Enhancing Arabic Maghribi Handwritten Text Recognition with RASAM 2: A Comprehensive Dataset and Benchmarking[*]

Chahan Vidal-Gorène[1,2,*,†], Clément Salah[3,†], Noëmie Lucas[4,†], Aliénor Decours-Perez[2,†] and Antoine Perrier[5]

[1]*École Nationale des chartes-Université PSL, Centre Jean-Mabillon, France*

[2]*Calfa, France*

[3]*Sorbonne Université (UMR 8167), Université de Lausanne (IHAR), France, Suisse*

[4]*University of Edinburgh, Scotland*

[5]*CNRS, Centre Jacques Berque, Maroc*

Abstract

Recent advancements in handwritten text recognition (HTR) for historical documents have demonstrated high performance on cursive Arabic scripts, achieving accuracy comparable to Latin scripts. The initial RASAM dataset, focused on three Arabic Maghribi manuscripts, facilitated rapid coverage of new documents via fine-tuning. However, HTR application for Arabic scripts remains constrained due to the vast diversity in spellings, ambiguities, and languages. To overcome these challenges, we present RASAM 2, an extended dataset with 3,750 lines from 15 manuscripts in the BULAC library, showcasing various hands, layouts, and texts in Arabic Maghribi script. RASAM 2 aims to establish a new benchmark for HTR model training for both Maghribi and Oriental scripts, covering text recognition and layout analysis. Preliminary experiments using a word-based CRNN approach indicate significant model versatility, with a nearly 40% reduction in Character Error Rate (CER) across new in-domain and out-of-domain manuscripts.

Keywords

dataset, Arabic scripts, handwritten text recognition, historical manuscripts

1. Introduction

In 2020, the Recognition and Analysis of Scripts in Arabic Maghrebi (RASAM) dataset was introduced to analyze and recognize handwritten Arabic documents, specifically focusing on Arabic Maghribi script manuscripts. This dataset demonstrated the feasibility of applying Handwritten Text Recognition (HTR) to Arabic Maghribi scripts, aiming for error rates comparable to

other non-Latin scripts. The initial dataset, RASAM 1, included 300 images from three Bibio-thèque des langues et civilisations (BULAC) manuscripts copied between 1734 and 1875, achieving promising results with an in-domain Character Error Rate (CER) of 4.8%.

However, the limited scope of RASAM 1 restricted its effectiveness in recognizing out-of-domain manuscripts, even those with similar contemporary scripts and themes (see Table 1). To overcome these limitations, we introduce RASAM 2, an expanded dataset comprising 3,750 lines from fifteen manuscripts, encompassing a broader range of themes and handwriting styles. RASAM 2 aims to provide a comprehensive reference for training HTR models for Arabic scripts, enhancing their robustness and applicability across diverse Arabic Maghribi and Oriental texts. This paper presents the technical details of RASAM 2, its composition, and the initial results of using a new word-based Convolutional Recurrent Neural Network (CRNN) approach, which shows significant improvement in model versatility and a substantial reduction in CER for both in-domain and out-of-domain manuscripts.

Table 1

Common limitations encountered with RASAM 1 and state-of-the-art HTR models of Arabic

BULAC.MS.ARA.1978	GT	RASAM 1 prediction
	يكون	يكوق

Commentary: The *nūn* is mistaken for a *qāf* (in both cases, a single dot subscribed).

BULAC.MS.ARA.1978	GT	RASAM 1 prediction
	البيع	الفيع

Commentary: The *fā* is confused with a *bā* (in both cases, a single point is subscribed).

BULAC.MS.ARA.1978	GT	RASAM 1 prediction
	حَاضر	حَاضو

Commentary: The *rā* is confused with a *wāw* (more or less open and long final).

BULAC.MS.ARA.1978	GT	RASAM 1 prediction
	بعدد	هدح

Commentary: The pair of letters *bā* and *'ayn* were confused with an *hā* (the subscript point of the *bā* was not spotted). The final *dāl* is confused with a *ḥā*, they may have a close ending.

BULAC.MS.ARA.1978	GT	RASAM 1 prediction
	الرمَان	الومَان

Commentary: The *rā* of *rummān* (pomegranates) became a *wāw*, both often very close realisations - a possible example of a food word unknown by the model.

BULAC.MS.ARA.1978	GT	RASAM 1 prediction
	الجوَاهر	الحوَاقصر

Commentary: The first subscribed point is misunderstood and the *ǧīm* of *ǧawāhir* (jewels or gems) is confused with an *ḥā*. The unusually wide realisation of the *ḥā* is mistaken for a *qāf* (the dot on the line below is mistakenly equated with this line) followed by a *ṣād*. The *rā* is well understood.

2. State-of-the-art datasets for Arabic scripts

The study of documents in Arabic constitutes a separate field within the handwritten text recognition and document analysis questions more generally, owing to the great diversity and variability they encompass, hence the workshops dedicated to this specific issue held at the last ICDAR and ICFHR conferences. The latest developments in HTR for Arabic have however demonstrated that the use of dedicated CRNN enables to overcome the issue of text recognition for these scripts, with CER below 5%, even below 3% in specific cases, with few training data [16, 9]. At this stage, these specialized models exceed the performance achieved by Transformers for Arabic, the latest results on Al-Soudani Maghrebi script achieving an average of 10% CER with large dataset[12]. The text detection is also effective on Arabic documents, for instance, the use of FCN [8] allows for a good text-line detection. For the semantic classification of contents, using a non-specialized U-net [15] outperforms the FCN results, which is notably facing problems in differentiating two close text regions of the same type, unlike U-net. Several open-ended questions remain, such as the processing of very cursive scripts, the issue of transcription and the ambiguity of diacritics, or the reading of abbreviations.

In recent years, numerous datasets have emerged in an attempt to overcome these different tasks. In the instance of non-historical documents, the IFN/ENIT dataset [14], focused on modern scripts and produced in a very restricted context, is an important point of reference, notably used for the automatic generation of handwritten lines [5]. Not designed for HTR purposes, the KHATT dataset offers a dataset in modern scripts with 1,000 different copyists[11], mainly intended for writer identification, as well as the QUWI and LAMIS-MSHD datasets[10, 4].

In the instance of historical documents, very specialized datasets exist, such as WAHD [1], dedicated to writer identification, or KERTAS [2], dedicated to manuscript dating. There exist datasets non-specialized on a specific Arabic script, such as HADARA80P [13] and VML-HD [7], notably for RASM2018 [3] comprised of scientific manuscripts from the Qatar Digital Library, or BADAM [8] focused on line detection in Arabic documents, particularly complex ones. More recently, the RASAM 1 dataset [16] targets Arabic Maghribi scripts, in contrast to RASM and BADAM, which focus on oriental scripts. It offers typical layouts and hands as representative of the common Maghribi production, selected for the purpose of quickly developing HTR models operable for both research and production. The dataset has since been extended within the scope of the TARIMA project, with 120 pages manually transcribed from 28 various Arabic Maghribi sources, including lithographs.[1] The dataset has been designed for fine-tuning tasks from RASAM 1. For the oriental scripts, we can also mention the Iskandar dataset from the Alexander Hackathon, focusing on 5 manuscripts of the Alexander romance in Middle Arabic.[2]

Together, these datasets are already covering a vast part of the production of documents in Arabic scripts (subject to their compatibility, see Table 2). Although the proof of concept is successful for text recognition, the challenge today is to increase the versatility of existing models by providing a greater variety of fully annotated and transcribed documents.

---

[1]https://github.com/calfa-co/tarima
[2]https://gitlab.huma-num.fr/lipa/iskandar

Table 2

Summary of the main existing datasets for Arabic historical documents. Different levels of annotation are offered, often partial, thus limiting data compatibility.

Dataset	Images	Focus	Annotation	Baseline	Region	Text	Format
Specialized datasets							
WAHD	43,976	Writer identification	-	-	-	-	NC
KERTAS	2,502	Manuscript dating	-	-	-	-	XML
HADARA80p	80	Word spotting	-	-	-	-	XML
VML-HD	680	Word spotting	-	-	-	-	Hadara XML
Datasets for Page Layout Analysis and HTR							
RASM2018	100	General	Full	yes	yes	yes	pageXML
BADAM	400	Layout	Partial	yes	no	no	pageXML
RASAM 1	300	Maghribi scripts	Full	yes	yes	yes	pageXML
TARIMA	120	Maghribi scripts	Full	yes	yes	yes	pageXML
Iskandar	297	Oriental scripts	Full	yes	yes	yes	pageXML

3. Dataset composition

3.1. Quantitative description

Summary: RASAM 2 dataset comprises 250 images from 15 different manuscripts. 3,750 lines in total have been transcribed, 250 lines by manuscript on average, regardless of the type (main text or marginal notes). It entails 5,702 annotated lines in total and focuses on Arabic Maghribi manuscripts (see Table 5 in appendix for the complete list of manuscripts). Its purpose is to extend the variety of cases encountered in RASAM 1, in order to provide a robust training basis for documents in Arabic scripts.

- **Dataset availability** (v.1.0): https://github.com/calfa-co/rasam-dataset.
- **License**: Apache2.0
- **Data format**: pageXML with Text regions and lines
- **Annotation tool**: Calfa Vision[3] [15]
- **Ontology for annotation**: SegmOnto [6]
- **Transcription guidelines**: Same as RASAM 1 (no missing *hamza* or diacritics added)

Methodology for data creation: The images have been randomly selected in the manuscripts to constitute a representative sample of the production, of the states of conservation, and of the handwriting quality. The images have been pre-annotated with the baseline and text region detection models trained on RASAM 1 and available within the project type "Arabic Manuscript (default)" on the annotation platform. Afterwards, the predictions have been manually checked by the participants during the hackathons. Transcription guidelines follow RASAM 1 recommendations [16].

---

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

The dataset holds 522,371 characters (divided in 54 classes) for a total of 93,855 words (divided in 22,027 classes). The ḍammatan and إٕ classes in particular are under-represented and are likely to be less encountered, and so less recognized in a character-based approach (see below Section 4). The words *waw* (و), *min/man* (من) and *fī* (في) are the most represented in the dataset, with 4,398; 2,246 and 2,189 occurrences respectively, a contrario the words *al-akhdūd* (الأخدود), *qaṭām* (قطام) and *laʿād* (لعَاد) are among the least represented (a single occurrence).

Table 3
Distribution of TextRegion types in RASAM 2 dataset (v1.0)

Manuscript	MainZone	MainZone: title	Margin TextZone	Margin TextZone: catchword	StampZone	TableZone
BULAC.MS.ARA.6	15	-	6	7	-	-
BULAC.MS.ARA.9	16	-	26	6	-	-
BULAC.MS.ARA.23	16	-	5	7	-	-
BULAC.MS.ARA.24	17	-	1	7	-	-
BULAC.MS.ARA.45b	16	-	46	7	-	-
BULAC.MS.ARA.65	13	-	6	6	-	1
BULAC.MS.ARA.1926	41	1	24	15	-	-
BULAC.MS.ARA.1936	20	-	41	8	-	-
BULAC.MS.ARA.1943	25	-	83	8	-	-
BULAC.MS.ARA.1944	35	-	43	13	2	-
BULAC.MS.ARA.1946	25	-	3	9	2	-
BULAC.MS.ARA.1947	18	1	28	7	-	-
BULAC.MS.ARA.1960	16	-	60	8	-	-
BULAC.MS.ARA.1982	25	1	9	16	-	-
BULAC.MS.ARA.1983	15	1	2	8	2	-
TOTAL	313	4	383	132	6	1

We retained four text regions and two annex regions for the semantic classification of contents:

- **MainZone**: the main text region of the document. This region can appear several times within a single page, when the text is segmented or in case of a multiple column layout;
- **MainZone:title**: text region located at the same level as the main text, for headings and stylized titles;
- **MarginTextZone**: marginal text region regardless of its location in the page;
- **MarginTextZone:catchword**: marginal text region corresponding to the catchwords, systematically under the main text region;

- **StampZone**: stamps present on the page;
- **TableZone**: region corresponding to a table.

A summary of the text regions distribution is given in Table 3.

3.2. Qualitative description

As outlined in the introduction, the aim of this new dataset is to enhance the versatility and robustness of RASAM 1 by training it on a wider variety of manuscripts in order to expand the base of its (1.) vocabulary, (2.) layouts and (3.) scripts. As a result, 15 manuscripts make up this new dataset.

Figure 1: Examples of complex layout. From left to right, first line: MS.ARA.6, MS.ARA.65, MS.ARA.1943, MS.ARA.1936; second line: MS.ARA.1947, MS.ARA.1926, MS.ARA.1960

(1.) Of the fifteen new manuscripts, five (1/3 of the corpus) have themes and/or vocabulary related to the first dataset. Like MS.ARA.1977 (RASAM 1), MS.ARA.1944 (RASAM 2) belongs to the historical genre; and like MS.ARA.609 (RASAM 1), the manuscripts MS.ARA.1936, 1943, 1960 and 1983 deal with Islamic law – with the difference that, on the one hand, the legal issues are not identical, which means that a new vocabulary has to be learned, and that, on the other hand, MS.ARA.1936 also includes Berber written in Arabic. The other ten manuscripts of the new dataset (the remaining 2/3 of the corpus) cover new themes, not yet dealt with by RASAM 1. In detail, MS.ARA.1947 is a classical Arabic literature text, MS.ARA.1926 a collection of litanies, MS.ARA.23, 24, 45b and 1982 cover vocabulary related to Arabic grammar and linguistics. MS.ARA.6, 9, 65 and 1946 consist of collections on various topics ranging from Islamic jurisprudence to Arabic grammar, including private correspondence and exegesis of Qur'anic verses. In addition, several manuscripts show significant variations in handwriting, particularly for the latter collections.

(2.) From the layout perspective, the RASAM 1 dataset already covered complex layouts: MS.ARA.609 integrated many tables within the body of the text and MS.ARA.1977 recorded many lines of poetry which traditionally are offset from the main text [16]. The RASAM 2 dataset intends to enhance the capabilities of the model in handling complex layouts. In detail (see Figure 1), the RASAM 2 dataset reinforces its capabilities in the treatment of poetry verses (MS.ARA.6), tables (MS.ARA.65) and marginal comments, whether they are aligned with the main text as in MS.ARA.1943, or rounded, or even inverted as in MS.ARA.1936. Moreover, the RASAM 2 dataset develops new skills, in particular in the identification of interlinear comments (MS.ARA.1947) or particularly stylised titles (MS.ARA.1926) as well as in the processing of more complex page layouts, notably with the presence of gap texts (MS.ARA.1960).
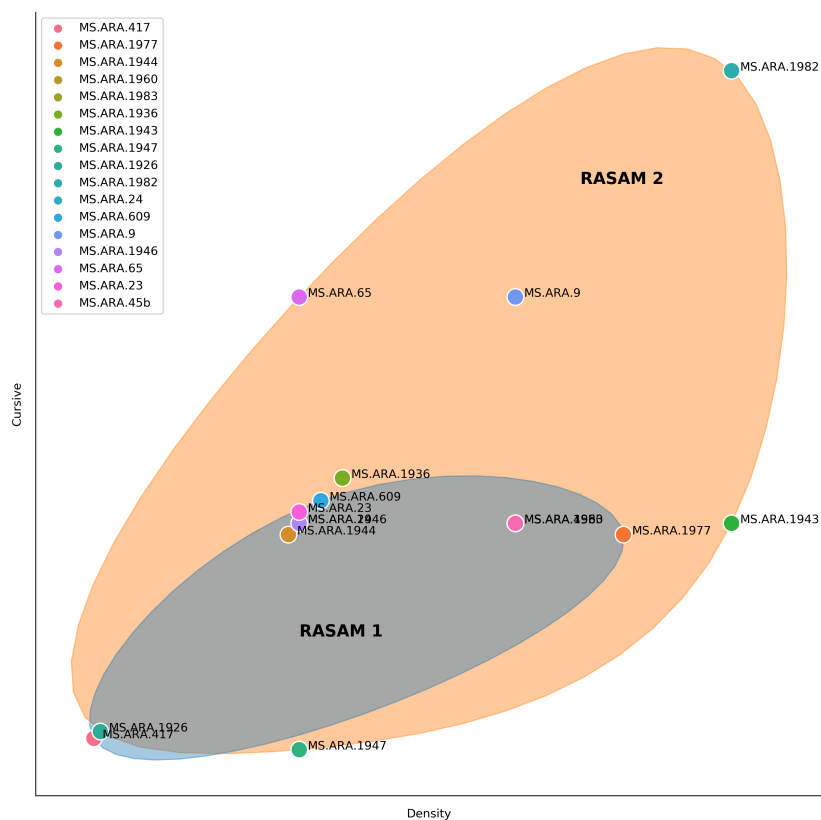
Figure 2: Representativity of the cursive and dense characteristics of RASAM 2 scripts in comparison with RASAM 1
We gave each manuscript a score out of 5 to characterize the cursiveness of the writing as well as the density of the text.

(3.) From a strictly palaeographic point of view, the RASAM 2 dataset intends to deal with a broader variety of hands. The emphasis has been placed on three points in particular. (a.) Firstly, particular interest has been given to the use of colors within these different manuscripts. Some recent experiments conducted on the basis of RASAM 1 show that the use of colors largely hinders the models' good recognition of characters [9]. Therefore, many manuscripts in the

RASAM 2 corpus aim at providing the model with many color realizations (see MS.ARA.1926 and MS.ARA.6 supra, where blue, green, red and yellow are used in particular). (b.) Secondly, RASAM 2 intends to be able to handle different text densities. RASAM 1 was indeed based on only 3 manuscripts which, although different from the density aspect [16], did not cover the multiple realizations of Arabic manuscripts in Arabic Maghribi scripts. In order to fill this gap, RASAM 2 is built on a broad continuum in terms of density from very airy manuscripts — such as MS.ARA.1926 with less than ten lines per page and less than ten words per line — to extremely dense manuscripts — such as MS.ARA.1982 with more than forty lines per page and slightly less than twenty words per line, or MS.ARA.1943 with thirty-five lines per page and more than twenty words per line. (c.) Finally, RASAM 2 covers a wider range of Arabic Maghribi scripts. The model is thus built from very careful and stylized, almost calligraphic hands following the example of MS.ARA.1926 (see below 6) or hands that are characterized by a wide amplitude of their final tails — see in particular the realization of the final *lām* in the word *qāla* of MS.ARA.6, 1926, 1946, 1947 (see Table 6 in appendix). Conversely, RASAM 2 also includes very cursive and crowded scripts, as is the case for MS.ARA.1943, 1982. In sum, and as schematically represented in Figure 2, RASAM 2 covers a wider reality of Arabic Maghribi hands. It leads to a pre-generic model for the treatment of Arabic Maghribi scripts, far exceeding the possibilities offered by RASAM 1, which was still only a proof of concept until then.

4. HTR of Arabic versatility experiments

4.1. Methodology

The latest developments in HTR for handwritten documents in Arabic scripts have shown that operating a word-based CRNN (where every word is considered as a different class to identify) outperforms a basic character-based CRNN (where each character is considered as a different class to identify) on documents with a steady lexicon (both in learning time and CER) [9]. This approach, despite being dependent on the targeted lexicon, relies on recognizing a word in context, which appears a more robust approach for cursive Arabic scripts) [9]. We hold onto this approach, which is a variation of the one implemented for RASAM [16]. Some under-represented word classes are in a few-shot learning situation. In this case, the word-based approach is based on context for predictions, and failing that relies on character recognition.

Lucas et al. have notably demonstrated that a fine-tuning strategy limited to 10 images (160 transcribed lines on average) for the Arabic Maghribi scripts, on the basis of a RASAM-trained model is sufficient to reach a CER below 10% and to shorten the transcription work [9].

We are taking this fine-tuning approach from the RASAM model and testing it on two samples: one in-domain sample, derived from RASAM 1 and RASAM 2, and one out-of-domain sample derived from manuscripts from Lucas et al. [9] (see Figure 3). The latter dataset is twice out-of-domain, with new scripts and new lexicon. We compare this new model with the one strictly trained on RASAM 1 (see Figure 4 and Table 4).
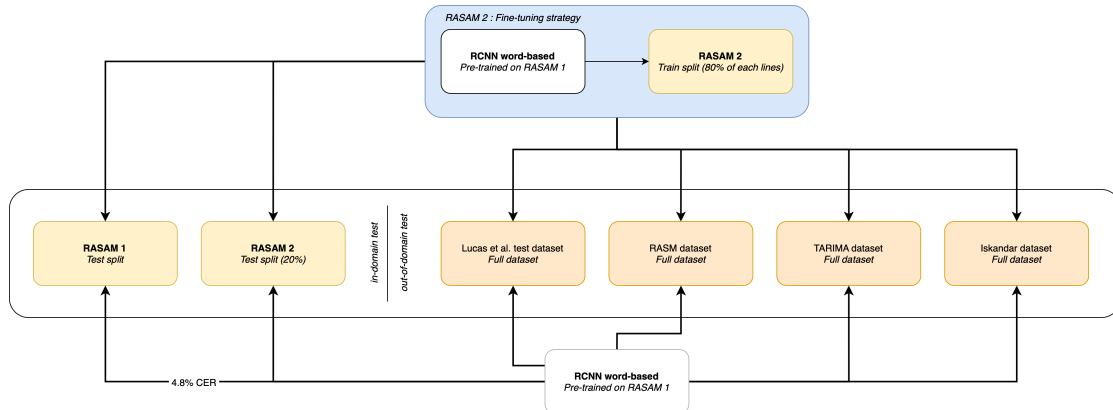
Figure 3: Experiments conducted on the new dataset and comparison with the RASAM 1 and RASAM 2 models

4.2. Results

Table 4 displays the average CER achieved by models trained on RASAM 1 and RASAM 2 in the in-domain and out-of-domain samples. Although RASAM 1 model evaluated on its original sample remains more efficient, owing to its high specialization, RASAM 2 model reaches a CER five times smaller on RASAM 2, and almost halves the CER obtained on out-of-domain documents. The lexical and visual diversity provided by RASAM 2, although relatively modest, allows the model to achieve an average CER comparable to state-of-the-art results obtained for Latin scripts, which benefit from significantly larger datasets (e.g., the CATMuS medieval dataset, which includes about 5 million characters).

4.2.1. Out-of-domain results (Maghribi scripts)

In out-of-domain documents but belonging to the same family of scripts as RASAM 1 and 2, such as the Arabic Maghribi scripts, RASAM 2 demonstrates notable efficiency, as evidenced in its application to TARIMA. Particularly noteworthy is its performance on Oriental scripts (RASM and Iskandar), where RASAM 2 not only outperforms RASAM 1 but also achieves significantly lower average CER scores (20.34 for RASM and 16.73 for Iskandar). These improved results not only enhance accuracy but also facilitate faster processing with minimal data requirements.

Table 4
Comparison of CER achieved on in-domain and out-of-domain samples. The outcome of RASAM 1 on RASAM 1 is drawn from the original article.

	in-domain test		out-of-domain test				
	RASAM 1	RASAM 2	RASAM 2	Lucas et al.	RASM	TARIMA	Iskandar
RASAM 1	4.8*	–	30.91	25.75	42.02	26.81	46.91
RASAM 2	5.50	6.79	–	16.38	20.34	9.70	16.73

Besides the versatility of RASAM 2 model, Figure 4 also shows its robustness with a very consistent CER per page and very little dispersion as in the case of RASAM 1. It is particularly visible on RASAM 2 dataset for which RASAM 1 model (out-of-domain test) reaches a CER between 11.67% (on the manuscript BULAC.MS.ARA.1982) and 48.80% (on the manuscript BULAC.MS.ARA.9).
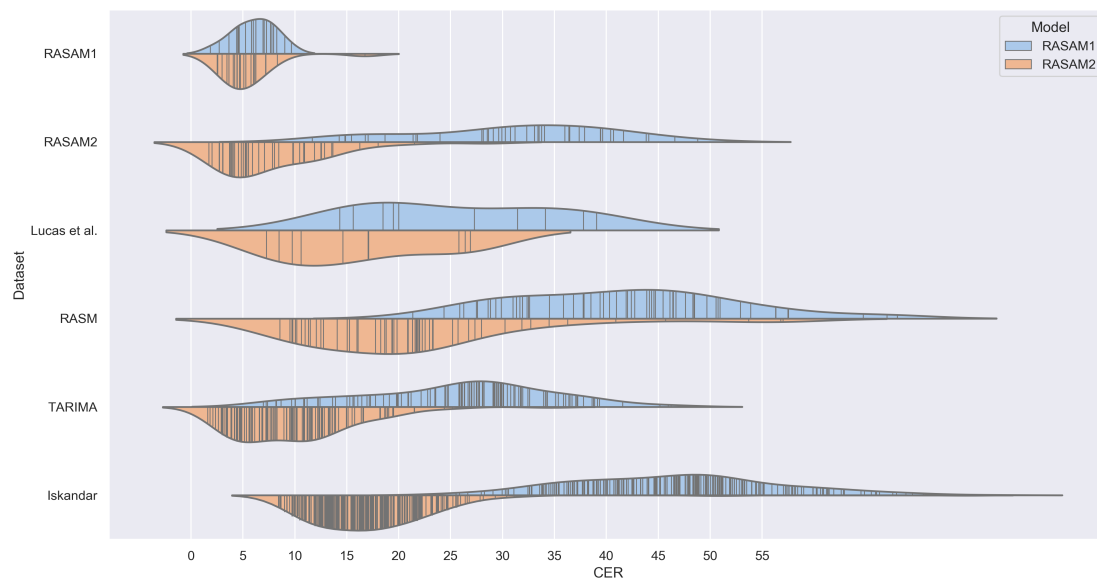
Figure 4: Distribution of the achieved CER on the three datasets: RASAM 1 (blue) and RASAM 2 (orange)

A contrario, the CER of RASAM 2 model ranges between 1.71% and 28.47% in an in-domain instance, and between 7.26% and 26.88% in an out-of-domain instance. The extreme values are therefore practically twice as small as those for RASAM 1. Thus, there remain pages for which our new model does not immediately succeed in producing workable outcome, for these pages, it will then be necessary to adopt a fine-tuning strategy, which should be fast.[4] The median observed in Figure 5 is 27.97% for RASAM 1 for out-of-domain documents, and is reduced to 15.83% for RASAM 2, hence a 42% decrease in the error rate.

Figure 5 presents the average CER for each manuscript. In the in-domain instance, several manuscripts have a CER of less than 5%: this is the case for the manuscripts BU-LAC.MS.ARA.1943 (3.43%), BULAC MS ARA 1977 (4.91%), BULAC. MS.ARA.1982 (3.26%), BU-LAC.MS.ARA.1983 (3.58%), and BULAC MS ARA 45b (3.20%). The BULAC.MS.ARA.1936 and BULAC.MS.ARA.1947 manuscripts, even if they largely benefit from the new model, retain a high CER, higher than 15% and up to 16.25% for the BULAC.MS.ARA .1936 (compared with the 46.47% CER achieved with RASAM 1, but which is out-of-domain).

---

[4]In Lucas et al., a CER of 3.23% was reached with a different split and a slightly redesigned architecture, based on a meta-word-based approach (in the context of a specialized in-domain model). It also shows in particular that for the manuscript BULAC.MS.ARA.1957, the initial CER of 30.46% (RASAM 1) is reduced to 21.8% after a fine-tuning of only 20 lines. Applied to the same manuscript (see Figure 5), RASAM 2 model obtains an initial CER of 25.5%[9].

In the out-of-domain instance, the gap between the results of RASAM 1 and RASAM 2 is narrower. If the manuscripts BULAC.MS.ARA.1922 (31.44% vs 26.38%) and BULAC.MS.ARA.1957 (35.95% vs 26.33%) retain a very high CER, the manuscripts BULAC.MS.ARA.1944 and BULAC.MS.ARA.1929 achieve a CER of 7.67% and 10.16%, better than the CER obtained in-domain for the manuscripts previously cited.

Despite the diversity of the TARIMA corpus, with both manuscripts and lithographs, the results remain very good. This is due to the proximity between the RASAM 1 & 2 dataset and the palaeographic characteristics of the TARIMA corpus, all of which are in Maghribi script.

4.2.2. Out-of-domain results (Oriental scripts)

Out-of-domain results (Oriental scripts) RASAM 2 also demonstrates significantly enhanced efficiency when applied to Oriental manuscripts, as illustrated by its performance with RASM and Iskandar. Its versatility is particularly evident in Iskandar, where the CER remains below 30%, with an average CER ranging between 8% and 20% (Fig. 4 and 5). Except for one manuscript (MS_Orient_A_02393), all the CER remain below 20% with RASAM 2. While RASM results exhibit some dispersion (albeit less than with RASAM 1), RASAM 2's perfor-
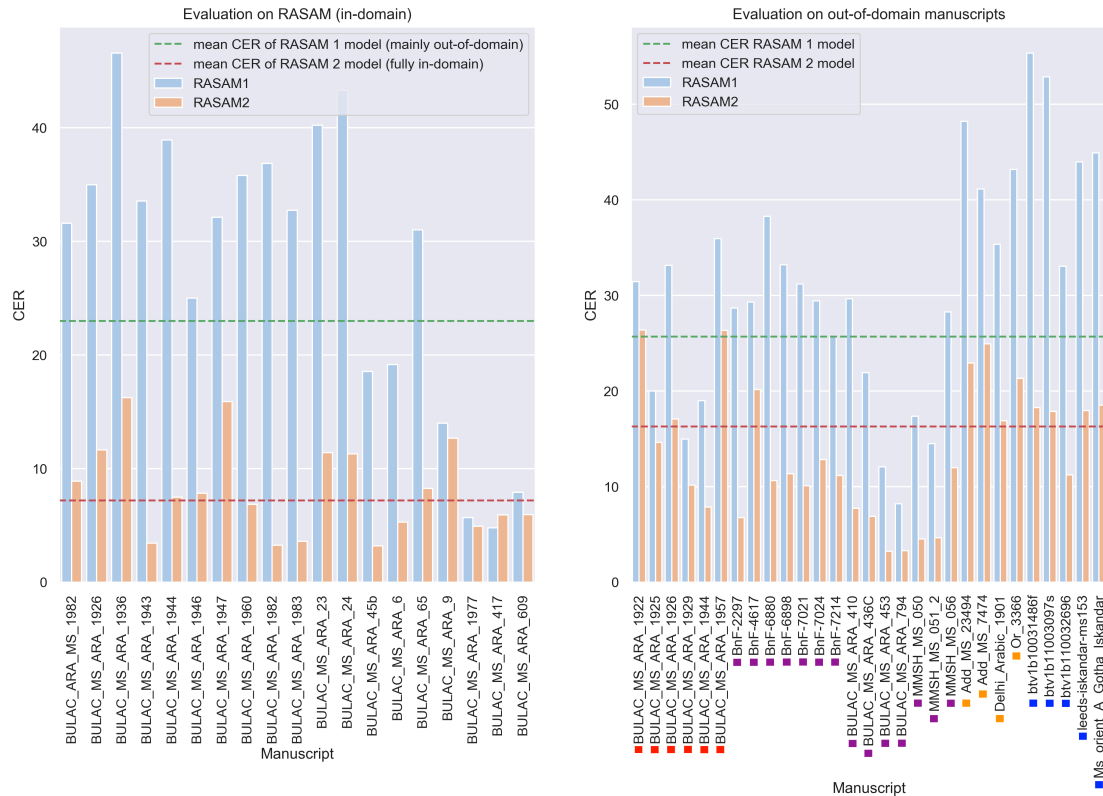
Figure 5: Distribution of CERs obtained by RASAM 1 (blue) and RASAM 2 (orange) for each in-domain and out-of-domain manuscript. For the out-of-domain evaluation, red dots refer to manuscripts from Lucas et al., purple dots to those from Tarima, orange dots from RASAM, and blue dots from Iskandar.

mance varies across the four manuscripts comprising the RASM dataset. Its highest result is observed in Dehli_Arabic_1901 (slightly above 16%), but none exceed 25%. The disparity in out-of-domain results between RASM and Iskandar likely arises from the difference in dataset adherence to RASAM guidelines. While Iskandar follows the RASAM guidelines, the RASM dataset diverges from them, which may explain the observed gap in CER results. For example, when the scribe omitted expected diacritics on certain letters, the transcriber left the letter without them, whereas the RASAM guidelines would have added the diacritics where necessary. This suggests that with minimal fine-tuning, RASAM 2 could readily adapt to various manuscripts, regardless of their script families.

4.3. Qualitative interpretation

RASAM 2 sets a new standard for the recognition of Arabic Maghribi scripts. Figure 5 shows that it nevertheless produces many more errors than the average on four in-domain and out-of-domain manuscripts, leading to an increase in the CER. Observation of the manuscripts (see Figure 6) reveals several situations where the CER decreases naturally.

Manuscript with vowel signs and numerous interlinear notes: This is the case of the manuscripts BULAC.MS.ARA.1936 and BULAC.MS.ARA.1957 for which we observe an important vocalization which is rarely present in these manuscripts. It leads, at this stage, to a greater ambiguity of the forms to be recognized, but is however not insurmountable: a specialized approach from RASAM shows for example that 160 lines are enough with a word-based approach to reach a CER of 10.41% for the manuscript BULAC.MS.ARA.1957 [9].

Variation in line color: This is a phenomenon already observed in RASAM 1 [16], with an over-representation of colored lines among lines with high CER. The MS.ARA.1947, which alternates blue and red lines (marginally present in training) is therefore penalized. Its CER drops to 6.56% without these lines.

Figure 6: Examples of complex layout. From left to right: BULAC.MS.ARA.1936 (RASAM 2 dataset, in-domain), BULAC.MS.ARA.1947 (RASAM 2 dataset, in-domain), BULAC.MS.ARA.1922 (Lucas et al., out-of-domain) and BULAC.MS.ARA.1957 (Lucas et al., out-of-domain)

5. Conclusion

In conclusion, the RASAM 2 dataset offers a high representativeness of Arabic Maghribi scripts. The word-based model trained on this dataset obtains very high in-domain and out-of-domain accuracies, achieving a 40-point CER reduction in all scenarios, which ensures an important coverage of Arabic Maghribi manuscript traditions. The dataset also demonstrates its versatility and can be easily fine-tuned on a new target, including Oriental scripts and new varieties of Arabic (Middle Arabic, Berber written in Arabic). In the future, we will study this transfer of RASAM models to other types of Arabic scripts, in particular Oriental ones. Additionally, we plan to conduct experiments using transformer-based models, as the critical mass of data for Arabic has now been reached, thanks to the RASAM team and all datasets produced within this scope. More generally, the datasets created in recent years around the RASAM team (TARIMA, Iskandar) have made it possible to create a set of open data decisive for the HTR of Arabic scripts.

Acknowledgments

References

[1] A. Abdelhaleem, A. Droby, A. Asi, M. Kassis, R. Al Asam, and J. El-sanaa. "Wahd: a database for writer identification of arabic historical documents". In: *2017 1st International workshop on arabic script analysis and recognition (ASAR)*. Ieee. 2017, pp. 64–68.

[2] K. Adam, A. Baig, S. Al-Maadeed, A. Bouridane, and S. El-Menshawy. "KERTAS: dataset for automatic dating of ancient Arabic manuscripts". In: *International Journal on Document Analysis and Recognition (IJDAR)* 21 (2018), pp. 283–290.

[3] C. Clausner, A. Antonacopoulos, N. Mcgregor, and D. Wilson-Nunn. "Icfhr 2018 competition on recognition of historical arabic scientific manuscripts–rasm2018". In: *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. Ieee. 2018, pp. 471–476.

[4] C. Djeddi, A. Gattal, L. Souici-Meslati, I. Siddiqi, Y. Chibani, and H. El Abed. "LAMIS-MSHD: A Multi-script Offline Handwriting Database". In: *2014 14th International Conference on Frontiers in Handwriting Recognition*. 2014, pp. 93–97. DOI: 10.1109/icfhr.2014.23.

[5] M. Eltay, A. Zidouri, I. Ahmad, and Y. Elarian. "Generative adversarial network based adaptive data augmentation for handwritten Arabic text recognition". In: *PeerJ Computer Science* 8 (2022), e861.

[6]   S. Gabay, J.-B. Camps, A. Pinche, and C. Jahan. "SegmOnto: common vocabulary and practices for analysing the layout of manuscripts (and more)". In: *1st International Workshop on Computational Paleography (IWCP ICDAR 2021)*. 2021.

[7]   M. Kassis, A. Abdalhaleem, A. Droby, R. Alaasam, and J. El-Sana. "Vml-hd: The historical arabic documents dataset for recognition systems". In: *2017 1st international workshop on Arabic script analysis and recognition (ASAR)*. Ieee. 2017, pp. 11–14.

[8]   B. Kiessling, D. S. B. Ezra, and M. T. Miller. "BADAM: a public dataset for baseline detection in Arabic-script manuscripts". In: *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing*. 2019, pp. 13–18.

[9]   N. Lucas, C. Salah, and C. Vidal-Gorène. "New Results for the Text Recognition of Arabic Maghribi Manuscripts - Managing an Under-resourced Script". 2022.

[10]   S. A. Maadeed, W. Ayouby, A. Hassaïne, and J. M. Aljaam. "QUWI: An Arabic and English Handwriting Dataset for Offline Writer Identification". In: *2012 International Conference on Frontiers in Handwriting Recognition*. 2012, pp. 746–751. DOI: 10.1109/icfhr.2012.256.

[11]   S. A. Mahmoud, I. Ahmad, W. G. Al-Khatib, M. Alshayeb, M. Tanvir Parvez, V. Märgner, and G. A. Fink. "KHATT: An open Arabic offline handwritten text database". In: *Pattern Recognition* 47.3 (2014), pp. 1096–1112. DOI: 10.1016/j.patcog.2013.08.009.

[12]   S. A. Maouloud, M. O. M. Dyla, and C. Ba. "Transformer-based Model For Handwritten Recognition Arabic Words Al-soudani Maghrebi Script". In: *Journal of Theoretical and Applied Information Technology* 101.24 (2023).

[13]   W. Pantke, M. Dennhardt, D. Fecker, V. Märgner, and T. Fingscheidt. "An historical handwritten arabic dataset for segmentation-free word spotting-hadara80p". In: *2014 14th International Conference on Frontiers in Handwriting Recognition*. Ieee. 2014, pp. 15–20.

[14]   M. Pechwitz, S. S. Maddouri, V. Märgner, N. Ellouze, H. Amiri, et al. "IFN/ENIT-database of handwritten Arabic words". In: *Proc. of CIFED*. Vol. 2. Citeseer. 2002, pp. 127–136.

[15]   C. Vidal-Gorène, B. Dupin, A. Decours-Perez, and T. Riccioli. "A Modular and Automated Annotation Platform for Handwritings: Evaluation on Under-Resourced Languages". In: *Document Analysis and Recognition – ICDAR 2021*. Ed. by J. Lladós, D. Lopresti, and S. Uchida. Cham: Springer International Publishing, 2021, pp. 507–522.

[16]   C. Vidal-Gorène, N. Lucas, C. Salah, A. Decours-Perez, and B. Dupin. "RASAM – A Dataset for the Recognition and Analysis of Scripts in Arabic Maghrebi". In: *Document Analysis and Recognition – ICDAR 2021 Workshops*. Ed. by E. H. Barney Smith and U. Pal. Cham: Springer International Publishing, 2021, pp. 265–281. DOI: 10.1007/978-3-030-86198-8_19.

A. Data availability

- **RASAM** 1 and 2 datasets: https://github.com/calfa-co/rasam-dataset
- **TARIMA** dataset: https://github.com/calfa-co/tarima
- **Iskandar** dataset: https://gitlab.huma-num.fr/lipa/iskandar

B. Paleographical features of RASAM 2 dataset

Table 6: Paleographical differences between manuscripts of RASAM 1, RASAM 2, RASM and Iskandar

	في	علَى	غير	الذي	التي	هذَا	هذه	هذه	ذلك	اَللّه	قَال
RASAM 1 (Magribi script)											
MS.ARA.417											
MS.ARA.609											
MS.ARA.1977											
RASAM 2 (Magribi script)											
MS.ARA.6											
MS.ARA.9											
MS.ARA.23											
MS.ARA.24											
MS.ARA.45b											
MS.ARA.65											
MS.ARA.1926											
MS.ARA.1936											
MS.ARA.1943											
MS.ARA.1944											

MS.ARA.1946									
MS.ARA.1947									
MS.ARA.1960									
MS.ARA.1982									
MS.ARA.1983									

RASM (Oriental script)

Add.MS.7474									
Add.MS.23494									
Dehli.Arabic.1901									
OR.3366									

Iskandar (Oriental script)

Orient.A.0238X									
Btv1b10031486f									
Btv1b11003097s									
Btv1b110032696									
Leeds.ms153									

Table 5
Composition of RASAM 2 dataset

Manuscript	Pages	Text Lines	Baseline	Layout	Text density	Conservation	Genre
BULAC.MS.ARA.6 https://bina.bulac.fr/s/bina/ark:/73193/b6q5p6	14	250	336	simple	low	good	Miscellaneous
BULAC.MS.ARA.9 https://bina.bulac.fr/s/bina/ark:/73193/bqnm44	14	250	448	simple	low	good	Miscellaneous
BULAC.MS.ARA.23 https://bina.bulac.fr/s/bina/ark:/73193/bnvxrc	14	250	350	simple	low	good	Grammar
BULAC.MS.ARA.24 https://bina.bulac.fr/s/bina/ark:/73193/bsn0x6	14	250	322	simple	low	good	Grammar
BULAC.MS.ARA.45b https://bina.bulac.fr/s/bina/ark:/73193/brv21m	12	250	312	medium	low	good	Grammar
BULAC.MS.ARA.65 https://bina.bulac.fr/s/bina/ark:/73193/bnckkp	12	250	324	simple	low	good	Miscellaneous
BULAC.MS.ARA.1926 https://bina.bulac.fr/s/bina/ark:/73193/b7d8bx	34	250	306	simple	very low	damaged	Litany
BULAC.MS.ARA.1936 https://bina.bulac.fr/s/bina/ark:/73193/bcc306	16	250	352	medium	low	damaged	Law
BULAC.MS.ARA.1943 https://bina.bulac.fr/s/bina/ark:/73193/b5x6qh	20	250	820	complex	high	damaged	Law
BULAC.MS.ARA.1944 https://bina.bulac.fr/s/bina/ark:/73193/bj10cm	34	250	260	simple	low	good	History
BULAC.MS.ARA.1946 https://bina.bulac.fr/s/bina/ark:/73193/b8pkg9	14	250	280	simple	low	good	Miscellaneous
BULAC.MS.ARA.1947 https://bina.bulac.fr/s/bina/ark:/73193/bdfnnt	13	250	323	complex	high	good	Literature
BULAC.MS.ARA.1960 https://bina.bulac.fr/s/bina/ark:/73193/bstrdn	13	250	325	complex	high	good	Law
BULAC.MS.ARA.1982 https://bina.bulac.fr/s/bina/ark:/73193/bvmdrp	14	250	560	simple	high	good	Grammar
BULAC.MS.ARA.1983 https://bina.bulac.fr/s/bina/ark:/73193/bz8x88	12	250	384	simple	high	good	Law
TOTAL	250	3,750	5,702	-	-	-	