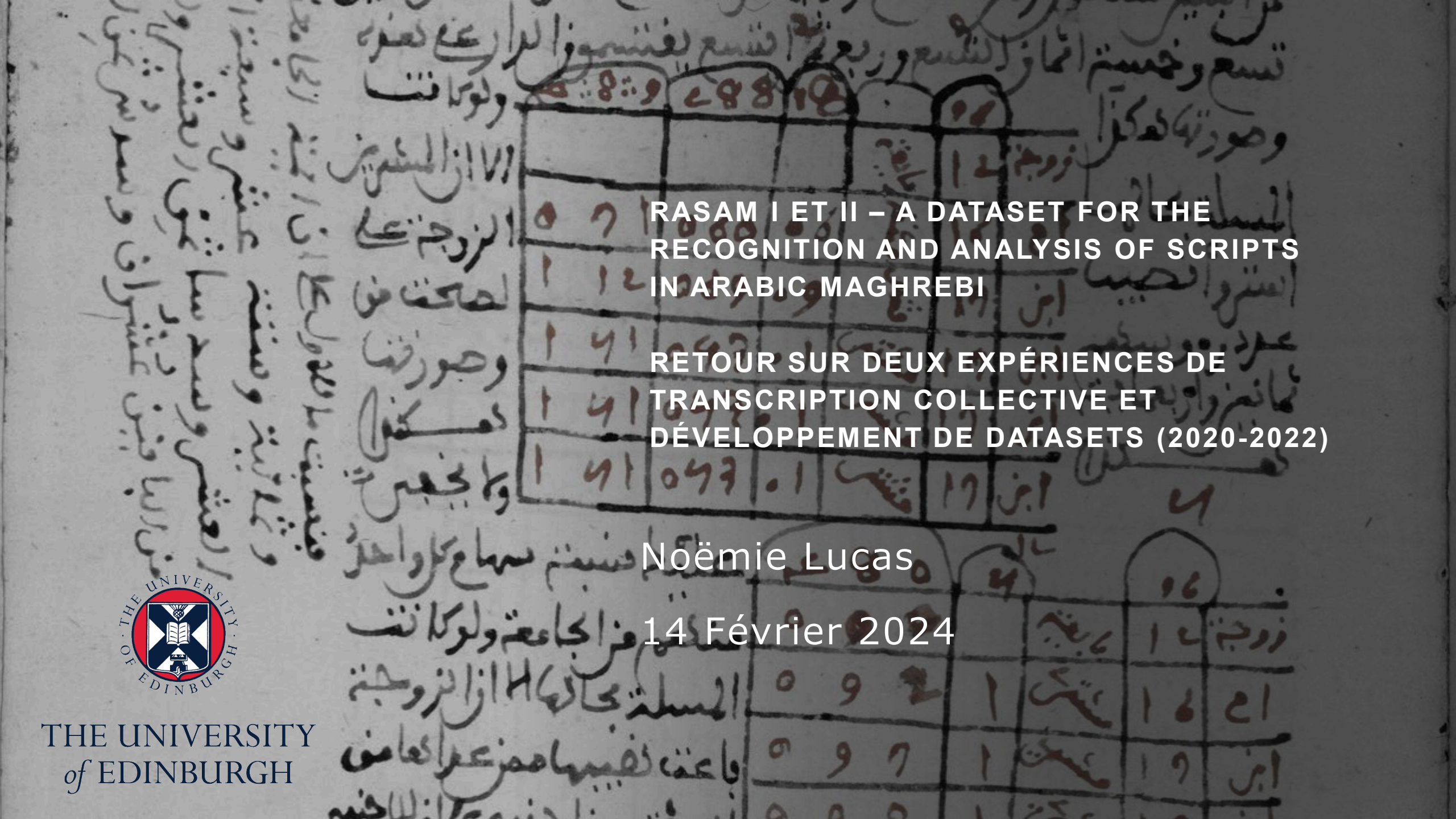


RASAM I ET II – A DATASET FOR THE RECOGNITION AND ANALYSIS OF SCRIPTS IN ARABIC MAGHREBI

RETOUR SUR DEUX EXPÉRIENCES DE TRANSCRIPTION COLLECTIVE ET DÉVELOPPEMENT DE DATASETS (2020-2022)

Noëmie Lucas

14 Février 2024

THE UNIVERSITY of EDINBURGH

Contexte

Moyen-Orient et Mondes Musulmans
Groupement d'Intérêt Scientifique

CALFA

BULAC
[הפריה] [大学] [γλώσσες] [ཤེས་རིག]
Bibliothèque universitaire des langues et civilisations

CONTEXTE

Objectifs de RASAM:

- Dimensions **technique** et **expérimentale**: HTR pour l'arabe manuscrit - preuve de concept
- Choix des écritures maghrébines - contexte de la **valorisation des collections patrimoniales** maghrébines de la **BULAC** et du travail en cours mené par le **GIS MOMM** sur le Maghreb et ses ressources
- **Formation** et Humanités numériques

INTELLIGENCE ARTIFICIELLE ET ÉCRITURE MAGHRÉBINE

Hackathon pour l'OCR de l'arabe manuscrit

Janvier-Avril 2021

Inalco CERMOM Centre de Recherches Moyen-Orient Méditerranée

cnrs

IRHT Institut de recherche et d'histoire des textes

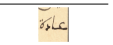
Moyen-Orient et Mondes Musulmans Groupement d'intérêt scientifique

BULAC

CALFA

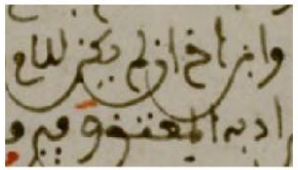
Plus d'informations

- Site de la Bulac
- Blog de Calfa
- Carnet philaranum

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 ب	 وركب	 لصاحب	 مغلوب
dāl *ḏāl*	**Isolated, median and final positions:** concave downstroke and final downward spur (*dāl kāfiyya*)	د د	 يزيد	 دنانير	 الديان
dāl *ḏāl*	**Final position:** marked semicircular descender, resembling the letters *rā’* and *zā’*	ر	 بعد	 فقد	 محمد
sīn *šīn* *ṣād* *ḍād* *qāf* *nū*	**Final position:** exaggerated semi-circular descenders, often described as ’swooping’ or ’plunging’, stretching below the following word	ن ن	 من ماله	 كان	 بن عبد
ṣād *ḍād* *ṭā’* *ẓā*	Oval or semi-circular body and lack of denticle	ط	 سبط	 الاصطلاح	 قسنطينة
‘ayn *ġayn*	**Initial position:** oversized curl	عـ	 عايشه (عاءشة)	 عم	 عادة
kāf	**Initial and median positions:** semicircle topped by a diagonal stroke	كـ كـ	 وكتب	 كذلك	 كثير
mīm	**Final and isolated positions:** long curved tail in two variants (concave or convex)	ل م	 ايام	 تقدم	 اسلام
hā’ *tā’* *marbūṭa*	**Isolated position:** drawn in the shape of a ’6’, sometimes inverted	٦ ه	 يذكره	 هاذه (هذه)	 اياه

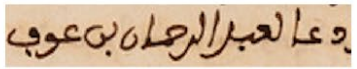
Graphie « maghrébine » / *ḫaṭṭ maġribī* : spécificités et enjeux des écritures arabes

1. Realization of the letter *nūn* (and *sīn*, *shīn*, *ṣad*, *ḍād*) in final position :

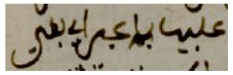


وابن اخ ان لم يكن للام

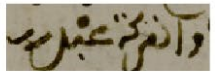
2. Confusion of some lettres (*dāl* / *rā’* ; *ḏāl* / *zayn*)



دعا لعبد الرحمان بن عوف

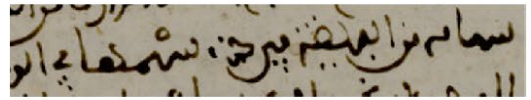


عليها بما عبر اي بقي



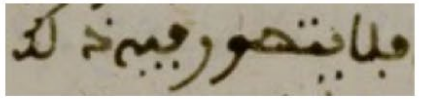
والتركة عبد

3. Same word made in different ways



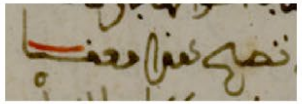
سهامه من الفريضة في جزء سهْمها في

4. Spacing between words

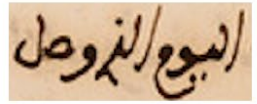


فلا يتصور فيه ذلك

5. Singular ligatures



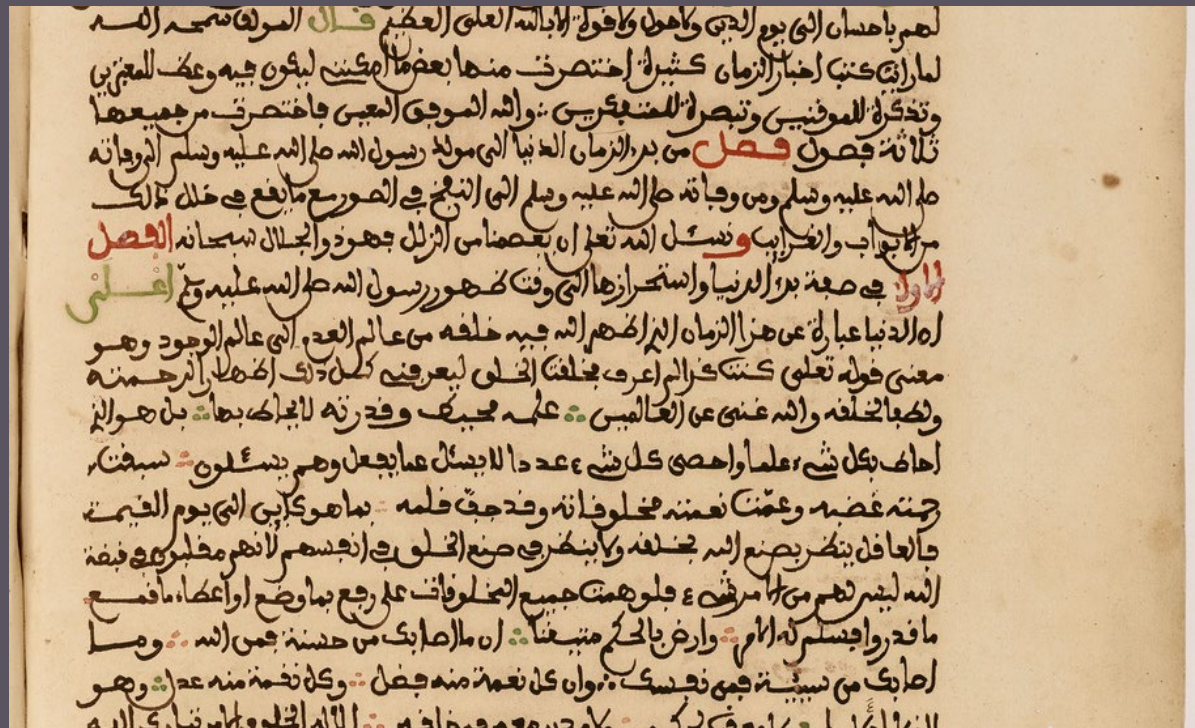
تصح هذا معنًا



اليوم الذي وصل

RASAM 1 (2020-2021)

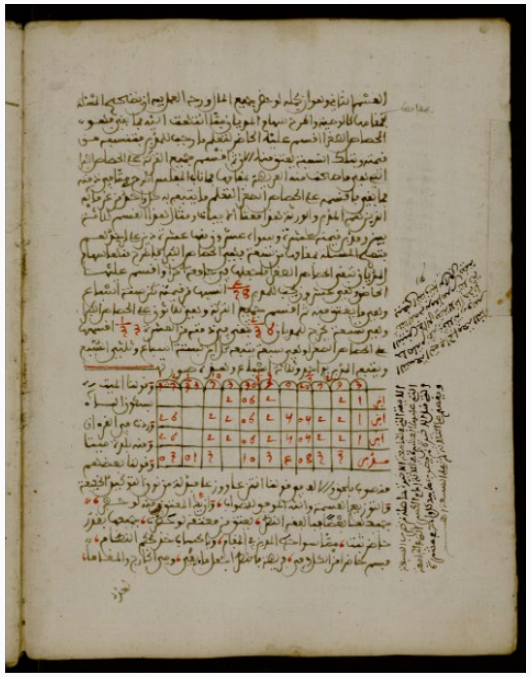
Séance de hackathon animée par Noëmie Lucas et Chahan Vidal-Gorène (Maxime Ruscio / BULAC).

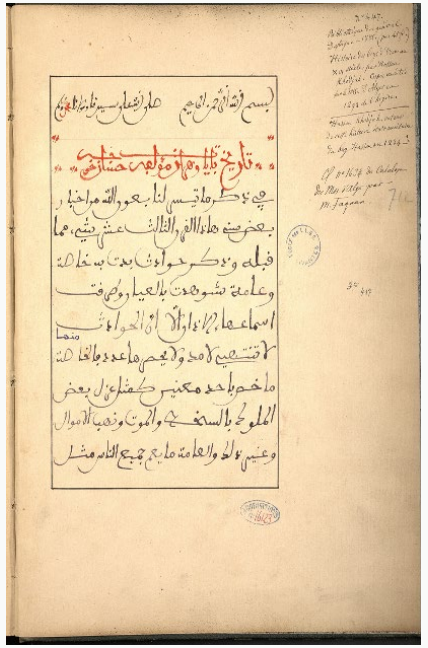


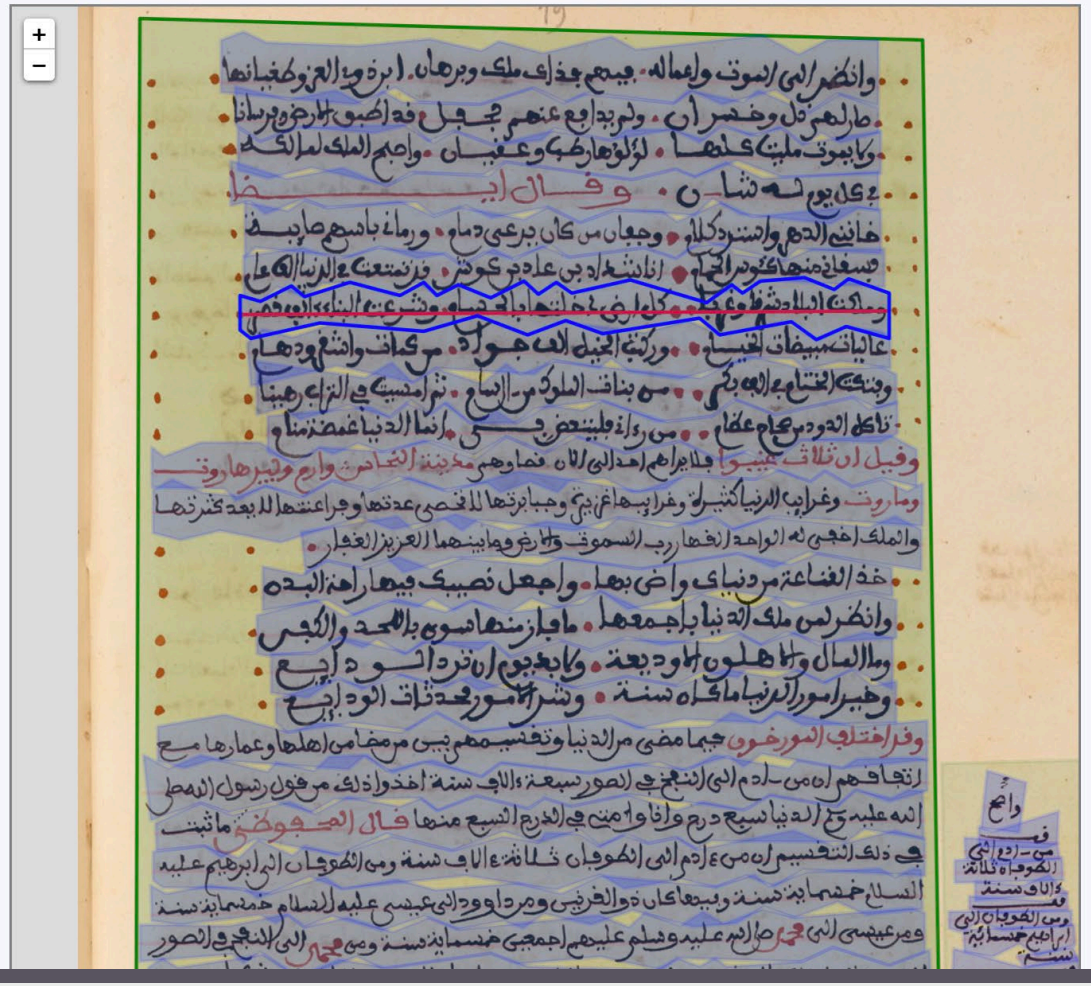
Corpus

- Styles d’écriture différents et courants
- Variété des themes limitée (histoire et droit)
- Variété des mises en page

- **Manuscrits issus des collections patrimoniales de la BULAC**
- **Deux (MS.ARA. 609 et MS.ARA.1977) étaient disponibles sur BINA https://bina.bulac.fr**
- **Le MS.ARA.417 a été numérisé à la demande**



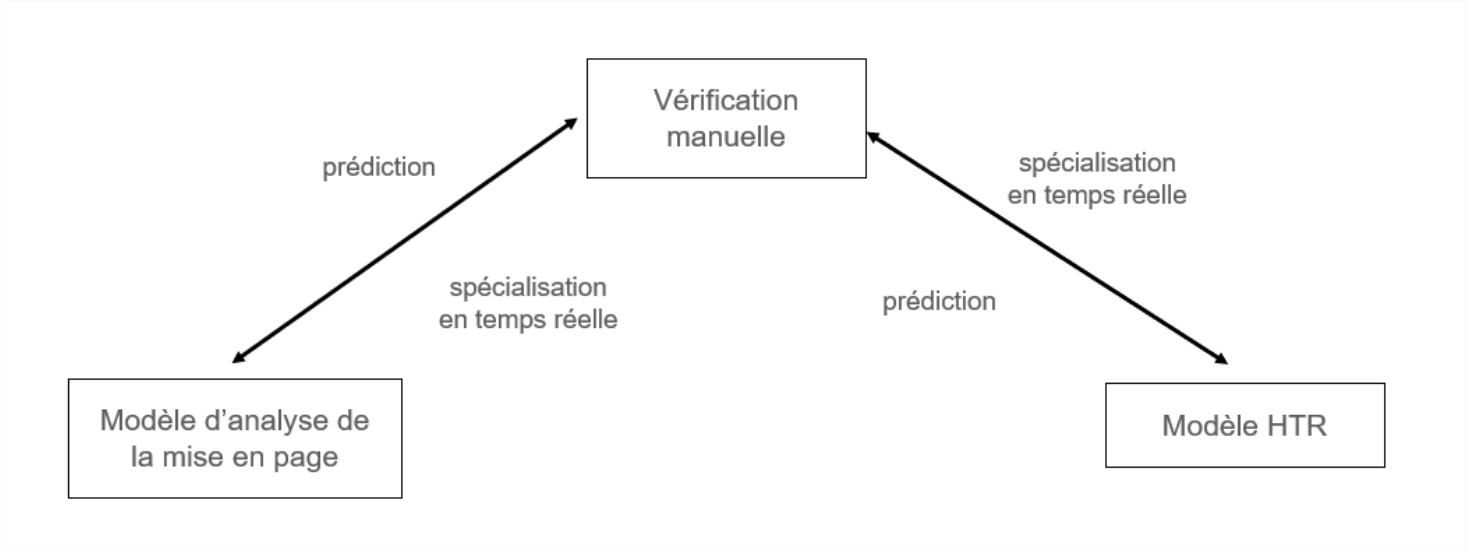




BULAC_MS_ARA_1977_0028

Save All

5	خانني الدهر واسترد كلام وجفان من كان يرعى دمام ورمانى باسهم صائبة	
6	فسقانى منها كئوس الحمام انا شداد بن عاد بن كوش قدد تمتعت في الدنيا الف عام	
7	وملكت البلاد شرقا وغربا كل ارض دخلتها بالحسام وشرعت البناء ءالف قصر	
8	عاليات مبيضات الخيام وركبت الخيل الف جواد من كمات واشقر ودهام	
9	وفتحت الختام في الف بكر من بنات الملوك من السام ثم امسيت في التراب رهينا	
10	تاكل الدود من صحاح عظام من رءانى فليتعض بى انما الدنيا غمضة منام	
11	يل ان ثلاث غيبوا فلا يراهم احد الى الان فصاو هى مدينة النحاس وارم وبير هاروت	
12	رائب الدنيا كثيرة وغرائبها غزيرة وجبابرتها لا تحصى عدتها وفراعنتها لا يعد كثرتها	
13	والملك اخفى له الواحد القهار رب السموت والارض وما بينهما العزيز الغفار	
14	خذ القناعة من دنياك واض بها واجعل نصيبك فيها راحة البده	
15	وانظر لمن ملك الدنيا باجمعها ما فاز منها سوه باللحد والكفر	

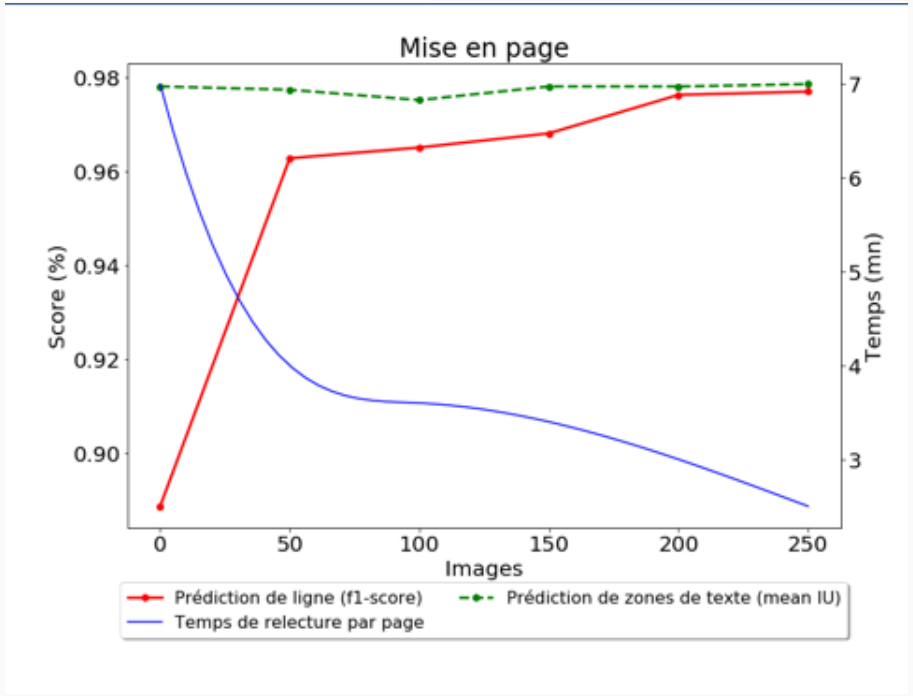


Déroulement du hackathon

Objectif : créer un dataset ouvert pour les écritures arabes maghribi représentatif et fonctionnel

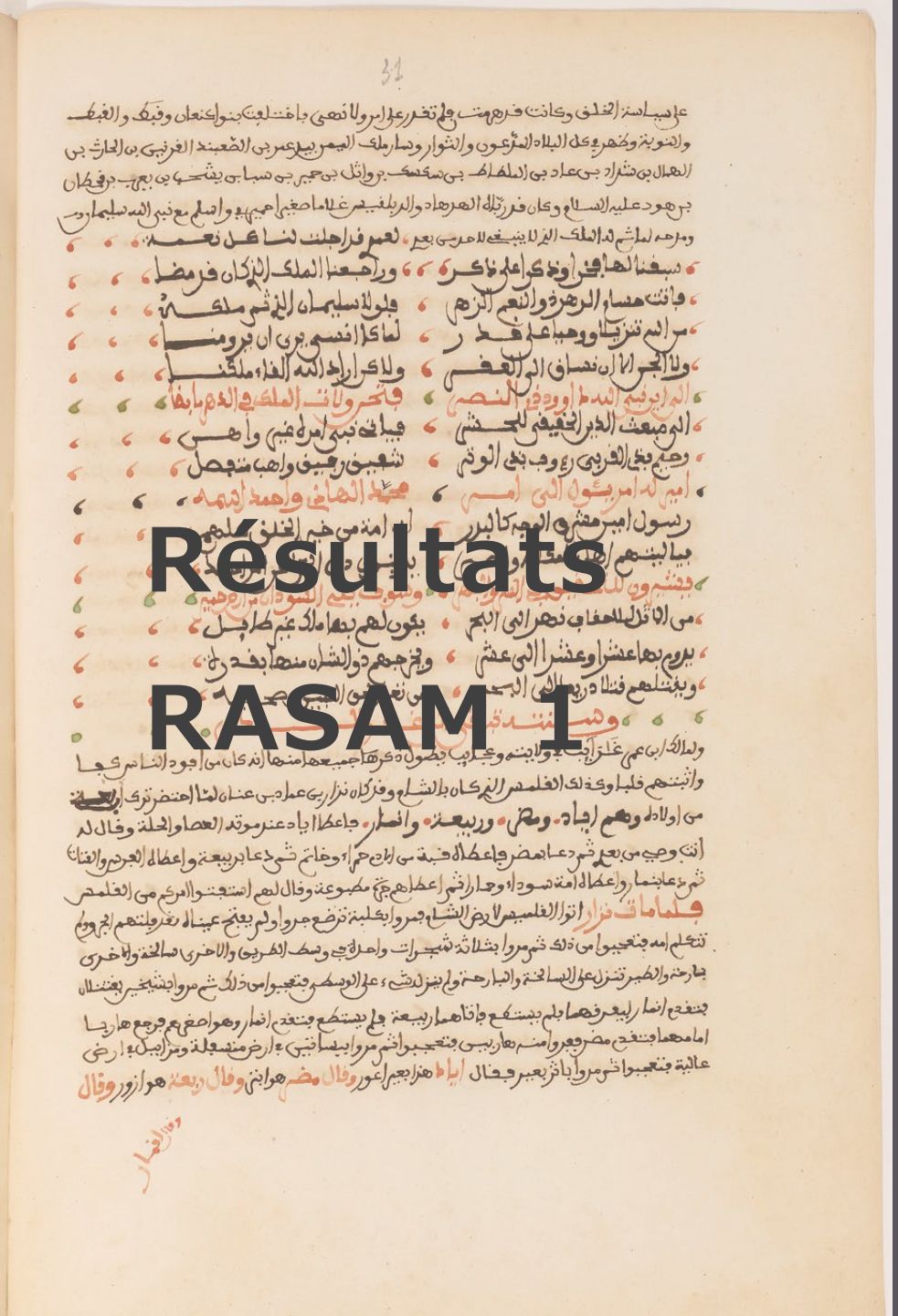
défaut

après 50 images relues



Déroulement

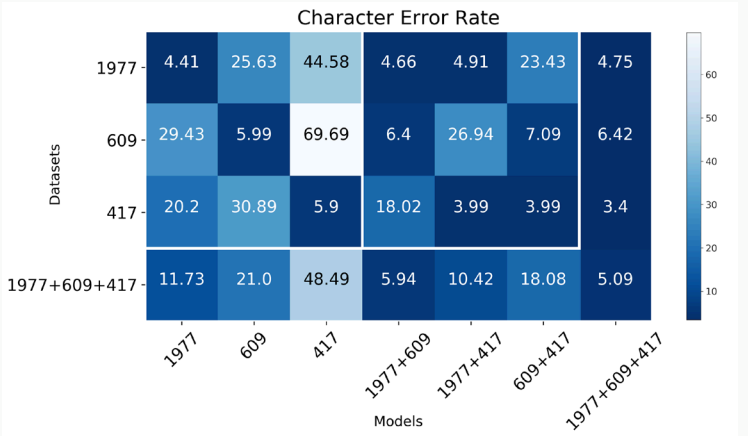
La spécialisation progressive des modèles a conduit à un gain de temps de **75%** dans la vérification de la mise en page.

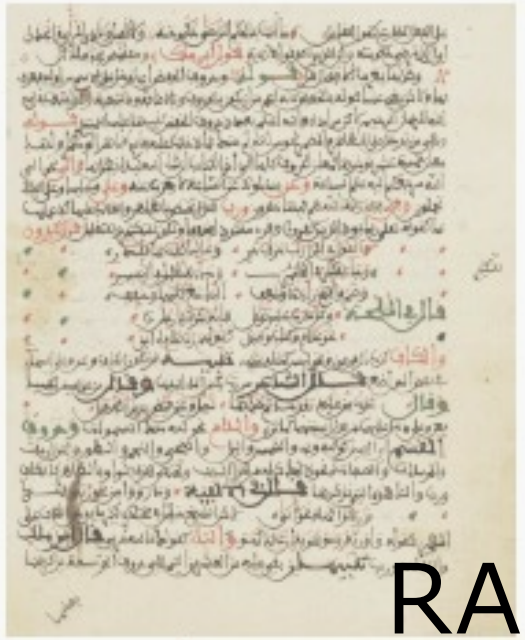
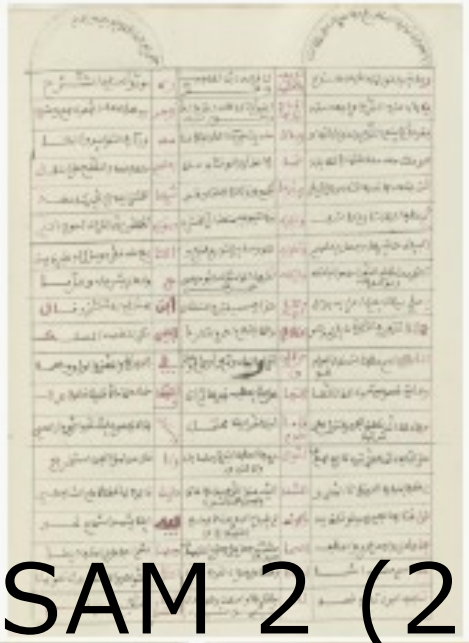


Résultats RASAM 1

- 300 pages annotées par 14 transcripteurs
- Dataset disponible sous licence CCBY : https://github.com/calfa-co/rasam-dataset
- Vidal-Gorène, C., Lucas, N., Salah, C., Decours-Perez, A., Dupin, B. “RASAM – A Dataset for the Recognition and Analysis of Scripts in Arabic Maghrebi,” in Document Analysis and Recognition – ICDAR 2021, edited by E.H. Barney Smith and U. Pal, 265-281. Springer: Cham, 2021.

CER FINAL IN-DOMAIN 4,8%



RASAM 2 (2021-2022)



Rasam1

- 3 ms, 300 pages
- Approche au mot
- Démonstration de faisabilité
- Enjeu des écritures arabes maghrébines
- Question mise en page

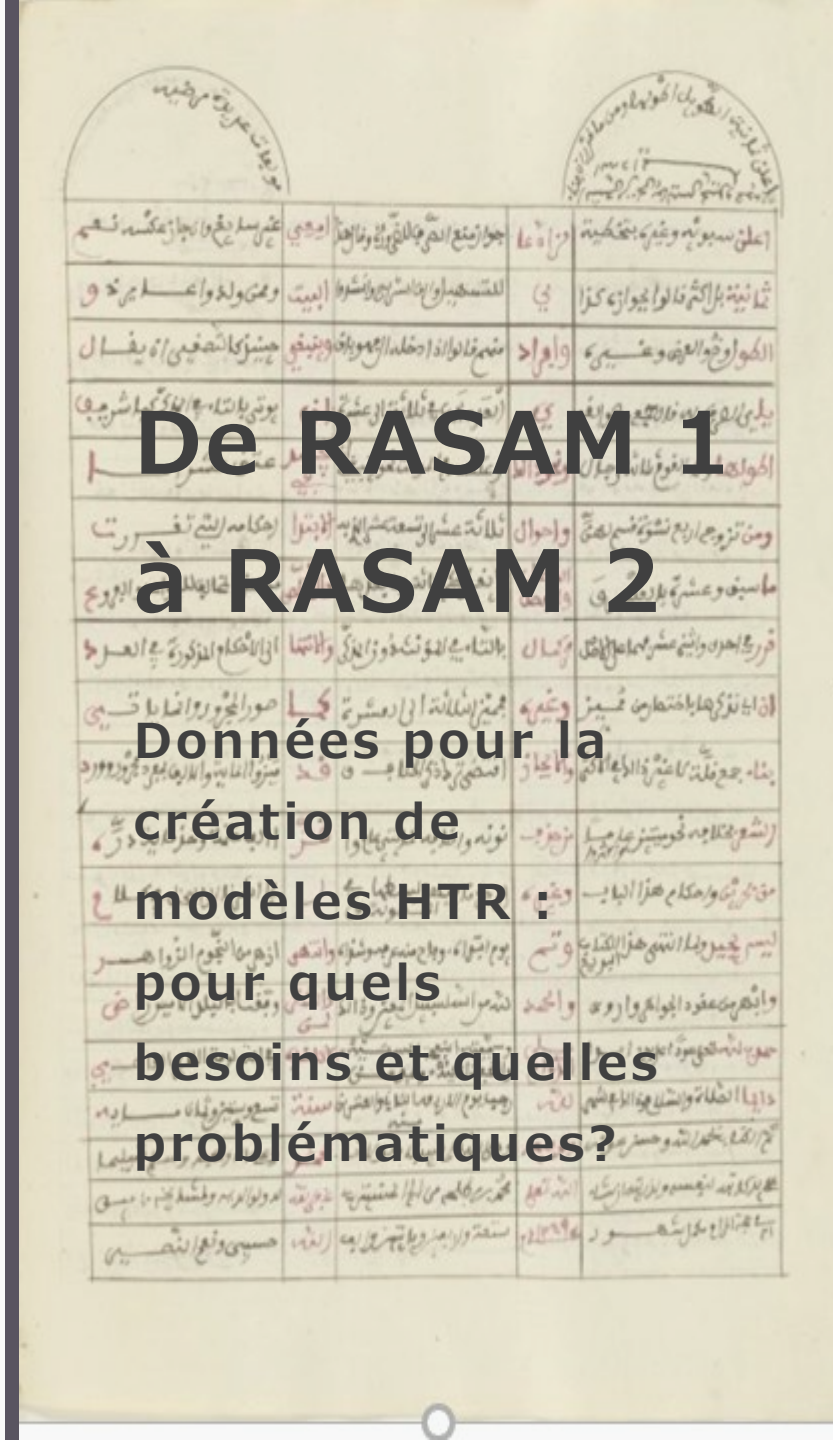
BASE SCIENTIFIQUE

Rasam 2

- 15 ms, 250 pages
- Renforcement de la polyvalence des modèles avec un hackathon (22 participants)
- Traitement HTR de ms de la Bulac
- Evaluation de la spécialisation: quels résultats? Quels coûts?

De RASAM 1 à RASAM 2

Données pour la création de modèles HTR : pour quels besoins et quelles problématiques?



Corpus

RASAM 2

BULAC.MS.ARA.6

BULAC.MS.ARA.9

BULAC.MS.ARA.23

BULAC.MS.ARA.24

BULAC.MS.ARA.45b

BULAC.MS.ARA.65

BULAC.MS.ARA.1926

BULAC.MS.ARA.1936

BULAC.MS.ARA.1943

BULAC.MS.ARA.1944

BULAC.MS.ARA.1946

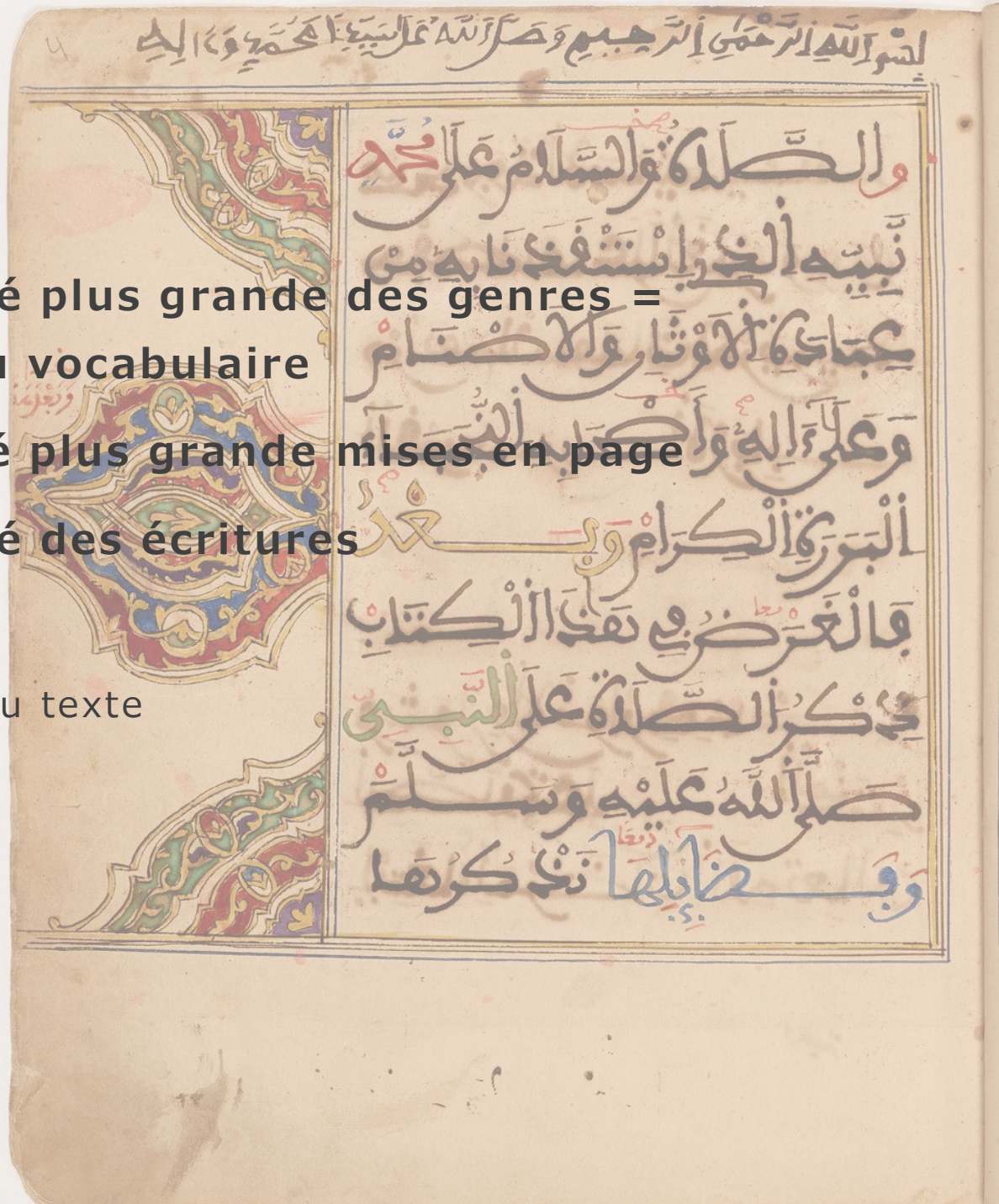
BULAC.MS.ARA.1947

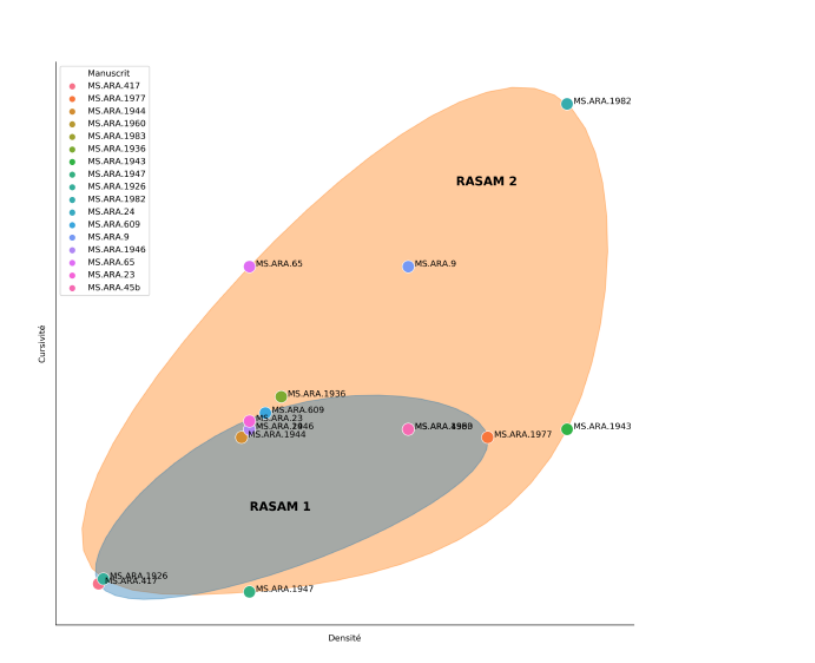
BULAC.MS.ARA.1960

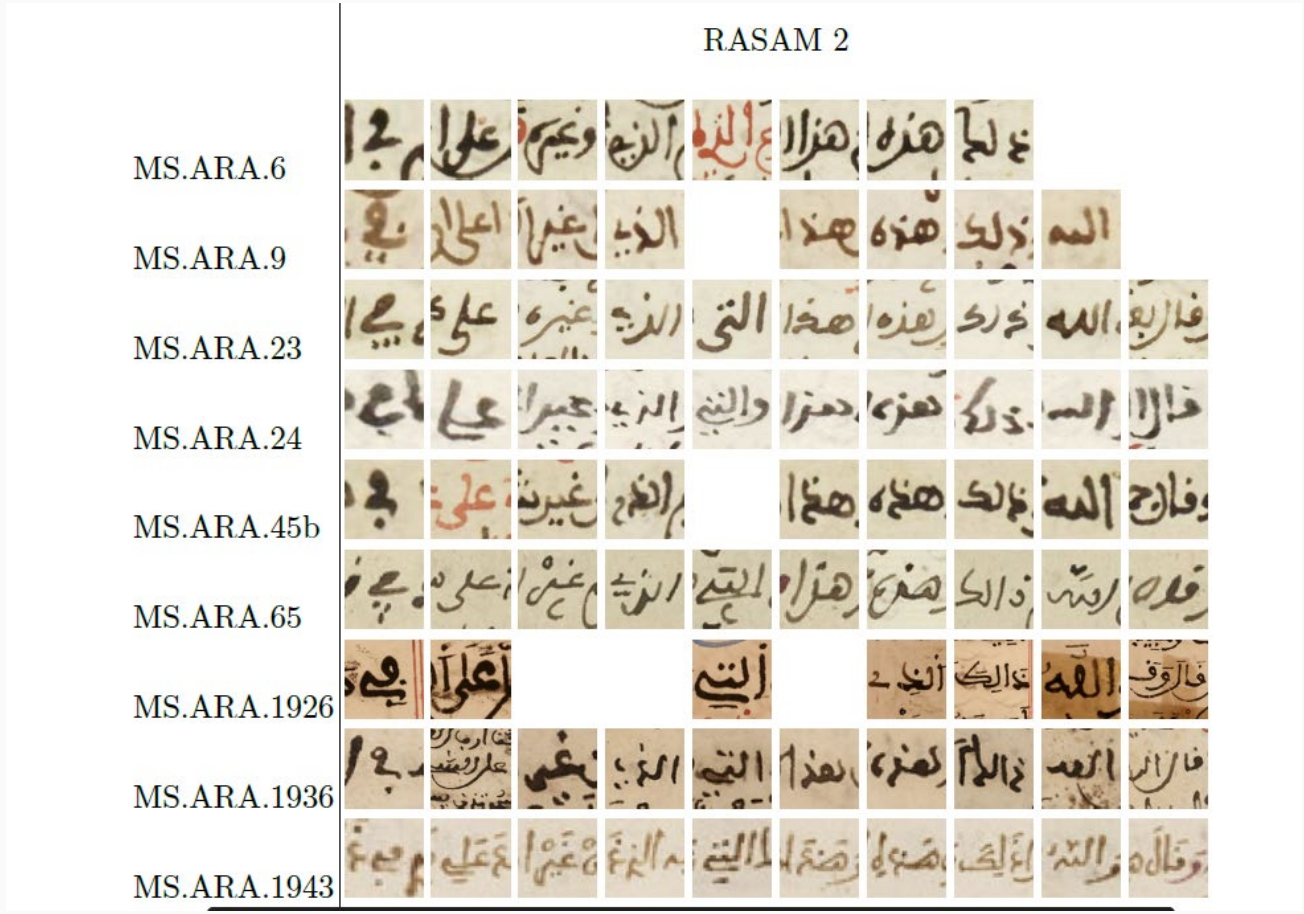
BULAC.MS.ARA.1982

BULAC.MS.ARA.1983

- **Diversité plus grande des genres = nouveau vocabulaire**
- **Pluralité plus grande mises en page**
- **Diversité des écritures**
- Couleur
- Densité du texte
- Style

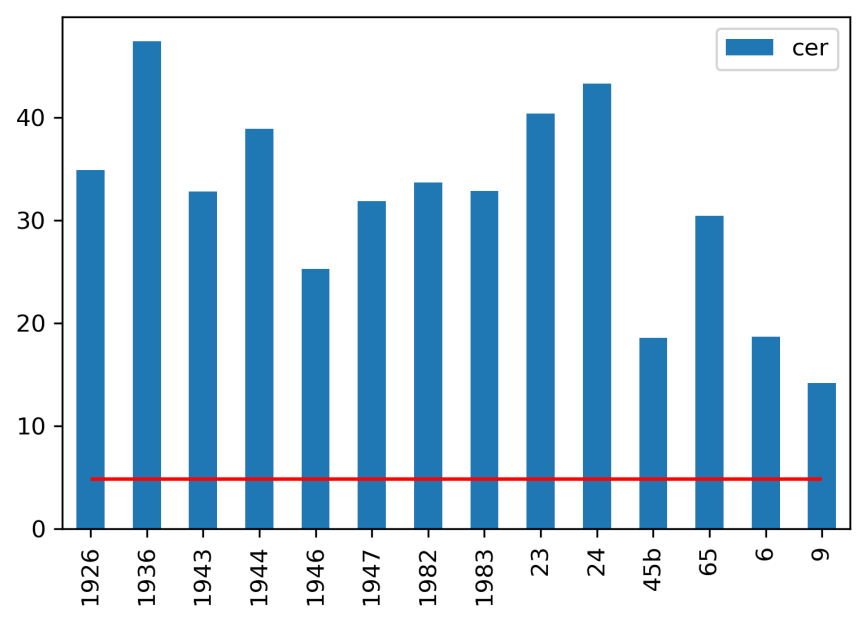



Manuscrit
MS.ARA.417
MS.ARA.1977
MS.ARA.1944
MS.ARA.1960
MS.ARA.1983
MS.ARA.1936
MS.ARA.1943
MS.ARA.1947
MS.ARA.1926
MS.ARA.1982
MS.ARA.24
MS.ARA.609
MS.ARA.9
MS.ARA.1946
MS.ARA.65
MS.ARA.23
MS.ARA.45b
RASAM 2
RASAM 1
MS.ARA.1982
MS.ARA.65
MS.ARA.9
MS.ARA.1936
MS.ARA.609
MS.ARA.23
MS.ARA.1946
MS.ARA.1944
MS.ARA.1977
MS.ARA.1943
MS.ARA.1926
MS.ARA.417
MS.ARA.1947
Cursivité
Densité


RASAM 2
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.6	18,68
MS.ARA.9	14,18
MS.ARA.23	40,34
MS.ARA.24	43,28
MS.ARA.45b	18,56
MS.ARA.65	0,44
MS.ARA.1926	34,86
MS.ARA.1936	47,39
MS.ARA.1943	32,78
MS.ARA.1944	38,86
MS.ARA.1946	25,28
MS.ARA.1947	31,83
MS.ARA.1982	33,67
MS.ARA.1983	32,84

Modèle RASAM1 appliqué aux manuscrits de RASAM2



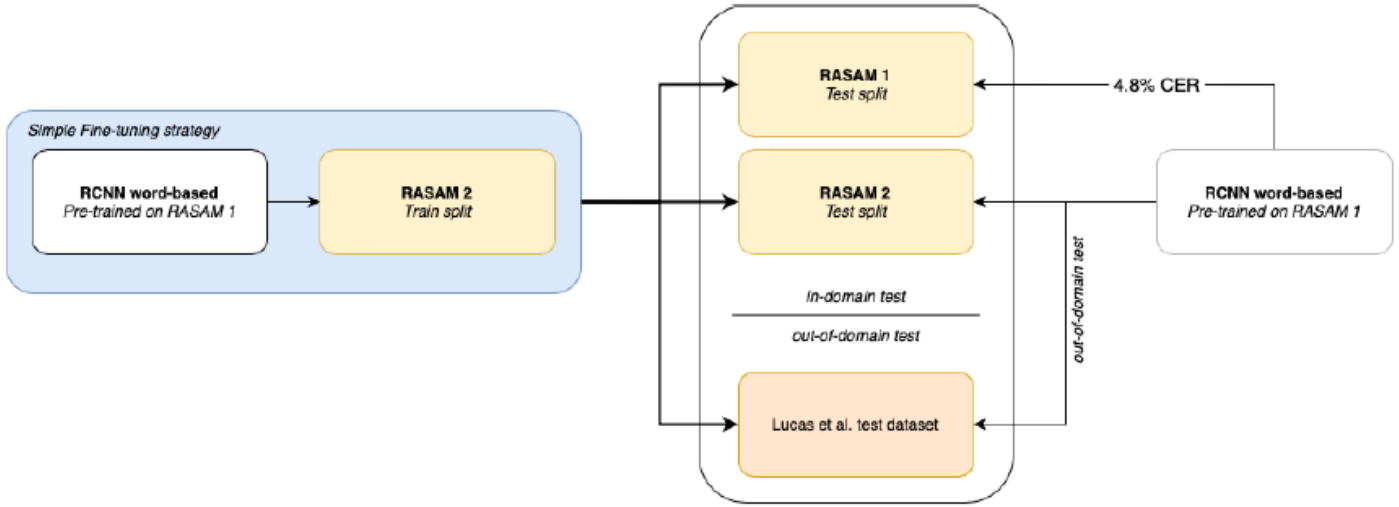


Fig. 3. Experiments conducted on the new dataset and comparison with the RASAM 1 and RASAM 2 models

=> Lucas, N., Salah, C., Vidal-Gorène, C.: *New Results for the Text Recognition of Arabic Maghribi Manuscripts - Managing an Under-resourced Script* (2022), https://hal-enc.archives-ouvertes.fr/hal-03874725 , working paper or preprint

De RASAM 1 à RASAM 2: Méthodologie appliquée

Tendance actuelle: RASAM1 + spécialisation avec 10-20 images = ~ 6 - 12% de CER

Résultats RASAM 2

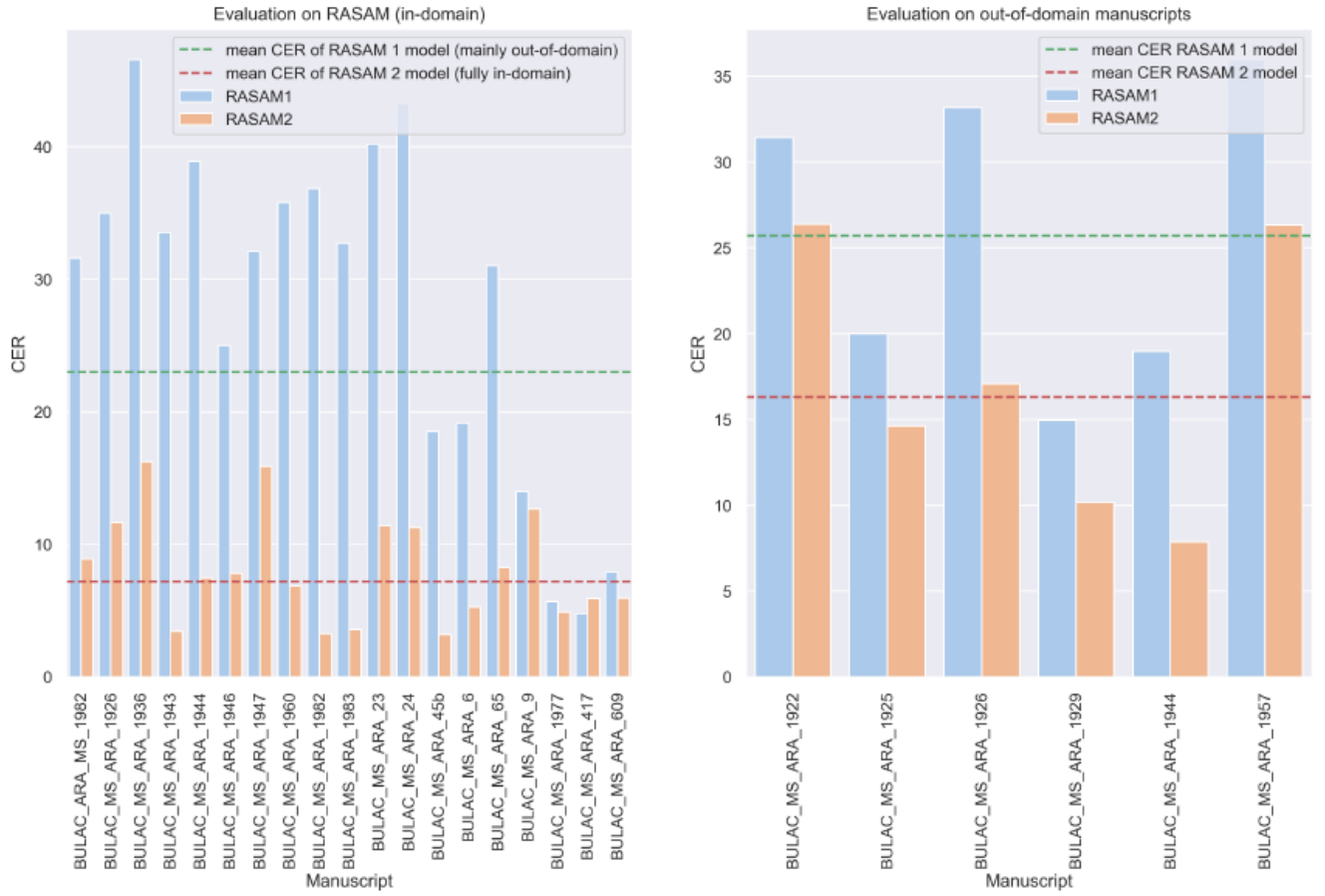
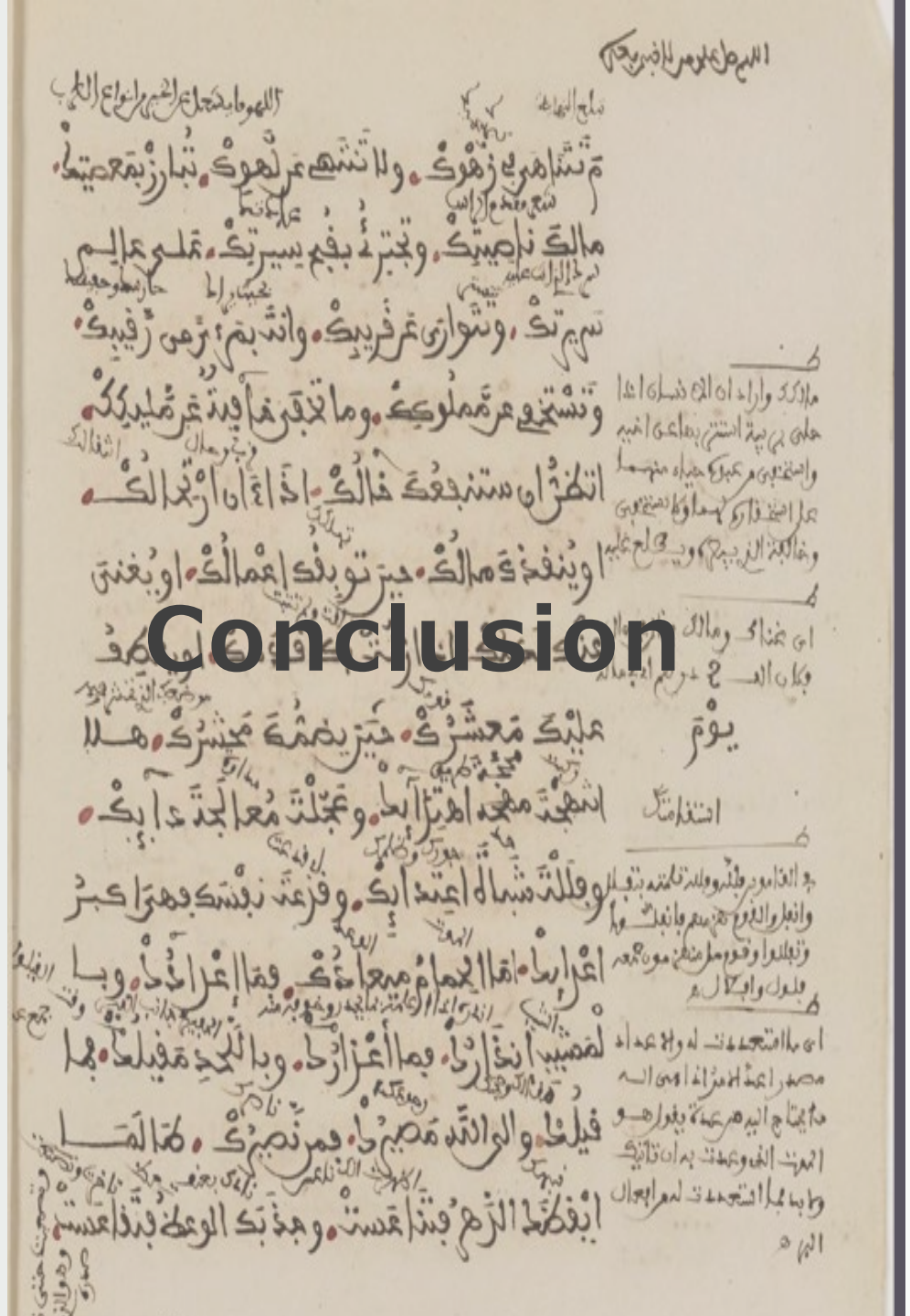


Fig. 5. Distribution of CERs obtained by RASAM 1 (in blue) and RASAM 2 (orange) for each in-domain and out-of-domain manuscrip

Résultats RASAM 2

- **Nouvelle tendance : RASAM2 + spécialisation avec 10 images = ~ 6 - 10% de CER. Gain de 50% dans l'annotation**
- Le jeux de données RASAM 2 dataset est représentatif de la variété des écritures arabes maghrébines manuscrites – Modèle plus polyvalent que RASAM1
- L'approche au mot démontre des performances in-domain et out-of-domains très satisfaisantes
- Le jeux de données constitue une base solide pour la spécialisation des modèles futures sur des manuscrits de ces graphies



Conclusion

- **Deux datasets disponibles:**
- RASAM 1: https://github.com/calfa-co/rasam-dataset
- RASAM 2: En cours de mise en ligne
- **Question de l'HTR des manuscrits arabes largement surmontées sous réserve d'un besoin identifié et d'un cahier des charges bien établi**
- **Pipeline désormais bien établi pour les écritures arabes**

Facteurs de réussite

1 - Les projets et l'initiative plus large étaient soutenus et pilotés par le Groupement d'intérêt scientifique « Moyen-Orient et mondes musulmans »

2- Le projet a bénéficié d'une collaboration (par le biais d'un partenariat) entre trois acteurs complémentaires : le Groupement d'intérêt scientifique MOMM d'une part, la BULAC d'autre part et Calfa.

3- Les objectifs avaient été clairement énoncés et précisés au départ.