Bidirectional AI-Powered Transliteration for Senegalese Languages: Bridging the Latin-Ajami Digraphia Gap

El Hadji Mamadou NGUER

Abstract- In Senegal and across West Africa, a significant linguistic di-graphia exists: local languages like Wolof are written in both a standard- ized Latin script and a supplemented Arabic script known as Ajami. This situation has created two separate ecosystems of knowledge and commu- nication, limiting information access for large portions of the population. While rule-based transliteration tools exist, they are often unidirectional (Latin to Ajami), limited in scope, and not based on modern artificial intelligence techniques. This paper presents a novel neural machine translation (NMT) approach to create a robust, context-aware, bidirectional transliteration model between Latin and Ajami scripts for Wolof. We propose a sequence-to-sequence (Seq2Seq) model with an attention mechanism, trained on a newly curated parallel corpus. Our model outperforms existing rule-based macros like Ajami70 in accuracy and handles complex linguistic phenomena such as gemination, nasalization, and vowel length more effectively. This work paves the way for seamless digital communication and knowledge sharing across script boundaries, directly contributing to digital inclusion.

Keywords-- Natural Language Processing, Transliteration, Digraphia, African Languages, Ajami, Neural Machine Translation, Senegal**.**

I. INTRODUCTION

The linguistic landscape of Senegal is characterized by a persistent digraphia [4]. National languages, with Wolof as the predominant vehicular language (spoken by over 80% of the population), are written in two distinct scripts: a standardized Latin alphabet, established by official decree, and a supplemented Arabic alphabet, known as Ajami or commonly named Wolofal [1]. The Latin script is predominantly used in official domains, ICT localization (e.g., Wolof web interfaces, software), and formal education. In contrast, the Ajami script thrives in informal yet vast ecosystems rooted in Quranic schools (daaras), used for daily commerce, literature, poetry, religious texts, and traditional medicine [9].

This digraphia has effectively created "two worlds that ignore each other" [9]. A large Ajami-literate population is excluded from digitally available knowledge resources written in Latin script, and vice-versa. This constitutes a major socio-economic barrier, hindering inclusive access to information, education, and modern communication tools (email, web, SMS).

Previous efforts to bridge this gap have focused on rule-based systems. The Ajami70 macro for Microsoft Word [12] offers fast Latin-to-Ajami conversion, and a Java-based application [7] provided early bidirectional capabilities, albeit without Harmonized Quranic Characters (HQC). These systems are valuable but suffer from limitations: they are often unidirectional, brittle to orthographic variations, difficult to maintain and extend, and lack the ability to learn from data and disambiguate context.

This paper argues for a paradigm shift from rule-based to data-driven, AI-powered transliteration. We present a bidirectional transliteration model based on a Neural Machine Translation (NMT) architecture, specifically a Recurrent Neural Network (RNN) based encoder-decoder model with an attention mechanism. This approach learns the complex mapping between the scripts directly from data, capturing subtle contextual rules and generalizing better to unseen words and variations.

The main contributions of this work are: 1. The formulation of the Latin-Ajami transliteration problem as a sequence-to-sequence learning task. 2. The creation of a curated parallel corpus of Latin and HQC Ajami texts for Wolof. 3. The design and implementation of a bidirectional NMT model for transliteration. 4. A comparative evaluation showing the superiority of the proposed NMT model over the existing rule-based Ajami70 macro in terms of accuracy and robustness.

II. RELATED WORK

The challenge of African language processing, including script conversion, has been documented [11]. For Senegalese languages, prior work has laid essential groundwork. Nguer2015LTC [9] extensively detailed the issues, challenges, and prospects of transliteration, highlighting the need for tools in text editors, web pages, emails, and SMS. Nguer2016TALAF [5] presented the Latin2Ajami algorithm, implemented as the efficient Ajami70 Word macro, which significantly improved processing speed over its predecessor, Ajami63. These rule-based systems operate primarily by using a handcrafted mapping table of Unicode characters and applying a set of pre-defined rules for diacritic placement (e.g., inserting sukun for a consonant without a vowel, shadda for gemination) [6]. While effective for straightforward conversions, they struggle with ambiguities and exceptions

Elhadji Mamadou NGUER (Author), Université Numérique Cheikh Hamidou KANE, Senegal

https://doi.org/10.17758/EARES23.F1225109

that are naturally handled by statistical or neural models. The Java-based transliterator [7] was an early attempt with a modular design but did not use the standardized HQC.

In the broader field, transliteration has been successfully tackled with NMT models, treating it as a character-level machine translation problem [2]. Seq2Seq models with attention have proven highly effective for transliteration between other script pairs (e.g., Cyrillic-Latin, Hindi-English) by learning alignments between input and output sequences. Our work is the first to apply this powerful framework to the Latin-Ajami digraphia problem in Senegal.

III. Methodology

A. Problem Formulation

We define transliteration as a character-level sequence transduction task. Given a source sequence of characters X=(x1,x2,...,xm) in script S1 (e.g., Latin), the model must learn to predict the target sequence Y=(y1,y2,...,yn) in script S2 (e.g., Ajami), where m and n may differ. The model should perform this mapping bidirectionally (S1↔S2).

B. Data Collection and Preprocessing

A significant challenge was acquiring a sufficient parallel corpus. We combined sources from: Religious texts (Quran, Bible) translated into Wolof in both scripts.

- Literary works and poetry available in dual scripts. The collaborative online Wolof dictionary project [10].
- Manual transliteration of selected Latin- text news articles into HQC Ajami by language experts.

The raw text was cleaned and normalized. The Ajami text was verified to comply with the Harmonized Quranic Characters (HQC) standard [8]. The final corpus was split into training (80%), validation (10%), and test (10%) sets.

C. Model Architecture

We employ an RNN-based encoder-decoder architecture with Bahdanau attention [3], which has become a standard for sequence-to-sequence tasks.

Encoder: A bidirectional Gated Recurrent Unit (GRU) layer. It processes the input sequence c_t (e.g., Latin characters) and encodes it into a sequence of hidden states h_i.

$$\overrightarrow{h_i} = GRU(x_i, \overrightarrow{h_{i-1}}), \quad \overleftarrow{h_i} = GRU(x_i, \overleftarrow{h_{i+1}})$$

$$h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$$

The final encoder output is the sequence of these concatenated hidden states.

Attention Mechanism: The attention layer calculates context vectors c_t for each decoding step t, allowing the decoder to focus on relevant parts of the input sequence.

$$\mathbf{c}_t = \sum_{i=1}^{m} \alpha_{t,i} \mathbf{h}_i$$

$$\alpha_{t,i} = \frac{\exp(score(\mathbf{s}_{t-1}, \mathbf{h}_i))}{\sum_{j=1}^{m} \exp(score(\mathbf{s}_{t-1}, \mathbf{h}_j))}$$

Where s_{t-1} is the previous decoder hidden state, and score is a feed forward network.

Decoder: Another GRU layer that generates the output sequence (Ajami characters) one token at a time, conditioned on the context vector from the attention mechanism and its previous state.

$$s_t = GRU\ (y_{t-1}, s_{t-1}, c_t)$$

The probability of the next character is computed by a softmax layer over the target vocabulary.

$$P(Y_t/Y_{<t},X)=softmax(W_o S_t+b_o)$$

We train two separate models: one for Latin-to-Ajami (*Lat2Aja*) and one for Ajami-to-Latin (*Aja2Lat*).

IV. Experiments and Results

A. Experimental Setup

We implemented our models using the TensorFlow library. Characters were embedded into 64-dimensional vectors. The GRU layers had 256 units each. Models were trained using the Adam optimizer with a learning rate of 0.001 and a batch size of 64. We used categorical cross-entropy as the loss function. Training was stopped early if the validation loss did not improve for 10 consecutive epochs.

We compared our NMT model (NMT-Translit) against the rule-based Ajami70 macro [12] (for Latin-to-Ajami only).

B. Evaluation Metrics

We used Character Error Rate (CER) and Word Error Rate (WER), standard metrics in transliteration and speech recognition, to evaluate performance quantitatively.

$$CER = \frac{Numberofcharacteroperations(insertions, deletions, substitutions)}{Totalnumberofcharactersintarget} \times 100\%$$

$$WER = \frac{Numberofincorrectwords}{Totalnumberofwords} \times 100\%$$

Lower scores indicate better performance.

C. Results and Analysis

Table 1 clearly shows that our proposed NMT model significantly outperforms the rule-based baseline, reducing CER and WER by more than half for the Latin-to-Ajami task. Furthermore, it provides high-quality Ajami-to-Latin transliteration, a feature absent in Ajami70.

https://doi.org/10.17758/EARES23.F1225109

TABLE I: PERFORMANCE COMPARISON ON THE TEST SET.

Direction	Model	CER(%)	WER (%)
Latin-to-Ajami	Ajami70(Rule-baed	8.4	22.1
	NMT-Translit(Ours)	3.7	11.8
Ajami-to-Latin	NMT-Translit(Ours)	2.9	9.2

Qualitative analysis revealed that the NMT model excelled where rule-based systems falter: - Contextual Disambiguation: Correctly handling characters like ’c’ (e.g., in car ’branch’ vs. proper names). - Handling OOV Words:

Better generalizing to words not seen in the training data by leveraging subword similarities. - Diacritic Prediction: More accurately placing sukuun and shadda based on learned context rather than rigid rules.

V. Conclusion And Perspectives

This paper addressed the critical challenge of digraphia in Senegal by introducing a novel Al-powered solution for bidirectional transliteration between Latin and Ajami scripts. We moved beyond existing rule-based approaches by formulating the problem within an NMT framework and training a robust Seq2Seq model with attention on a curated parallel corpus.

Our results demonstrate a substantial improvement in accuracy over the state-of-the-art rule-based system, effectively reducing the error rate by more than 50%. This model is a crucial step towards breaking down the digital and informational barriers between the Latin-literate and Ajami-literate communities in Senegal.

Future work will focus on several areas: 1. Expanding Language Coverage: Adapting the model to other Senegalese languages (Pulaar, Seereer, etc.). 2. Developing Integrated Tools: Implementing the model as a web service API and plugins for popular platforms (e.g., web browsers with realtime page transliteration, email clients, mobile keyboards). 3. Active Learning for Data Collection: Deploying the model in a human-in-the-loop system to continuously collect correction data from users to improve itself and expand the corpus. 4. Exploring Advanced Architectures: Investigating the use of Transformer models [13], which have recently set new standards in NMT, potentially offering further gains in accuracy and efficiency.

This research underscores the potential of AI and NLP technologies to address sociolinguistic challenges and promote digital inclusion for underrepresented languages and scripts.

References

[1] Décret n° 2005-992 du 21 octobre 2005, relatif à l'orthographe et la séparation des mots en wolof (2005), http://www.jo.gouv.sn/spip.php?article4802

[2] Abbas Malik, M.: Methods and tools for weak problems of translation (2010), english. <tel-00502192>

[3] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR abs/1409.0473 (2015), https://arxiv.org/abs/1409.0473

[4] Cissé, M.: Ecrits et écriture en afrique de l'ouest. Revue électronique internationale de sciences du langage Sudlangues (6) (2006), https://www.scribd.com/document/664937900/doc-135-1

[5] Fall, E.h.M., Nguer, E.h.M., Bao Diop, S., Khoule, M., Mangeot, M., Cisse, M.T.: Digraphie des langues ouest africaines: Latin2ajami un algorithme de translit- tération automatique. In: Actes de la conférence conjointe JEP-TALN-RECITAL 2016, volume 11: TALAF. pp. 62-73 (2016)

[6] Fall, E.h.M.: Transliteration automatique latin-ajami d'un document texte wolof. Master's thesis, Université Gaston Berger, Sénégal (2015)

[7] Gueye, S.T., Fall, T.M.G.: La transliteration automatique wolof wolofal en java. Master's thesis, Université Gaston Berger, Sénégal (2011)

[8] LO, C., Fall, E.h.M.: Les Caractères Coraniques Harmonisées. 2e édition edn. (2010)

[9] Nguer, E.h.M., Bao-Diop, S., Fall, Y.A., Khoule, M.: Digraph of senegal's local languages: issues, challenges and prospects of their transliteration. In: LTC 2015 (2015)

[10] Nguer, E.h.M., Khoulé, M., Thiam, M.N., Mbaye, B.T., Thiaré, O., Cissé, M.T., Mangeot, M.: Dictionnaires wolof en ligne état de l'art et perspectives. In: CNRIA 2014 (2014)

[11] Osborn, D.Z.: Les langues africaines et la technologie de l'information et de com- munication: localiser le futur? (2005)

[12] Paul-timothy: Ajami edition tools for senegal's local languages (2015), http://currah.download/pages/ajamisenegal/index.html

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N.. Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neu- ral Information Processing Systems 30: Annual Conference on Neural In- formation Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998-6008 (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf