A Dataset of Wolof Ajami Manuscripts for HTR and OCR

Oreen Yousuf[1], **Elhadji Djibril Diagne**[2], **Christian Høgel**[4]
Beáta Megyesi[3], **Joakim Nivre**[1]
[1]Uppsala University, Uppsala, Sweden
[2]Murid Islamic Community in America, Inc. (MICA, Inc.), New York, USA
[3]Stockholm University, Stockholm, Sweden
[4]Lund University, Lund, Sweden
oreen.yousuf@lingfil.uu.se

Abstract

We present the first ever dataset of manually segmented and transcribed Ajami manuscripts written in Wolof. The term Ajami refers to modified Arabic-script orthographies used to transcribe African languages. Handwritten text recognition (HTR) and optical character recognition (OCR) models for Arabic-script languages perform poorly on African languages written in Ajami orthographies because these languages are not represented in the pre-training data of the models. This leads to recognition models being unable to extract unique Arabic-script letters and ubiquitous diacritics used in African languages, and struggling to adapt to various calligraphy styles used across Africa. We release the following as an open-source dataset: an ALTO formatting of high-quality images of handwritten and printed, 20th–century Wolof manuscripts; manual segmentation (region and line); and manual transcriptions. We extend our contribution by evaluating several Arabic-script recognition models intended for historical manuscripts and find they produce character error rates (CER) of 61–81%. Transcriptions produced by the evaluated recognition models, as well as a keyboard to transcribe Wolof Ajami manuscripts, are released as well. The digitally transcribed text in the dataset can also be utilized for various natural language processing (NLP) and historical linguistic tasks.

Keywords: handwritten text recognition, optical character recognition, benchmark dataset, Ajami, Wolof

1. Introduction

Wolof (ISO 639-3: *wol*) is a low-resource language spoken by about 18 million people in Senegal, Gambia, and Mauritania. In recent years Wolof has increasingly been included in various natural language processing (NLP) works (Adelani et al., 2022a,b; Dossou et al., 2022; Ogundepo et al., 2023; Dione et al., 2023). Handwritten text recognition (HTR) and optical character recognition (OCR) is a branch of computer vision and pattern recognition that aims to automatically and accurately extract text from images. The digitized text can then be a valuable resource for various NLP tasks (Boroş et al., 2020; Ignat et al., 2022; Lopresti, 2008; Van Strien et al., 2020). Developing gold-standard segmentation and transcription HTR/OCR datasets is crucial to perform accurate recognition for any language. Models trained for several NLP and document analysis tasks perform poorly on low-resource languages, including many African languages (Adebara et al., 2025; Costa-Jussà et al., 2022), partly due to the lack of high-quality training data for specialized needs – including HTR/OCR (Belay et al., 2020; Oni and Asahiah, 2020).

For these reasons, it is necessary to develop HTR datasets for both modern and historical documents written in African languages. Recent work has been conducted to curate historical datasets for select East and West African-language manuscripts (Belay et al., 2024; Yousuf et al., 2025). Before European colonization, several West African languages were written in *Ajami* (Ngom, 2016; Ngom et al., 2023), modified Arabic-script orthographies, that differed from the "standard" Arabic alphabet by including some unique letters and diacritics. Tashkīl are Arabic-script diacritics used for vocalization, lack of vocalization, or gemination. They are ubiquitous in several African-language manuscripts and necessary to read the text. In contrast, tashkīl are optionally written in the most prominent Arabic-script languages in Asia (e.g., Arabic, Persian, Urdu). This is directly reflected in the lack of diacritics in the pre-training data for recognition models, which leads to high character error rates (CER) for Ajami-text recognition (Yousuf et al., 2025).

2. Dataset and Curation

We have curated 19 20th-century Wolof Ajami manuscripts covering 4,743 lines and 313 pages in total. The longest manuscript in the dataset is 1,151 lines (76 pages), while the shortest is 31 lines (3 pages). The dataset has an average of 249.6 lines and 16.5 pages per manuscript. While historical West African manuscripts covered numerous subjects (e.g., legal and scientific texts, history, literature, etc.) (Ngom, 2017), our dataset consists of 9 prose manuscripts and 10 poetry manuscripts. The provenance of all manuscripts is Touba, Sene-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3234–3239
11–16 May 2026.

Manuscript	Domain	Pages	Lines	Type
Càntug Murid (A Murid Expression of Gratitude)	Poetry	17	255	Handwritten
Diskuuru Sëriñ Abdul Ahad Mbàkke 1: Tabaski 1972 (Speech by Shaykh Abdul Ahad Mbakke 1: Tabaski 1972)	Prose	3	31	Printed
Diskuuru Sëriñ Abdul Ahad Mbàkke 2: Korite 1971-1972 (Speech by Shaykh Abdul Ahad Mbakke 2: Korite 1971-1972)	Prose	9	89	Printed
Manāfiu l-Muslim (A Healing Manual for Muslims)	Prose	19	270	Printed
Marsiya Sëriñ Masamba Mbakke (Elegy for Sëriñ Masamba Mbakke)	Poetry	14	299	Handwritten
Nahju Qaḍā'i l-Ḥāji (Path to the Satisfaction of Needs)	Poetry	17	241	Printed
Qasidak Wolofalu Maam Jaaratul Laahi (A Wolof Poem Praising Maam Jaaratul Laahi)	Poetry	25	272	Printed
Sëriñ Fallu (Bàkk wi) (Song of the Champion)	Poetry	30	365	Handwritten
Soxna Asta Waalo Mbakke (Biography of Lady Asta Waalo Mbakke)	Prose	3	38	Printed
Soxna Aysatu Mbakke-Kajoor (Biography of Lady Aysatu Mbakke-Kajoor)	Prose	4	55	Printed
Soxna Faati Ja Mbakke (Biography of Lady Faati Ja Mbakke)	Prose	3	50	Printed
Soxna Maam Jaaratul Laahi Buso (Biography of Lady Maam Jaara Buso)	Prose	5	81	Printed
Ubbiteg Fooraas bi (Water Tower Inaugural Speech)	Prose	10	111	Printed
Wolofalu Jumaa ji (Poem of the Mosque of Touba)	Poetry	19	541	Handwritten
Wolofalu Màggal gu Njëkk gi (Poem of the First Màggal)	Poetry	7	105	Handwritten
Xarbaaxi Yonnen bi (Miracles of the Prophet)	Poetry	12	186	Handwritten
Xareb Badar bu Njëkk ba (The First Battle of Badr)	Poetry	8	127	Handwritten
Yeneen mbindum Serigne Mor Kayre (Other Writings of Serigne Mor Kayre)	Poetry	76	1151	Handwritten
Yoonu Murid (The Murid Way)	Prose	32	478	Printed

Table 1: Wolof Ajami manuscript dataset details.

gal. There are 8 handwritten manuscripts and 11 printed manuscripts, all in the Maghrebi calligraphy style. We have sourced our manuscripts from Boston University's Ajami collection, specifically from the "*The Four Languages*" and "*African Ajami Library*" repositories. The dataset is detailed in Table 1.

We release our dataset (Yousuf et al., 2026) in an ALTO format (Analyzed Layout and Text Object). This ALTO formatted dataset contains and links the high-quality manuscript images, the polygon coordinates (i.e., bounding boxes) for the manual region and line segmentation, line order, and manual transcriptions.

We follow a strict annotation protocol to create this Wolof Ajami dataset. We use eScriptorium (Kiessling et al., 2019) for our experiments, which is a framework to digitally transcribe and create recognition tools for historical documents. The platform can be used to catalog manuscripts, create manual or automatic segmentation and transcription data, and import and fine-tune recognition models. Each manuscript has first been transcribed by a philologist with expertise in Wolof Ajami and then independently reviewed for error correction afterwards.

Creating digital transcriptions of historical materials can often require specialized input methods, as historical orthographies are sometimes not readily available on modern computer keyboards. As the

Model	Trained on	Training Language(s)	Calligraphy/Typeface
MS Mellon Print	Print	Arabic, Persian, Ottoman Turkish, Urdu	Naskh, Nasta'liq
MS Pretrained	Print	Arabic, Persian, Ottoman Turkish, Urdu	Naskh, Nasta'liq
Generalized	Handwritten	Arabic, Persian, Ottoman Turkish, Urud	Naskh, Nasta'liq Maghrebi
Mellon Ottoturk Print	Print	Ottoman Turkish	Naskh, Nasta'liq
Mellon Print	Print	Arabic	Naskh, Tahoma
Pretrained Print	Print	Arabic, Persian, Ottoman Turkish, Urdu	Naskh, Nasta'liq
Urdu Print	Print	Urdu	Nasta'liq, Naskh

Table 2: Suite of OpenITI recognition models for historical Arabic-script languages.

Latin script was imposed on Wolof speakers during the colonial era, there is no accessible keyboard to type in the historical Wolof Ajami orthography. Although there may be existing online Ajami keyboards, oftentimes these keyboards are designed for modern Ajami orthographies that were standardized in the past few decades rather than traditional orthographies. To rectify this, we have created a Keyman[1] keyboard with specifications that reflect the character inventory needed to transcribe historical Wolof manuscripts. Keyman Developer[2] is a free and open-source keyboarding platform that can be used to create an input method for any language and writing system (Showalter, 2016; Santos and Harrigan, 2020). We also release this historical Wolof Ajami-manuscript keyboard as open-source (Yousuf, 2026).

3. Baseline Experiments

We extend our work by evaluating existing Arabic-script HTR/OCR models on Wolof Ajami data. We select a suite of recognition models trained on historical manuscripts created by the Open Islamicate Texts Initiative (OpenITI),[3] a multi-institutional research group that studies Islamic texts in different languages (e.g., Arabic, Persian, Urdu, Ottoman Turkish) (Miller et al., 2018; Romanov et al., 2019).

[1] https://keyman.com/
[2] https://keyman.com/developer/
[3] https://openiti.org/

Model	CER
MS Mellon Print	0.625
MS Pretrained	0.668
Generalized	0.671
Mellon Ottoturk Print	0.768
Mellon Print	0.770
Pretrained Print	0.791
Urdu Print	0.822

Table 3: Historical Arabic-script recognition model baselines for Wolof Ajami manuscript dataset.

These models are detailed in Table 2. As these models are trained on diverse languages, including standard Arabic, and diverse calligraphy styles in a low-resource setting (Kiessling et al., 2024) they are appropriate for our studies. CER was used as our evaluation metric.

3.1. Results and Analysis

As seen in Table 3, we observe high CERs when these existing recognition models for historical Arabic-script manuscripts attempt to transcribe Wolof Ajami manuscripts. We find MS Mellon Print performs the best when transcribing Wolof text, with a CER of 0.625. It is unsurprising that the Urdu-exclusive model (Urdu Print) performs the worst on our dataset as Urdu, spoken mainly in Pakistan, is almost always written in the Nasta'liq calligraphy style – which is non-existent in West African

Model	Transcription
Ground Truth	بُوىْ سِلَتْدُوكْ زِيتُوكْ بُرُومْ جِزِ مِلُومْ كُلَىْ مِلُومْبِرْ لِكْ رَجْاَىْ سُلُومْ
MS Mellon Print	تووم سبهوشّزْيتوةٔ بوم يزعلوم املنيملوم يزالتخروأتّ سلوم
MS Pretrained	بوى سبه وكر شود بوم بزطوم مكلنى فلوم بر لكريم ان سلوم
Generalized	يوي سبتا وكريتوك بروم بر ملوم كلي فلوم يز لكريم أني سلوم
Mellon Ottoturk Print	توتمسبه و زرتوةٔ بوم يزطوم محلنى ملوم بزللريوأتمسلويم
Mellon Print	تووم سبهوشّزْيتوةٔ بوم يزعلوم املنيملوم يزالتخروأتّ سلوم
Pretrained Print	وسوموعروم ولوم حلى طوم ولوسلوع
Urdu Print	بوتى ٣٨ و" زبؤ بروم بز طوم "لنى علوم بزلتے ہ وائى سلومم

Figure 1: Ground truth and automatic transcriptions of a line from "*Yeneen mbindum Serigne Mor Kayre*".

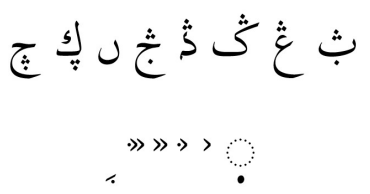

Figure 2: Arabic-script letters that are used in Wolof, but not used in prominent Arabic-script Asian languages like Arabic, Persian, etc.

manuscript cultures. Conversely, the Generalized model is the only model trained on the Maghrebi calligraphy style, which is the sole style our Wolof dataset is comprised of, yet it performs worse than MS Mellon Print and MS Pretrained – both of which were not exposed to the Maghrebi style during pre-training. There is a noticeable ~0.10 CER drop between the Generalized and Ottoturk model, effectively dividing the top three and bottom four performing models.

Unsurprisingly, our evaluated models struggle to recognize the majority of Arabic-script diacritics. Although these diacritics exist in the languages included in these models' pre-training, they are scarely used in practice. Furthermore, the models are also unable to recognize unique letters and diacritics that exist in the Wolof Ajami orthography (Figure 2) and are not represented in the training data for these models. A qualitative example is shown in Figure 1.

We qualitatively observe in many examples that the top three models better recognize the general structure of the ground truth lines. While the transcriptions of the MS Mellon Print, MS Pretrained, and Generalized models are barren of diacritics and are unable to recognize unique Wolof characters, they do regularly produce words of similar length as those found in the ground truth line. Conversely, much longer words are produced by the Mellon Ottoturk Print and Mellon Print models.

Letters in the Arabic script often have dots above or below the base grapheme (known as i'jām). These differ from tashkīl by being part of the letter itself and are mandatory in writing. However, the visual distinction between i'jām and tashkīl is a noted and prominent challenge in Arabic-script recognition (Faizullah et al., 2023; Kasem et al., 2023). We also observe that the evaluated recognition models often incorrectly recognize an Arabic letter without i'jām as a similar character *with* i'jām. However, the heavy use of tashkīl on letters without i'jām have further led models to simply produce letters with i'jām instead. Lutf et al. (2014) have focused on this diacritic challenge in Arabic-script recognition for Arabic. The more prevalent use of diacritics in Ajami orthographies warrants furthering this research in future work.

4. Conclusion

In this paper we release the first ever handwritten text recognition and optical character recognition dataset for Wolof Ajami manuscripts. Our open-source dataset contains high-quality images of 17 manuscripts, manual segmentation for the region and line level, and diplomatic transcriptions provided by experts. We also release a specialized Wolof Ajami keyboard as open-source for the ben-

efit of future researchers. The dataset is not only limited to HTR/OCR as researchers can solely utilize the digitized text for various natural language processing tasks.

We have extended our work by evaluating OpenITI's suite of Arabic-script recognition models specifically for historical manuscripts on our dataset. We find that these existing models produce high character error rates of 61–81% when trying to transcribe Wolof Ajami text. These models perform poorly due to a lack of appropriate training data for Wolof Ajami orthography. The heavy use of Arabic-script diacritics and unique Arabic-script letters in Wolof, which are absent from the models' pre-training languages (i.e., Arabic, Persian, Urdu, Ottoman Turkish), contribute to the models' poor performance.

We hope that this dataset enhances recognition models for diverse languages and inspires the creation of more HTR/OCR datasets for the dozens of other African languages with Ajami manuscript traditions (Mumin, 2014).

5. Acknowledgments

We would like to sadly acknowledge the passing of Elhadji Djibril Diagne. His extensive knowledge of Wolof linguistics and manuscripts was pivotal to this work, and he will be dearly missed by both his loved ones and colleagues in the field.

This work has been financed in part by The Swedish Graduate School of Digital Philology through the Swedish Research Council (grant 2022-06343); and by Riksbankens Jubileumsfond, grant M24-0028: Echoes of History: Analysis and Decipherment of Historical Writings (DESCRYPT).

6. Bibliographical References

Ife Adebara, Hawau Olamide Toyin, Nahom Tesfu Ghebremichael, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2025. Where are we? evaluating llm performance on african languages. *arXiv preprint arXiv:2502.19582*.

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, et al. 2022a. A few thousand translations go a long way! leveraging pre-trained models for african news translation. *arXiv preprint arXiv:2205.02022*.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O Alabi, Shamsuddeen H Muhammad, Peter Nabende, et al. 2022b. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. *arXiv preprint arXiv:2210.12391*.

Birhanu Belay, Tewodros Habtegebrial, Million Meshesha, Marcus Liwicki, Gebeyehu Belay, and Didier Stricker. 2020. Amharic ocr: an end-to-end learning. *Applied Sciences*, 10(3):1117.

Birhanu Hailu Belay, Isabelle Guyon, Tadele Mengiste, Bezawork Tilahun, Marcus Liwicki, Tesfa Tegegne, and Romain Egele. 2024. A historical handwritten dataset for ethiopic ocr with baseline models and human-level performance. In *International Conference on Document Analysis and Recognition*, pages 23–38. Springer.

Emanuela Boroş, Ahmed Hamdi, Elvys Linhares Pontes, Luis-Adrián Cabrera-Diego, Jose G Moreno, Nicolas Sidere, and Antoine Doucet. 2020. Alleviating digitization errors in named entity recognition for historical documents. In *Proceedings of the 24th conference on computational natural language learning*, pages 431–441.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Cheikh M Bamba Dione, David Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, et al. 2023. Masakhapos: Part-of-speech tagging for typologically diverse african languages. *arXiv preprint arXiv:2305.13989*.

Bonaventure FP Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Chinenye Emezue. 2022. Afrolm: A self-active learning-based multilingual pretrained language model for 23 african languages. *arXiv preprint arXiv:2211.03263*.

Safiullah Faizullah, Muhammad Sohaib Ayub, Sajid Hussain, and Muhammad Asad Khan. 2023. A survey of ocr in arabic language: applications, techniques, and challenges. *Applied Sciences*, 13(7):4584.

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022. Ocr improves machine translation for low-resource languages. *arXiv preprint arXiv:2202.13274*.

Mahmoud SalahEldin Kasem, Mohamed Mahmoud, and Hyun-Soo Kang. 2023. Advancements and challenges in arabic optical character recognition: A comprehensive survey. *ACM Computing Surveys*.

Benjamin Kiessling, Gennady Kurin, Matthew Thomas Miller, and Kader Smail. 2024. Advances and limitations in open source arabic-script ocr: A case study. *arXiv preprint arXiv:2402.10943*.

Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra. 2019. escriptorium: an open source platform for historical document analysis. In *2019 international conference on document analysis and recognition workshops (icdarw)*, volume 2, pages 19–19. IEEE.

Daniel Lopresti. 2008. Optical character recognition errors and their effects on natural language processing. In *Proceedings of the second workshop on Analytics for Noisy Unstructured Text Data*, pages 9–16.

Mohammed Lutf, Xinge You, Yiu-ming Cheung, and CL Philip Chen. 2014. Arabic font recognition based on diacritics features. *Pattern Recognition*, 47(2):672–684.

Matthew Thomas Miller, Maxim G Romanov, and Sarah Bowen Savant. 2018. Digitizing the textual heritage of the premodern islamicate world: Principles and plans. *International Journal of Middle East Studies*, 50(1):103–109.

Meikal Mumin. 2014. The arabic script in africa: understudied literacy. In *The Arabic Script in Africa*, pages 41–76. Brill.

Fallou Ngom. 2016. *Muslims beyond the Arab world: The odyssey of ajami and the Muridiyya*. Oxford University Press.

Fallou Ngom. 2017. West african manuscripts in arabic and african languages and digital preservation. In *Oxford research encyclopedia of African history*.

Fallou Ngom, Daivi Rodima-Taylor, and David Robinson. 2023. ajamī literacies of africa: The hausa, fula, mandinka, and wolof traditions. *Islamic Africa*, 14(2):119–143.

Odunayo Ogundepo, Tajuddeen R Gwadabe, Clara E Rivera, Jonathan H Clark, Sebastian Ruder, David Ifeoluwa Adelani, Bonaventure FP Dossou, Abdou Aziz Diop, Claytone Sikasote, Gilles Hacheme, et al. 2023. Afriqa: Cross-lingual open-retrieval question answering for african languages. *arXiv preprint arXiv:2305.06897*.

Olalekan Joseph Oni and Franklin Oladiipo Asahiah. 2020. Computational modelling of an optical character recognition system for yorùbá printed text images. *Scientific African*, 9:e00415.

Maxim Romanov, Matthew Thomas Miller, Sarah Bowen Savant, and Masoumeh Seydi. 2019. Open islamicate texts initiative: A machine-readable corpus of texts produced in the premodern islamicate world (poster). In *Digital Humanities 2019 Conference Papers (9-12 July 2019). Utrecht University, 2019*.

Eddie Antonio Santos and Atticus Harrigan. 2020. Design and evaluation of a smartphone keyboard for plains cree syllabics. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 88–96.

Esther H Showalter. 2016. Mobile device keyboard customization for a newly constructed orthography of a rural west african language. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development*, pages 1–4.

Daniel Van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the impact of ocr quality on downstream nlp tasks.

Oreen Yousuf, Abdulmalik Aminu, Musa Salih Muhammad, Bashir Usman, Mustapha Kurfi Hashim, Joakim Nivre, Beáta Megyesi, and Christian Høgel. 2025. A handwritten text recognition dataset for ajami manuscripts in fulfulde and hausa. In *International Conference on Document Analysis and Recognition*, pages 620–637. Springer.

7. Language Resource References

Yousuf, Oreen. 2026. *Wolof Ajami Keyboard*. Keyman keyboard to transcribe Wolof Ajami manuscripts. The keyboard is designed specifically to transcribe the Wolof Ajami manuscripts sourced from Boston University used in this paper: https://github.com/oyousuf/Ajami-Manuscript-Keyboards.

Yousuf, Oreen and Diagne, Elhadji Djibril and Nivre, Joakim and Megyesi, Beáta and Høgel, Christian. 2026. *Wolof Ajami Handwritten Text Recognition Dataset*. Handwritten and Printed Wolof Ajami manuscripts for HTR and OCR: https://doi.org/10.5281/zenodo.15691685.