

The Best of Both Worlds: Exploring Wolofal in the Context of NLP

Ngoc Tan Le¹, Ali Mijiyawa¹, Abdoulat Leye¹, Alla Lo², Elhadji M. Nguer³, Fatiha Sadat¹

¹Université du Québec à Montréal-Canada, ²Université Numérique Cheikh Hamidou KANE-Sénégal, ³Université Gaston Berger - Saint-Louis-Sénégal

leye.abdoulat@courrier.uqam.ca, mijiyawa.ali@courrier.uqam.ca,
alla.lo@university.sn, elhadjimamadou.nguer@unchk.edu.sn,
Correspondence: le.ngoc_tan@uqam.ca, sadat.fatiha@uqam.ca

Abstract

This paper examines the three writing systems used for the Wolof language: the Latin script, the Ajami script (Wolofal), and the Garay script. Although the Latin alphabet is now the official standard for writing Wolof in Senegal, Garay and Ajami still play an important cultural and religious role, especially the latter. This article focuses specifically on Ajami, a system based on the Arabic script, and describes its history, its use, and its modern writings. We also analyze the challenges and prospects of these systems from the perspective of language preservation.

1 Introduction

African languages represent a unique linguistic and cultural diversity, and many use alternative scripts to the Latin alphabet, such as Ajami, a script derived from Arabic. The use of Ajami for African languages, including Wolof or Hausa - the rich African cultural heritage of the Hausa community (Gee, 2005), is of particular importance because of its role in the transmission of religious, cultural, and educational knowledge (Sane, 2010).

Hausa, one of Africa's most widely spoken languages, has a rich linguistic heritage that extends beyond its contemporary use of the Latin script. Historically, Hausa has also been written in the Ajami script, a modified form of Arabic adapted to represent Hausa's unique phonological system (Inuwa-Dutse, 2023; Adamu, 2023). The origins of Hausa Ajami can be traced to the introduction of Islam in the region during the 14th century, which brought Arabic literacy to the Hausa-speaking communities (Philips, 2004). Over time, this script was adapted to accommodate sounds specific to Hausa, enabling scholars and communities to document their language and culture while maintaining ties to Islamic traditions.

Ajami served as a vital tool for education, religious scholarship, and administrative purposes

across Wolof-speaking or Hausa-speaking regions. Despite its historical prominence, the use of Ajami has faced challenges, including regional orthographic variations and a lack of standardization (Bondarev, 2019; Library, 2020). Today, with the advent of natural language processing (NLP), Hausa Ajami presents an opportunity for linguistic innovation (Abdullahi, 2022). By leveraging its structural similarities to Arabic, researchers aim to enhance digital applications such as machine translation, text recognition, and language modeling, ensuring that this culturally significant script remains relevant in modern technological contexts. However, the lack of resources and suitable tools hinders the development of NLP for Ajami texts.

This article aims to analyze the writing systems of the Wolof language, particularly the Ajami script, its cultural and religious importance, and recent progress. This document discusses Ajami in Wolof, from a Natural Language Processing (NLP) perspective, examining the specific challenges of analyzing this writing system and assessing recent research and developments.

2 Linguistic Context and Writing Forms of Wolof

Wolof, the lingua franca of Senegal and the mother tongue of many speakers, has a long tradition of writing in Ajami (Cissé and Sadat, 2003). This adaptation of the Arabic, Wolof, alphabet allows the sound of local languages to be represented, but it also includes specific morphological and phonological adaptations (Ngom, 2004).

Wolof is a language spoken in Senegal (over 80% of the Senegalese population), the Gambia, and Mauritania (Cissé and Sadat, 2024). It belongs to the Atlantic branch of the Niger-Congo languages (Ngom, 2000). Historically, there are three writing systems for transcribing Wolof. The Latin alphabet based on modern school characters,

the Ajami script or Wolofal (Ngom, 2016) based on the Arabic alphabet and widely used in religious circles, and the Garay transcription system invented in 1961 by Assane Faye and inspired by African linguistic characteristics (Everson, 2016). The use of the Latin alphabet is now dominant in educational systems and administrative documents, but it remains largely reserved for the schooled population.

The Garay script, an Indigenous and minority writing systems according to the Atlas of Endangered Alphabets¹, was invented by Assane Faye of Senegāl in 1961, for writing the Wolof language (James, 2012). It runs right-to-left like Arabic, and some of the shapes are reminiscent of Arabic (James, 2012). The Garay script takes into account the phonetic characteristics of Wolof and aims to offer a transcription system used mainly in artistic and community contexts.

Finally, Ajami is a legacy of the Islamization of West Africa and remains at the heart of the religious and cultural practices of the Wolof people educated in Quranic schools (Ngom and Kurfi, 2017).

The Ajami script has been used historically to disseminate religious, poetic, and educational writings, but its modern usage is limited compared to the Latin alphabet, which today dominates publishing and education. However, the research, in terms of NLP applied to Ajami texts remains limited (Vydrin, 2014) or even non-existent, largely due to the challenges of data collection and digitization (Ngom et al., 2023).

Few studies have focused on automatic text processing in Ajami. The NLP literature on Wolof is growing, but it mainly focuses on the Latin script. Some research projects, notably in Senegal, have started to digitize and annotate corpora in Ajami, but the data remains limited. The absence of standardized linguistic resources, such as lexicons, grammars and parallel corpora, constitutes a major obstacle to the advancement of NLP for Wolof in Ajami.

3 Linguistic features

Wolof has distinctive linguistic features that differentiate it from neighboring languages. Here are some key elements of its linguistic structure:

¹<https://www.endangeredalphabets.net/>

3.1 Phonology

The Wolof language is characterized by simple consonants and vowels, but it also includes nasal sounds and consonants borrowed from other languages. It does not distinguish tones (as in some African languages) but uses a particular prosody (Souag, 2010).

3.2 Morphology

Wolof is an agglutinative language, which means that words are often formed by combining several morphemes. Personal pronouns, for example, can be modified according to the grammatical situation and social context. Unlike French, Wolof does not have grammatical genders (masculine/feminine) (McLaughlin, 2017).

3.3 Syntax

The basic syntactic structure of Wolof is Subject-Verb-Object (SVO), as in French. However, the word order can vary depending on the desired stress or accentuation (Torrence, 2013). The structure of Wolof sentences reflects a simple grammar, but its rich vocabulary allows for great expressiveness (Dione, 2021).

3.4 Vocabulary

Wolof has incorporated several lexical borrowings from French, Arabic and neighboring languages, such as Tarifiyt Berber, a variant of Amazigh language (Thiam, 2020) due to colonial, religious and commercial influences. For example, terms related to Islam and trade are often borrowed from Arabic.

4 Analysis of the Three Writing Systems

4.1 Latin alphabet

Since its standardization in the 1970s, the Latin alphabet has become the official writing system of the Wolof language, particularly in legal, educational, and administrative documents (Diagne, 2018). For examples:

- Bët (eye) / Bëtt (to pierce)
- Dég (thorn) / Dégg (to hear)

Its presence in digital resources, notably the translation of websites and software into Wolof, along with a significantly reduced illiteracy rate, contributes to its increasing use by new generations. It is written from left to right like French, English, etc. However, the Latin script remains unfamiliar to

most of the rural population, who rather use Ajami in their daily communications (Ngom, 2020).

4.2 Garay alphabet

The Garay script, which was designed by Assane Faye in 1961, is a transcription system adapted to the African phonetic characteristics of Wolof (Pandey, 2011). Garay, in the Figure 1, is written from right to left and includes special features such as nasalized letters and variations of uppercase and lowercase letters (Everson, 2016). Although Garay represents a significant innovation for the Wolof language, its usage remains limited due to a lack of institutional support and relative complexity.

	CONSONANTS						VOWELS			
	INITIAL	NON-INIT.	INITIAL	NON-INIT.	INITIAL	NON-INIT.	PROVISIONAL IDENTIFICATION			
[a] (1)	Ⲁ	ⲁ	w (8)	Ⲃ	ⲃ	y (40)	Ⲅ	ⲅ	a	/
c (2)	Ⲇ	ⲇ	l (9)	Ⲉ	ⲉ	t (50)	Ⲋ	ⲋ	e	'
m (3)	Ⲍ	ⲍ	g (10)	Ⲏ	ⲏ	r (60)	Ⲑ	ⲑ	e	ʹ
k (4)	Ⲓ	ⲓ	ng	Ⲕ	ⲕ	n (70)	Ⲗ	ⲗ	o	ʹʹ
b (5)	ⲙ	Ⲛ	v	ⲛ	Ⲝ	f (80)	Ⲟ	ⲟ	i	\
mb	Ⲡ	ⲡ	d (20)	Ⲣ	ⲣ	n (90)	Ⲥ	ⲥ	o	ʹ
j (6)	ⲧ	ⲩ	nd	ⲫ	Ⲭ	p (100)	ⲭ	Ⲯ	o	ʹʹʹ
nj	ⲯ	ⲱ	x (30)	ⲳ	Ⲵ	DIACRITICS			u	ʹʹ
s (7)	ⲵ	Ⲷ	h	Ⲹ	ⲹ	long vowel (postscript)	ⲿ			
						zero vowel (postscript)	Ⲿ	ⲿ	ⲿ	ⲿ
						double consonant (superscript)	ⲿ	ⲿ	ⲿ	ⲿ
NUMERALS 1 1 2 3 4 5 6 7 8 9 10										

Figure 1: Script of Assane Faye’s alphabet for the Wolof (Everson, 2016)

4.3 Ajami alphabet (Wolofal)

The history of the Ajami script, or Wolofal, in the Figure 2, dates back to the introduction of Arabic writing in West Africa through Quranic schools called "daara" since the earliest contacts between the local population and Arab-Muslim culture in the 8th and 9th centuries CE (Cissé, 2010). In Senegal, Ajami is used by speakers who have learned to read and write Arabic but do not necessarily know French or the Latin alphabet (Ngom, 2016).

5 Challenges of NLP about Ajami for Wolof

5.1 Linguistic Complexity and Lack of Standardization

Ajami for Wolof presents dialectal and stylistic variations, complicating normalization and automatic recognition (Gauthier et al., 2016). Each writer may have preferences in the use of certain letters or diacritics to transcribe specific sounds of Wolof that have no direct equivalent in Arabic.

5.2 Lack of Corpus and Language Models

NLP relies heavily on large annotated corpora, which are nonexistent for Wolof in Ajami. To address the lack of annotated corpora for Ajami beyond manual efforts, we propose two approaches such as community-driven methods and transfer learning from Arabic.

First, the community-driven approach involves developing a methodology for crowdsourcing annotations. Second, transfer learning from Arabic offers a technological solution to augment Ajami NLP capabilities. Pretrained Arabic language models can serve as a foundation and be adapted to Ajami-specific applications. This involves fine-tuning these models on Ajami text corpora to enhance their understanding of the script and its unique linguistic features. Additionally, domain adaptation techniques can be applied to bridge the structural and linguistic differences between Arabic and Ajami, ensuring the transfer is both efficient and effective.

Furthermore, current optical character recognition (OCR) systems for Arabic scripts are not easily adaptable to Ajami due to typographical divergences and specific diacritics (Nguer et al., 2020). OCR technologies are essential for digitizing ancient handwritten texts in Ajami (Mangeot and Sadat, 2014). Recent research in deep learning has shown progress for other languages using non-Latin scripts, but adjustments are needed to account for the particularities of Ajami (Mbaye and Diallo, 2023).

6 NLP advances and developing techniques for Wolof in Ajami

6.1 Transliteration models

To overcome the barrier between users of the Latin script and users of the Ajami script, several efforts have been launched to develop transliteration tools for this digraphy of the language. The Latin2Ajami algorithm (Fall et al., 2016) is an automatic transliteration tool that allows converting text written in Latin Wolof into Ajami script. The Latin2Ajami algorithm was used to evaluate a cutting-edge transliteration solution, which addresses the digraphy (the use of multiple writing systems for one language). Let’s take one example, of transliteration with Latin2Ajami, of the word "sañ-sañ", which means "right" or "freedom" in Wolof:

- Writing in Latin alphabet: sañ-sañ

- Transliteration in Ajami: سانسان

The algorithm performs this conversion by using a phonetic correspondence between the graphemes of the Latin alphabet and the characters of Ajami, independently of the pronunciation. The operation is reversible and therefore depends on the target writing system, but not on the language (Fall et al., 2016).

The scalability of the Latin2Ajami tool when working with larger datasets or addressing dialectal differences in Wolof involves several considerations such as data quality and diversity. To handle dialectal differences, this tool could face the phonetic variations, and the ambiguity in grapheme mapping due to the rich morphology of Wolof language. We suggest a plausible solution which consists of incorporating dialectal data into the training set and developing region-specific transliteration rules or a dialect-adaptive mode or hybrid systems.

6.2 Deep learning models

Some researchers are exploring deep learning models to train OCR systems for Ajami, using transfer learning approaches from models pre-trained on standard Arabic texts. Techniques such as convolutional neural networks (CNNs) and recurrent networks (RNNs) can be useful, but require specific data for Wolof in Ajami (Ignat et al., 2022; Cissé and Sadat, 2024).

6.3 Annotated Corpus Development

Manual annotation of existing Ajami texts is an essential step for the development of NLP models. Crowdsourcing, with native speakers and linguistic experts, can accelerate this process, especially for the transcription and translation of Ajami texts into Latinized Wolof or other working languages (Couty et al., 1968; Diagne, 2023).

6.4 Perspectives for Multilingual Models

Integrating multilingual models, such as BERT (Devlin, 2018) or XLM-R (Conneau, 2019), could provide an interesting solution by learning shared representations between different scripts. This may pave the way for models capable of processing both Ajami and Latinized Wolof, facilitating translation and information retrieval tasks (Dione et al., 2022; Cissé and Sadat, 2024).

7 The social and cultural role of Wolof

Wolof plays a central role in Senegalese culture and reflects the history, values and beliefs of the

people. Indeed, it is often the language of orality and tradition, conveying tales, proverbs and popular poems. Music, particularly mbalax (a Senegalese musical genre) and rap, is widely performed in Wolof and conveys social messages (Couty et al., 1968; Ngom, 2000).

In daily life, Wolof allows for social and linguistic cohesion between populations of different ethnic groups in Senegal. Its use in the media and in politics reinforces its importance in public spaces, thus contributing to the diffusion of Wolof culture beyond national borders (Cissé, 2010).

8 Preservation and contemporary challenges

Wolof faces several challenges in the context of globalization and the dominance of European languages such as French. The lack of written documentation and teaching materials is a barrier to the structured teaching of this language, especially for future generations (Hashimi, 2020).

However, efforts are being made for its preservation and standardization. Cultural and linguistic organizations in Senegal are working to create educational programs in Wolof, while the media and social networks contribute to its visibility and valorization. These initiatives aim to strengthen the use of Wolof not only in homes, but also in formal education and administration (Sinatti, 2014).

9 Conclusion and Perspectives

This paper presents a first study on the three writing systems used for Wolof. Ajami remains an important cultural medium for Senegalese Muslim communities, while Latin and Garay contribute to diversifying options for linguistic preservation, even though nowadays the Latin script is the most used due to public schools that have reduced illiteracy rates, enabling the majority of the Senegalese population to read and write the Wolof language.

The Wolof language, rich in culture and history, remains a pillar of Senegalese and West African identity. It plays a fundamental role in society, both as a language of communication and as an expression of cultural identity. Although faced with contemporary challenges, Wolof benefits from initiatives aimed at preserving and promoting this language. Recognition of its heritage value and support for its teaching could ensure its sustainability for future generations.

References

- Aminu Aliyu Abdullahi. 2022. Deep neural networks for the recognition of hausa ajami script. *Nigerian Journal of Computing, Engineering and Technology (NIJOCET)*, 1(2):106–111.
- Abdalla Uba Adamu. 2023. The gutenbergr principle: Hausa digital alàrammà and ajamization of knowledge.
- Dmitry Bondarev. 2019. Standardisation et variantes régionales de l’ajami haoussa. *Transliteration Studies*, 5:45–67.
- Momar Cissé. 2010. Parole chantée et communication sociale chez les wolof du sénégal.
- Thierno Ibrahima Cissé and Fatiha Sadat. 2024. Advancing language diversity and inclusion: Towards a neural network-based spell checker and correction for wolof. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages@ LREC-COLING 2024*, pages 140–151.
- Thierno Ibrahima Cissé and Fatiha Sadat. 2003. Automatic spell checker and correction for under-represented spoken languages: Case study on wolof.
- Thierno Ibrahima Cissé and Fatiha Sadat. 2024. Advancing language diversity and inclusion: Towards a neural network-based spell checker and correction for wolof.
- A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Philippe Couty, Jean Copans, and BK Dia. 1968. Contes wolof du baol. *Paris, ORSTOM*, page 1967.
- Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Abibatou Diagne. 2018. *La terminologie wolof dans une perspective de traduction et de combinatoire lexicale restreinte*. Ph.D. thesis, Université de Lyon.
- Abibatou Diagne. 2023. Constitution et usage de corpus pour langues terminologiquement peu documentées. *Littérature, Langues et Linguistique*, (14).
- Cheikh M Bamba Dione. 2021. Multilingual dependency parsing for low-resource african languages: Case studies on bambara, wolof, and yoruba. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 84–92.
- Cheikh M Bamba Dione, Alla Lo, Elhadji Mamadou Nguer, and Sileyé Ba. 2022. Low-resource neural machine translation: Benchmarking state-of-the-art transformer for wolof<-> french. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6654–6661.
- Michael Everson. 2016. Proposal for encoding the garay script in the smp of the ucs.
- El hadji M. Fall, El hadji Mamadou NGUER, Sokhna Bao Diop, Mouhamadou KHOULE, Mathieu MANGEOT, and Mame T. Cisse. 2016. Digraphie des langues ouest africaines: Latin2ajami: un algorithme de translittération automatique. In *Atelier Traitement Automatique des Langues Africaines TALAf 2016, conférence JEP-TALN-RECITAL 2016*.
- Elodie Gauthier, Laurent Besacier, and Sylvie Voisin. 2016. Speed perturbation and vowel duration modeling for asr in hausa and wolof languages. In *Inter-speech 2016*.
- Quintin Gee. 2005. Review of script displays of african languages by current software. *New review of hypermedia and multimedia*, 11(2):247–255.
- AO Hashimi. 2020. Ajami tradition in non-islamic society: The roles of ajami-arabic scripts in keeping records and documentation. *NIU Journal of Humanities*, 5(2):373–379.
- Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022. Ocr improves machine translation for low-resource languages. *arXiv preprint arXiv:2202.13274*.
- Isa Inuwa-Dutse. 2023. The first large scale collection of diverse hausa language datasets. *Proceedings of AfricaNLP workshop at ICLR2023*.
- Ian James. 2012. Garay script for wolof.
- British Library. 2020. *Hausa and Ajami Manuscripts: A Historical Perspective*. British Library Publishing, London.
- Mathieu Mangeot and Fatiha Sadat. 2014. Actes de l’atelier sur le traitement automatique des langues africaines talaf 2014.
- Derguene Mbaye and Moussa Diallo. 2023. Beqi: Revitalize the senegalese wolof language with a robust spelling corrector. *arXiv preprint arXiv:2305.08518*.
- Fiona McLaughlin. 2017. Ajami writing practices in atlantic-speaking africa. *The Atlantic Languages*, pages 1–30.
- Fallou Ngom. 2000. Sociolinguistic motivations of lexical borrowings in senegal.
- Fallou Ngom. 2004. Ethnic identity and linguistic hybridization in senegal.
- Fallou Ngom. 2016. *Muslims beyond the Arab world: The odyssey of ajami and the Muridiyya*. Oxford University Press.
- Fallou Ngom. 2020. Ajami scripts in the senegalese speech community. *Journal of Arabic and Islamic Studies*.

Fallou Ngom and Mustapha H. Kurfi. 2017. *Ajamization of Islam in Africa*, volume 8. Islamic Africa.

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.

Elhadji Mamadou Nguer, Alla Lo, Cheikh M Bamba Dione, Sileye O Ba, and Moussa Lo. 2020. Sencorpus: A french-wolof parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2803–2811.

Anshuman Pandey. 2011. Introducing the wolof alphabet of assane faye.

John Edward Philips. 2004. Hausa in the twentieth century: An overview. *Sudanic Africa*, 15:55–84.

Sokhna Sane. 2010. Decolonization and questions of language: The case of senegal. *HAGAR Studies in Culture, Polity and Identities*, 9(2):181–87.

Giulia Sinatti. 2014. 11 masculinities and intersectionality in migration: transnational wolof migrants negotiating manhood and gendered family roles. *Migration, gender and social justice: Perspectives on human insecurity*, pages 215–226.

Lameen Souag. 2010. Ajami in west africa. *Afrikanistik online*, 2010.

Alioune Badara Thiam. 2020. Typology of linguistic borrowing in the wolof language. pages 145–159.

Harold Torrence. 2013. The clause structure of wolof.

Valentin Vydrin. 2014. Ajami script for the mande languages valentin vydrin.

Appendix

Ajami Alphabet (Wolofal)

Phonetic Sounds	Arabic Orthography	Wolofal Orthography
1. [ʔ], [a:]	ا	ا
2. [b]	ب	ب
3. [t]	ت	ت
4. [θ]	ث	X
5. [dʒ], [j]	ج	چ
6. [h]	ح	ح
7. [x]	خ	خ
8. [d]	د	د
9. [ð]	ذ	X
10. [r]	ر	ر
11. [z]	ز	X
12. [s]	س	س
13. [ʃ]	ش	X
14. [sʰ]	ص	X
15. [dʰ]	ط	X
16. [tʰ]	ظ	X
17. [zʰ]	ظ	X
18. [ʕ]	ع	X
19. [y]	غ	X
20. [f]	ف، فـ	ف، فـ
21. [q]	ق، قـ	ق، قـ
22. [k]	ك	ك
23. [l]	ل	ل
24. [m]	م	م
25. [n]	ن	ن
26. [h]	ه	ه
27. [w], [u:]	و	و
28. [j], [i:]	ي	ي
29. [ʔ]	ء	ء ⁴

Figure 2: Arabic orthography, Wolofal orthography and their corresponding IPA symbols (Ngom, 2016)