Opportunities and Challenges of Natural Language Processing for Low-Resource Senegalese Languages in Social Science Research

Derguene Mbaye * , Tatiana D. P. Mbengue , Madoune R. Seye , Moussa Diallo , Mamadou L. Ndiaye , Dimitri S. Adjanohoun , Djiby Sow , Cheikh S. Wade , Jean-Claude B. Munyaka , Jerome Chenal

Posted Date: 15 January 2026

doi: 10.20944/preprints202601.1124.v1

Keywords: natural language processing; low-resource African languages; Senegalese languages; computational social science

Article

Opportunities and Challenges of Natural Language Processing for Low-Resource SENEGALESE Languages in Social Science Research

Derguene Mbaye [1,*], Tatiana D. P. Mbengue [2], Madoune R. Seye [1], Moussa Diallo [1], Mamadou L. Ndiaye [1], Dimitri S. Adjanohoun [2], Djiby Sow [2], Cheikh S. Wade [2], Jean-Claude B. Munyaka [3] and Jerome Chenal [3]

[1] Polytechnic School (ESP), Dakar, Senegal
[2] Gaston Berger University (UGB), Saint Louis, Senegal
[3] Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland
* Correspondence: mbayederguene@hotmail.fr

Abstract

Natural Language Processing (NLP) is rapidly transforming research methodologies across disciplines, yet African languages remain largely underrepresented in this technological shift. This paper provides the first comprehensive overview of NLP progress and challenges for the six national languages officially recognized by the Senegalese Constitution: `Wolof`, `Pulaar`, `Sérère`, `Diola`, `Mandingue`, and `Soninké`. We synthesize linguistic, sociotechnical, and infrastructural factors that shape their digital readiness and identify gaps in data, tools, and benchmarks. Building on existing initiatives and research works, we analyze ongoing efforts in text normalization, machine translation, and speech processing. We also provide a centralized GitHub repository that compiles publicly accessible resources for a range of NLP tasks across these languages, designed to facilitate collaboration and reproducibility. A special focus is devoted to the application of NLP to the `social sciences`, where multilingual transcription, translation, and retrieval pipelines can significantly enhance the efficiency and inclusiveness of field research. The paper concludes by outlining a roadmap toward sustainable, community-centered NLP ecosystems for Senegalese languages, emphasizing ethical data governance, open resources, and interdisciplinary collaboration.

Keywords: natural language processing; low-resource african languages; senegalese languages; computational social science

1. Introduction

Natural Language Processing (NLP) has emerged as a transformative field within artificial intelligence, enabling machines to process and understand human language at scale. In recent years, its applications have profoundly influenced research across disciplines, from computational linguistics and digital humanities to sociology and political science. However, the vast majority of NLP advances have been concentrated on a small set of `high-resource` languages, leaving most African (`low-ressource`) languages under-represented in both datasets and algorithmic development [1]. However, the term "low resource" can cover several dimensions and is not limited to language alone: it can also refer to `domains` or `tasks` for which little data is available, even when the language in question is rich in resources. This is particularly evident in [2], where the concept of "low resource" is defined according to three distinct aspects:

- The availability of **task-specific** annotations ;
- The existence of **unannotated** texts in the language ;
- The presence of **auxiliary** data.

 doi:10.20944/preprints202601.1124.v1

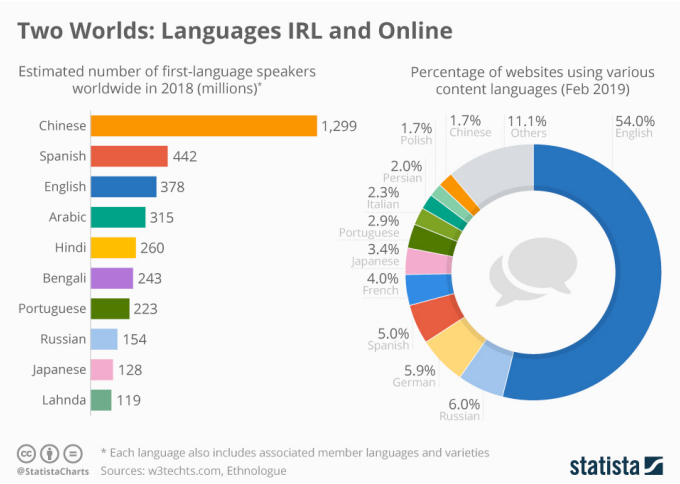

Figure 1. Contrast between the world's most spoken languages and their representation online.

As illustrated in [3], the majority of African languages fit this definition, which complicates the work of researchers, and contributes to their under-representation in Natural Language Processing (NLP) research [4]. Evaluation results from the `SAHARA` benchmark [5], which assesses 517 African languages across multiple NLP tasks, reveal a pronounced performance gap: English consistently ranks among the highest-performing languages, whereas many widely spoken African languages, such as **Fulfulde** (Fula or Pulaar), **Wolof**, Hausa, Oromo, and Kinyarwanda, systematically underperform across reasoning, generation, and classification tasks. These widening linguistic inequities in AI capabilities is particularly alarming as a new type of digital divide[1] is emerging, which now concerns `the extent to which languages are represented and processed by AI systems` [6].

Senegal, a multilingual nation with over twenty languages, constitutionally recognizes six national languages[2]: **Wolof**, **Pulaar**, **Sérère**, **Diola**, **Mandingue**, and **Soninké**, as central to its cultural and civic identity. Although a minority, there are also populations of Arabic (Afro-asiatic) speakers, including those who speak `Hassaniyya` (Mauritanian dialect of Arabic), as well as `Levantine` and `Moroccan dialects`. `Portuguese Creole` is also spoken in some parts of Casamance, and in Dakar by immigrant and migrant populations from the Cape Verde islands and Casamance respectively [7]. However, `French` is the dominant European language in Senegal, being recognized as the official language and the one used in education [7]. In this article, we will only focus on the 06 main local languages recognized as `national` languages and leave the other ones for future research.

Despite their societal importance and widespread use in daily communication and media, these languages remain largely excluded from the digital and scientific landscape of NLP. This gap poses a dual challenge: the risk of technological marginalization of major linguistic communities, and the missed opportunity to harness NLP for advancing locally grounded research, especially in the social sciences.

Social science research in Senegal relies heavily on qualitative methods, including interviews, focus groups, and ethnographic recordings [8] often multilingual and resource-intensive to transcribe, translate, and analyze. The integration of NLP pipelines into these workflows could dramatically improve efficiency, accessibility, and analytic depth. Yet, realizing this potential requires robust linguistic resources, interoperable tools, and sustainable community infrastructures. Linguistic inequity in AI constitutes a structural issue rather than a marginal fairness concern, as it directly shapes access to reliable information, the ability to challenge decisions, and `meaningful participation in democratic processes` [6]. This paper therefore seeks to **(i)** provide a systematic overview of existing NLP research and resources for Senegalese `national` languages, **(ii)** identify structural and methodological challenges impeding progress, and **(iii)** explore the opportunities of applying NLP

[1] Traditionally been framed in terms of access to internet connectivity.
[2] Presidential Decree No 71-566 of May 21, 1971.

 doi:10.20944/preprints202601.1124.v1

to the social sciences. Drawing inspiration from the `Ethiopian NLP` ecosystem [9], we adapt this comparative framework to the Senegalese context. In addition, we introduce a centralized and openly accessible repository on GitHub[3] that compiles existing datasets, benchmarks, and tools available for these languages. The repository is designed as a living resource to be periodically expanded through community contributions. Our objective is to map existing efforts, identify critical research gaps, and encourage the development of sustainable, inclusive NLP research for Senegal's national languages.

2. Sociolinguistic and Linguistic Features

The Senegalese linguistic landscape is characterized by considerable diversity, with approximately 30 national languages, most of which belong to the `Niger-Congo language family` [10]. Among these, 06 languages: **Wolof**, **Pulaar (Fula)**, **Sereer**, **Diola (Joola)**, **Mandinka (Malinké)**, and **Soninke**; occupy a particularly prominent position. As early as 1971, Presidential Decree No. `71-566 of May 21, 1971` formally recognized these languages as **"national languages"**, while French retains the status of the country's "official language". These languages have officially standardized orthographies recognized and enforced by the Senegalese state, making them, in principle, suitable for integration into the educational system. However, this integration remains very weak, further work remaining necessary, particularly with regard to pedagogical standardization and the production of appropriate teaching materials [10]. From a demographic perspective, only `Wolof`, `Pulaar`, `Serer`, and `Mandinka` are spoken by more than one million speakers[4] as illustrated in Figure 2. However, many of these languages, especially Pulaar and Mandinka, are characterized by substantial dialectal variation, which poses additional challenges for linguistic standardization and computational processing.

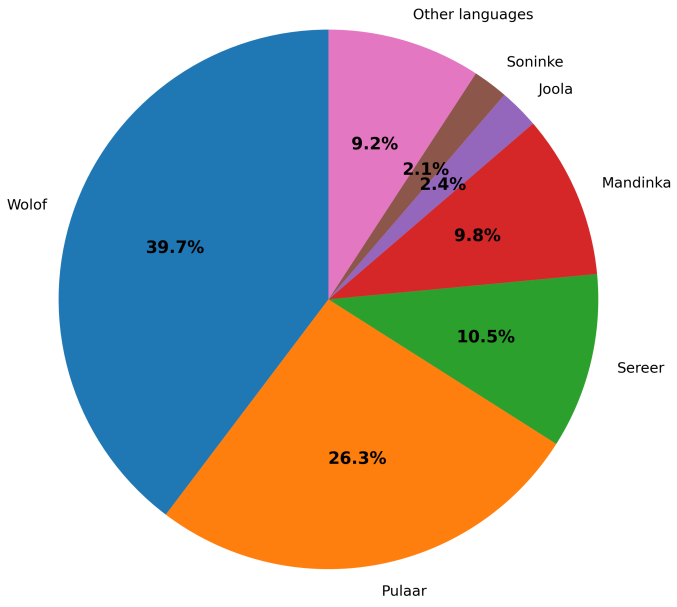

Figure 2. Proportion of speakers of the main local languages in terms of percentages [10].

2.1. Language Digraphs

Although the `Latin script` is the most widely used, some languages such as Wolof, Pulaar, Mandinka and Soninké also have an Arabic-based alphabet called `Ajami` [12]. It results from the early Islamization of the major Muslim ethnic groups in the country, and remains an important means of written communication among people who are illiterate in French and who have attended Quranic schools [13]. This phenomenon has thus created two distinct worlds that do not converge, and each

[3] https://github.com/DerXter/State-of-NLP-Research-in-Senegal
[4] Senegal's population is approximately 18.5 million people, with recent estimates [11].

 doi:10.20944/preprints202601.1124.v1

writing system has its own applications. While the Ajami script is mainly used in religious contexts and traditional medicine, the Latin script is widely used in the digital world, particularly for the localization of online platforms [14].

The history of the Ajami script has been explored in [15], as well as its use and its modern writings. Authors analyzed the challenges and prospects of these systems from the perspective of language preservation and highlighted the potential of this script to represent an important instrument for literacy and digital inclusion in sub-Saharan Africa. Therefore, considerable efforts have been made in order to promote its scriptural rehabilitation through `transliteration`[5]. Challenges in Latin-Ajami transliteration have been explored in [14] with a purpose of involving people using the Arabic alphabet within a collaborative dictionnary project. The creation of `Latin2Ajami`: a transliteration algorithm from latin towards Ajami, has been studied in [16] with an approach based on a correspondence table, whose data comes from an external editable file. The `AjamiXTranslit` project went further by offering a collaborative data collection platform for native speakers and a publicly available multilingual corpus of Latin–Ajami text pairs along with annotated manuscripts [12]. The authors also introduced automatic transliteration and optical character recognition (OCR) models adapted to the graphic diversity of Ajami. Although small in size (the largest of the Senegalese languages < 70 rows), this is the corpus with the widest coverage of Senegalese languages (Wolof, Pulaar, and Soninké) in Ajami script.

2.2. Language Overview

While most of the senegalese languages belong to the Niger–Congo language family, the majority of them belong either to the `West Atlantic group` or `the Mandinka group` [7]. They encompass a wide range of phonological, morphological, and orthographic systems and are more localized in different areas in Senegal, as illustrated in the Figure 3. This section outlines the main sociolinguistic profiles of these dominant languages in Senegal as well as linguistic characteristics relevant to NLP development.

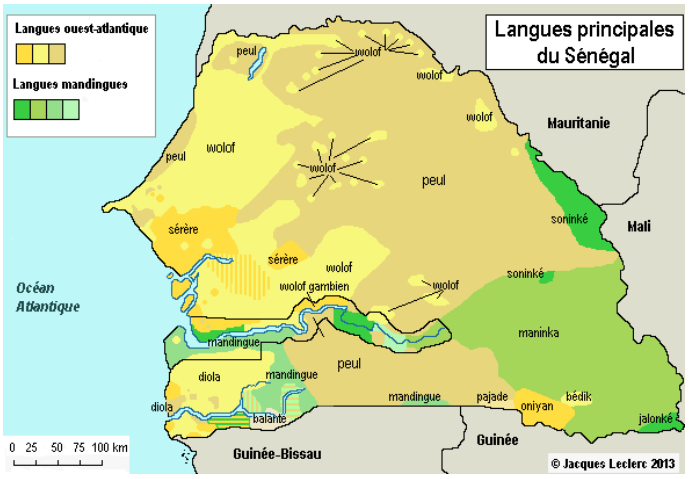

Figure 3. Main Senegalese languages and their locations in the country [10].

2.2.1. Wolof

Wolof serves as the dominant lingua franca of Senegal, spoken by more than 80% of the population `either as a first or second language`. It belongs to the Atlantic branch of the Niger–Congo family and exhibits rich morphophonemic alternations, vowel harmony, and extensive use of nasalization. Orthographic variation remains a challenge, especially around diacritics such as *ñ* and *ŋ* ("ng"). Due to its sociolinguistic reach and early digitalization efforts, Wolof has become the most resourced Senegalese language in NLP research to date as illustrated in Tab 1. There are two main geograph-

[5] Process of representing a word, phrase, or text in a different script or writing system.

 doi:10.20944/preprints202601.1124.v1

ical varieties of Wolof: one spoken in Senegal, and the other in Gambia [17]. Although Wolof speakers understand each other, Senegalese Wolof and Gambian Wolof are considered two distinct languages, each with its own ISO 639-3 language code: WOL for Senegalese Wolof and WOF for the Gambian one [18]. Wolof has also been adapted to the Latin alphabet, despite having a long written tradition using the Arabic alphabet in the form of Ajami (or Wolofal) [15]. Despite showing remarkably little variation accross dialects, the major contemporary Wolof dialectal divide is between rural and urban varieties. The latter has extensively borrowed from French as the result of language contact [7] which led to the code-switching phenomenon, making it more challenging to process using computer tools [19].

2.2.2. Pulaar / Fula

Pulaar, also known as **Fula**[6] or Fulfulde, is part of the Atlantic family and is spoken across West and Central Africa. In Senegal, it is mainly concentrated in the Fouta Toro region (cf Fig 3) whose dialect being the most dominant one across several mutually intelligible varieties [7]. Pulaar exhibits complex noun class systems, consisting of more than 20 classes in some dialects of Pulaar, a large set of verbal extensions, and morphologically conditioned consonant mutation.

Other dialects spoken in Senegal include the Casamance dialects such as Fulakunda and Fulaadu, as well as the Fuuta Jalon dialect spoken by the substantial population of Guineans living in Senegal [7]. Its transnational presence makes it an ideal candidate for regional NLP initiatives, although orthographic harmonization across borders remains incomplete.

2.2.3. Sérère (Sereer)

Sérère, belongs to the northern branch of the Atlantic family of the Niger-Congo phylum, which makes it related to Wolof and especially Fula [20]. The different varieties of Serer are spoken by more than 1 million people in Senegal and Gambia (cf Fig 2), but it is important to note that this term also refers to populations in eastern Senegal who are culturally similar but speak Cangin languages[7]. Recognized as one of the national languages of Senegal, it has an official writing system based on the variety known as Seereer Siin, meaning "Sereer from the Sine region" between the **Petite Côte** south of Dakar and Gambia, which has become a kind of “standard” Sereer [20]. Despite its significant speaker base, it remains severely underrepresented in digital corpora. Documentation efforts are growing through community and academic collaborations, but resources for NLP applications remain still minimal (cf Tab 1).

2.2.4. Diola (Joola)

The Joola group comprises several dialects such as Joola Foñi (dominant dialect) or Kujamaat Joola, as well as Boulouf, Fogny and Kasa [10]. It belongs to the Bak branch of Atlantic [7] and is primarily spoken in the Casamance region (cf Fig 3). It shares certain typological features with Wolof, Pulaar and Sereer (Pulaar’s closest language), and are unusual among Niger-Congo languages in that they are not tonal [7]. Joola’s morphosyntactic diversity make it particularly challenging for corpus alignment, and limited orthographic standardization and dialectal fragmentation also contribute to data scarcity. It is rarely used in writing but rather on local radio stations, limiting its use to oral communication in everyday life [21]. However, ongoing linguistic documentation projects are beginning to fill the gap.

2.2.5. Mandingue (Mandinka)

A member of the Mande language family, Mandingue is widely spoken in Senegal’s southern regions and across neighboring countries (cf Fig 3). About half of Mandinka speakers live in Gambia, where Mandinka is the dominant language nationwide [22]. With Wolof, Pulaar and Sereer, they

[6] This variant is usually used in NLP research papers.
[7] https://en.wikipedia.org/wiki/Cangin_languages

 doi:10.20944/preprints202601.1124.v1

represent the dominant languages in Senegal in terms of numbers of speakers (cf Fig 2). The most salient features of Mandinka are very similar to those of other Manding languages (particularly **Bambara**) such as **(1)** a `tonal` system based on the opposition between high and low tones, **(2)** the virtual absence of syllables ending in a consonant, with the exception of syllables ending in the nasal η, **(3)** a very limited range of morphological inflection, **(4)** the absence of grammatical gender and, **(5)** an extremely rigid word order [22]. Its relatively stable morphology and regional presence, offer potential for `transfer learning`[8] from related Mande languages such as `Bambara`, in which substantial NLP resources are being developed [23].

2.2.6. Soninké

Soninké, one of the oldest written languages in West Africa, belongs to the `Mande` family and retains strong oral traditions. It plays a key role in historical, cultural, and economic exchanges across `Mali`, `Mauritania`, and `Senegal` [24]. Soninké is a **tonal language**, in which each syllable is characterized by a musical pitch, either high (´) or low (‘), and represents somewhat of an outlier, as it has little mutual intelligibility with other Mande languages [7]. However, the differences between the various dialectal varieties `within Soninke` are relatively minor and do not hinder mutual intelligibility [24]. Moreover, here is no standard form of Soninke, nor is there a dialectal variety of this language that is recognized as more dominant than the others [24]. While limited digital resources exist, Soninké's cross-border usage and structured morphology make it a promising target for multilingual modeling.

3. Low-Resource Context and Data Availability

Senegal's NLP ecosystem remains in an early but dynamic stage of development. While data scarcity and resource fragmentation continue to constrain progress, significant institutional and community-driven initiatives are emerging to bridge this gap. The establishment of the **national AI strategy** has marked a pivotal step toward building national capacity in artificial intelligence and data governance [25]. From an **academic** standpoint, several institutions have contributed to the early stages of resource development and linguistic documentation, supporting foundational research and training in machine learning and natural language processing [26]. **Community-driven movements** have also been instrumental. **GalsenAI** and **Masakhane** have spearheaded open, collaborative data collection and multilingual modeling across African languages [27]. These networks have facilitated the creation of various datasets, contributing to the first generation of open benchmarks for low-resource NLP. **Private companies** and **start-ups** also contribute to the development of NLP resources [28]. Table 1 summarizes the publicly known datasets, corpora, and tools relevant to Senegalese national languages, highlighting `the current imbalance between Wolof and other local languages`. A separate GitHub repository[9] has also been set up to facilitate the tracking of datasets continuously produced in these languages. It also includes additional `data sources`.

4. Current NLP Efforts and Tasks

Senegal's participation in African and global NLP initiatives is steadily growing, with empirical studies and open datasets focusing on the country's main national languages. This section synthesizes the current state of progress across major NLP tasks that form the core building blocks for future applied research, particularly in multilingual and interdisciplinary contexts.

4.1. Parsing & Tokenization

Parsing is a foundational component of modern NLP, enabling systems to read, generate, and interpret text with unprecedented accuracy. It consists of analyzing the grammatical structure of a `sentence` or `text`, and identifies the roles of words such as `nouns`, `verbs`, and `adjectives`, and maps

[8] A technique in machine learning (ML) in which knowledge learned from a task/language, is re-used in order to boost performance on a related task/language.

[9] https://github.com/WolofProcessing/online_wolof_data.

 doi:10.20944/preprints202601.1124.v1

Table 1. Publicly available `datasets` and `tools` for Senegalese **national** languages covering Machine Translation, Token Classification, QA/Instruct/Pre-Training, Sentence Classification, Automatic Speech Recognition, Speech Synthesis, Keyword Spotting, Language Identification, Morphological features tagging and alignment evaluation, and Masked Language Modeling.

Language	Text corpora	Speech corpora	Existing tools
Wolof	OPUS, FLORES-200, NTREX, LORELEI, MAFAND-MT, SMOL, MADLAD-400, MasakhaNER, Masakha-POS Universal Dependencies, AfriQA, AYA, AfriWOZ-1.0, Belebele, FineWeb2, AWOFRO, WolBanking77, Masakhane-NLU, AjamiXTranslit	AI4D-Urban, ALFFA, Wol-Banking77, Waxal, Fleurs, Kallaama, AI4D-Baamtu, Keyword Spotting	Wolof keyboards, Wolof library, Stanza, MorphScore, Common Voice, Dvoice, AfroLID,GlotLID
Pulaar	AjamiXTranslit, FineWeb2, Fulani-English Pair Dataset, MADLAD-400, SMOL	Kallaama, Keyword Spotting	-
Sereer	-	Kallaama, Keyword Spotting	-
Joola	-	Keyword Spotting	-
Mandinka	-	Keyword Spotting	-
Soninké	AjamiXTranslit	Keyword Spotting	-

 doi:10.20944/preprints202601.1124.v1

the relationships that link them together [29]. Parsing generally assumes that the input has already been tokenized i.e. broken down into tokens (words, subwords, or symbols). Traditional syntactic parsing (dependencies, constituency) for examples, relies on Word-level tokenization to know the structured units to attach syntactic roles to [30].

The design and implementation of a morphological analyzer for Wolof has been reported in [31] in order to obtain a linguistically motivated tool using finite-state techniques. As a foundational step toward an LFG-based[10] computational grammar for Wolof, the authors introduced a newly constructed Finite-State Transducer (FST) for the language and presented experimental evaluations assessing the analyzer's performance across several statistical metrics. To cope with the challenging clitics[11] phenomenon in Wolof, [32] adopted a non-transformational approach grounded in LFG to avoid verb-movement rules and preserves lexical integrity. The study includes an implementation showing how LFG effectively captures the behavior of Wolof clitics in order to demonstrate its practicality. Therefore, a finite-state transducer (FST) designed to tokenize and normalize natural-text input for a large-scale Wolof LFG grammar has been presented in [33]. An initial language-independent tokenizer proved insufficient as issues with multiword expressions, clitics, and normalization emerged. Integrating FST components resolved these problems, enabling the grammar to handle free text more effectively and improving overall parsing performance. [34] described an LFG-based parsing system for Wolof that handles major grammatical constructions. The system relies on FST for tokenization and morphological analysis, supported by lexicons and robust parsing strategies such as fragmenting and skimming. The system demonstrated solid parsing coverage on real Wolof data and achieved competitive accuracy after manual disambiguation, with strong precision and an overall balanced performance. [35] presented the development of a multilingual parallel treebank spanning 10 languages including Wolof from 06 language families. Built using deep LFG grammars created within the ParGram project, the system produces highly parallel syntactic analyses across languages, that form the foundation of a richly annotated treebank. The analyses capture a wide range of linguistic phenomena and all produced resources are publicly accessible. To accelerate the LFG parsing process, [36] introduced an efficient method that includes a Constraint Grammar (CG) parser into a probabilistic context-free grammar. Experiments demonstrated substantial gains in efficiency and robustness when annotating Wolof running text. Authors then presented a set of techniques for handling ambiguity in LFG parsing of Wolof in [37] by addressing both theoretical and practical concerns. The study explored multiple avenues to build a large-scale Wolof grammar and developed strategies that enhance the grammar's efficiency, robustness, and coverage. The first Universal Dependencies (UD) resource within the Northern Atlantic branch of the Niger-Congo family has been presented in [38]. Various challenges related to word segmentation for tokenization and the mapping of Part-of-speech (PoS) tags (cf Section 4.2), morphological features, and dependency relations to existing Wolof annotation conventions are reported. Several characteristic constructions are also outlined as a basis for discussing broader UD guidelines. This work has had a huge impact, enabling Wolof to be supported in NLP tools such as Stanza [39], or in larger studies such as [40], which evaluated Morphological Alignment of Tokenizers in 70 Languages. Authors introduced a tokenizer evaluation metric named Morphscore, which assesses tokenizers morphological alignment and help fill the performance gap between agglutinative languages like Wolof and fusional languages like English. A systematic method for deriving Universal Dependencies from LFG structures has been presented [41]. Several challenges encountered by existing algorithms when applied to Wolof are discussed, along with the strategies adopted to address them. Evaluation results indicated that the approach achieved high accuracy and represented a clear improvement over earlier conversion methods. [42] leveraged multilingual embeddings and UD treebanks from both high-resource and low-resource languages including Wolof, to introduce a syntactic knowledge transfer method which allows to predict a wide range of UD annotations and

[10] Lexical Functional Grammar.
[11] A morpheme that has syntactic characteristics of a word, but depends phonologically on another word or phrase.

 doi:10.20944/preprints202601.1124.v1

dependency trees. xperiments showed that combining high-resource languages with low-resource languages in contemporary contact, leads to better performance than pairing unrelated languages.

4.2. Token Classification

Token classification is a natural language understanding (NLU) task where a specific label is assigned to each token (word or sub-word) in a text. It is used for tasks such as `Named Entity Recognition (NER)` to identify names, dates, and places, and `Part-of-Speech (POS)` tagging to categorize words as nouns, verbs, adjectives, and so on [43].

Regarding NER, [44] introduced the first large, publicly available, high-quality dataset in 10 African languages including Wolof[12]. Authors train and evaluate multiple NER models and conducted an extensive empirical evaluation of state-of-the-art methods across both supervised and transfer learning settings. Similarly, very little work has been done in POS tagging in Senegalese languages. The design of a part-of-speech-tagset for Wolof has been reported in [46] alongside with an efficient process of creating a semi-automatically annotated gold standard. Authors leveraged available lexical resources and used purpose-built heuristic tools for stemming and guessing of word forms. They evaluated afterwards the performance of state-of-the-art statistical PoS taggers on the collected data, and briefly summarize cross-lingual projection experiments utilizing the parallel corpus data. `AfricaPOS`, the largest part-of-speech (POS) dataset for 20 typologically diverse African languages including Wolof has been introduced in [47]. Researchers conducted extensive POS baseline experiments using both Conditional Random Field (CRF) and several multilingual pre-trained language models and discussed the challenges in annotating POS for these languages using the universal dependencies (UD) guidelines. All ressources (data, code, and models) has been released to inspire future research in NLP on these languages.

4.3. Text Classification

Unlike token classification which assigns a label to each token, text classification assigns a single label to an entire sentence. It is a core task in NLP used for applications like `spam filtering`, `sentiment analysis`, and `news categorization` to organize and analyze large volumes of unstructured text [48].

4.3.1. Sentiment Analysis

Sentiment analysis (also known as opinion mining) is the process of using NLP to identify and extract subjective information from text to determine the author's emotional tone as `positive`, `negative`, or `neutral` [49]. It is very helpful to understand public opinion on products, services, and brands by analyzing data like customer reviews, support tickets, and social media comments. [50] conducted a survey on the term weighting schemes as it represents the crucial step for representing the documents in a suitable way for classification algorithms. Authors proposed an efficient term weighting scheme that provide better classification accuracy than `TF-IDF` [51] and `IF-IGM` [52]. As code-switching is quite common in Senegalese languages with the presence of French [53], an extended lexicon with French and Wolof words and expressions used in both languages was proposed in [54]. Researchers introduced a sentiment scoring algorithm named `FWLSA-score`[13], that uses word similarity to address the spelling problem, and classifies reviews as positive or negative based on the polarity of the words or expressions. An improvement over the FWLSA-score has subsequently been studied in [55] with word-level trigrams (list of consecutive three letters) in order to improve its effectiveness on verbs conjugation and words declination in both languages. However, relying on word-level similarity and trigrams to map inflected/derived forms to a base form, are complex and limited, especially for morphologically complex Wolof and French words written using the French alphabet. To advance bilingual French-Wolof sentiment analysis, the authors in [56] proposed `(i)` a new French-Wolof dictionary and dictionary-based lemmatizer to accurately handle morphology, and `(ii)` a novel

[12] The coverage has subsequently been extended to 20 African languages in [45] with no additional Senegalese languages.
[13] FWLSA = French and Wolof Lexicon-based Sentiment Analysis.

 doi:10.20944/preprints202601.1124.v1

Markov Model-based method for identifying context-dependent sentiment words. To address the complexity related to ambiguity in Senegalese on-line press comments, [57] suggested an opinion lexicon with Wolof, French and urban language words and expressions, to process these types of data. Authors therefore aim to pave the way for the development of tools for processing local languages. Another approach, based on graph structures has been studied in [58] to address the same challenge. Researchers proposed a dictionary-based lemmatizer and an graph algorithm to model the relationship between French and Wolof opinion words. COMFO, a multilingual corpus (French, English, and Urban Wolof) for opinion mining, has been introduced in [59] to facilitate the exploration of supervised learning algorithms for sentiment classification. The authors detailed the corpus collection process, covering the data source, preparation, and the lexical-based annotation approach used. In [60] we collected a set of X (formely Twitter) and Facebook comments related to the youth's perception about the mobile internet costs in Senegal and applied sentiment analysis to gather their general feelings. We leveraged a multilingual language model based on XLM-Roberta [61] and pre-trained on nearly 200 million Tweets across some 30 languages (including French). Domain-specific model (in this case, social media) is more effective than its general counterpart when it comes to refining task-specific multilingual Language Models [62].

4.3.2. Hate Speech Detection

The immense growth and public nature of social media, where any content can be posted and reach millions, necessitates automated methods for identifying inappropriate content. Among these, the detection of hate speech is crucial, despite its complex and subjective definition [63]. Given the scale of social media, the systems used to detect hate speech must be highly accurate, effective, and efficient which is especially challenging in low-resource languages. Researchers in [64] investigated hate speech detection in low-resource languages through the lens of KeyWord Spotting (KWS) (cf Section 4.7.3). The main objective is to search through an audio corpus for a pre-determined set of keywords indicative of hate speech. Their findings suggested that KWS using multilingual Acoustic Word Embeddings (AWEs) is a promising approach for quickly implementing hate speech detection in a new unseen language (here Wolof and Swahili) if resources are severely limited. To contribute to lowering this ressource barrier, [65] introduced AWOFORO: the first open annotated corpus of 3510 tweets in code mixed Wolof. Authors performed an exploratory analysis of the corpus and validated the annotations using Cohen's Kappa measures. A comparative study of machine learning models for abusive message detection has been presented in [66] focusing on code-mixed[14] data in Wolof and French languages. Authors introduced a meticulously annotated dataset of 2,022 tweets, which were manually classified as abusive or non-abusive. They also conducted extensive experiments, comparing the performance of various machine learning and deep learning algorithms on this dataset. [67] introduced AbuseBERTWoFr, the first model for abusive message detection on Wolof-French code-mixed tweets, trained on a large dataset of nearly 145k tweets. Researchers evaluated the model's performance on a corpus of +2k code-mixed tweets, and then compared its results against state-of-the-art language models.

4.3.3. Intent Classification

In Task-Oriented Dialogue (ToD) systems (cf Section 4.6), Natural Language Understanding (NLU) is essential for identifying the user's main goals and information [68]. NLU is typically split into two sub-tasks:

- Intent classification which consists of assigning one or more goal labels to the entire utterance [69];
- Slot filling in which specific values are extracted from the utterance to populate predefined information slots [70].

[14] The use of two or more languages in the same sentence.

 doi:10.20944/preprints202601.1124.v1

Although operating at different levels (token vs. sentence level), they are generally performed as a `joint task` to maximize performance in both simultaneously [71]. Similarly to other NLP tasks, existing large-scale benchmarks often omit low-resource languages and tend to heavily rely on English translations, which results in a predominant focus on Western-centric concepts. To mitigate this limit, INJONGO, a multicultural open-source benchmark dataset for 16 African languages including Wolof, has been introduced in [72]. Authors covered 05 domains[15] and performed several supervised fine-tuning experiments with multilingual encoders and Large Language Model (LLM) prompting. [73] introduced an intent classification dataset consisting of around 10k customer service queries (Bank and Transport) from the `Banking77` dataset [74] translated to French and Wolof. Authors evaluated different pre-trained models in `zero-shot` and `few-shot` settings and reported promising results. However, these results may be biased due to labeling errors that were discovered in the original `Banking77` dataset [75]. Furthermore, it has been observed that translationese[16] often exhibits features such as `stylistic ones` that are different from text written directly in the original language and thus can mislead model training [76]. This is a major problem in African language datasets, which are mainly based on translations of existing corpora [77].

To date, this is the only papers that specifically addresses intent classification in African languages, including at least one Senegalese language. This task is also studied in the works presented in Section 4.6, in the ToD context.

4.4. Lexicons & Spell Checking

Writing clearly and accurately can be challenging, especially for non-native speakers, as there are often many ways to express the same idea. A single spelling error (unexpected word form) significantly hinders readability and processing. In applications like Natural Language Processing (NLP), unnormalized, incorrectly spelled, or poorly digitized text severely diminishes its informational value [78]. To overcome the writer's constraints of time and proficiency, Automatic Spelling Correction (ASC) is deployed to locate misspelled words and generate a ranked set of potential replacements. Several approaches have therefore been studied to solve the problem of automatic spell checking. The study conducted in [79] divides these approaches into three categories:

- Those based on `expert rules`;
- Those incorporating a `context model` that allows candidate corrections to be reorganized;
- Those that learn `error patterns from a training dataset`.

Although significant progress has been made in the field of spell checking for low-resource languages, little work has been done specifically for Wolof. As part of the African Language-French Dictionaries (DiLAF) project, several dictionaries covering five other African languages in addition to Wolof have been developed [80] whose online publication was presented in [81]. The implementation of a spell checker for Wolof has been studied in [82], using a lexical approach based on a French-Wolof dictionary [83] and a Wolof morphological analyzer [31]. However, this work did not go as far as implementing a functional spell checker and was limited to a review of existing methods based on expert rules and context models based on n-gram language models. Furthermore, at the time of writing, all the dictionaries developed in [81] are available online, apart from the Wolof dictionary[17]. This absence prevents the exploration of dictionary-based approaches, even though these latter present several limitations:

- The `maintenance complexity` due to the rapid increase in the number of rules and the increasing difficulty of updates ;
- The dependence on the size of the dictionary ;
- The lack of linguistic context awareness.

[15] Banking, Home, Travel, Utility, and Kitchen & Dining.
[16] Texts translated by either humans or machines.
[17] http://pagesperso.ls2n.fr/~enguehard-c/DiLAF/index.php

 doi:10.20944/preprints202601.1124.v1

A proper Wolof spell checker has been proposed in [84] and relies on a combination of trie data structures, dynamic programming, and weighted Levenshtein distance to generate suggestions for misspelled words. The authors created new linguistic resources for Wolof i.e., a lexicon and a corpus of misspelled words, using a semi-automatic approach that combines manual and automatic annotation methods. However, the correction techniques described therein focus exclusively on the word level and do not take into account the context in which it appears. The integration of a context model, usually an n-gram language model [85], nevertheless allows contextual information to be included based on the history of previous words. However, this approach remains limited, as it only takes into account the immediate context preceding a word. Although additional classifiers can be used to overcome these limitations [79], the use of neural networks allows for the integration of a broader context, taking into account the words on both sides of the target word. Thus, deep learning with neural networks with attention [86] is a promising approach, which has already been studied for spell checking in various languages. This approach addresses this task by modeling spelling correction as a `translation task` from misspelled (noisy) text to well-spelled (correct) text and shows promising results. However, it requires a parallel corpus of noisy data on the one hand and correct data on the other hand, whereas languages like Wolof are low-resource languages and might not have such a corpus. In [53], we introduced BEQI: an efficient way to address this constraint by generating synthetic data based on regex[18] and seed data scraped on social media. We presented sequence-to-sequence models based on LSTM and Transformers for spelling correction in Wolof and evaluated these models in three different scenarios depending on the subwording method applied to the data. The work in [87] followed the same direction by leveraging transformer models and neural networks for word correction and spelling in Wolof. Authors also introduced a model trained on a parallel corpus consisting of misspelled sentences and their error-free counterparts and optimized the model to translate error-prone text into accurate sentences.

Text normalization is a foundational step in text processing under-resourced languages, especially those with inconsistent orthographies. A centralized and up-to-date lexical database is therefore essential for defining a common reference system. The iBaatukaay project [88] has been initiated as a collaborative project whose objective was to design a collaborative multilingual lexical database on the web for African languages, particularly Senegalese ones. Any expert in the field (lexicographers, linguists, etc.) could contribute via Internet, and the data could be downloaded free of charge from the platform. The project presented 25 indigenous senegalese languages, three of which were chosen for the project's launch: Wolof, Pulaar, and Bambara. Nevertheless, the `Institut Fondamental d'Afrique Noire (IFAN)` and the `École Supérieure Polytechnique (ESP)` went further and launched SENTERMINO [89] : a terminology platform aimed at centralizing, harmonizing, and providing scientific and technical terminology that has been validated and adapted to national languages. Such a platform facilitates the production of scientific and educational content in national languages, which improves the availability of data in local languages online. It also harmonizes the use of terms referring to the same concepts, which mitigates code-switching [19] and reduces the vocabulary of NLP systems to maximize their performance [90].

4.5. Machine Translation

A machine translation (MT) system converts a text sequence (or audio source) from a source language into the same sequence in a target language. For a long time, statistical machine translation (SMT) systems [91] were the dominant approach before the emergence of neural machine translation (NMT) systems [92], which have gradually achieved increasingly higher performance.

However, the quality of these systems has always been closely linked to the amount of data used in their design [93]. Therefore, `NTREX-128`, a dataset for machine translation evaluation from English into a total of 128 target languages has been released in [94], comprising around 2k sentences for each language including Wolof. [95] open-sourced `SMOL (Set of Maximal Overall Leverage)`, a

[18] A string of characters that defines a search pattern for matching text

 doi:10.20944/preprints202601.1124.v1

suite of 6.1M tokens training data that has been translated into 124 (and growing) under-resourced languages (125 language pairs including Wolof and Pulaar), including many for which there exist no previous public resources. These initiatives are very important as they contribute to unlock machine translation for low-resource languages. Thus, the most advanced machine translation systems have been developed with sequence-to-sequence models exploiting the attention mechanism [86] as well as the Transformer architecture [96]. Neural machine translation for low-resource languages (LRL-NMT) has been the subject of extensive research within the community, and various approaches have been studied. An overview of LRL-NMT work has been provided in [97], along with a set of recommendations for optimizing the design of translation systems based on the configuration of the language data (size, type of datasets, and available computational resources). Despite the substantial work carried out in neural machine translation in low-resource languages, very few local studies have specifically targeted Senegalese languages. To our knowledge, the first studies that have specifically explored Wolof-French machine translation systems are those presented in [98], where the authors introduced a corpus of 70,000 Wolof-French parallel sentences used to develop Word Embedding models [99] as well as translation models [100] based on the LSTM architecture [101]. However, the results presented in [98] were evaluated in terms of Accuracy, which complicates the effective assessment of the translation quality of their systems. We addressed this gap in [102], where we proposed an LSTM-based machine translation system that is evaluated using the BLEU metric [103]. BLEU is commonly used in the evaluation of NMT systems and offers a better correlation with human evaluations than Accuracy [104]. [100] also used the BLEU metric, but their results were biased due to a significant overlap between the training, validation, and test sets; an issue known as DATA LEAKAGE. [105] indicate that approximately 60% of the sentences in the test set were also found in the training data, which greatly overestimated the model's capabilities. Data leakage is one of the pitfalls associated with the adoption of machine learning methods, leading to failures in terms of validity, reproducibility, and generalization [106]. [107] provides a set of best practices for avoiding these errors, ranging from what to do before building the model, to how to build reliable models, evaluate them robustly, compare models fairly, and report results. The corpus initially introduced in [98] has thereafter been subsequently expanded to 83,000 sentences in [105], enabling the training of two neural machine translation systems for the French→Wolof and Wolof→French directions, based on the TRANSFORMER architecture [96]. A translation platform named SENTEKKI [108] has subsequently been set up based on this model, to allow the public to interact with the system via a web interface when deployed. A RestFul Web Service[19] was also developed, enabling other applications to integrate translation features, which is very important for the effective inclusion of our local languages. However, the authors did not mention the deployment of the platform, and no URL was provided to access it. It is also interesting to note that none of the French-Wolof datasets mentioned above have been made publicly available to date. This makes reproducibility difficult and hinders the progress of local work on these languages. This lack of openness could be explained by the still low level of research funding in Senegal and more broadly across the African continent (less than 1% of GDP).

Pre-trained multilingual translation models are also an interesting direction which have shown promising performance in supporting low-resource languages. They enable information sharing between similar languages, which significantly improves the translation of these language pairs, as studied in [109]. Much work has therefore been done in this direction, leading to the development of a wide range of multilingual translation models that include at least, the Wolof language. Such models have been developed in [110], where the authors leveraged existing pre-trained models to design translation systems for 16 low-resource African languages. Meta (formerly Facebook) introduced M2M-100 [111] as the first multilingual machine translation model capable of translating between any pair of 100 languages without relying on English data. A distilled version of M2M100 named SMALL-100 was introduced in [112] with the particularity of being `3.6 times smaller and 4.3 times faster`

[19] A service that uses the principles of REST (Representational State Transfer) to allow applications to communicate over the web using standard HTTP methods.

 doi:10.20944/preprints202601.1124.v1

at inference while having equivalent performance. The Wolof↔French dataset introduced in [102] has been expanded to **175,000 sentences** and then used in [113] to fine-tune the SMALL-100 model, achieving a BLEU score of `26.38`. This is to date the largest locally created French↔Wolof corpus (not openly available too). The work on M2M100 has also subsequently been expanded by META upon in the NO LANGUAGE LEFT BEHIND project[20] offering a state-of-the-art model capable of translating 200 languages into each other [114]. In 2024, the well-known translation platform GOOGLE TRANSLATE[21] expanded its support for underrepresented languages thanks to its PALM 2 large language model [115] to 110 additional languages, including Wolof [116]. The `DEEPL TRANSLATE`[22] platform, considered as the main alternative to Google Translate (although having less language support), now also offers Wolof [117]. This is a major step forward as DeepL is often more accurate for nuanced translations and represents a better choice for professional use [118]. At the local level, players such as `Baamtu Technologies`[23], a pioneering AI company in Senegal [28], and more recently `LAfricaMobile`[24], were already offering proprietary machine translation systems in local languages. The arrival of these new major stakeholders therefore represents both an opportunity and a threat, forcing local players to quickly reinvent themselves [**?**].

4.6. Question Answering and Dialogue Systems

Dialogue systems allow users to interact in natural language, via **text** or **voice**, in order to perform specific tasks (e.g. make reservations, obtain information, order a service) or to engage in open conversation (chatbots). Modern practice frequently combines `task-oriented paradigms` (driven by frames/slots) and `open conversation` in the same assistant, such as Siri, Alexa, and Google Assistant, or more recently, with the rise of GENERATIVE AI and LLMS, products such as ChatGPT, Gemini, and Claude. Despite technological advances, dialogue modeling remains a major challenge. Conversational agents must be able to handle interactions on a wide range of topics, provide relevant responses, and adapt to varied linguistic and cultural contexts [119]. In addition, dialogic phenomena (multi-turns, initiative, grounding, corrections) require a large volume of training data, robust and adaptive architectures [120] and issues like ethics and algorithmic biases raise concerns about the accessibility and neutrality of models. LLMs are currently the de facto backbone of modern chatbots, and as with major advances in NLP and AI in general, low-resource languages remain underserved. [121] introduced a human-translated benchmark dataset for 17 typologically diverse low-resource African languages, called IROKOBENCH. It covers three tasks: `Natural Language Inference`, `Mathematical Reasoning`, and `Multi-choice Knowledge-based QA`. Their evaluation of 10 open and 06 proprietary large language models (LLMs) in zero-shot, few-shot, and translate-test settings showed a significant performance gap between high-resource languages (English/French) and the African languages. Wolof is among the lowest-performing languages in the evaluation, mainly due to the small amount of publicly available data across the web (< 50 million characters) [122] and its generally poor quality [123]. Furthermore, `given that low-resource languages are not all equally low in resources` [2], one phenomenon that remains understudied is the gap between African languages themselves. Across 42 supported African languages and 23 available public data sets, [124] identifed 04 languages (`Amharic`, `Swahili`, `Afrikaans`, and `Malagasy`) that are always treated, while there is over 98% of unsupported African languages. This inequality also extends to the scripts used by these languages and can have various causes beyond the lack of data, such as `tokenization biases`, `high computational costs`, and `evaluation issues`. Although tedious, costly, and time-consuming, data collection remains nevertheless a major stake in improving the representativeness of African languages on the global AI map. Therefore, [125] introduced the first high-quality dialogue datasets for six

[20] https://ai.facebook.com/research/no-language-left-behind/
[21] https://translate.google.com/?sl=fr&tl=wo&op=translate
[22] https://www.deepl.com/en/translator
[23] https://baamtu.com/
[24] https://lafricamobile.com/en/produit-tts/

 doi:10.20944/preprints202601.1124.v1

African languages: Swahili, Wolof, Hausa, Nigerian Pidgin English, Kinyarwanda, and Yorùbá. The corpus consists of 1,500 turns each, which has been translated from a portion of the English Multi-Domain MultiWOZ dataset [126] to enable the creation of dialogue agents for African languages. AFRIQA, the first cross-lingual Question Answering (QA) dataset with a focus on African languages has been proposed in [127] laying the foundation for research on QA systems for one of the most linguistically diverse regions in the world. AFRIQA includes +12,000 XOR QA examples across 10 African languages including Wolof. A multiple-choice machine reading comprehension (MRC) dataset spanning 122 language variants including Wolof, has been introduced in [128]. Built from parallel passages in FLORES-200 [114], the dataset enables controlled cross-lingual evaluation and reveals that, despite strong cross-lingual transfer in English-centric LLMs, smaller multilingual masked language models trained on balanced data exhibit broader language understanding. However, the best strategy to decide which cross-lingual data to include during training still remains an open question. Authors in [129] analyzed transfer strategies between 263 different languages (including Wolof), from 33 language families across 03 tasks: **POS tagging**, **Dependency parsing**, and **Topic classification**. They showed that beyond the definition of the concept of linguistic similarity, its effect on transfer performance depends mainly on the **NLP task** at hand, and the **(mono- or multilingual) input representations**. Regarding the latter, the authors in [130] evaluated 05 embedding similarity metrics across 03 multilingual models trained on African languages (including Wolof). The actual transfer performance has been then measured on 03 downstream tasks: **NER**, **PoS-tagging**, and **Sentiment Analysis** for a total of 816 transfer experiments. The experiments showed promising transfer prediction capabilities for NER and POS, with comparable predictive power to URIEL linguistic typology.

Beyond cross-lingual transfer, LLMs also need to be fine-tuned on instruction tuning data[25], in order to make them more engaging in conversations, and enable them to properly follow the given instructions. This is particularly challenging given that this type of data is much more expensive to collect, as it requires human instructions and annotations [131]. COHERELABS[26] responded to this challenge by launching the AYA project: a year-long participatory research initiative that brought together nearly 3,000 participants from over 100 different countries [77]. This made it possible to collect the largest native speaker instruction dataset with a total of 204K human-curated prompt-response pairs written by native speakers in 65 languages. Wolof is one of the languages that was not included in the initial list but was subsequently added thanks to the involvement of the GALSENAI community[27] [132]. This shows the importance of local AI communities in shaping African AI ecosystems, and GalsenAI is one of the pioneers that has played a major role in this regard in Senegal [28]. Despite these efforts, sufficient data had not been collected in time, which led to the exclusion of Wolof from the training of the ensuing AYA models [133]. A similar situation also happened during the AfriqueLLM project, whereby all languages for which a minimum of 90 million tokens could not be collected (e.g. Wolof), were excluded from the training corpus [134]. This highlights the importance of creating sufficiently large and clean pretraining corpora in order to improve the representativeness of national languages in LLM projects. In this regard, [122] introduced MADLAD-400, a manually audited, general domain 3T token monolingual dataset based on CommonCrawl[28], spanning 419 languages including Wolof and Pulaar. The authors discussed the limitations revealed by self-auditing MADLAD-400 (with poor quality Wolof crawled data), and the role data auditing had in the dataset creation process. To overcome this quality issue, [135] introduced a new pre-training dataset curation pipeline based on FineWeb [136] **that can be automatically adapted to support any language**. They then leveraged this pipeline to create FineWeb2, a new 20 terabyte (5 billion document) dataset covering over 1000

[25] Input-output pairs that teach AI models to follow instructions, with each data point typically containing an instruction, an input query, and the expected output response.

[26] https://cohere.com/research

[27] Senegalese Artificial Intelligence community with thousands of members across Senegal, Africa, and the diaspora.

[28] https://commoncrawl.org/

 doi:10.20944/preprints202601.1124.v1

languages including Wolof and Pulaar. They also showed that models trained on language-specific corpora `produced by this pipeline`, perform better than those trained on other public web-based multilingual datasets. [137] introduced a novel self-active learning framework to pre-train a Language Model from scratch on 23 African languages including Wolof named AFROLM. With a dataset `14x smaller` than existing baselines, AFROLM outperforms many multilingual pretrained language models like AfriBERTa [138], XLMR-base [61] and mBERT [139]; on various NLP downstream tasks such as `Named Entity Recognition`, `Text Classification`, and `Sentiment Analysis`. To further enhance African language coverage in language models, researchers in [140] introduced `SERENGETI`, a set of massively multilingual language model that covers 517 African languages and language varieties including Wolof and Pulaar. They evaluated the models on 08 natural language understanding tasks across 20 datasets, and showed through meaningful comparisons, how SERENGETI model excels and acquire new SOTA.

In spite of all these challenges, conversational agents represent nevertheless a strategic lever for socioeconomic development, especially in the Senegalese context. Integrating automatic speech and language processing into systems adapted to local realities could significantly improve access to digital services in key sectors such as:

- **Commerce:** Automation of customer services, intelligent recommendations, optimization of sales processes ;
- **Healthcare:** Digital medical assistance, easier access to health information, AI-assisted preliminary diagnosis ;
- **Banking and fintech:** Simplification of transactions, user support for mobile banking services ;
- **Education:** Access to online learning in local languages, interactive tutorials via educational chatbots, support for digital literacy.

The rise of dialogue systems in African languages is therefore a major opportunity to promote digital inclusion and ensure equitable access to information and communication technologies on the continent. This is particularly evident in the release of AWA, introduced as the first AI assistant capable of conversing fluently in Wolof [141]. Its announcement sparked unprecedented enthusiasm, highlighting the potential of such systems to have a direct and lasting impact on citizens' lives. However, proof of concept for such a Wolof conversational agent was explored in [142] well before the advent of AWA. Researchers opted for a modular architecture[29] to design a conversationnal agent that provides information to the customers of a telecommunications provider. This approach is very common in the design of Task-oriented Dialogue Systems [68] and has long been used in the majority of chatbots deployed in industry prior to the recent rise of LLMs [143]. To overcome the limitations associated with manual collection of synthetic data in [142], [113] proposed a more scalable approach based on the `translate-train` paradigm. It is a general training approach to multilingual tasks where the key idea is to use the translator of the target language to generate training data in order to mitigate the gap between the source and target languages [76]. Researchers proposed thus a chatbot generation engine based on the Rasa framework [144] and a robust methodology for projecting annotations onto the Wolof language using an in-house machine translation system. Researchers in [145] proposed an educational intelligent chatbot to improve literacy regarding the Von Willebrand disease (VWD)[30] in Senegal. Their system is also based on the modular architecture with an `Automatic Speech Recognition (ASR)` system that converts spoken inputs into text which is then processed through the `Natural Language Understanding (NLU)` module[31] to identify user intent across VWD-related themes. The chatbot generates appropriates responses via pre-defined templates containing culturally relevant information, and continuously improves its accuracy through a feedback loop that analyzes user interaction. `Baamtu Technologies` took a similar approach when setting up the

[29] Approach, in which the dialog system is broken down into a pipeline of different sub-tasks.
[30] A lifelong bleeding condition that makes it hard for blood to clot.
[31] Authors used `certainly`, a proprietary conversational AI platform to train an NLU model.

 doi:10.20944/preprints202601.1124.v1

SaytuTension chatbot, which aims to raise awareness about high blood pressure, a very common condition in Senegal [**?**]. These initiatives demonstrate a growing interest in designing practical applications based on these technologies for the common good.

Although more data-efficient than LLMs, this modular approach nevertheless suffers from `component isolation` i.e. each module must be optimized separately, and `error cascading`[32] which make the maintenance of the overall system very tedious [68]. Systems such as AWA remain therefore quite promising for the local language inclusion, offering better interaction fluency and better scalability. Since AWA is a closed system, an `open source` alternative named OOLEL and based on QWEN 2.5 [146], was subsequently released in [147]. It is one of the very first open source language models in Wolof that combines state-of-the-art AI technology with deep Wolof linguistic expertise.

4.7. Speech Processing

Speech processing is a field that analyzes and manipulates human speech signals using digital technology [148]. It involves a range of tasks like `Automatic Speech Recognition (ASR)` (converting speech to text), `Speech Synthesis` (text-to-speech), and `Speaker recognition`, and is used in applications such as `voice assistants` and transcription services. As with other NLP tasks, speech processing has also been revolutionized in recent years by deep learning approaches. These approaches are highly data-intensive, and much work has focused on data collection and more efficient data approaches.

4.7.1. Automatic Speech Recognition

Automatic transcription of speech by any speaker in any environment is still far from solved, but ASR technology has matured to the point where it is now viable for many practical tasks [119]. [149] conducted a theoretical study in the areas of speech recognition in the Wolof language and presented the results obtained from their implementation with the `Julius` Speech Recognition [**?**] open source software. Researchers leveraged an approach based on Hidden Markov Models (HMM) [150] with a language model that consists of a word pronunciation dictionary and a syntactic constraint. In a favourable context for the development of a market for voice technologies in African languages, the `ALFFA` project has been launched in [151] to conduct fundamentals research on speech analysis (language phonetic and linguistic description, dialectology) and develop efficient speech technologies (ASR and TTS) for African languages. Authors described their achievements after 18 months of project and presented a multilingual calculator prototype in several African languages (including Wolof, Hausa and accented French) leveraging `Kaldi` [152] for the ASR and the proprietary `Voxygen` [33] engine for the TTS. To cope with the vowel length contrast issue[34] in languages like Wolof and Hausa, [153], proposed a vowel length contrast modeling with contrasted and non length-contrasted CD-DNN-HMM[35] models for ASR. They also used the Kaldi toolkit to train the Wolof model on 18,000 recorded utterances representing 21.3 hours of signal, and thus introduced the first large vocabulary continuous speech recognition system ever developed for the Wolof language. This work has subsequently been expanded to address a wider range of challenges faced by these languages, such as the `small amount of transcribed speech`, `written language normalization issues`, `limited text resources available for language modeling`, as well as `specific features (tones, morphology, etc.)`. Therefore, in addition to vowel length contrast modeling, data augmentation techniques through speed perturbation have been explored in [154] to overcome the lack of resources. Researchers also developed ASR systems for Hausa and Wolof and made them openly available to the research community. They then leveraged the new research opportunities brought by growing digital archives and improving text and speech algorithms, to further explore automatic analysis approaches of the vowel length contrast phenomenon

[32] Series of failures in a system of interconnected parts, where the failure of one component triggers the failure of others.
[33] https://www.voxygen.fr/home
[34] Two versions (short/ long) of a same vowel that exist in the phoneme inventory of the language.
[35] Gaussian Mixture Model, Deep Neural Networks and Hidden Markov model.

 doi:10.20944/preprints202601.1124.v1

in [155]. Authors introduced multiple features to make a fine evaluation of the degree of length contrast under different factors and showed their abilities to porperly highlight a variety of contrast degrees for each vowel considered. Still in the direction of data augmentation, [156] presented the first systematic assessment of large-scale synthetic voice corpora for African ASR. Researchers showed that the `generated synthetic voice data could be created for less than 1% of the cost of collecting real human data`, while holding potential to complement this human data in creating and improving ASR models for African languages.

The ALFFA project has been completed afterwards in [18] where the researchers presents the data collected and ASR systems developped for 04 sub-saharan african languages (Swahili, Hausa, Amharic and Wolof). They illustrated their methodology by making a focus on Wolof for which they designed one of the first ASR systems ever built in this language and trained on 18k recorded utterances representing more than 21h of signal. All data and scripts had been made available online and this dataset had a huge impact on the local NLP ecosystem, allowing a wide range of actors to experiment ASR systems in Wolof as in [64,157]. Therefore, [158] conducted an extensive review of the state of the art in speech recognition in order to offer a comprehensive overview of the most recent and relevant developments in speech recognition. Researchers explored technological advances, cutting-edge algorithmic models, deep learning methodologies, and persistent challenges that drive research such as `low-resource languages`, multilingual models and innovation in this constantly evolving field. Different approaches that address the low-resource property of African languages in speech recognition have also been studied in [159]. The authors leveraged `self-supervised multilingual pretrained models` to introduce monolingual baselines and multilingual systems that are evaluated across 07 African languages including Wolof. They explored several multilingual strategies in order to mitigate language confusion and lexical ambiguity, and demonstrated that incorporating language-aware mechanisms, improves multilingual ASR performance while reducing reliance on external language identification. To help build foundational digital resources for African languages, the `AI4D - African Language Program`[36] has been launched with 03 main objectives:

- **Incentivise** the crowd-sourcing, collection and curation of language datasets through an online quantitative and qualitative challenge ;
- **Support** research fellows for a period of 3-4 months to create datasets annotated for NLP tasks ;
- **Host** competitive Machine Learning challenges on the basis of these datasets.

It is within this context that Baamtu Technologies launched the `AI4D Baamtu Datamation Automatic Speech Recognition in WOLOF` hackathon[37] bringing together more than 200 participants and leading to the design of an ASR model achieving a `Word Error Rate` (WER) of `0.110` [160]. This initiative has strengthened interest in the ecosystem and reinforced Baamtu Technologies' position as a local leader in NLP for local languages, particularly with its `Kàllaama` suite[38]. Similarly to machine translation, tech giants are also interested in speech processing in under-resourced languages. `Meta` (formerly Facebook) has notably launched the `Massively Multilingual Speech (MMS)` project in [161] that increases the number of supported languages by 10-40x, depending on the task. The core of this work involved creating a new dataset from readings of publicly available religious texts and effectively using `self-supervised learning` [162]. This effort resulted in pre-trained `wav2vec 2.0` models [163] for thousands of languages, a single multilingual speech recognition and synthesis models each covering 1,107 languages, as well as a language identification model for more than 4k languages. They recently introduced `Omnilingual ASR` [164], the first large-scale ASR system designed for extensibility, which allows communities to introduce unserved languages `with only a handful of data samples`. This model scales self-supervised pre-training to 7B parameters to learn robust speech representations, and introduces an encoder-decoder architecture designed for zero-shot

[36] https://k4all.org/project/language-dataset-fellowship/
[37] https://zindi.africa/competitions/ai4d-baamtu-datamation-automatic-speech-recognition-in-wolof
[38] https://www.youtube.com/watch?v=P5PRgugOu8o

 doi:10.20944/preprints202601.1124.v1

generalization, leveraging a LLM-inspired decoder. Omnilingual ASR expands coverage to over 1,600 languages including Wolof (Senegalese and Gambian) and Fula, the largest such effort to date, including over 500 never before served by ASR. This kind of initiatives drastically reduce the entry barrier to developing advanced language processing models in these languages, which encourages local players such as `LAfricaMobile`[39] and `Lengo` [165] to offer Wolof ASR for different use cases. Even `Orange`[40], which was relatively inactive at the beginning of the local AI ecosystem's emergence [28], is now taking a keen interest in local languages. They recently announced their partnership with `OpenAI` and `Meta`, to fine-tune Large Language models (LLMs) to understand regional languages in Africa that today are not understood by any GenAI model, with an initial focus on Wolof and Pulaar [**?**]. This follows the advent of `Whisper` models [166] developed by OpenAI, which show promising performance when fine-tuned on Wolof data as in [167]. The researchers focused on `Maternal and Reproductive Health` and collected 250 essential healthcare keywords that has been expanded to 750 real-world phrases, and translated them into Wolof and Hausa. They used a Whisper model initially fine-tuned on Wolof [168] to adapt it to the medical domain via the `LoRA (Low-Rank Adaptation)` approach [169], which requires fewer computational resources. This is particularly relevant as domains such as **healthcare** typically suffer from a `double resource scarcity`, where there is both a lack of language and domain data [2].

Beyond proprietary projects, `Orange` is also active in the open source community with the introduction of the first self-supervised multilingual speech model trained exclusively on African speech [170]. They pre-trained an `HuBERT-based` model [171] on nearly 60k hours of unlabeled speech segments, in 21 languages and dialects spoken in sub-Saharan Africa (including Wolof and Pulaar) [172]. The evaluation on the SSA subset of the `FLEURS-102` dataset [173] demontrated competitive results. This confirms the findings in [174] which showed that combining under-resourced languages that share similar linguistic and phonetic characteristics (Wolof, Swahili and Fongbe), enhances the quality of features extracted for each language individually. [175] went further by curating around 1.4 TB of raw Wolof speech and filtered it down to 860 hours of high-quality spontaneous audio, using a multi-stage pipeline (source separation, diarization, VAD, quality filtering). They performed continued pretraining [176] of HuBERT on this Wolof data, which allows to improve ASR performance over both the original Meta/Hubert model [171] and Orange/HuBERT [170], `while using far less language-specific compute`. [73] introduced the first Wolof spoken intent classification dataset consisting of more than 04 hours of spoken customer service queries. Researchers fine-tuned multilingual pre-trained ASR models and conducted extensive evaluations. Although the dataset remains small, it nevertheless constitutes the first initiative paving the way for research in `Spoken Language Understanding` in local languages (cf Section 4.7.3).

4.7.2. Speech Synthesis

The modern task of text-to-speech or TTS, also called speech synthesis, is exactly the reverse of ASR: `to map text to an acoustic waveform` [119]. It is used in applications like spoken language models that interact with people or for reading text out loud, greatly improving inclusion in information access. [149] conducted a theoretical study in Wolof speech synthesis and implemented a basic TTS system with the festival speech tool [177]. Beyond models, authors also created different lexicons and knowledge bases of phonetic, acoustic and lingistic features in order to introduce other languages. As part of the `Cracking the Language Barrier for a Multilingual Africa` project[41], the first open Wolof TTS dataset has been introduced in [**?**]. It contains recordings from two natif Wolof actors (a male and female voice). Each actor recorded more than 20,000 sentences for approximately 19 hours of recordings for the female voice and 22 hours for the male voice. Having identified artifacts in the textual data as well as poor recording quality, the GalsenAI community proposed a cleaned

[39] https://lafricamobile.com/en/produit-stt/
[40] Leading telecom operator in Senegal.
[41] https://k4all.org/project/building-wolof-text-to-speech-system/

 doi:10.20944/preprints202601.1124.v1

version of this dataset in [178]. They extracted the `female voice`, denoised it and enhanced it with the `Resemble Enhance` library [179] and removed emojis, special characters as well as Arabic and Russian characters. They also removed audios judged not qualitative enough, reducing its size to around 18h40mn of high-quality recordings. They subsequently trained a TTS model based on `xTTS-V2` [180] on the cleaned dataset and openly released it for public use [181]. This initiative has inspired other local stakeholders such as Concree[42], which introduced `AdiaTTS` [182] based on the `ParlerTTS` model [183,184]. Their model is trained on 40 hours of Wolof speech data, which has not been published. These initiatives are generating real enthusiasm for Wolof speech synthesis systems, and the 2025 CNRIA[43] Demo Paper Award was even given to a demo on this topic [185]. Given the difficulty of accessing high-quality data for languages with limited resources, [186] explored efficient corpus creation and sharing approaches as well as the deployment of TTS systems. They thus demonstratd it was possible to develop synthesizers that generate intelligible speech with only 25 minutes of created speech, even when recorded in suboptimal environments. Speech data, code, and trained models for 12 African languages including Wolof was subsequently released for future research. `LAfricaMobile`[44] offers the widest language coverage in its TTS in Senegal with Wolof, Bambara, Dioula, Lingala, Hausa, and Fulfulde. As a private company, all of its models and data are kept private.

4.7.3. Spoken Dialog Systems

A spoken dialogue model refers to a dialogue system capable of generating intelligent verbal responses based on the input speech [187]. It has two essential components that do not exist in a written text dialog system: a `speech recognizer` and a `text-to-speech` module. Spoken Dialog Systems represent thus one of the most direct methods of `human-computer interaction (HCI)`[45], and has tremendous potential in the African context, given that African languages are spoken more than they are written [188]. The proof of concept presented in [142] and highlighted in Section 4.6 is the first Wolof dialog system documented in research. Authors used an in-house speech recognition model trained on 132 hours of data (1/3 gold data, 2/3 synthetic data), which is based on `Kaldi` [152]. The researchers used pre-recorded messages instead of a speech synthesis system to return the output to the user. The introduction of voice features also increases the complexity of the architecture by adding additional components, exacerbating the problem of cascading errors (cf Section 4.6). One way to address this issue is to replace the ASR and the NLU components by a single one, that directly maps the audio input to the corresponding intent (SLU) [70]. Therefore, [73] introduced the first Wolof spoken intent classification dataset consisting of more than 4 hours of spoken customer service queries. Researchers explored speech recognition models and leave room for research in `spoken intent classification` in Wolof by openly releasing the dataset. In [141], voice features are also presented as part of the AWA dialogue system, but no information was shared about the ASR and TTS used.

Another approach is to design an entirely end-to-end architecture, that directly processes the audio input to produce audio output. Although the definition of `Speech LLMs` remains non-standardized in current research [189], this concept can be defined as `speech-to-speech generation tasks` [190]. In this way, [175] introduced the first speech language model for Wolof alongside with a 860 hours spontaneous and high-quality unsupervised speech dataset. Authors highlighted the effectiveness of continued pretraining [176], which allows to reuse the compute already invested in the base model. The GALSENAI community introduced the first `Keyword spotting` (KWS)[46] dataset that covers all the 06 senegalese languages presented in this survey [191]. They extended the Speech commands dataset [192] that included a limited vocabulary composed of around twenty common words at

[42] https://concree.com/
[43] Conference on Research in Computer Science and its Applications
[44] https://lafricamobile.com/en/produit-tts/
[45] Study and design of how people and computers interact with each other, focusing on creating user-friendly and effective technology interfaces.
[46] The task of learning to detect spoken keywords.

 doi:10.20944/preprints202601.1124.v1

its core i.e. digits from zero to nine, and seventeen words that would be useful as commands in IoT or Robotics applications. The development of keyword spotting models has been explored as part of a dedicated hackathon[47], showcasing the project while training young people in the fundamentals of machine learning. Researchers in [64] proposed an alternative to the conventional KWS approach that involves transcribing the audio corpus with an automatic speech recognition (ASR) system and then searching for keywords in the output. They introduced an ASR-free approach by extending the query-by-example (QbE) methodology with multilingual acoustic word embeddings (AWEs) and compare their effectiveness w.r.t "classical" ASR-based methods. In controlled experiments on Wolof and Swahili where training and test data are from the same domain, their results showed that an ASR model trained on just five minutes of data outperforms the AWE approach.

While sharing some similarities with spoken dialog systems, command and control speech systems can be further distinguished from them, as they are able to respond to requests but do not attempt to maintain continuity over time.

5. Case Study: NLP Pipelines for the Social Sciences

Natural Language Processing has immense potential to transform methodologies in social science, particularly in multilingual African contexts such as Senegal. NLP tools can reduce the time, cost, and cognitive burden of qualitative research while broadening access to linguistic data collected in national languages (cf Section 4.3.1). Field-based social research in Senegal typically relies on extensive interviews, focus groups, and participant observations conducted in multiple languages, often Wolof and French [193,194]. Manual transcription and translation of these data are among the most time-consuming phases of the research process. To prevent this step from becoming an obstacle to data exploitation, researchers generally include the costs of transcribing interviews conducted in national languages in their budgets. Furthermore, to optimize efficiency, interviews conducted in a local language are transcribed by young researchers (doctoral students and linguistics students) who understand the language of the respondents. The challenges researchers encounter in transcribing survey data into local languages "forces" them to favor thematic analysis over lexical analysis, which requires manual transcription[48]. Integrating speech recognition (ASR) and machine translation (MT) systems into the workflow, can drastically reduce transcription time and costs, while preserving multilingual fidelity [195]. The interviews that have already been transcribed could have served as an interesting corpus for ASR training, but the personal data they contain makes it difficult to collect them. Automatic alignment and metadata tagging (cf Section 4.2) can also further facilitate corpus analysis, making it easier to identify recurrent themes or sentiments across respondents and regions (cf Section 4.3.1).

However, automating qualitative analysis introduces ethical and privacy challenges, particularly when dealing with sensitive or identifiable data. The study in [196] offers a comprehensive synthesis of how NLP techniques are increasingly reshaping social science research. It outlines a three-layer framework, from preprocessing and representing unstructured text, to extracting semantic information, and finally applying these insights to sociological and political analyses. The review highlights how classical methods (e.g., dictionaries, topic models, supervised classifiers) and modern deep-learning approaches (e.g., contextual embeddings, transformers, large language models) enable researchers to work at unprecedented scale while uncovering complex patterns. These patterns can be related to bias, culture, political behavior, online polarization, and collective action. The authors also identify key methodological challenges, including representativeness, model bias, and interpretability. They also discussed future directions, especially the growing role of large language models (LLMs) in annotation, research design, and simulation. Although LLMs now play a central role in tasks such as large-scale text coding [197], synthetic data generation [198], and the simulation of human judgments [199], researchers in [200] argued that, their deployment raises unresolved concerns

[47] https://www.kaggle.com/competitions/indabaxsenegal2023-wolof-language-keyword-spotting
[48] Feedback shared by a small group of local social science researchers we interviewed.

 doi:10.20944/preprints202601.1124.v1

regarding reliability, validity, replicability, and model drift. Recent methodological work in computational social science has therefore emphasized the need for rigorous frameworks when using LLMs for empirical research [201]. To address these issues, the authors in [200] outlined a set of best practices, such as `transparent reporting of model specifications`, `systematic validation against human-coded ground truth`, and `explicit handling of nondeterminism`, that are essential to ensure credible scientific inference.

These works are particularly relevant for research on low-resource languages such as Wolof, where methodological opacity or unvalidated model behavior could disproportionately distort scientific findings (cf Section 4.5). By foregrounding replicability and the careful evaluation of model outputs, this primer offers a transferable roadmap for responsibly integrating NLP tools into linguistic and socio-cultural analyses involving African languages.

6. Discussions & Perspectives

6.1. Challenges

Despite growing momentum, the development of NLP for Senegalese national languages faces multiple interrelated challenges spanning data, methodology, infrastructure, and governance.

Data Scarcity and Quality: The most critical limitation remains the scarcity of high-quality, standardized linguistic data. Existing corpora are fragmented across institutions and often lack sufficient size, metadata, or representativeness. Many textual sources are informal or domain-specific (e.g., religious texts or social media posts), introducing stylistic bias. For speech, recordings are limited in both duration and speaker diversity, while annotation is slow due to the lack of trained linguists in the different languages and **funding**. Overall, the progress to date underscores both the rapid advances and the persistent inequalities across languages. While Wolof has benefited from relatively larger datasets and visibility in continental benchmarks, other national languages remain at the exploratory stage. Continued investment in data curation, open evaluation, and multi-institutional collaboration will be critical to ensure balanced development across Senegal's linguistic spectrum.

Linguistic Complexity and Orthographic Variation: The structural diversity of Senegal's languages presents unique modeling challenges. Tonal contrasts, morphophonemic alternations, and agglutinative morphology affect tokenization and subword segmentation. In addition, orthographic conventions are still evolving and inconsistently encoded across digital platforms. The lack of standard transliteration systems further complicates data integration across dialects and scripts, and reinforces the dominance of Wolof over other local languages.

Limited Computational Infrastructure: Sustained progress requires access to modern compute infrastructure and cloud services for model training and evaluation. While the government invested in a national compute infrastructure [202], the latter struggles to be made operational [203] and most local universities still rely on limited or outdated hardware. Dependence on foreign hosting services raises additional issues of data sovereignty and sustainability.

Capacity Building and Skill Gaps: The availability of trained researchers and engineers in NLP remains limited. Although institutions and local universities are expanding training programs, the scale of need outpaces current capacity. Bridging this gap demands long-term investment in curriculum development, mentorship, and community-based learning initiatives.

Visibility of local search Local research also lacks visibility, with the vast majority of articles not being published in major conferences. To date, [73] is the only research work that has been published at **NeurIPS**, which is one of the major conferences in Artificial Intelligence. Researchers would therefore benefit from targeting conferences and journals such as NeurIPS, AAAI, ICLR, ICML, IJCAI, InterSpeech, EMNLP, ACL, TACL, COLING, etc.

Ethical, Legal, and Governance Challenges: Ethical and governance frameworks for language data remain underdeveloped. Questions of consent, cultural ownership, and intellectual property are critical for corpora derived from oral traditions and social research. Ensuring compliance

 doi:10.20944/preprints202601.1124.v1

with both CDP[49] and SODAV[50] principles requires institutional coordination. Without such alignment, the risk of data misuse or extractive research practices persists. Therefore, the UNESCO AI Readiness Assessment Methodology (RAM) [204] report highlighted the urgent need for a strong, centralized AI and data governance body, with technical sectoral committees and ethical oversight.

6.2. Opportunities and Future Directions

While Senegal's AI ecosystem remains nascent, the convergence of political will [25,205–207], community engagement [208–210], and academic expertise [211] provides fertile ground for sustainable growth. This section outlines several key opportunities and future directions that could accelerate progress across linguistic, technical, and social dimensions.

Regional Collaboration and Cross-Lingual Transfer: Given the linguistic continuity between Senegalese and neighboring West African languages, regional collaboration represents a major opportunity. Shared resources across Wolof, Pulaar, Mandingue, and Soninké communities could enable the development of multilingual models with stronger generalization, and thus reduce the gap between Wolof and the other local languages. Partnerships with initiatives such as Masakhane[51], EthioNLP[52], Mak-AI[53] and Sunbird AI[54] can enhance interoperability and ensure the inclusion of underrepresented languages in continental benchmarks. As an example, a historic milestone for African AI Research has been achieved for the first time, by a team from InstaDeep[55] and Stellenbosch University, by earning an Oral presentation at NeurIPS[56] [212], one of the most prestigious AI conference in the world [213]. As of today, the continent contributes less than 3% to the global AI market [214], and these initiatives highlight the critical need to strengthen the key ingredient: Partnerships. Public-private partnerships and cross-border collaborations have therefore a huge potential to accelerate AI adoption and fuel a generation of African innovators.

Open Data, Tools, and Infrastructure: Expanding the availability of open data and reproducible tools is essential for building capacity and trust. Many master's level projects and theses are also carried out on Senegalese languages, but remain difficult to gather due to their lack of visibility. The centralized GitHub repository introduced in this work can serve as a living catalogue of Senegalese NLP resources, facilitating transparency and reuse. Institutional support and community contributions could further extend this infrastructure through GPU and open data grants, national cloud storage, and public data portals. Data collection should also be included in training curricula, as in this AMMI program[57], in which each student (or duo of students) had to record 02 hours of speech in their native languages. The development of multilingual and multitask benchmarks such as XTREME-UP [215] is also important in order to facilitate the evaluation of models on these languages and foster their inclusion.

Human Capital and Interdisciplinary Training: Bridging the skills gap requires long-term investment in education. Embedding NLP and data science curricula within local universities, especially in linguistics curricula, can produce a new generation of computational linguists fluent in both the technical and cultural dimensions of Senegalese languages. Interdisciplinary programs linking computer science, linguistics, and the social sciences would ensure that technology development remains responsive to local research needs.

Applied Research and Societal Impact: Beyond academic exploration, NLP technologies can yield tangible benefits for governance, education, and cultural preservation. In the social

[49] Senegalese Committee for the Protection of Sensitive Data.
[50] Senegalese Agency for Copyright and Related Rights.
[51] https://www.masakhane.io/
[52] https://ethionlp.github.io/
[53] https://air.ug/
[54] https://sunbird.ai/
[55] https://instadeep.com/
[56] Neural Information Processing Systems.
[57] https://github.com/besacier/AMMIcourse

 doi:10.20944/preprints202601.1124.v1

sciences, language models can accelerate ethnographic analysis, making research outputs more inclusive and timely. CLAD[58] and IFAN[59], for example, have tremendous potential to launch the first national applied research laboratory in NLP for Senegalese languages and initiate projects such as NaijaVoices[60] to equip our local languages. This will allow for the promotion of **computational social science** and modernize the research approaches of local social science researchers.

Toward a Sustainable and Ethical Ecosystem: Sustainability depends on continuous coordination among policymakers, researchers, and communities. Embedding ethical data governance into the national AI framework will help safeguard cultural heritage while promoting equitable innovation. Long-term public–private partnerships and regional alliances will be key to scaling these initiatives. Senegal thus has the opportunity to become a continental reference for inclusive, ethical, and locally grounded NLP development.

7. Conclusion

This paper has presented the first comprehensive synthesis of Natural Language Processing initiatives and resources for the six national languages of Senegal: Wolof, Pulaar, Sérère, Diola, Mandingue, and Soninké; while situating them within the broader context of the social sciences. We have shown that, although Senegal's NLP ecosystem is still emerging, it is underpinned by strong institutional, academic, and community foundations. Current local and regional initiatives are nevertheless catalyzing a shift from isolated experiments to a coordinated, nationally anchored framework. The publicly available datasets, models, and tools are released in a centralized GitHub repository[61].

The integration of NLP into the social sciences offers transformative potential: automating transcription, translation, and thematic analysis can significantly enhance research productivity and inclusiveness. Yet this potential will only be realized through sustained investment in open data, interdisciplinary training, and ethical governance. The establishment of shared infrastructures (computational, linguistic, and institutional) will ensure that NLP development in Senegal reflects both global best practices and local realities. Continued collaboration between linguists, computer scientists, policymakers, and local communities will be essential to build a fair and enduring digital future for all Senegalese languages.

The findings of this survey provide a foundation for future research on NLP for Senegalese languages. They offer researchers a means of identifying priority areas for further investigation and formulating research agendas that directly address the challenges and opportunities highlighted by the survey. In future work, we plan to explore a qualitative evaluation of existing datasets in Senegalese languages as well as prospects for implementing benchmarks in various NLP tasks. Platforms such as WhatsApp also offer unprecedented opportunities to conduct large-scale data collection campaigns as in [216] and could be the subject of further study.

Acknowledgments: This project has been funded by Fondation Botnar.

References

1. Nekoto, W.; Marivate, V.; Matsila, T.; Fasubaa, T.; Fagbohungbe, T.; Akinola, S.O.; Muhammad, S.; Kabongo Kabenamualu, S.; Osei, S.; Sackey, F.; et al. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020; Cohn, T.; He, Y.; Liu, Y., Eds., Online, 2020; pp. 2144–2160. https://doi.org/10.18653/v1/2020.findings-emnlp.195.
2. Hedderich, M.; Lange, L.; Adel, H.; Strötgen, J.; Klakow, D. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In Proceedings of the Proceedings of the 2021 Conference

[58] https://clad.ucad.sn/
[59] https://ifan.ucad.sn/
[60] https://naijavoices.com/
[61] https://github.com/DerXter/State-of-NLP-Research-in-Senegal

 doi:10.20944/preprints202601.1124.v1

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 01 2021, pp. 2545–2568. https://doi.org/10.18653/v1/2021.naacl-main.201.

3. Adebara, I.; Abdul-Mageed, M. Towards Afrocentric NLP for African Languages: Where We Are and Where We Can Go. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 2022; pp. 3814–3841. https://doi.org/10.18653/v1/2022.acl-long.265.
4. Esch, D.V.; Lucassen, T.; Ruder, S.; Caswell, I.; Rivera, C.E. Writing System and Speaker Metadata for 2,800+ Language Varieties. In Proceedings of the Proceedings of the Language Resources and Evaluation Conference, Marseille, France, 2022; pp. 5035–5046.
5. Adebara, I.; Toyin, H.O.; Ghebremichael, N.T.; Elmadany, A.A.; Abdul-Mageed, M. Where Are We? Evaluating LLM Performance on African Languages. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 32704–32731. https://doi.org/10.18653/v1/2025.acl-long.1572.
6. Olofsson, S. When AI Can't Understand Your Language, Democracy Breaks Down. https://www.techpolicy.press/when-ai-cant-understand-your-language-democracy-breaks-down-/, 2025. Accessed: 2025-12-17.
7. Simpson, A. *Language and National Identity in Africa*; Oxford University Press, 2008. https://doi.org/10.1093/oso/9780199286744.001.0001.
8. Dimé, M. Reflux des solidarités intergénérationnelles en contexte de précarité à Dakar. *Gérontologie et société* **2019**, *41*, 85–98. https://doi.org/10.3917/gs1.158.0085.
9. Tonja, A.L.; Belay, T.D.; Azime, I.A.; Ayele, A.A.; Mehamed, M.A.; Kolesnikova, O.; Yimam, S.M. Natural Language Processing in Ethiopian Languages: Current State, Challenges, and Opportunities, 2023, [arXiv:cs.CL/2303.14406].
10. Leclerc, J. « Senegal » dans *L'aménagement linguistique dans le monde*, Québec, CEFAN, Université Laval. http://www.axl.cefan.ulaval.ca/afrique/senegal.htm, 2015. Accessed: 2025-12-16.
11. World Bank. Population, total - Senegal. https://data.worldbank.org/indicator/SP.POP.TOTL?locations=SN, 2024. Accessed: 2025-12-16.
12. Ouzerrout, S.; Saadallah, I. Réhabiliter l'écriture Ajami : un levier technologique pour l'alphabétisation en Afrique. In Proceedings of the Actes des 18e Rencontres Jeunes Chercheurs en RI (RJCRI) et 27ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL); Bechet, F.; Chifu, A.G.; Pinel-sauvagnat, K.; Favre, B.; Maes, E.; Nurbakova, D., Eds., Marseille, France, 6 2025; pp. 253–267.
13. Ngom, F. Ajami Scripts in the Senegalese Speech Community. *Journal of Arabic and Islamic Studies* **2017**, *10*, 1–23. https://doi.org/10.5617/jais.4599.
14. Nguer, E.M.; Bao, D.S.; Fall, Y.A.; Khoule, M. Digraph of Senegal s local languages: issues, challenges and prospects of their transliteration, 2020, [arXiv:cs.CL/2005.02325].
15. Le, N.T.; Mijiyawa, A.; Leye, A.; Sadat, F. The Best of Both Worlds: Exploring Wolofal in the Context of NLP. In Proceedings of the Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script; El-Haj, M., Ed., Abu Dhabi, UAE, 2025; pp. 1–6.
16. Fall, E.M.; hadji M. Nguer, E.; Sokhna, B.D.; Khoule, M.; Mangeot, M.; Cisse, M.T. Digraphie des langues ouest africaines : Latin2Ajami : un algorithme de translitteration automatique, 2020, [arXiv:cs.CL/2005.02827].
17. Eberhard, D.; Simons, G.; Fennig, C. *Ethnologue: Languages of the World, 22nd Edition*; SIL International, 2019.
18. Gauthier, E.; Besacier, L.; Voisin, S.; Melese, M.; Elingui, U.P. Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: a Case Study of Wolof. In Proceedings of the Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16); Calzolari, N.; Choukri, K.; Declerck, T.; Goggi, S.; Grobelnik, M.; Maegaard, B.; Mariani, J.; Mazo, H.; Moreno, A.; Odijk, J.; et al., Eds., Portorož, Slovenia, 2016; pp. 3863–3867.
19. Çetinoğlu, Ö.; Schulz, S.; Vu, N.T. Challenges of Computational Processing of Code-Switching. In Proceedings of the Proceedings of the Second Workshop on Computational Approaches to Code Switching, Austin, Texas, 2016; pp. 1–11. https://doi.org/10.18653/v1/W16-5801.
20. Kihm, A. Le sérère (seereer siin). working paper or preprint.
21. Creissels, D. Le joola fooñi. working paper or preprint.
22. Creissels, D. Le mandinka (màndì■kàká■ò). working paper or preprint.
23. Tapo, A.A.; Assogba, K.; Homan, C.M.; Rafique, M.M.; Zampieri, M. Bayelemabaga: Creating Resources for Bambara NLP. In Proceedings of the Proceedings of the 2025 Conference of the Nations of the Americas

 doi:10.20944/preprints202601.1124.v1

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers); Chiruzzo, L.; Ritter, A.; Wang, L., Eds., Albuquerque, New Mexico, 2025; pp. 12060–12070. https://doi.org/10.18653/v1/2025.naacl-long.602.

24. Creissels, D.; Ismael, D. LE SONINKÉ[quelques contrastes pertinents pour l'acquisition du Français Langue Seconde par des locuteurs du soninké], 2016. LE SONINKÉ[quelques contrastes pertinents pour l'acquisition du Français Langue Seconde par des locuteurs du soninké].
25. Scharff, C.; Gaikhe, O.; Tretyakov, J.; Shah, N.K.; Scharff, E. AI Strategies: A Review of Selected National and Intergovernmental Approaches. In Proceedings of the Research and Innovation Forum 2024; Visvizi, A.; Troisi, O.; Corvello, V.; Grimaldi, M., Eds., Cham, 2026; pp. 3–17.
26. Ndiaye, S. NLP and Some Research Results in Senegal. In Proceedings of the Mathematics of Computer Science, Cybersecurity and Artificial Intelligence; Gueye, C.T.; Ngom, P.; Diop, I., Eds., Cham, 2024; pp. 21–29.
27. Grallet, G. *Pionniers: Voyage aux frontières de l'intelligence artificielle*; Bernard Grasset: Paris, France, 2025. Release date: 05/11/2025.
28. Heng, S.; Tsilionis, K.; Scharff, C.; Wautelet, Y. Understanding AI ecosystems in the Global South: The cases of Senegal and Cambodia. *International Journal of Information Management* **2022**, *64*, 102454. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2021.102454.
29. Sinha, Y.R. What is Parsing in NLP: Its Types and Techniques. https://intellipaat.com/blog/what-is-parsing-in-nlp/, 2025. Accessed: 2025-12-05.
30. Jaiswal, S. Natural Language Processing – Dependency Parsing. https://towardsdatascience.com/natural-language-processing-dependency-parsing-cf094bbbe3f7/, 2021. Accessed: 2025-12-05.
31. Dione, C.M.B. A Morphological Analyzer For Wolof Using Finite-State Techniques. In Proceedings of the Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 2012; pp. 894–901.
32. Dione, C.B., Handling Wolof clitics in LFG. In *Challenging Clitics*; Salvesen, C.M.; Helland, H.P., Eds.; John Benjamins Publishing Company, 2013; chapter Handling Wolof clitics in LFG, pp. 87–118. https://doi.org/doi:10.1075/la.206.04dio.
33. Dione, C.M.B. Finite-State Tokenization for a Deep Wolof LFG Grammar. *Bergen Language and Linguistics Studies* **2017**, *8*. https://doi.org/10.15845/bells.v8i1.1340.
34. Dione, C.M.B. Implementation and Evaluation of an LFG-based Parser for Wolof. In Proceedings of the Proceedings of the Twelfth Language Resources and Evaluation Conference; Calzolari, N.; Béchet, F.; Blache, P.; Choukri, K.; Cieri, C.; Declerck, T.; Goggi, S.; Isahara, H.; Maegaard, B.; Mariani, J.; et al., Eds., Marseille, France, 2020; pp. 5128–5136.
35. Sulger, S.; Butt, M.; King, T.H.; Meurer, P.; Laczkó, T.; Rákosi, G.; Dione, C.B.; Dyvik, H.; Rosén, V.; De Smedt, K.; et al. ParGramBank: The ParGram Parallel Treebank. In Proceedings of the Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Schuetze, H.; Fung, P.; Poesio, M., Eds., Sofia, Bulgaria, 2013; pp. 550–560.
36. Dione, C.M.B. Pruning the Search Space of the Wolof LFG Grammar Using a Probabilistic and a Constraint Grammar Parser. In Proceedings of the Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14); Calzolari, N.; Choukri, K.; Declerck, T.; Loftsson, H.; Maegaard, B.; Mariani, J.; Moreno, A.; Odijk, J.; Piperidis, S., Eds., Reykjavik, Iceland, 2014; pp. 2863–2870.
37. Dione, B. LFG parse disambiguation for Wolof. *Journal of Language Modelling* **2014**, *2*, 105. https://doi.org/10.15398/jlm.v2i1.81.
38. Dione, C.B. Developing Universal Dependencies for Wolof. In Proceedings of the Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019); Rademaker, A.; Tyers, F., Eds., Paris, France, 2019; pp. 12–23. https://doi.org/10.18653/v1/W19-8003.
39. Qi, P.; Zhang, Y.; Zhang, Y.; Bolton, J.; Manning, C.D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, 2020, [arXiv:cs.CL/2003.07082].
40. Arnett, C.; Hudspeth, M.; O'Connor, B. Evaluating Morphological Alignment of Tokenizers in 70 Languages, 2025, [arXiv:cs.CL/2507.06378].
41. Dione, C.M.B. From LFG To UD: A Combined Approach. In Proceedings of the Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020); de Marneffe, M.C.; de Lhoneux, M.; Nivre, J.; Schuster, S., Eds., Barcelona, Spain (Online), 2020; pp. 57–66.
42. Dione, C.M.B. Multilingual Dependency Parsing for Low-Resource African Languages: Case Studies on Bambara, Wolof, and Yoruba. In Proceedings of the Proceedings of the 17th International Conference on

 doi:10.20944/preprints202601.1124.v1

Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021); Oepen, S.; Sagae, K.; Tsarfaty, R.; Bouma, G.; Seddah, D.; Zeman, D., Eds., Online, 2021; pp. 84–92. https://doi.org/10.18653/v1/2021.iwpt-1.9.

43. HuggingFace. Token Classification. https://huggingface.co/tasks/token-classification, 2024. Accessed: 2025-11-26.
44. Adelani, D.I.; Abbott, J.; Neubig, G.; D'souza, D.; Kreutzer, J.; Lignos, C.; Palen-Michel, C.; Buzaaba, H.; Rijhwani, S.; Ruder, S.; et al. MasakhaNER: Named Entity Recognition for African Languages. *Transactions of the Association for Computational Linguistics* **2021**, *9*, 1116–1131. https://doi.org/10.1162/tacl_a_00416.
45. Adelani, D.I.; Neubig, G.; Ruder, S.; Rijhwani, S.; Beukman, M.; Palen-Michel, C.; Lignos, C.; Alabi, J.O.; Muhammad, S.H.; Nabende, P.; et al. MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing; Goldberg, Y.; Kozareva, Z.; Zhang, Y., Eds., Abu Dhabi, United Arab Emirates, 2022; pp. 4488–4508. https://doi.org/10.18653/v1/2022.emnlp-main.298.
46. Dione, C.M.B.; Kuhn, J.; Zarrieß, S. Design and Development of Part-of-Speech-Tagging Resources for Wolof (Niger-Congo, spoken in Senegal). In Proceedings of the Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10); Calzolari, N.; Choukri, K.; Maegaard, B.; Mariani, J.; Odijk, J.; Piperidis, S.; Rosner, M.; Tapias, D., Eds., Valletta, Malta, 2010.
47. Dione, C.M.B.; Adelani, D.I.; Nabende, P.; Alabi, J.; Sindane, T.; Buzaaba, H.; Muhammad, S.H.; Emezue, C.C.; Ogayo, P.; Aremu, A.; et al. MasakhaPOS: Part-of-Speech Tagging for Typologically Diverse African languages. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Rogers, A.; Boyd-Graber, J.; Okazaki, N., Eds., Toronto, Canada, 2023; pp. 10883–10900. https://doi.org/10.18653/v1/2023.acl-long.609.
48. Geitgey, A. Text Classification is Your New Secret Weapon. https://medium.com/@ageitgey/text-classification-is-your-new-secret-weapon-7ca4fad15788, 2018. Accessed: 2025-11-27.
49. Amazon, A. What is Sentiment Analysis? https://aws.amazon.com/what-is/sentiment-analysis/, 2024. Accessed: 2025-11-27.
50. Kandé, D.; Marone, R.M.; Ndiaye, S.; Camara, F. A Novel Term Weighting Scheme Model. In Proceedings of the Proceedings of the 4th International Conference on Frontiers of Educational Technologies, New York, NY, USA, 2018; ICFET '18, p. 92–96. https://doi.org/10.1145/3233347.3233374.
51. Das, M.; K., S.; Alphonse, P.J.A. A Comparative Study on TF-IDF feature Weighting Method and its Analysis using Unstructured Dataset, 2023, [arXiv:cs.CL/2308.04037].
52. Chen, K.; Zhang, Z.; Long, J.; Zhang, H. Turning from TF-IDF to TF-IGM for term weighting in text classification. *Expert Syst. Appl.* **2016**, *66*, 245–260. https://doi.org/10.1016/j.eswa.2016.09.009.
53. Mbaye, D.; Diallo, M. Beqi: Revitalize the Senegalese Wolof Language with a Robust Spelling Corrector. In Proceedings of the Innovations and Interdisciplinary Solutions for Underserved Areas; Cheikh M. F. Kebe, K.; Gueye, A.; Ndiaye, A.; Sene, N.A.; Maiga, A.S., Eds., Cham, 2025; pp. 311–325.
54. Kandé, D.; Camara, F.; Ndiaye, S.; Guirassy, F.M.L. FWLSA-score: French and Wolof Lexicon-based for Sentiment Analysis. In Proceedings of the 2019 5th International Conference on Information Management (ICIM), 2019, pp. 215–220. https://doi.org/10.1109/INFOMAN.2019.8714667.
55. Samb, S.M.K.; Kandé, D.; Camara, F.; Ndiaye, S. Improved Bilingual Sentiment Analysis Lexicon Using Word-level Trigram. In Proceedings of the 2019 IEEE 5th International Conference on Computer and Communications (ICCC), 2019, pp. 112–119. https://doi.org/10.1109/ICCC47050.2019.9064223.
56. Sarr, A.D.; Kandé, D.; Camara, F. Markov Model for French-Wolof Text Analysis. In Proceedings of the 2023 International Conference on Communications, Computing and Artificial Intelligence (CCCAI), 2023, pp. 29–34. https://doi.org/10.1109/CCCAI59026.2023.00014.
57. Faty, L.; Ndiaye, M.; Sarr, E.N.; Sall, O.; Mbaye, S.N.; Landu, T.T.; Birregah, B.; Bousso, M.; Toure, F. SenOpinion: A New Lexicon for Opinion Tagging in Senegalese News Comments. In Proceedings of the 2020 15th Iberian Conference on Information Systems and Technologies (CISTI), 2020, pp. 1–6. https://doi.org/10.23919/CISTI49556.2020.9140887.
58. Sarr, A.D.; Kandé, D.; Camara, F. A lexicon-based sentiment analysis approach using a graph structure for modeling relationships between opinion words in French and Wolof corpora. In Proceedings of the Proceedings of the 2024 2nd International Conference on Communications, Computing and Artificial Intelligence, New York, NY, USA, 2024; CCCAI '24, p. 71–76. https://doi.org/10.1145/3676581.3676594.
59. Faty, L.; Drame, K.; Ngor Sarr, E.; Ndiaye, M.; Dia, Y.; Sall, O. COMFO : Corpus Multilingue pour la Fouille d'Opinions (COMFO: Multilingual Corpus for Opinion Mining). In Proceedings of the Actes de la 29e

 doi:10.20944/preprints202601.1124.v1

Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale; Estève, Y.; Jiménez, T.; Parcollet, T.; Zanon Boito, M., Eds., Avignon, France, 6 2022; pp. 297–304.

60. Mbaye, D.; Seye, M.R.; Diallo, M.; Ndiaye, M.L.; Sow, D.; Adjanohoun, D.S.; Mbengue, T.; Wade, C.S.; Pablo, D.R.; Munyaka, J.C.B.; et al. Sentiment Analysis on the Young People's Perception About the Mobile Internet Costs in Senegal. In Proceedings of the Proceedings of Tenth International Congress on Information and Communication Technology; Yang, X.S.; Sherratt, R.S.; Dey, N.; Joshi, A., Eds., Singapore, 2026; pp. 201–217.
61. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics; Jurafsky, D.; Chai, J.; Schluter, N.; Tetreault, J., Eds., Online, 2020; pp. 8440–8451. https://doi.org/10.18653/v1/2020.acl-main.747.
62. Barbieri, F.; Anke, L.E.; Camacho-Collados, J. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond, 2022, [arXiv:cs.CL/2104.12250].
63. Malik, J.S.; Qiao, H.; Pang, G.; van den Hengel, A. Deep Learning for Hate Speech Detection: A Comparative Study, 2023, [arXiv:cs.CL/2202.09517].
64. Jacobs, C.; Rakotonirina, N.C.; Chimoto, E.A.; Bassett, B.A.; Kamper, H. Towards hate speech detection in low-resource languages: Comparing ASR to acoustic word embeddings on Wolof and Swahili, 2023, [arXiv:cs.CL/2306.00410].
65. Ndao, I.; Dramé, K.; Sambe, G.; Diallo, G. Annotated tweet data of mixed Wolof-French for detecting obnoxious messages. *Data in Brief* **2025**, *60*, 111500. https://doi.org/https://doi.org/10.1016/j.dib.2025.111500.
66. Ndao, I.; Dramé, K.; Sambe, G.; Diallo, G. Comparative Study of Machine Learning Models for the Detection of Abusive Messages: Case of Wolof-French Codes Mixing Data. In Proceedings of the Innovations and Interdisciplinary Solutions for Underserved Areas; Cheikh M. F. Kebe, K.; Gueye, A.; Ndiaye, A.; Sene, N.A.; Maiga, A.S., Eds., Cham, 2025; pp. 252–263.
67. Ndao, I.; Dramé, K.; Sambe, G.; Diallo, G. AbuseBERT-WoFr: refined BERT model for detecting abusive messages on tweets mixing Wolof-French codes. In Proceedings of the Proceedings of Digital Avenues for Low-Resource Languages of Sub-Saharan Africa (DASSA 2025); Melatagia Yonta, P.; Nouvel, D.; Valentin, S., Eds., Yaoundé, Cameroon, 2025; p. 42 p. Source Agritrop Cirad (https://agritrop.cirad.fr/614384/).
68. Razumovskaia, E.; Glavaš, G.; Majewska, O.; Ponti, E.M.; Korhonen, A.; Vulić, I. Crossing the Conversational Chasm: A Primer on Natural Language Processing for Multilingual Task-Oriented Dialogue Systems, 2022, [arXiv:cs.CL/2104.08570].
69. Ravuri, S.; Stolcke, A. Recurrent Neural Network and LSTM Models for Lexical Utterance Classification. In Proceedings of the Proc. Interspeech. ISCA - International Speech Communication Association, September 2015, pp. 135–139.
70. Mesnil, G.; Dauphin, Y.; Yao, K.; Bengio, Y.; Deng, L.; Hakkani-Tur, D.; He, X.; Heck, L.; Tur, G.; Yu, D.; et al. Using Recurrent Neural Networks for Slot Filling in Spoken Language Understanding. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **2015**, *23*, 530–539. https://doi.org/10.1109/TASLP.2014.2383614.
71. Weld, H.; Huang, X.; Long, S.; Poon, J.; Han, S.C. A survey of joint intent detection and slot-filling models in natural language understanding, 2021, [arXiv:cs.CL/2101.08091].
72. Yu, H.; Alabi, J.O.; Bukula, A.; Zhuang, J.Y.; Lee, E.S.A.; Guge, T.K.; Azime, I.A.; Buzaaba, H.; Sibanda, B.K.; Kalipe, G.K.; et al. INJONGO: A Multicultural Intent Detection and Slot-filling Dataset for 16 African Languages, 2025, [arXiv:cs.CL/2502.09814].
73. Kandji, A.K.; Precioso, F.; Ba, C.; Ndiaye, S.; Ndione, A. WolBanking77: Wolof Banking Speech Intent Classification Dataset, 2025, [arXiv:cs.CL/2509.19271].
74. Casanueva, I.; Temčinas, T.; Gerz, D.; Henderson, M.; Vulić, I. Efficient Intent Detection with Dual Sentence Encoders. In Proceedings of the Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI; Wen, T.H.; Celikyilmaz, A.; Yu, Z.; Papangelis, A.; Eric, M.; Kumar, A.; Casanueva, I.; Shah, R., Eds., Online, 2020; pp. 38–45. https://doi.org/10.18653/v1/2020.nlp4convai-1.5.
75. Ying, C.; Thomas, S. Label Errors in BANKING77. In Proceedings of the Proceedings of the Third Workshop on Insights from Negative Results in NLP; Tafreshi, S.; Sedoc, J.; Rogers, A.; Drozd, A.; Rumshisky, A.; Akula, A., Eds., Dublin, Ireland, 2022; pp. 139–143. https://doi.org/10.18653/v1/2022.insights-1.19.
76. Yu, S.; Sun, Q.; Zhang, H.; Jiang, J. Translate-Train Embracing Translationese Artifacts. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume

 doi:10.20944/preprints202601.1124.v1

2: Short Papers); Muresan, S.; Nakov, P.; Villavicencio, A., Eds., Dublin, Ireland, 2022; pp. 362–370. https://doi.org/10.18653/v1/2022.acl-short.40.

77. Singh, S.; Vargus, F.; D'souza, D.; Karlsson, B.F.; Mahendiran, A.; Ko, W.Y.; Shandilya, H.; Patel, J.; Mataciunas, D.; O'Mahony, L.; et al. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 11521–11567. https://doi.org/10.18653/v1/2024.acl-long.620.
78. Hládek, D.; Staš, J.; Pleva, M. Survey of Automatic Spelling Correction. *Electronics* **2020**, *9*. https://doi.org/10.3390/electronics9101670.
79. Hládek, D.; Staš, J.; Pleva, M. Survey of Automatic Spelling Correction. *Electronics* **2020**, *9*, 1670. https://doi.org/10.3390/electronics9101670.
80. Mangeot, M.; Enguehard, C. DILAF : des dictionnaires africains en ligne et une méthodologie. In Proceedings of the Francophonie et Langues Nationales, Dakar, Senegal, 2014.
81. Mbodj, C.; Enguehard, C. Production et mise en ligne d'un dictionnaire électronique du wolof. *JEP-TALN-RECITAL* **2015**, *2*.
82. Lo, A.; Nguer, E.H.M.; Abdoulaye, N.; Dione, C.B.; Mangeot, M.; Khoule, M.; Bao-Diop, S.; Cissé, M.T. Correction orthographique pour la langue wolof : état de l'art et perspectives. In Proceedings of the JEP-TALN-RECITAL 2016: Traitement Automatique des Langues Africaines TALAF 2016, Paris, France, 2016.
83. Khoule, M.; Mangeot, M.; Nguer, E.H.M.; Cissé, M.T. iBaatukaay : un projet de base lexicale multilingue contributive sur le web à structure pivot pour les langues africaines notamment sénégalaises. In Proceedings of the hal-02054921 , version 1, 2016.
84. Cissé, T.I.; Sadat, F. Automatic Spell Checker and Correction for Under-represented Spoken Languages: Case Study on Wolof. In Proceedings of the Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023); Mabuya, R.; Mthobela, D.; Setaka, M.; Van Zaanen, M., Eds., Dubrovnik, Croatia, 2023; pp. 1–10. https://doi.org/10.18653/v1/2023.rail-1.1.
85. Goodman, J. The State of the Art in Language Modeling. In Proceedings of the Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Tutorials - Volume 5, USA, 2003; NAACL-Tutorials '03, p. 4. https://doi.org/10.3115/1075168.1075172.
86. Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. *ArXiv* **2014**, *1409*.
87. Cissé, T.I.; Sadat, F. Advancing Language Diversity and Inclusion: Towards a Neural Network-based Spell Checker and Correction for Wolof. In Proceedings of the Proceedings of the Fifth Workshop on Resources for African Indigenous Languages @ LREC-COLING 2024; Mabuya, R.; Matfunjwa, M.; Setaka, M.; van Zaanen, M., Eds., Torino, Italia, 2024; pp. 140–151.
88. Khoulé, M.; Mangeot, M.; Nguer, M. Manipulation de dictionnaires d'origines diverses pour des langues peu dotées : la méthodologie iBaatukaay. In Proceedings of the Traitement Automatique des Langues Africaines 2018, Grenoble, France, 2018.
89. IFAN. Sentermino, première plateforme terminologique du Sénégal. https://ifan.ucad.sn/sentermino-premiere-plateforme-terminologique-du-senegal/, 2025. Accessed: 2025-11-24.
90. Outeirinho, D.B.; Otero, P.G.; de Dios-Flores, I.; Campos, J.R.P. Exploring the effects of vocabulary size in neural machine translation: Galician as a target language. In Proceedings of the Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1; Gamallo, P.; Claro, D.; Teixeira, A.; Real, L.; Garcia, M.; Oliveira, H.G.; Amaro, R., Eds., Santiago de Compostela, Galicia/Spain, 2024; pp. 600–604.
91. Koehn, P. *Statistical Machine Translation*; Cambridge University Press, 2009. https://doi.org/10.1017/CBO9780511815829.
92. Luong, T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. In Proceedings of the Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015; pp. 1412–1421. https://doi.org/10.18653/v1/D15-1166.
93. Koehn, P.; Knowles, R. Six Challenges for Neural Machine Translation. In Proceedings of the Proceedings of the First Workshop on Neural Machine Translation, Vancouver, 2017; pp. 28–39. https://doi.org/10.18653/v1/W17-3204.

 doi:10.20944/preprints202601.1124.v1

94. Federmann, C.; Kocmi, T.; Xin, Y. NTREX-128 – News Test References for MT Evaluation of 128 Languages. In Proceedings of the Proceedings of the First Workshop on Scaling Up Multilingual Evaluation; Ahuja, K.; Anastasopoulos, A.; Patra, B.; Neubig, G.; Choudhury, M.; Dandapat, S.; Sitaram, S.; Chaudhary, V., Eds., Online, 2022; pp. 21–24. https://doi.org/10.18653/v1/2022.sumeval-1.4.
95. Caswell, I.; Nielsen, E.; Luo, J.; Cherry, C.; Kovacs, G.; Shemtov, H.; Talukdar, P.; Tewari, D.; Diane, B.M.; Diane, D.; et al. SMOL: Professionally translated parallel data for 115 under-represented languages, 2025, [arXiv:cs.CL/2502.12301].
96. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.u.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems; Guyon, I.; Luxburg, U.V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; Garnett, R., Eds. Curran Associates, Inc., 2017, Vol. 30.
97. Ranathunga, S.; Lee, E.S.A.; Skenduli, M.P.; Shekhar, R.; Alam, M.; Kaur, R. Neural Machine Translation for Low-Resource Languages: A Survey, 2021, [arXiv:cs.CL/2106.15115].
98. Nguer, E.M.; Lo, A.; Dione, C.M.B.; Ba, S.O.; Lo, M. SENCORPUS: A French-Wolof Parallel Corpus. In Proceedings of the Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 2020; pp. 2803–2811.
99. Lo, A.; Dione, C.M.B.; Nguer, E.M.; Ba, S.O.; Lo, M. Building Word Representations for Wolof Using Neural Networks. In Proceedings of the Innovations and Interdisciplinary Solutions for Underserved Areas; Thorn, J.P.R.; Gueye, A.; Hejnowicz, A.P., Eds., Cham, 2020; pp. 274–286.
100. Alla, L.; Bamba, D.C.; Mamadou, N.E.; Ba, B.S.O.; Moussa, L. Using LSTM to Translate French to Senegalese Local Languages: Wolof as a Case Study, 2020, [arXiv:cs.CL/2004.13840].
101. Hochreiter, S.; Schmidhuber, J. Long Short-term Memory. *Neural computation* **1997**, *9*, 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.
102. Mbaye, D.; Diallo, M.; Diop, T.I. Low-Resourced Machine Translation for Senegalese Wolof Language. In Proceedings of the Proceedings of Eighth International Congress on Information and Communication Technology; Yang, X.S.; Sherratt, R.S.; Dey, N.; Joshi, A., Eds., Singapore, 2024; pp. 243–255.
103. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002; pp. 311–318. https://doi.org/10.3115/1073083.1073135.
104. Jónsson, H.P.; Símonarson, H.B.; Snæbjarnarson, V.; Steingrímsson, S.; Loftsson, H. Experimenting with Different Machine Translation Models in Medium-Resource Settings. In Proceedings of the Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings, Berlin, Heidelberg, 2020; p. 95–103. https://doi.org/10.1007/978-3-030-58323-1_10.
105. Dione, C.M.B.; Lo, A.; Nguer, E.M.; Ba, S. Low-resource Neural Machine Translation: Benchmarking State-of-the-art Transformer for Wolof<->French. In Proceedings of the Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 2022; pp. 6654–6661.
106. Kapoor, S.; Cantrell, E.; Peng, K.; Pham, T.H.; Bail, C.A.; Gundersen, O.E.; Hofman, J.M.; Hullman, J.; Lones, M.A.; Malik, M.M.; et al. REFORMS: Reporting Standards for Machine Learning Based Science, 2023, [arXiv:cs.LG/2308.07832].
107. Lones, M.A. How to avoid machine learning pitfalls: a guide for academic researchers, 2024, [arXiv:cs.LG/2108.02497]. https://doi.org/https://doi.org/10.1016/j.patter.2024.101046.
108. Lo, A.; Nguer, E.M.; Ba, S.O.; Dione, C.M.B.; Lo, M. SenTekki: Online Platform and Restful Web Service for Translation Between Wolof and French. In Proceedings of the Innovations and Interdisciplinary Solutions for Underserved Areas; Mambo, A.D.; Gueye, A.; Bassioni, G., Eds., Cham, 2022; pp. 290–298.
109. Arivazhagan, N.; Bapna, A.; Firat, O.; Lepikhin, D.; Johnson, M.; Krikun, M.; Chen, M.X.; Cao, Y.; Foster, G.; Cherry, C.; et al. Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges, 2019, [arXiv:cs.CL/1907.05019].
110. Adelani, D.; Alabi, J.; Fan, A.; Kreutzer, J.; Shen, X.; Reid, M.; Ruiter, D.; Klakow, D.; Nabende, P.; Chang, E.; et al. A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, United States, 2022; pp. 3053–3070. https://doi.org/10.18653/v1/2022.naacl-main.223.
111. Fan, A.; Bhosale, S.; Schwenk, H.; Ma, Z.; El-Kishky, A.; Goyal, S.; Baines, M.; Celebi, O.; Wenzek, G.; Chaudhary, V.; et al. Beyond English-Centric Multilingual Machine Translation, 2020, [arXiv:cs.CL/2010.11125].

 doi:10.20944/preprints202601.1124.v1

112. Mohammadshahi, A.; Nikoulina, V.; Berard, A.; Brun, C.; Henderson, J.; Besacier, L. SMaLL-100: Introducing Shallow Multilingual Machine Translation Model for Low-Resource Languages, 2022, [arXiv:cs.CL/2210.11621].
113. Mbaye, D.; Diallo, M. Task-Oriented Dialog Systems for the Senegalese Wolof Language. In Proceedings of the Proceedings of the 31st International Conference on Computational Linguistics; Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B.D.; Schockaert, S., Eds., Abu Dhabi, UAE, 2025; pp. 4803–4812.
114. Team, N.; Costa-jussà, M.R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; et al. No Language Left Behind: Scaling Human-Centered Machine Translation, 2022, [arXiv:cs.CL/2207.04672].
115. Anil, R.; Dai, A.M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. PaLM 2 Technical Report, 2023, [arXiv:cs.CL/2305.10403].
116. Caswell, I. 110 new languages are coming to Google Translate. https://blog.google/products/translate/google-translate-new-languages-2024/, 2024. Accessed: 2025-11-23.
117. DeepL. DeepL Translator languages. https://support.deepl.com/hc/en-us/articles/360019925219-DeepL-Translator-languages, 2025. Accessed: 2025-11-28.
118. Smartling. How Accurate is DeepL? https://www.smartling.com/blog/how-accurate-is-deepl, 2024. Accessed: 2025-11-28.
119. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed.; Prentice Hall, 2025. Online manuscript released January 12, 2025.
120. Young, S.; Gašić, M.; Thomson, B.; Williams, J.D. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proceedings of the IEEE* **2013**, *101*, 1160–1179. https://doi.org/10.1109/JPROC.2012.2225812.
121. Adelani, D.I.; Ojo, J.; Azime, I.A.; Zhuang, J.Y.; Alabi, J.O.; He, X.; Ochieng, M.; Hooker, S.; Bukula, A.; Lee, E.S.A.; et al. IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models, 2024, [arXiv:cs.CL/2406.03368].
122. Kudugunta, S.; Caswell, I.; Zhang, B.; Garcia, X.; Choquette-Choo, C.A.; Lee, K.; Xin, D.; Kusupati, A.; Stella, R.; Bapna, A.; et al. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset, 2023, [arXiv:cs.CL/2309.04662].
123. Kreutzer, J.; Caswell, I.; Wang, L.; Wahab, A.; van Esch, D.; Ulzii-Orshikh, N.; Tapo, A.; Subramani, N.; Sokolov, A.; Sikasote, C.; et al. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics* **2022**, *10*, 50–72, [https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00447/109285/Quality-at-a-Glance-An-Audit-of-Web-Crawled]. https://doi.org/10.1162/tacl_a_00447.
124. Hussen, K.Y.; Sewunetie, W.T.; Ayele, A.A.; Imam, S.H.; Muhammad, S.H.; Yimam, S.M. The State of Large Language Models for African Languages: Progress and Challenges, 2025, [arXiv:cs.AI/2506.02280].
125. Adewumi, T.; Adeyemi, M.; Anuoluwapo, A.; Peters, B.; Buzaaba, H.; Samuel, O.; Rufai, A.M.; Ajibade, B.; Gwadabe, T.; Traore, M.M.K.; et al. AfriWOZ: Corpus for Exploiting Cross-Lingual Transferability for Generation of Dialogues in Low-Resource, African Languages, 2022, [arXiv:cs.CL/2204.08083].
126. Budzianowski, P.; Wen, T.H.; Tseng, B.H.; Casanueva, I.; Ultes, S.; Ramadan, O.; Gašić, M. MultiWOZ – A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling, 2020, [arXiv:cs.CL/1810.00278].
127. Ogundepo, O.; Gwadabe, T.R.; Rivera, C.E.; Clark, J.H.; Ruder, S.; Adelani, D.I.; Dossou, B.F.P.; DIOP, A.A.; Sikasote, C.; Hacheme, G.; et al. AfriQA: Cross-lingual Open-Retrieval Question Answering for African Languages, 2023, [arXiv:cs.CL/2305.06897].
128. Bandarkar, L.; Liang, D.; Muller, B.; Artetxe, M.; Shukla, S.N.; Husa, D.; Goyal, N.; Krishnan, A.; Zettlemoyer, L.; Khabsa, M. The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Ku, L.W.; Martins, A.; Srikumar, V., Eds., Bangkok, Thailand, 2024; pp. 749–775. https://doi.org/10.18653/v1/2024.acl-long.44.
129. Blaschke, V.; Fedzechkina, M.; Ter Hoeve, M. Analyzing the Effect of Linguistic Similarity on Cross-Lingual Transfer: Tasks and Experimental Setups Matter. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025; Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., Vienna, Austria, 2025; pp. 8653–8684. https://doi.org/10.18653/v1/2025.findings-acl.454.
130. Idris, T.K.; Mitra, P.; Eiselen, R. Can Embedding Similarity Predict Cross-Lingual Transfer? A Systematic Study on African Languages, 2026, [arXiv:cs.CL/2601.03168].

 doi:10.20944/preprints202601.1124.v1

131. Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback, 2022, [arXiv:cs.CL/2203.02155].
132. CohereForAI. The Aya Movement at a glance: Accelerating multilingual AI through open science. https://cohere.com/research/aya/aya-at-a-glance.pdf, 2024. Accessed: 2025-11-25.
133. Üstün, A.; Aryabumi, V.; Yong, Z.X.; Ko, W.Y.; D'souza, D.; Onilude, G.; Bhandari, N.; Singh, S.; Ooi, H.L.; Kayid, A.; et al. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model, 2024, [arXiv:cs.CL/2402.07827].
134. Yu, H.; Xu, T.; Hedderich, M.A.; Hamidouche, W.; Zamir, S.W.; Adelani, D.I. AfriqueLLM: How Data Mixing and Model Architecture Impact Continued Pre-training for African Languages, 2026, [arXiv:cs.CL/2601.06395].
135. Penedo, G.; Kydlíček, H.; Sabolčec, V.; Messmer, B.; Foroutan, N.; Kargaran, A.H.; Raffel, C.; Jaggi, M.; Werra, L.V.; Wolf, T. FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language, 2025, [arXiv:cs.CL/2506.20920].
136. Penedo, G.; Kydlíček, H.; allal, L.B.; Lozhkov, A.; Mitchell, M.; Raffel, C.; Werra, L.V.; Wolf, T. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale, 2024, [arXiv:cs.CL/2406.17557].
137. Dossou, B.F.P.; Tonja, A.L.; Yousuf, O.; Osei, S.; Oppong, A.; Shode, I.; Awoyomi, O.O.; Emezue, C. AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages. In Proceedings of the Proceedings of the Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP); Fan, A.; Gurevych, I.; Hou, Y.; Kozareva, Z.; Luccioni, S.; Sadat Moosavi, N.; Ravi, S.; Kim, G.; Schwartz, R.; Rücklé, A., Eds., Abu Dhabi, United Arab Emirates (Hybrid), 2022; pp. 52–64. https://doi.org/10.18653/v1/2022.sustainlp-1.11.
138. Ogueji, K.; Zhu, Y.; Lin, J. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages. In Proceedings of the Proceedings of the 1st Workshop on Multilingual Representation Learning; Ataman, D.; Birch, A.; Conneau, A.; Firat, O.; Ruder, S.; Sahin, G.G., Eds., Punta Cana, Dominican Republic, 2021; pp. 116–126. https://doi.org/10.18653/v1/2021.mrl-1.11.
139. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); Burstein, J.; Doran, C.; Solorio, T., Eds., Minneapolis, Minnesota, 2019; pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.
140. Adebara, I.; Elmadany, A.; Abdul-Mageed, M.; Alcoba Inciarte, A. SERENGETI: Massively Multilingual Language Models for Africa. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023; Rogers, A.; Boyd-Graber, J.; Okazaki, N., Eds., Toronto, Canada, 2023; pp. 1498–1537. https://doi.org/10.18653/v1/2023.findings-acl.97.
141. TRT Afrika. AWA: Senegalese start-up's AI muse speaks in Wolof. https://www.trtafrika.com/english/article/18244712, 2024. Accessed: 2025-11-25.
142. Gauthier, E.; Séga Wade, P.; Moudenc, T.; Collen, P.; De Neef, E.; Ba, O.; Khoyane Cama, N.; Bamba Kebe, A.; Aissatou Gningue, N.; Mendo'O Aristide, T. Preuve de concept d'un bot vocal dialoguant en wolof (Proof-of-Concept of a Voicebot Speaking Wolof). In Proceedings of the Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale; Estève, Y.; Jiménez, T.; Parcollet, T.; Zanon Boito, M., Eds., Avignon, France, 6 2022; pp. 403–412.
143. Kang, Y.; Zhang, Y.; Kummerfeld, J.K.; Tang, L.; Mars, J. Data Collection for Dialogue System: A Startup Perspective. In Proceedings of the Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers); Bangalore, S.; Chu-Carroll, J.; Li, Y., Eds., New Orleans - Louisiana, 2018; pp. 33–40. https://doi.org/10.18653/v1/N18-3005.
144. Bocklisch, T.; Faulkner, J.; Pawlowski, N.; Nichol, A. Rasa: Open Source Language Understanding and Dialogue Management, 2017, [arXiv:cs.CL/1712.05181].
145. Akuthota, K.S.; Kishor Kumar Reddy, C.; Shuaib, M.; Alam, S.; Alshanketi, F.; Kyatham, A.R. A Comprehensive Von Willebrand Disease Awareness and Support Chatbot for Senegalese Communities. In Proceedings of the 2025 International Conference on Information Networking (ICOIN), 2025, pp. 714–719. https://doi.org/10.1109/ICOIN63865.2025.10992905.
146. Qwen.; :.; Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; et al. Qwen2.5 Technical Report, 2025, [arXiv:cs.CL/2412.15115].

 doi:10.20944/preprints202601.1124.v1

147. Sy, Y.; Doucoure, D. Oolel: A High-Performing Open LLM for Wolof. https://huggingface.co/soynade-research/Oolel-v0.1, 2024. Accessed: 2025-11-25.
148. Mehrish, A.; Majumder, N.; Bhardwaj, R.; Mihalcea, R.; Poria, S. A Review of Deep Learning Techniques for Speech Processing, 2023, [arXiv:eess.AS/2305.00359].
149. Tamgno, J.K.; Elingui, P.U.; Mendo'o, A.T.; Richomme, M.; Lishou, C.; Obono, S.D.O. Speech Recognition and Text-to-Speech Solution for Vernacular Languages: Free Software and Community Involvement to Develop Voice Services. In Proceedings of the ICDT 2011: The Sixth International Conference on Digital Telecommunications. IARIA, 2011, pp. 56–63.
150. Kakade, S.M.; Krishnamurthy, A.; Mahajan, G.; Zhang, C. Learning Hidden Markov Models Using Conditional Samples, 2024, [arXiv:cs.LG/2302.14753].
151. Besacier, L.; Gauthier, E.; Mangeot, M.; Bretier, P.; Bagshaw, P.; Rosec, O.; Moudenc, T.; Pellegrino, F.; Voisin, S.; Marsico, E.; et al. Speech Technologies for African Languages: Example of a Multilingual Calculator for Education. In Proceedings of the Interspeech 2015 (short demo paper), Dresden, Germany, 2015.
152. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.K.; Hannemann, M.; Motlícek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi Speech Recognition Toolkit. In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, 2011.
153. Gauthier, E.; Besacier, L.; Voisin, S. Automatic Speech Recognition for African Languages with Vowel Length Contrast. *Procedia Computer Science* **2016**, *81*, 136–143. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia, https://doi.org/https://doi.org/10.1016/j.procs.2016.04.041.
154. Gauthier, E.; Besacier, L.; Voisin, S. Speed perturbation and vowel duration modeling for ASR in Hausa and Wolof languages. In Proceedings of the Interspeech 2016 proceedings, San-Francisco, United States, 2016.
155. Gauthier, E.; Besacier, L.; Voisin, S. Machine Assisted Analysis of Vowel Length Contrasts in Wolof, 2017, [arXiv:cs.CL/1706.00465].
156. DeRenzi, B.; Dixon, A.; Farhi, M.; Resch, C. Synthetic Voice Data for Automatic Speech Recognition in African Languages, 2025. https://doi.org/10.48550/arXiv.2507.17578.
157. SpeechBrain. speechbrain/asr-wav2vec2-dvoice-wolof: ASR model for Wolof. https://huggingface.co/speechbrain/asr-wav2vec2-dvoice-wolof, 2024. Accessed: 2025-11-29.
158. Kandji, A.K.; Ba, C.; Ndiaye, S. State-of-the-Art Review on Recent Trends in Automatic Speech Recognition. In Proceedings of the Emerging Technologies for Developing Countries; Masinde, M.; Möbs, S.; Bagula, A., Eds., Cham, 2024; pp. 185–203.
159. Abdou Mohamed, N.; Allak, A.; Gaanoun, K.; et al. Multilingual Speech Recognition Initiative for African Languages. *International Journal of Data Science and Analytics* **2025**, *20*, 3513–3528. https://doi.org/10.1007/s41060-024-00677-9.
160. Siminyu, K.; Kalipe, G.; Orlic, D.; Abbott, J.; Marivate, V.; Freshia, S.; Sibal, P.; Neupane, B.; Adelani, D.I.; Taylor, A.; et al. AI4D – African Language Program, 2021, [arXiv:cs.CL/2104.02516].
161. Pratap, V.; Tjandra, A.; Shi, B.; Tomasello, P.; Babu, A.; Kundu, S.; Elkahky, A.; Ni, Z.; Vyas, A.; Fazel-Zarandi, M.; et al. Scaling Speech Technology to 1,000+ Languages, 2023, [arXiv:cs.CL/2305.13516].
162. Ericsson, L.; Gouk, H.; Loy, C.C.; Hospedales, T.M. Self-Supervised Representation Learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine* **2022**, *39*, 42–62. https://doi.org/10.1109/msp.2021.3134634.
163. Baevski, A.; Zhou, H.; Mohamed, A.; Auli, M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, 2020, [arXiv:cs.CL/2006.11477].
164. Omnilingual ASR team.; Keren, G.; Kozhevnikov, A.; Meng, Y.; Ropers, C.; Setzler, M.; Wang, S.; Adebara, I.; Auli, M.; Balioglu, C.; et al. Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages, 2025, [arXiv:cs.CL/2511.09690].
165. Velluet, Q. Senegalese startup Lengo brings AI to informal retailers. *The Africa Report* **2024**. Accessed: 2025-11-29.
166. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision, 2022, [arXiv:eess.AS/2212.04356].
167. Basilwango, E.; Oche Ankeli, D.B.; Le Beux, Y. Fine-Tuning Automatic Speech Recognition Models for Wolof and Hausa in the Domain of Maternal and Reproductive Health. Deep Learning Indaba Poster session 1: African Datasets, IndabaX, Publications and General posters nos. 60 – 190, https://drive.google.com/file/d/1Qv8Y7SV0oSJoWjktggdDOoHaXBGAFSAt/, 2025. Accessed: 2025-12-18.

 doi:10.20944/preprints202601.1124.v1

168. Diallo, S. Whosper-large: A Multilingual ASR Model for Wolof with Enhanced Code-Switching Capabilities, 2025.
169. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models, 2021, [arXiv:cs.CL/2106.09685].
170. Caubrière, A.; Gauthier, E. Africa-Centric Self-Supervised Pre-Training for Multilingual Speech Representation in a Sub-Saharan Context, 2024, [arXiv:cs.CL/2404.02000].
171. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, 2021, [arXiv:cs.CL/2106.07447].
172. Caubrière, A.; Gauthier, E. Représentation de la parole multilingue par apprentissage auto-supervisé dans un contexte subsaharien. In Proceedings of the Actes des 35èmes Journées d'Études sur la Parole; Balaguer, M.; Bendahman, N.; Ho-dac, L.M.; Mauclair, J.; G Moreno, J.; Pinquier, J., Eds., Toulouse, France, 7 2024; pp. 163–172.
173. Conneau, A.; Ma, M.; Khanuja, S.; Zhang, Y.; Axelrod, V.; Dalmia, S.; Riesa, J.; Rivera, C.; Bapna, A. FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech, 2022, [arXiv:cs.CL/2205.12446].
174. Djionang Pindoh, P.; Melatagia Yonta, P. Self-supervised and multilingual learning applied to the Wolof, Swahili and Fongbe. *Revue Africaine de Recherche en Informatique et Mathématiques Appliquées* **2025**, *42*. soumission à Episciences, https://doi.org/10.46298/arima.13416.
175. Sy, Y.; Doucouré, D.; Cerisara, C.; Illina, I. Speech Language Models for Under-Represented Languages: Insights from Wolof, 2025, [arXiv:cs.CL/2509.15362].
176. Parmar, J.; Satheesh, S.; Patwary, M.; Shoeybi, M.; Catanzaro, B. Reuse, Don't Retrain: A Recipe for Continued Pretraining of Language Models, 2024, [arXiv:cs.CL/2407.07263].
177. Black, A.W.; Taylor, P.; Caley, R. *Festival Speech Synthesis System: System Documentation, edition 2.4, for Festival version 2.4.0*, 2014. Accessed: 2025-11-29.
178. GalsenAI-Lab. Anta: Wolof female-voice Text-to-Speech dataset. https://huggingface.co/datasets/galsenai/anta_women_tts, 2024.
179. Resemble-AI. Introducing Resemble Enhance: Open Source Speech Super Resolution AI Model. https://www.resemble.ai/introducing-resemble-enhance/, 2023. Accessed: 2025-11-29.
180. Casanova, E.; Davis, K.; Gölge, E.; Göknar, G.; Gulea, I.; Hart, L.; Aljafari, A.; Meyer, J.; Morais, R.; Olayemi, S.; et al. XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model, 2024, [arXiv:eess.AS/2406.04904].
181. GalsenAI-Lab. galsenai/xTTS-v2-wolof. https://huggingface.co/galsenai/xTTS-v2-wolof, 2024.
182. Concree. Adia TTS. https://huggingface.co/CONCREE/Adia_TTS, 2025.
183. Lacombe, Y.; Srivastav, V.; Gandhi, S. Parler-TTS. https://github.com/huggingface/parler-tts, 2024.
184. Lyth, D.; King, S. Natural language guidance of high-fidelity text-to-speech with synthetic annotations, 2024, [arXiv:cs.SD/2402.01912].
185. Mbaye, M. TTS-WOLOF : Building Inclusive Voice AI for African Languages – The Wolof Case. https://ascii.org.sn/index.php/cnria-2025, 2025. Accessed: 2025-11-29.
186. Ogayo, P.; Neubig, G.; Black, A.W. Building African Voices, 2022, [arXiv:cs.CL/2207.00688].
187. Ji, S.; Chen, Y.; Fang, M.; Zuo, J.; Lu, J.; Wang, H.; Jiang, Z.; Zhou, L.; Liu, S.; Cheng, X.; et al. WavChat: A Survey of Spoken Dialogue Models, 2024, [arXiv:eess.AS/2411.13577].
188. Fihlani, P. Lost in translation - How Africa is trying to close the AI language gap. https://www.bbc.com/news/articles/crkzgkkpx0lo, 2025. Accessed: 2025-11-28.
189. Peng, J.; Wang, Y.; Li, B.; Guo, Y.; Wang, H.; Fang, Y.; Xi, Y.; Li, H.; Li, X.; Zhang, K.; et al. A Survey on Speech Large Language Models for Understanding, 2025, [arXiv:eess.AS/2410.18908].
190. Arora, S.; Chang, K.W.; Chien, C.M.; Peng, Y.; Wu, H.; Adi, Y.; Dupoux, E.; Lee, H.Y.; Livescu, K.; Watanabe, S. On The Landscape of Spoken Language Models: A Comprehensive Survey, 2025, [arXiv:cs.CL/2504.08528].
191. GalsenAI. Keyword Spotting with African Languages. https://k4all.org/project/keyword-spotting-with-african-languages/, 2019. Accessed: 2025-11-28.
192. Warden, P. Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition, 2018, [arXiv:cs.CL/1804.03209].
193. Adjanohoun, D.S.; Mbengue, T.P.; Sow, D.; Wade, C.S.; Seye, M.R.; Mbaye, D.; Diallo, M.; Ndiaye, M.L.; De-Roulet, P.; Munyaka, J.C.B.; et al. The digital policies in the face of access and usage inequalities among young people in intermediate cities in Senegal: The case of Saint-Louis. *Journal of Infrastructure, Policy and Development* **2025**, *9*, 10899. https://doi.org/10.24294/jipd10899.
194. Sow, D.; Adjanohoun, D.; Mbengue, T.; Wade, C.; Seye, M.; Mbaye, D.; Diallo, M.; Ndiaye, M.; De Roulet, P.; Munyaka Baraka, J.C.; et al. DIGITAL INCLUSION AND YOUTH PARTICIPATION IN URBAN GOVER-

 doi:10.20944/preprints202601.1124.v1

NANCE IN SUB-SAHARAN AFRICA: THE CASE OF SAINT-LOUIS, SENEGAL. *GLOBAL JOURNAL FOR RESEARCH ANALYSIS* **2025**, pp. 50–55. https://doi.org/10.36106/gjra/0103770.

195. Yang, J.; Hussein, A.; Wiesner, M.; Khudanpur, S. JHU IWSLT 2022 Dialect Speech Translation System Description. In Proceedings of the Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022); Salesky, E.; Federico, M.; Costa-jussà, M., Eds., Dublin, Ireland (in-person and online), 2022; pp. 319–326. https://doi.org/10.18653/v1/2022.iwslt-1.29.
196. Hou, Y.; Huang, J. Natural language processing for social science research: A comprehensive review. *Chinese Journal of Sociology* **2025**, *11*, 121–157, [https://doi.org/10.1177/2057150X241306780]. https://doi.org/10.1177/2057150X241306780.
197. Rathje, S.; Mirea, D.M.; Sucholutsky, I.; Marjieh, R.; Robertson, C.E.; Bavel, J.J.V. GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences* **2024**, *121*, e2308950121, [https://www.pnas.org/doi/pdf/10.1073/pnas.2308950121]. https://doi.org/10.1073/pnas.2308950121.
198. Gilardi, F.; Alizadeh, M.; Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* **2023**, *120*. https://doi.org/10.1073/pnas.2305016120.
199. Chiang, C.H.; yi Lee, H. Can Large Language Models Be an Alternative to Human Evaluations?, 2023, [arXiv:cs.CL/2305.01937].
200. Abdurahman, S.; Ziabari, A.S.; Moore, A.K.; Bartels, D.M.; Dehghani, M. A Primer for Evaluating Large Language Models in Social-Science Research. *Advances in Methods and Practices in Psychological Science* **2025**, *8*, 25152459251325174, [https://doi.org/10.1177/25152459251325174]. https://doi.org/10.1177/25152459251325174.
201. Suri, G.; Slater, L.R.; Ziaee, A.; Nguyen, M. Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5, 2023, [arXiv:cs.AI/2305.04400].
202. Ecofin. Atos to deliver Senegal's supercomputer in the coming months. https://www.ecofinagency.com/telecom/1501-39520-atos-to-deliver-senegal-s-supercomputer-in-the-coming-months, 2019. Accessed: 2025-12-14.
203. Socialnetlink. Macky Sall exige la mise en service du Supercalculateur non fonctionnel depuis son acquisition à hauteur de 15 millions d'euros. https://www.socialnetlink.org/2022/01/27/macky-sall-exige-la-mise-en-service-du-supercalculateur-non-fonctionnel-depuis-son-acquisition-a-hauteur-de-15-millions-deuros/, 2022. Accessed: 2025-12-14.
204. UNESCO. Global AI Ethics and Governance Observatory. https://www.unesco.org/ethics-ai/en/senegal, 2025. Accessed: 2025-12-14.
205. Ecofin. Le Sénégal dévoile une stratégie à 206 millions de dollars pour numériser l'éducation. https://www.agenceecofin.com/actualites-numerique/0701-124683-le-senegal-devoile-une-strategie-a-206-millions-de-dollars-pour-numeriser-l-education, 2025. Accessed: 2025-12-14.
206. RTS. Éducation: Le MEN lance un vaste programme d'intégration de l'intelligence artificielle à l'école. https://www.rts.sn/actualite/detail/a-la-une/education-le-men-lance-un-vaste-programme-dintegration-de-lintelligence-artificielle-a-lecole, 2025. Accessed: 2025-12-14.
207. Presidency. Senegal Signs a $10 Million Strategic Partnership with the Gates Foundation to Accelerate the Technological New Deal. https://www.presidence.sn/en/actualites/senegal-signs-a-10-million-strategic-partnership-with-the-gates-foundation-to-accelerate-the-technological-new-deal-1, 2024. Accessed: 2025-12-14.
208. Gueye, O. Le Sénégal en force au Sommet mondial pour l'Action sur l'Intelligence Artificielle à Paris. https://letechobservateur.sn/le-senegal-en-force-au-sommet-mondial-pour-laction-sur-lintelligence-artificielle-a-paris/, 2025. Accessed: 2025-12-14.
209. Sikiru, R. From Senegal with Insights: Learnings from Deep Learning Indaba 2024. https://medium.com/@rasheedatsikiru/from-senegal-with-insights-learnings-from-deep-learning-indaba-2024-def6b334b596, 2024. Accessed: 2025-12-14.
210. Fall, A. SALTIS 2025: Africa codes its future in Dakar. https://www.seneweb.com/en/news/Technologie/saltis-2025-lafrique-code-son-futur-a-dakar_n_464812.html, 2025. Accessed: 2025-12-14.
211. Africa Tech Review. Senegal bolsters AI development with launch of ALIVE and DiCentre4AI laboratories. https://techreviewafrica.com/news/1997/senegal-bolsters-ai-development-with-launch-of-alive-and-dicentre4ai-laboratories, 2025. Accessed: 2025-12-14.
212. Chalumeau, F.; Rajaonarivonivelomanantsoa, D.; de Kock, R.; Formanek, C.; Abramowitz, S.; Mahjoub, O.; Khlifi, W.; Toit, S.D.; Nessir, L.B.; Shabe, R.; et al. Breaking the Performance Ceiling in Reinforcement Learning requires Inference Strategies, 2025, [arXiv:cs.LG/2505.21236].

 doi:10.20944/preprints202601.1124.v1

213. Google Scholar. Top publications. https://scholar.google.es/citations?view_op=top_venues&hl=en&vq=eng_artificialintelligence, 2025. Accessed: 2025-12-18.
214. UNESCO. AI's potential for Africa development and prosperity. https://www.unesco.org/en/articles/ais-potential-africa-development-and-prosperity, 2025. Accessed: 2025-12-14.
215. Ruder, S.; Clark, J.; Gutkin, A.; Kale, M.; Ma, M.; Nicosia, M.; Rijhwani, S.; Riley, P.; Sarr, J.M.; Wang, X.; et al. XTREME-UP: A User-Centric Scarce-Data Benchmark for Under-Represented Languages. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, p. 1856–1884. https://doi.org/10.18653/v1/2023.findings-emnlp.125.
216. Ahesi. Using WhatsApp for Speech Dataset Building in Ghana. https://ashesi-org.github.io/dataset/nlp/ai/whatsapp/speech/2022/05/16/using-whatsapp-speech-dataset.html, 2022. Accessed: 2025-12-14.