 3, no 4). There remain overlaps between the polygons of lines (see figure 3, no 2), but the HTR results have demonstrated that these overlaps have very few impact on HTR predictions (see *infra* 4.2).

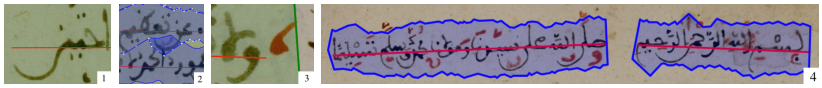


Fig. 3. Special cases for baselines and polygons annotation (MS.ARA.609 and MS.ARA.1977)

3.2 Specifications for transcription

A common framework for the text input was defined to preserve the uniformity of the transcriptions of each participant. The aim is to achieve a HTR producing predictions as close as possible to the original text, and thus to offer a complete transcription and allow for a big panel of editorial choices. The dataset comprises 54 classes, whose detail we give in table 2.

Table 2. Letters distribution in RASAM dataset (v1.1)

space	88,674	ت	12,386	ص	3,957	fatḥa	371	3	16
ا	66,996	ف	12,064	خ	3,880	sukūn	357	4	12
ل	48,324	د	9,508	ذ	3,500	šadda	344	6	10
م	26,608	ق	9,219	ش	2,924	ؤ	121	9	10
و	25,606	ك	8,818	ض	2,675	#	100	5	9
ن	22,378	س	8,793	ط	2,271	أ	91	7	8
ي	22,105	ة	6,658	ز	2,096	fatḥat^{an}	78	8	7
ه	19,213	ح	6,522	ء	1,696	kasra	44	0	3
ر	16,559	ى	6,418	غ	1,278	ḍamma	31	ḍammat^{an}	2
ب	15,956	ج	5,343	ظ	720	2	18	إ	2
ع	13,970	ث	4,312	ئ	648	1	16		

We have notably realized transcriptions that restore the spaces in Arabic, even when there are visually no discernable spaces in the manuscript. In view of the great variety of character morphologies in the Maghrebi manuscripts scripts, we favored the word-based approach instead of the character-based approach where the word separation is managed in post-processing [5].

Table 3. Spelling conventions (examples from MS.ARA.609 and MS.ARA.1977)

	Example
Confusion between *ḍād* and *ẓā'*	 تحفض where the copist should have written تحفظ
ṭā' marbūṭa in final position	 الميت / الية
haḏā/haḏihi	 هاذه / هاذا
In case of an erased character	 Although we can guess that it may be: كالرجل, the transcription was addressing what can be actually identified, here: كالرج

In order to remain as close as possible to the text, the transcription follows the spellings habits of the copyist, even when they depart from the norm of standard Arabic. Hence, the frequent confusion, in particular in the MS.ARA.609, between *ḍād* and *ẓā*, and between *ṣād* and *ṭā'* have been retained. For instance, for the transcription of *ṭā'*, that could be spelled as *tā' marbūṭa*, or the other way around, the misspell was kept. Furthermore, the demonstratives *hāḏā*, *hāḏihi*, and in some instances *ḏālika*, that are spelled in modern Arabic with a defective form or with a dagger *alif*, are often spelled with their archaic form with a medial *alif*: we maintain in our transcriptions the spelling of this *alif* (see table 3).

Table 4. Some realizations of the *hamza* in MS.ARA.609 and MS.ARA.1977

ان – *hamza* in أن or إن is not present	 فان  انه  ان
Alif madda has been transcribed as it was done	 ءادم where we would have written آدم
When the *hamza* was not supported, we respected the way it was done	 اسراءيل and not اسرائيل

The punctuation, when present, has not been transcribed. In the event of vocalization signs or *šadda* (ـّـ), the participants were free to transcribe or not, the priority was on the characters and words. When not understood, the char-

acter was to be replaced by the sign #. When impossible to read because of an alteration of the manuscript, nothing was to be transcribed (see table 3).

Two particular cases caught the attention of the transcribers and were much debated: the *hamza* (ء) and its different forms and the diacritic signs of the *yā'*. The main rule adopted for the *hamza* was the following: if there is no *hamza* do not add one and transcribe the *hamza* as it is written in other cases (see table 4).

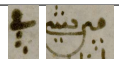
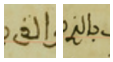
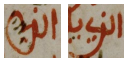
However, in the MS.ARA.609 and in the MS.ARA.417, the copyists used a singular *hamza* shape but consistent. In the MS.ARA.609, the *hamza* was written as a full *sukūn*. In these cases, we have considered that it was the way the copyist realized the letter (see table 5).

Table 5. Specific realizations of the *hamza* in MS.ARA.609



In the cases where the *hamza* was written below the line, with or without the diacritics of the ي, the *hamza* was drawn on *alif maqṣūra*. The issue of transcription of the ي was raised for the MS.ARA.1977 manuscript in particular, in which the copyist only wrote diacritics in rare instances. It was decided not to correct and to transcribe as close as possible to the text with only three exceptions: for the preposition في and the relative pronouns التي / الذي which, in the Maghrebi script, sometimes form some kind of glyphs (see table 6).

Table 6. Exceptions regarding *yā'* and its diacritics

	MS.ARA.1977	MS.ARA.609	MS.ARA.417
في			
التي			
الذي			

4 Evaluation of the crowdsourcing campaign and HTR models

The annotations have been realized with the Calfa Vision platform[14] [14], which incorporates – besides the online collaborative work on an image in real time – models for layout analysis and HTR prediction. These models are automatically assessed and fine-tuned, according to the corrections given by the contributor to the project, in order to speed up the checking task for the next images.

[14] https://vision.calfa.fr

4.1 General considerations and implementation protocol

The crowdsourcing campaign gathered 14 participants from January to April 2021, divided into three projects. The contributors were paired: one annotated and the other checked. Three main tasks were set to annotate each image.

1. Annotation and check of the layout analysis. A text-region (polygon) and a baseline (polyline) detection is automatically carried out beforehand. A first team is entrusted with the verification of the predictions and when necessary with the correction of the polygons and polylines shapes. The specifications to follow are defined in part 3.1.
2. Transcription of the text. Once the layout verified, the page is manually transcribed according to the specifications described in part 3.2. Some pre-annotations are realized in a second phase, once the HTR models are sufficiently precise (see table 9).
3. Extraction of the lines with a surrounding polygon for each transcribed line.

After each task, the images are re-assigned, to enable cross-check and to smooth the annotation habits. When all three tasks are completed, a comprehensive verification is carried out by a team of administrators.

4.2 Benefits of fine-tuning and transfer learning for a under-resourced language

The layout analysis models, provided by Calfa Vision for a project of handwritten documents annotation, are trained with an extensive dataset of various handwritten documents both medievals and recents [14]. A first assessment has been realized on BADAM with 0.9132% precision and 0.8575% recall [14].

Layout analysis and baseline models : After each proofreading of predictions, models are evaluated. The automatic re-training has been processed with a batch of 50 verified images from the three manuscripts. Evaluation is carried out on the remaining images to annotate. The batch of 50 was compiled to encompass a wide range of assessed layouts.

Table 7. Fine-tuning of Calfa Vision models for baseline prediction (v1.0)

Model	Precision (%)	Recall (%)	F1-score (%)
Default	0.8886	0.9522	0.9193
Model 1 (Default + batch 1)	0.9627	0.9720	0.9673
Model 2 (Model 1 + batch 2)	0.9650	0.9716	0.9683
Model 3 (Model 2 + batch 3)	0.9680	0.9756	0.9718
Model 4 (Model 3 + batch 4)	0.9762	0.9694	0.9728
Model 5 (Model 4 + batch 5)	0.9769	0.9700	0.9734

We use the metric of the cBAD competition [8] and implemented on Calfa Vision [14]. From the first fine-tuning, we notice a significant increase in the

model ability to correctly predict baselines on the dataset images. We also notice a steady improvement for the precision and the F1-score. The already high recall has little variation throughout the fine-tuning, with a slight dip when the batch is predominantly comprised of very curved lines. The results displayed in table 7 constitute a baseline for the HTR ability to identify the lines of text in common Arabic Maghrebi scripts manuscripts. The very high score achieved by the first model demonstrates a strong ability to rapidly reach quality fine-tuning for a under-resourced language (see figure 4).

At the region level, we evaluate the relevance with an Intersection over Union (IoU) metric. The default model is already convincing to identify the area of the main text, but without distinction between the main text, catchwords and the marginal notes. Table and numbering are not considered.

Table 8. Fine-tuning of Calfa Vision models for text-region prediction (v1.0)

	average IoU (%)				
Model	T	M	C	Tab	N
Default	0.9780	-	-	-	-
Model 1 (Default + batch 1)	0.9673	0.2177	0.3221	0.0866	0.0095
Model 2 (Model 1 + batch 2)	0.9751	0.3809	0.4068	0.1823	0.0213
Model 3 (Model 2 + batch 3)	0.9720	0.3617	0.6197	0.1900	0.1285
Model 4 (Model 3 + batch 4)	0.9680	0.5528	0.7772	0.2737	0.1826
Model 5 (Model 4 + batch 5)	0.9685	0.6268	0.8853	0.2813	0.1219

The fine-tuning triggers mechanically a decrease in the score of main text identification, but also the rapid inclusion of the other classes (see figure 4). Concerning the text-regions, the latest model achieves an accuracy of 0,8534% on average. The margins and the catchwords are sometimes very close to the main text 1 which makes it difficult to distinguish from the main text. As for the tables and numbering, their very low distribution and unequal division in batches result in lower scores. The relevance of the numbering class may also be questioned based on the outcomes[15]. We achieve similar result with the creation from scratch of a model with the whole dataset. We nevertheless notice a quick integration of the new text-regions in the models (see figure 4).

At the polygon of lines level, with the same metric, we measure a global relevance of 94%, no matter the curve of the line. The main difficulty encountered concerns the diacritics, which sometimes are not encompassed by the polygon and must be manually corrected. An adjustment of the seam carve has occurred as soon as the batch 20 to better manage the line height. The polygons verification constitutes nevertheless the most time-consuming task. For the first proofreading task of layout, baseline and polygon predictions, the time saved amounts to 75%. On average, the full process (predictions and proofreading) takes 7 min. for an image with the default model, bringing down to 4 min. for

[15] Numbering class is not kept in the v1.1 of the dataset, for which we notice a 9% gain in average for identification of catchword and table classes.

an image from the first model. The time required is down to 3.2 min. for the model 3, then to 2.5 min. for the last model. At this point, the proofreading can be confined to the curved lines of the marginal notes. Results are summarized on figure 4.

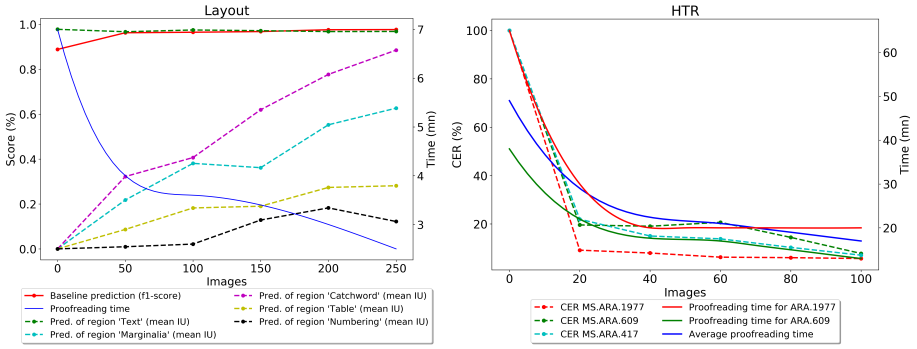


Fig. 4. Evolution of fine-tuning and effects on proofreading time for layout analysis and transcription (v1.0)

HTR models : Two types of models have been created and evaluated with the default architecture proposed by Calfa Vision. The first are HTR models specific to a project, thus specialized on a manuscript to accompany the transcription. Here, we have chosen batches of 20 corrected images.
(i) Models dedicated to a project : Each manuscript has its own difficulties and a specific number of lines (see *supra* 2.3). We have measured the learning ability of HTR models on each manuscript to go along with the transcription (see table 9). The models are trained incrementally as new transcriptions are checked and evaluated on the following folios of the project.

Table 9. Evolution of the CER for dedicated HTR models (v1.0)

	CER (%)		
Model	MS.ARA.1977	MS.ARA.609	MS.ARA.417
Model 1 (batch 1)	9.17	19.69	21.96
Model 2 (Model 1 + batch 2)	7.99	19.07	15.03
Model 3 (Model 2 + batch 3)	6.28	20.68	13.85
Model 4 (Model 3 + batch 4)	6.08	14.46	10.32
Final (Model 4 + batch 5)	5.71	7.80	7.10

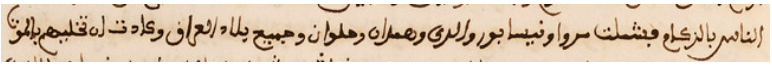
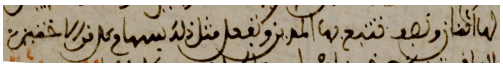
The MS.ARA.609 and MS.ARA.417 have fewer text lines, thus the CER stays high until the batch 4[16]. However, we observe a significant gain in the

[16] With better polygons (dataset v1.1), the CER decreases more quickly (16.6 for batch 1, then 15.87, 13.67, 11.52, and finally 6.67 for the last batch).

annotation upon application of the first model, with an average transcription time cut from 49 min. for an unassisted transcription, to 29 min. as of the first batch and 21 min. for the last model: hence an average gain of 42% (see figure 4). In detail, the gain is 56% for ARA.1977 (with extrema of 1H15 unassisted and 20 min. with a model) and 45% for ARA.609 (with extrema of 45 min. unassisted and 13 min. with a model). Word separation, HTR classical issue, is accurate at 80,4% for each specialized model.

Most character-level prediction errors seem to be more about the characters in the initial or final position in the word. Among the most frequent errors, the final *nūn* can be confused with the *rā'* or the *zāy*, as well as the *dāl* and the *rā'* or the *ḍād* and the *ḥā'*. Furthermore, we observe difficulties in identifying hyphenations between words, leading to misidentification of words. It should also be noted that when several characters with superscript or subscript diacritics follow each other (e.g. a sequence *tā'*, *nūn*, *šīn* or a sequence *bā'*, *yā'*, *fā'*), the prediction of this sequence of letters is frequently incorrect and random, but adding more context with mixed models show a significant improvement of predictions in these cases (see figure 5).

Table 10. Example of predictions on MS.ARA.1977 and 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 بالموت
	
Pred	لها اثنان ونصف تتبع بها المدين ويفعل مثل ذلك سهام كل من الاختينة
GT	لها اثنان ونصف تتبع بها المدين ويفعل مثل ذلك بسهام كل من الاختين

(ii) HTR models for Arabic Maghrebi scripts : We have also measured the relevance of transfer learning in-between manuscripts. Some transcription campaigns have in deed progressed more quickly than others and to re-purpose a specialized model for another manuscript is proving beneficial. Results of transfer learning are described in the figure 5. For each model, we have used 80% of data for training and 20% for testing. Four mixed models have been evaluated.

The confusion matrix highlights the big discrepancies between the three manuscripts, and no specialized model can achieve a CER below 20% on the other manuscripts. Manuscripts display a wide variety of text density, loop shapes and diacritics management that could affect transfer benefits. These limitations also lead to very different shapes of polygons. But in contrast, we observe a much higher convergence of mixed models. The MS.ARA.1977 and MS.ARA.417 manuscripts benefit more from this transfer than MS.ARA.609 which presents specific difficulties. CER of MS.ARA.417 is below 4% with mixed models. If we observe no real gain for CER on each manuscript, the transfer favors greater robustness for word separation with a gain of 8.54% on average, that



Fig. 5. Impact of transfer learning on CER (v1.0)

is consistent with the word-based approach. Experiments also show an ability to correct manual transcriptions misprint. The complete model demonstrates great versatility for these scripts with an average CER of 4.8% Moreover, first experiments on v1.1 tend to indicate that a precise polygonization of lines, with all diacritics included, is necessary with few data (see footnote 16). At the word-level however, gain remains marginal with a larger dataset.

In practice, the editorial choice to manually transcribe the majority of the dataset was made to limit possible typos. For an annotation project, this annotation approach with mixed models is deemed effective and will be implemented in future work, transfer showing a good and fast specialization with a slight fine-tuning.

5 Conclusion

We are presenting a new dataset comprising 300 annotated pages of Arabic Maghrebi script manuscripts of the BULAC. The chosen manuscripts display various layouts, sometimes very complex, with diverse deteriorations. The selected scripts encompass a representative panel of the handwritten production in Maghrebi scripts, in order to foster the emergence of robust HTR systems for these writings. Our work takes part in the commitment of the French scientific community towards the studies of Maghreb and for the promotion of the Maghrebi archives and manuscripts. Though the norms of transcription may be subject to evolution, our evaluations attest already a good recognition of these scripts, with a CER of 4.8% for the three manuscripts on average. The Arabic scripts and Arabic Maghrebi scripts in particular raise several difficulties for their layout processing and their recognition. We demonstrate that a crowdsourcing approach incorporating automatic fine-tuning and transfer learning is a successful strategy for data creation for under-resourced languages. It achieves similar results to those of the state of the art for manuscripts in Latin scripts.

Future work will focus on evaluating the versatility of this dataset and the HTR capabilities for Maghrebi Arabic scripts.

References

1. Abdelhaleem, A., Droby, A., Asi, A., Kassis, M., Asam, R.A., El-sanaa, J.: WAHD: A database for writer identification of Arabic historical documents. In: 2017 1st International Workshop on Arabic Script Analysis and Recognition. pp. 64–68 (2017)
2. Adam, K., Baig, A., Al-Maadeed, S., Bouridane, A., El-Menshawy, S.: KERTAS: dataset for automatic dating of ancient Arabic manuscripts. International Journal on Document Analysis and Recognition (IJDAR) **21**(4), 283–290 (2018)
3. Ben Azzouza, N.: Les corans de l'occident musulman médiéval : état des recherches et nouvelles perspectives. Perspectives **2**, 104–130 (2017)
4. Bongianino, U.: The Origins and Developments of Maghribī Rounds Scripts, Arabic Paleography in the Islamic West (4th/10th-6th/12th centuries). Ph.D. thesis, University of Oxford (2017)
5. Camps, J.B., Vidal-Gorène, C., Vernet, M.: Handling Heavily Abbreviated Manuscripts: HTR engines vs text normalisation approaches (2021), accepted for IWCP workshop of ICDAR 2021
6. Clausner, C., Antonacopoulos, A., Mcgregor, N., Wilson-Nunn, D.: ICFHR 2018 Competition on Recognition of Historical Arabic Scientific Manuscripts – RASM2018. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 471–476 (2018)
7. Clérice, T.: Evaluating Deep Learning Methods for Word Segmentation of Scripta Continua Texts in Old French and Latin. Journal of Data Mining & Digital Humanities **2020** (2020), https://jdmdh.episciences.org/6264
8. Diem, M., Kleber, F., Sablatnig, R., Gatos, B.: cBAD: ICDAR2019 Competition on Baseline Detection. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1494–1498 (2019)
9. Kassis, M., Abdalhaleem, A., Droby, A., Alaasam, R., El-Sana, J.: VML-HD: The historical Arabic documents dataset for recognition systems. In: 2017 1st International Workshop on Arabic Script Analysis and Recognition. pp. 11–14 (2017)
10. Kiessling, B., Ezra, D.S.B., Miller, M.T.: BADAM: A Public Dataset for Baseline Detection in Arabic-Script Manuscripts. In: Proceedings of the 5th International Workshop on Historical Document Imaging and Processing. p. 13–18. HIP '19, Association for Computing Machinery (2019)
11. Milo, T., Martínez, A.G.: A New Strategy for Arabic OCR: Archigraphemes, Letter Blocks, Script Grammar, and shape synthesis. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage. pp. 93–96. DATeCH2019, Association for Computing Machinery, New York, NY, USA (2019)
12. Pantke, W., Dennhardt, M., Fecker, D., Märgner, V., Fingscheidt, T.: An Historical Handwritten Arabic Dataset for Segmentation-Free Word Spotting - HADARA80P. In: 2014 14th International Conference on Frontiers in Handwriting Recognition. pp. 15–20 (2014)
13. Van Den Boogert, N.: Some notes on maghribi script. Manuscripts of the Middle East **4**, 30–43 (1989)
14. Vidal-Gorène, C., Dupin, B., Decours-Perez, A., Riccioli, T.: A Modular and Automated Annotation Platform for Handwritings: Evaluation on Under-resourced Languages (2021), accepted for ICDAR 2021 main conference