Towards Afrocentric NLP for African Languages: Where We Are and Where We Can Go

Ife Adebara
Deep Learning and
Natural Language Processing Group
The University of British Columbia
ife.adebara@ubc.ca

Muhammad Abdul-Mageed
Deep Learning and
Natural Language Processing Group
The University of British Columbia
muhammad.mageed@ubc.ca

Abstract

Aligning with ACL 2022 special Theme on “Language Diversity: from Low Resource to Endangered Languages", we discuss the major linguistic and sociopolitical challenges facing development of NLP technologies for African languages. Situating African languages in a typological framework, we discuss how the particulars of these languages can be harnessed. To facilitate future research, we also highlight current efforts, communities, venues, datasets, and tools. Our main objective is to motivate and advocate for an Afrocentric approach to technology development. With this in mind, we recommend *what* technologies to build and *how* to build, evaluate, and deploy them based on the needs of local African communities.

1 Introduction

Language is the foundation on which communication rests, allowing us to share ideas and interact with one another. Cultures are built on this foundation. We cannot understand, nurture, or help a culture thrive without understanding and nurturing the language carrying it. Language, in turn, is incubated and evolved by culture (Fourie, 1995). Each culture is thus naturally best expressed using the language in which it evolved, which encodes knowledge about people, their traditions, wisdom, environment, and how they interact with the sum of the concepts that belong to their own culture. Technology is an element of culture that arguably both *shapes* and *is shaped by* it. Technology interacts in complex ways with other elements of culture such as gender, race, and class. Natural language processing (NLP) technologies are no exception, and play an increasingly important role in today’s world. Modern NLP technologies, however, have primarily been developed in Western societies. As such, they often function within contexts of values, norms, and beliefs that reflect these societies and serve their needs. On the other hand, the very methods employed to develop most of these technologies and the knowledge on which they rest also derive from the same Western-Centric approaches. This poses challenges to the extension and use of these technologies in communities with different social fabrics that speak different languages. The scale of this problem is huge, because the majority of the world’s 7000+ living languages (Eberhard et al., 2021) are not NLP-supported. Apart from perhaps two dozens of popular languages, most languages of the world are under-resourced, indigenous, and/or endangered. Most African languages fall within this category and are the focus of this paper (Figure 1).

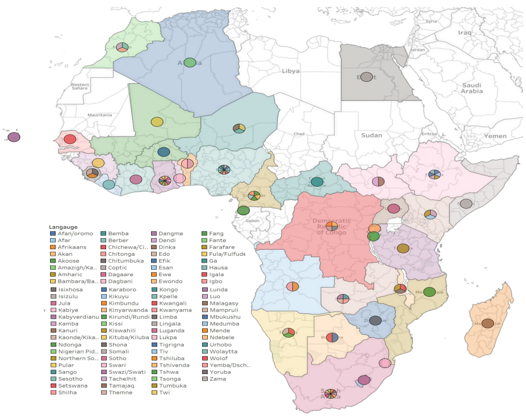

Figure 1: African languages discussed in this paper. A high quality version is in Figure F.1 (Appendix).

Our goal is to discuss the major linguistic and

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3814 - 3841
May 22-27, 2022 ©2022 Association for Computational Linguistics

sociopolitical challenges facing development of NLP technologies for African languages.[1] In doing so, we both motivate and advocate for an Afrocentric approach to technology development where *what* technologies to build and *how* to build, evaluate, and deploy them arise from the needs of local African communities. We start by typologically situating African languages and providing illustrating examples as to what makes them challenging from a computational linguistics perspective (§ 2). Next, we discuss consequences of the literacy situation in Africa on NLP (§ 3). We then further explain why the classical binary approach to technology development of *feature engineering* vs. *end-to-end* solutions familiar to most NLP researchers is not ideal for the African context (§ 4). We follow by data quality (§ 5). To facilitate future work, we also point to ongoing community efforts, venues, and datasets (§ 6). We conclude in § 7.

2 Why Typology Matters

Although it has been argued that the best way to achieve cross-linguistically useful NLP is to leverage findings of typological research (Bender, 2016), most NLP work remains Indo-Eurocentric in terms of algorithms for preprocessing, training, and evaluation. This is a mismatch to the fact that every NLP approach requires either explicit or implicit representative linguistic knowledge (O'Horan et al., 2016; Ponti et al., 2019; Bender, 2016). Knowledge of linguistic typology can indeed be very useful for both language-specific and language-independent NLP (O'Horan et al., 2016), including for African languages. This knowledge can be useful for determining which *languages* may be treated together (e.g., in multilingual models) and/or which *methods* are best suited for a language-specific task (e.g., a method can be deemed potentially useful if it has been applied successfully on a language with a similar typology). To illustrate what typological information can concretely mean for African languages, it may be useful here to list a number of the most notable typological features prevalent in African languages across several language families including *Afro-Asiatic, Austronesian, Niger-Congo, Nilo-Saharan, Indo-European* and *Creole*. These features include use of tone, open syllables, vowel harmony, splitting verbs, serial verb construction, reduplication, use of very few or no adjectives (closed class of adjectives (Segerer, 2008)), and a large number of ideophones. We will discuss three of these features which we judge as largely absent from most of the top 10 NLP-popular languages.[2] We provide a list indicating presence of one or more of these features in over 100 African languages in Appendix Table B.1.

2.1 Tone

Phonemic tone is characteristic of many African languages, with $\sim 80\%$ of these languages being tone languages (Hyman, 2003; Creissels et al., 2008). This includes most languages of the Niger-Congo family, except Swahili, Wolof, Serer, Cangin, and Fulani which are not tone languages. All Nilotic and Khoisan languages and many Afroasiatic languages are also tonal. A smaller number of languages, including Somali and many Bantu languages, are tonal accent languages, in which a distinctive or demarcative accent is expressed by a toneme of high pitch (Clements and Rialland, 2007).

Tone can occur both at the lexical and grammatical levels. Lexical tones are a difference in pitch that distinguishes one lexeme from another. In Yorùbá, for instance, lexical tone is responsible for the differences in meaning in the following: `igbá` (*calabash*, *basket*), `igba` (*200*), `ìgbà` (*time*), `ìgbá` (*garden egg*), and `igbà` (*rope*). Grammatical tone, on the other hand, distinguishes one grammatical category from another. In Akan, a language with both lexical and grammatical tone, grammatical tone distinguishes habitual and stative verbs as in: `Ama dá ha` *'Ama sleeps here'* and `Ama dà ha` *'Ama is sleeping here'*. Grammatical tone is also used to indicate case in some Bantu languages (Creissels et al., 2008; König et al., 2008), as a definite marker, for inflectional or

[1] We do not cover Arabic since it is spread in both Africa and Asia, and has a sizeable NLP community.

[2] We use the language diversity index of Joshi et al. (2020) to select the top 10 languages.

derivational purposes, or to code spatial relations (Creissels et al., 2008).

Two approaches have been adopted in the orthographies of African tone languages: *no tone marking* or *tone marking*. **No Tone Marking.** Hausa, spoken in Niger and Nigeria, has grammatical tone but adopts a no tone marking approach in its orthography. This results in ambiguities that may not be resolved in context as in `jáá tàfí` 'He went', `jáà tàfí` 'He may go', and `jà tàfí` 'He should go' (Cahill, 2019). It is worth mentioning that no tone marking makes little difference in tone languages with few minimal pairs. NLP systems designed for a tone language without tone marking may therefore suffer from issues with ambiguity, if contextual information is not adequate for disambiguation or if many minimal pairs exist in the language.

Tone Marking. Languages that mark tone may adopt a *shallow* marking (Yorùbá) or *deep* marking approach (Cahill, 2019; Bird, 1999a) by using diacritics, punctuation marks, or letters to indicate tone (Cahill, 2019). A shallow marking approach uses the surface level tone after phonological rules (such as assimilation) that change the representation of tones have been applied. The implication of this type of approach is that the same word will have different tone representations in different contexts. In a low-resource scenario, therefore, each word will have fewer occurrences and some contexts may not be seen in training data (Bird, 1999b) (i.e., data sparsity). For languages that adopt a deep marking approach, a word would have the same tone, *orthographically*, in every context. However, the speech token representing the same word will vary, thus creating ambiguity at the speech front. Although adopting a shallow or deep marking approach may not have significant implications on languages with few tone phonological rules, the degree of shallow-to-deep marking may increase ambiguity for languages with many phonological rules (Bird, 1999b,a). Tone-marking can also be *partial* or *exhaustive*. **Partial Tone-marking.** Some African languages such as Yorùbá adopt a partial tone marking approach with diacritics. Yorùbá has three distinctive tones - high, mid and low tones - but only represents the high tone with the acute symbol and the low tone with the grave symbol in its orthography. The mid tone is not marked and vowels without diacritics unambiguously indicate the presence of the mid tone. Rangi, a language spoken in Tanzania, marks only high tone on nouns while Akoose, a language spoken in Cameroon, marks high tone and contour tones but leaves low tones unmarked (Cahill, 2019). Karaboro, spoken in Burkina faso, marks grammatical tones in plurals using a word final hyphen as in: `sààpjé` 'Rabbit' and `sàápjé-` 'Rabbits'. **Exhaustive Tone-marking.** In exhaustive tone-marking, every tone bearing unit is orthographically marked for tone as in Dschang, spoken in Cameroon (Bird, 1999b).

Furthermore, a higher number of distinctive tones increases ambiguity. In Dan, a language with five distinctive tones, the following can occur: $\texttt{gba}^1$ *(caterpillar)*, $\texttt{gba}^2$ *(shelter)*, $\texttt{gba}^3$ *(fine)*, $\texttt{gba}^4$ *(roof)*, and $\texttt{gba}^5$ *(antelope)* (Clements and Rialland, 2008). For another example, Yorùbá has three distinctive tones where each monosyllabic sequence of sounds can have up to three pitch contrasts and a bi-syllabic can have 2^3 pitch contrasts.[3]

Recommendations. (1) For speech applications, there exists a plethora of unexplored research questions to answer with regard to the implication of tone on *text-to-speech* and *text-free speech processing* (Lakhotia et al., 2021). We therefore call for empirical studies that investigate the influence of tones in text-to-speech, text-free speech processing, and universal speech processing (Yang et al., 2021). Since tone is absent in Indo-European languages where most recent speech work is situated, we expect this to be a fruitful direction. **(2)** For text applications, tone will be relevant for *natural language understanding (NLU)* tasks including but not limited to *part of speech tagging (POS)*, *text classification*, and *natural language generation (NLG)* tasks such as *machine translation*. For many of these applications, it is not clear how tone would interact

[3]However, some phonological rules can restrict the occurrence of certain combinations and there may be lexical gaps. For instance, the high tone occurs only in marked consonant-initial words.

with system performance. For example, we do not know where to keep and where to remove tone (if at all). For example, we find that while removing tone has negligible impact on Bambara→English MT, it has significant negative impact on Yorùbá→English (see Table A.2 in Appendix). We also do not necessarily know what the best ways to encode (and decode) tone information are. **(3)** For work involving languages with shallow tone marking at the orthographic level, we recommend budgeting for collection and preparation (e.g., annotation) for sizeable datasets (to alleviate data sparsity). In absence of large datasets, knowledge of the finite phonological rules of a language can also be exploited for generating data for downstream tasks. **(4)** Orthographic conventions should not be taken as a good indication of the functional load (i.e., information load) of tone in a language, for there are many non-linguistic (e.g., political) reasons for employing a particular orthographic convention (Cahill, 2011). Hence, NLP researchers should do due diligence as to understanding how tone works in a given language. **(5)** Punctuation marks may be tone indicators, and care needs to be taken on how these are pre-processed.

2.2 Vowel Harmony

Vowel harmony is a phonological pattern in which vowels within a given domain agree in properties such as tongue position or lip rounding (Hyman, 2003). It restricts the possibilities of vowels that can co-occur (Archangeli and Pulleyblank, 2007). Different languages adopt different types of vowel harmony. Three types of vowel harmony that are unique to African languages have been recorded in the literature (Clements and Rialland, 2007): (i) advanced tongue root (ATR) harmony, (ii) cross height ATR harmony, and (iii) reduced ATR harmony. ATR harmony occurs when some vowels have the $[-ATR]$ feature and others have the $[+ATR]$ feature. Within a word, all non-low vowels agree in $[+ATR]$ or $[-ATR]$ features. With cross height ATR, $[+ATR]$ in mid vowels require $[+ATR]$ in high vowels and vice versa. The reduced ATR, on the other hand, occurs in languages with only one mid vowel and $[-ATR]$ mid and high vowels shift to $[+ATR]$ in the context of $[+ATR]$ high vowels (Clements and Rialland, 2007).

Recommendations. **(1)** Since vowel harmony is largely absent in most Indo-European languages, knowledge of vowel harmony is currently underexplored in NLP. Such a knowledge can be useful for tasks such as *POS tagging* since tokens with the same part of speech tend to have similar harmonies. **(2)** *Automatic spelling checkers* can also exploit information about vowel harmony since certain co-occurrences of vowels are barred by phonological rules of vowel harmony.

2.3 Serial Verb Constructions

Serial verb constructions (SVC) involve two or more verbs that combine as a whole without any indication of dependency or any conjunction between them (Creissels et al., 2008; Déchaine, 2008). Languages with SVC use serial verbs to encode events that are usually encoded as single verbs in Indo-European languages. This poses a unique problem when creating/evaluating *cross-lingual embeddings* and in applications such as *dictionary creation*. For instance translating from English to Yorùbá, we have the following examples: *borrow* - 'Gbà àwìn (receive credit)', *believe* - 'Gbà gbọ́ (receive hear)', *pinch* - 'Já l' ééékáná (cut with fingernails)' so that a single English verb is a serial verb in Yorùbá. When these words are used in sentences, they may have intervening words as in: Gbà á l' áawìn (receive 3SG-O on credit) *'borrow it'*, Gbà á gbọ́ (receive 3SG-O hear) *'believe it'*, Já a l' ééékáná (cut 3SG-O with fingernails) *'pinch it'*. In Africa, serial verb constructions are very common in Kwa (e.g. Ewe) and Western Benue-Congo languages (e.g. Yorùbá). They have also been recognized in the North Khoisan language !Xun.

Recommendations. **(1)** Given how pervasive word embedding models are in most NLP applications, we recommend investigating how embeddings accounting for SVC can be developed. Similarly, SVC will have bearings in how (cross-lingual) embeddings are evaluated. For example, researchers may need to create dictionaries customized to African languages. **(2)** For *POS tagging*, decisions need to be made on what approach to take in treating such constructions. **(3)** Research investigating the extent to which SVC affects performance across different tasks needs to be explored. For example, this can be valuable for *parsing* and *MT*.

3 No Literacy, No NLP

NLP for high resource languages (HRL) benefits from the level of literacy NLP researchers have in these languages. Most researchers usually have literacy beyond high school in one or more of the languages they work on. In Africa, however, with very complex multilingual societies, many educated Africans cannot read nor write their Indigenous languages.[4] These people do not have basic linguistic knowledge in their languages either. For example, many people do not know which words are nouns or verbs (Cahill, 2001). For context, more than $2,000$ languages have been reported in Africa - about $1/3$ of all the languages in the world (Hammarström, 2018) - making many African communities truly multilingual. As a result, it is not uncommon for a child to be exposed to multiple Indigenous languages before reaching school age. This is especially the case in families where the father, mother, and grandparents all speak different languages (which may, in turn, be different from the languages spoken in the communities they live in). People who receive formal education - the sole way people become literate - thus attain only partial literacy in one or more African language(s) which may not even be their mother tongues. Many others have no knowledge of any Indigenous language, and are only literate in a foreign language (Cahill, 2001; Ouane and Glanz, 2010).

As seen in Table C.1 in the Appendix, out of the 56 countries in Africa, only 17 countries have an Indigenous language as a national language (although in 14 of these 17 countries, a foreign language is the main official language). Furthermore, the countries that give any official status to Indigenous languages, tend to restrict such a status to those languages belonging to majority speakers.[5] For example, in Nigeria, only three out of 512 languages are officially recognized as regional languages; Ghana uses 10 of its 73 Indigenous languages as institutional languages; Swahili is the only official Indigenous language in Tanzania out of 118 others; 12 of 61 languages in Kenya have some official status; only 12 of 20 Indigenous languages in South Africa are institutional languages. This challenging situation is the result of poor language policies, which we now turn to.

Language policy. Language policy determines which languages are used in education, media, commerce, and almost every domain controlled by government. With most Africans educated in English, French, Portuguese or majority African languages, most African languages (those without any official status) are rarely used or used only at home (Petzell, 2012; Foster, 2021; Ouane and Glanz, 2010). In countries where an Indigenous language has official status, governments and implementing bodies only pay lip service to these policies (Kaschula and Kretzer, 2019). In addition, lack of trained personnel and adequate educational resources in Indigenous languages, as well as rarity of teachers sufficiently proficient to offer Indigenous language courses, make policies difficult to implement (Trudell, 2018; Kaschula and Kretzer, 2019). Furthermore, in many schools, Indigenous languages are referred to as vernaculars and are prohibited. Violation usually attracts fines, and even corporal punishment in some cases. English and other foreign languages remain the prerequisite for scientific and technological development, and a key to social prestige and power. Students who do not pass examinations in these foreign languages cannot continue studying beyond elementary school (Foster, 2021; Petzell, 2012; Mohr, 2018). Effect of these currently implemented policies is visible in the NLP situation of African languages. Languages officially recognized within their countries have more resources and tools for NLP than those that do not. For instance, all African languages with a diversity index (Joshi et al.,

[4] We use *Indigenous* languages to refer to languages native to Africa.

[5] It is worth mentioning that some of the excluded languages have millions of L1 speakers.

2020) greater than zero are either official national, regional, or educational languages or are languages of wider communication (Eberhard et al., 2021). We provide more details about available resources of different types (labelled, unlabelled, parallel, and raw) and tools in Section F (Appendix).

Recommendations. Partial and lack of literacy or knowledge of Indigenous languages has significant negative impacts on NLP in African languages. Therefore, **(1)** we include in our concept of a grand challenge the development of language policies that facilitate literacy in Indigenous African languages. Literacy improvement takes time, and policies that teach Indigenous languages only for brief periods in elementary school need to be reformed. **(2)** We also recommend the implementation of policies that require use of Indigenous languages in media, government, and other domains. **(3)** Adequate funding needs to be allocated to develop pedagogical materials, train teachers, and provide teaching aids in order to facilitate the implementation of these policies. *Simply put, without improvement of literacy in African languages, we do not see a flourishing future for African NLP.*

4 A Tale of Two Approaches

There are two main approaches for developing NLP systems. We discuss each of these *vis-a-vis* the situation for African languages here, giving relevant recommendations.

Feature engineering. Feature engineering requires domain knowledge, which is lacking for many African languages due to the aforementioned literacy situation. This negatively impacts use of written African languages in many domains of human endeavor, let alone NLP research. Weak literacy simply means unavailability and inaccessibility of linguists, annotators, language experts, and computational linguists with expertise in African languages. It also manifests itself in lack of grammars, primers, teaching aids, and dictionaries (Cahill, 2011). As it turns out, grammatical information is either fully lacking or under-documented for almost half of Africa's languages. This makes Africa the second *least known* continent (after Oceania, dominated by the New Guinea area) (Güldemann, 2018). In Appendix Table F.2, we list available linguistic resources for all African languages we could trace.

Deep Learning Approaches. A major bottleneck in the development of end-to-end deep learning NLP systems for African languages is the paucity of machine-readable data (Adda et al., 2016). Deep learning systems for high-resource languages are usually fed ever-growing amounts of data that are abundant online and via several other avenues in today's connected society. Without these type of (interactive) data, it is challenging to develop NLP models for real-world use. In particular, models that are endowed with the implicit and explicit knowledge embedded in language are hard to build (at least by current technologies) without large volumes of data derived from diverse contexts. Many African languages lack the environment from which these types of machine-readable data can be collected. Social media, which is a venue for data collection for many high-resource languages, are often not widely used for African languages. In fact, most Africans post to social media in foreign languages rather than in Indigenous African languages (Malatji, 2019).[6] One reason behind this issue is unavailability of keyboards for Indigenous languages. Most keyboards, for example, do not support symbols for representing tone and some other grammatical features. Partial or complete lack of writing literacy is another reason. A third reason is related to the lack access to smart machines and internet connectivity.

Furthermore, countries such as Nigeria where official status is given to a handful of Indigenous languages, still document official activities in foreign language exclusively. Media organizations that often read the news in a foreign language as well as local languages also archive *only* the English news and discard those in Indigenous languages. All such practices stifle opportunities for developing large datasets for African languages, effectively causing African NLP to lag behind. If archived, data for many Indigenous African languages

[6]https://www.talkwalker.com/quick-search.

can facilitate development across a wide host of speech and language tasks, including text-to-speech and machine translation. Collectively, these compounded issues mean there are only few (and often smaller) online communities that contribute to web fora, Wikipedias, and other platforms where data are growing in large-to-massive amounts for high-resource languages. This is evident in the diversity index for African languages offered by Joshi et al. (2020).

According to Joshi et al. (2020) who summarized the digital status and 'richness' of languages in the context of data availability, 542 African languages are `left-behinds`. That is, these languages have exceptionally limited resources that will make it probably impossible to lift them up in the digital space. A total of 26 African languages are `scraping-bys` and are in a better position than the `left-behinds`. However, even these are said to require organized awareness and strong data collection effort with most of these languages having no labelled datasets. Only nine African languages are in the `hopefuls` category, with a small set of labeled datasets, researchers, and language support communities. A single African language (i.e., Afrikaans) is in the `rising-stars` category with a strong web presence and a thriving cultural community online (although with insufficient efforts in labeled data collection). We offer a summary of the diversity index for 578 African languages in Table F.6 in the Appendix.

Recommendations. **(1)** We recommend that daily engagements in education, commerce, media, and government which are otherwise archived only in foreign languages (see Table C.1), be archived in Indigenous languages as well. These would comprise valuable sources of labelled and unlabelled machine-readable data for NLP, let alone painting a more equitable and representative picture of African languages. **(2)** Humans and machines complement each other's strengths, so we recommend stronger interactions between NLP experts and theoretical linguists or knowledgeable native speakers when developing resources and models for African languages. **(3)** Funding should also be allocated to theoretical linguists and language experts, along with machine learning and NLP experts, to aid this work. **(3)** For African languages with available linguistic research, it has been found that certain POS, morphological, named entity, and dependency information can be accurately retrieved *automatically* by using tone, vowel harmony, or even syllable structure patterns (Adegbola, 2016). These approaches may aid faster development of POS taggers, lemmatizers, NER, or even dependency parsers. **(4)** When developing NLP pipelines for African languages, removal of numbers and non-alphanumeric symbols should be approached with caution. This should especially be the case for languages with insufficient research as to the functions played by these symbols, and would help avoid making any irrecoverable issues in the data. **(5)** The most effective ways for building pipelines for African languages remains an under-explored area of research. We therefore call for empirical studies that investigate development of viable pipelines. **(6)** We emphasize the need to respect user consent, data sovereignty, wishes of local communities, and other important issues such as privacy while carrying out any collection or archival effort (Sutherland, 2018; Daigle, 2021; Makulilo, 2012). This is to prevent the predatory use of data collected from local communities including monitoring or controlling local peoples, censorship, and other surveillance activities. Properly handling data mitigates physical, financial, and other security risks that poor data practices expose local communities to (Turianskyi, 2018) and must also be prioritized. We now further discuss issues around data quality.

5 Garbage in, Garbage out

A manual evaluation of 205 datasets involving African languages such as those in CCAligned (El-Kishky et al., 2020), ParaCrawl (Bañón et al., 2020; Esplà-Gomis et al., 2019), WikiMatrix (Schwenk et al., 2021), OSCAR (Ortiz Suárez et al., 2020), and mC4 (Xue et al., 2021) show that at least 15 corpora were completely erroneous, a significant fraction contained less than 50% of correct data, and 82 corpora were mislabelled or used ambiguous language codes (Kreutzer et al., 2021). The inaccuracy is due to a lack, or poor quality of language identification tools, dictionaries, and text pre-processing piplelines, for many low resource languages including African languages represented in these datasets. Furthermore, available resources are rarely evaluated especially when crawled as part of a multilingual dataset. Furthermore, Alabi et al. (2020) find that, fastText embeddings for Yorùbá has an estimated 135K out of 150K words belonging to other languages such as English, French, and Arabic. New embedding models created by Alabi et al. (2020) with a curated high quality dataset outperform the off-the-shelf fastText embeddings even though the curated set has fewer words. Results of these few studies paint a gloomy picture for most current multilingual datasets involving African languages, and models derived from them.

Inconsistent orthographies also contribute to the data quality problem (Martinus and Abbott, 2019). In many cases, orthographies may not be standardized and will have significant spelling and punctuation variations across different domains. In some cases where standard orthographies exist, word lists or dictionaries do not necessarily represent the standardized orthography. Using Hausa as an example, all commercially published books and nearly all Hausa language newspapers use the standard romanized orthography. Standard romanized orthography is written without tones or any indication of vowel length (Schuh and Yalwa, 1993). The orthography used in grammars, dictionaries, and pedagogical documents on the other hand, indicate tone and vowel length (Schuh and Yalwa, 1993). Furthermore, languages that have standard orthographies may also suffer from inconsistencies when orthographic conventions are not adhered to (Olúmúyìwá, 2013). This is evident in the methods and practices for content archiving of many African languages on the web. For example, all VOA websites, omit tones for African languages whose standard orthographies require tone diacritics. BBC also does not adhere to the orthographic conventions for Yorùbá texts except in the headlines, JW.org also does the same for some African languages.

Apart from the aforementioned issues, lack of constant and systematic use of African languages in contexts such as governance, law, technology, science, and education prevents African languages from expanding in vocabulary to accommodate new concepts that have become important parts of conversation elsewhere. As a result, it is not uncommon to have large amounts of foreign words in a dataset which are not adapted to the phonological or orthographic structure of the target African language. Furthermore, terminologies continue to be employed inconsistently and spelt differently in many African venues.

To provide a concrete example of the data quality problem for African languages, we perform a manual evaluation of Flores-101 dataset (Goyal et al., 2021; Guzmán et al., 2019b) for Yorùbá. We find the following: **(1)** 5.29% spelling errors **(2)** 2.7% inconsistent spellings **(3)** 1.2% borrowed words not adapted to the orthographic conventions of target language and **(4)** 12.4% incorrect tone marks. Detailed information is in Appendix G.

It is important to mention that a single error in assignment of diacritics, for instance, can result in significant semantic and syntactic differences in texts. The implication of inconsistencies in orthography is hence enormous for low resource African languages. Such inconsistencies worsen the issue of *data sparsity*: when different spellings of the same word are employed, or when tone or other grammatical features are inconsistently marked, the same 'word' will have many more surface forms than what it actually should. Data sparsity can in turn aggravate the situation for any work involving training with data from different domains (e.g., in domain adaptation). That is, reliability of models trained with erroneous data from a source domain will be diminished while transferring into a target domain. Orthographic inconsistencies also affect results of search engines (Choroś, 2005) in that these engines would not recognize the relationship between a diacritized text and its undiacritized counterparts (Asubiaro, 2014; Olúmúyìwá, 2013). Again, this results in difficulty retrieving resources for many African languages. To optimize search for African languages that involve diacritics, some users employ normalized text which in turn further creates a mismatch between web documents and other standard offline documents (e.g., books) for many African languages.

Recommendations. **(1)** We recommend developing language identification tools that cover African languages. **(2)** Development of dictionaries or even extended word lists will also help the community ensure data quality. **(3)** Manual inspection of sizeable samples of multilingual datasets should also continue to be prioritized. **(4)** We also suggest orchestrated efforts to enforce consistency in orthography for the various languages. **(5)** Linguistic rules may be appropriate for developing automatic data cleaning and pre-processing, but development of any such rules should be carried out carefully. We now briefly highlight community efforts invested in developing skills, datasets, and tools in the African NLP space.

6 Communities and Resources

The majority of existing resources for NLP are the initiative of various non-governmental organizations determined to develop datasets and tools for African languages. We list some of these efforts for NLP, but also within the larger contexts of artificial intelligence. We focus on communities and venues here and list recent funding initiatives in Table D.1 (Appendix).

Workshops. As far as we know, there are two main venues in the form of workshops supporting NLP for African languages, and African AI. These are AfricanNLP and BlackInAI. We provide details about these venues in Appendix E.

Communities. Masakhane, Black in AI, Deep Learning Indaba, Knowledge 4 All Foundation Ltd (K4A), Zindi and ALTI are some of the active communities for research on NLP for African languages. More information about each of these communities is in Section D.

Resources. The religious domain is currently the major source of data for a large number of African languages. Top amongst religious resources is the Bible corpus (available in over $1,000$ languages of Africa (Resnik et al., 1999; McCarthy et al., 2020a)) and the JW300 website (with data for ~ 100 low-resource African languages). Religious sources are constantly updated with new data from the same languages and new languages are often added, making these sources increasingly useful. One issue of these datasets is that, although they are parallel, they may not be sentence aligned. Regardless, these resources remain significantly inadequate. Most other data available for African languages are raw and unlabelled. Still, these can be useful in many applications (e.g., in training word embeddings or language models, for backtranslation). We provide more details about available resources (labelled, unlabelled, and raw) and tools in Appendix F.

Recommendations. **(1)** To achieve Afrocentric NLP, we recommend active interactions between differently existing communities, as well as encouraging new regional and thematically-defined communities. **(2)** We recommend extending these communities beyond AI, NLP, and machine learning to involve theoretical linguists, anthropologists, sociologists, field workers, and other scholars and practitioners with interest in African languages. **(3)** We believe ACL and other similar organizations should continue to prioritize work on low-resource languages by securing dedicated tracks in their publication and dissemination venues.

7 Discussion and Conclusion

We discussed major challenges facing development of NLP technologies for African languages. One of the most important recommendations we would like to emphasize is to *prioritize African NLP work based on the needs of African communities*. For example, we believe development for data and tools for improving health and education should be a priority. We also caution against extractive practices, and encourage creation of opportunities, contexts, and venues for work on African languages and advocacy for reclaiming African language policies. In addition, data literacy and issues around data sovereignty and privacy should remain of highest importance. We highlighted various communities and venues here that we think should continue to be supported.

Acknowledgements

We gratefully acknowledge support from the Natural Sciences and ENgineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research COuncil of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation of Innovation (CFI; 3771), Compute Canada (CC),[7] and UBC ARC-Sockeye.[8] Any opinions, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of NSERC, SSHRC, CFI, CC or UBC ARC-Sockeye. We thank Rose-Marie Déchaine for helpful discussions.

References

Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem, Tewodros Abebe, Wondimagegnhue Tsegaye, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. 2018. Parallel corpora for bilingual English-Ethiopian languages statistical machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3102–3111, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélene Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, et al. 2016. Breaking the unwritten language barrier: The bulb project. *Procedia Computer Science*, 81:8–14.

Ife Adebara, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2021. Translating the Unseen? Yorùbá→ English MT in Low-Resource, Morphologically-Unmarked Settings. *arXiv preprint*.

Tunde Adegbola. 2016. Pattern-based unsupervised induction of yorùbá morphology. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 599–604, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

[7]https://www.computecanada.cacomputecanada
[8]https://arc.ubc.ca/ubc-arc-sockeye

David I Adelani, Dana Ruiter, Jesujoba O Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Awokoya, and Cristina España-Bonet. 2021a. Menyo-20k: A multi-domain English-Yorùbá corpus for machine translation and domain adaptation. *arXiv preprint arXiv:2103.08647*.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021b. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Dana Ruiter, Jesujoba O Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Awokoya, and C Espana-Bonet. 2021c. The effect of domain and diacritics in Yorùbá–English neural machine translation. *Proceedings of Machine Translation Summit (Research Track). European Association for Machine Translation*.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Orevaoghene Ahia and Kelechi Ogueji. 2020. Towards supervised and unsupervised neural machine translation baselines for Nigerian Pidgin. *arXiv preprint arXiv:2003.12660*.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina Espana-Bonet.

2020. Massive vs. curated embeddings for low-resourced languages: The case of Yorùbá and Twi. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2754–2762.

Antonios Anastasopoulos, Ondřej Bojar, Jacob Bremerman, Roldano Cattoni, Maha Elbayad, Marcello Federico, Xutai Ma, Satoshi Nakamura, Matteo Negri, Jan Niehues, Juan Pino, Elizabeth Salesky, Sebastian Stüker, Katsuhito Sudoh, Marco Turchi, Alexander Waibel, Changhan Wang, and Matthew Wiesner. 2021. FINDINGS OF THE IWSLT 2021 EVALUATION CAMPAIGN. In *Proceedings of the 18th International Conference on Spoken Language Translation (IWSLT 2021)*, pages 1–29, Bangkok, Thailand (online). Association for Computational Linguistics.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Diana Archangeli and Douglas Pulleyblank. 2007. Harmony. *The Cambridge handbook of phonology*, pages 353–378.

Franklin Oladiipo Asahiah, Odetunji Ajadi Odejobi, and Emmanuel Rotimi Adagunodo. 2017. Restoring tone-marks in standard Yorùbá electronic text: improved model. *Computer Science*, 18.

Toluwase Victor Asubiaro. 2014. Effects of diacritics on web search engines' performance for retrieval of Yorùbá documents. *Journal of Library & Information Studies*, 12(1).

Israel Abebe Azime and Nebil Mohammed. 2021. An Amharic news text classification dataset.

Felix Banda. 2009. Critical perspectives on language planning and policy in Africa: Accounting for the notion of multilingualism. *Stellenbosch Papers in Linguistics PLUS*, 38:1–11.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Emily M Bender. 2016. Linguistic typology in natural language processing. *Linguistic Typology*, 20(3):645–660.

Steven Bird. 1999a. Strategies for representing tone in African writing systems.

Steven Bird. 1999b. When marking tone reduces fluency: An orthography experiment in Cameroon. *Language and Speech*, 42(1):83–115.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Dörthe Bühmann and Barbara Trudell. 2008. Mother tongue matters: Local language as a key to effective learning. *France: UNESCO*.

William E Bull. 1955. The use of vernacular languages in education.

Michael Cahill. 2019. Tone, orthographies, and phonological depth in African languages. *African linguistics across the disciplines*, page 103.

Mike Cahill. 2001. Avoiding tone marks: a remnant of English education? *PUB TYPE JOURNAL CIT EDRS PRICE DESCRIPTORS*, 27:15.

Mike Cahill. 2011. Non-linguistic factors in orthographies. *Linguistic Society of America Annual Meeting, Pittsburgh, Jan. 6-9, 2011 Symposium on Developing Orthographies for Unwritten Languages*.

Kazimierz Choroś. 2005. Testing the effectiveness of retrieval to queries using Polish words with diacritics. In *International Atlantic Web Intelligence Conference*, pages 101–106. Springer.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49(2):375–395.

G. N. Clements and Annie Rialland. 2007. *Africa as a phonological area*, Cambridge Approaches to Language Contact, page 36–85. Cambridge University Press.

George N Clements and Annie Rialland. 2008. Africa as a phonological area. *A linguistic geography of Africa*, 36:85.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Denis Creissels, Gerrit J Dimmendaal, Zygmunt Frajzyngier, and Christa König. 2008. Africa as a morphosyntactic area. *A linguistic geography of Africa*, 86150.

Jia Cui, Xiaodong Cui, Bhuvana Ramabhadran, Janice Kim, Brian Kingsbury, Jonathan Mamou, Lidia Mangu, Michael Picheny, Tara Sainath, and Abhinav Sethy. 2013. Developing speech recognition systems for corpus indexing under the IARPA Babel program. *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pages 6753–6757.

Brian Daigle. 2021. Data protection laws in Africa: A Pan-African survey and noted trends. *J. Int'l Com. & Econ.*, page 1.

Marelie Davel, Etienne Barnard, Charl van Heerden, Febe Wet, and Jaco Badenhorst. 2014. The nchlt speech corpus of the South African languages. *Spoken Language Technologies for Under-resourced Languages (SLTU'14)At: St Petersburg, Russia*.

Rose-Marie Déchaine. 2008. Serial verb constructions. In *Syntax*, pages 799–825. De Gruyter Mouton.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bonaventure F. P. Dossou and Chris Chinenye Emezue. 2021. OkwuGbé: End-to-end speech recognition for Fon and Igbo. In *Proceedings of the Fifth Workshop on Widening Natural Language Processing*, pages 1–4, Punta Cana, Dominican Republic. Association for Computational Linguistics.

David M Eberhard, F Simons Gary, and Charles D Fennig (eds). 2021. Ethnologue: Languages of the world. *Twenty-fourth edition*, Dallas, Texas: SIL International.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics.

Chris Chinenye Emezue and Femi Pancrace Bonaventure Dossou. 2020. FFR v1.1: Fon-French neural machine translation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 83–87, Seattle, USA. Association for Computational Linguistics.

Miquel Esplà-Gomis, Mikel L Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. 2019. Paracrawl: Web-scale parallel corpora for the languages of the EU. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 118–119.

Ignatius Ezeani, Paul Rayson, Ikechukwu Onyenwe, Chinedu Uchechukwu, and Mark Hepple. 2020. Igbo-English machine translation: An evaluation benchmark. *arXiv preprint arXiv:2004.00648*.

Danny S Foster. 2021. *Language of Instruction in Rural Tanzania: A Critical Analysis of Parents' Discursive Practices and Valued Linguistic Capabilities*. Ph.D. thesis, University of Bristol.

Chayma Fourati, Abir Messaoudi, and Hatem Haddad. 2020. Tunizi: a Tunisian Arabizi sentiment analysis dataset. *arXiv preprint arXiv:2004.14303*.

PJ Fourie. 1995. Introduction to communication: Course book 3: Communication and the production of meaning.

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. 2016. Collecting resources in sub-Saharan African languages for automatic speech recognition: a case study of Wolof. In *Proceedings of the*

Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 3863–3867, Portorož, Slovenia. European Language Resources Association (ELRA).

Andargachew Mekonnen Gezmu, Andreas Nürnberger, and Tesfaye Bayu Bati. 2021. Extended parallel corpus for Amharic-English machine translation. *arXiv preprint arXiv:2104.03543*.

Andargachew Mekonnen Gezmu, Andreas Nürnberger, and Binyam Ephrem Seyoum. 2018. Portable spelling corrector for a less-resourced language: Amharic. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Pierre Godard, Gilles Adda, Martine Adda-Decker, Juan Benjumea, Laurent Besacier, Jamison Cooper-Leavitt, Guy-Noel Kouarata, Lori Lamel, Hélène Maynard, Markus Mueller, Annie Rialland, Sebastian Stueker, François Yvon, and Marcely Zanon-Boito. 2018. A very low resource language speech corpus for computational language documentation experiments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019a. Two new evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-Eglish. *arXiv preprint arXiv:1902.01382*.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019b. Two new evaluation datasets for low-resource machine translation: Nepali-English and Sinhala-English. *arXiv preprint arXiv:1902.01382*.

Tom Güldemann, editor. 2018. *The Languages and Linguistics of Africa*. De Gruyter Mouton.

Asmelash Teka Hadgu, Adam Beaudoin, and Abel Aregawi. 2020. Evaluating Amharic machine translation. *arXiv preprint arXiv:2003.14386*.

Harald Hammarström. 2018. 1. a survey of African languages. In *The languages and linguistics of Africa*, pages 1–57. De Gruyter Mouton.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2580–2591, Online. Association for Computational Linguistics.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Larry M Hyman. 2003. African languages and phonological theory. *Glot International*, 7(6):153–163.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Russell H. Kaschula and Michael M. Kretzer. 2019. The politics of language education in Africa.

Mary Khejeri. 2014. Teachers' attitudes towards the use of mother tongue as a language of instruction in lower primary schools in hamisi district, kenya. *International journal of Humanities and social science*, 4(1):75–85.

Christa König, CRKN MiL Collection, and Ebook Central. 2008. *Case in Africa*. Oxford University Press, New York;Oxford;.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz

Suárez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv preprint arXiv:2103.12028*.

Surafel M Lakew, Matteo Negri, and Marco Turchi. 2020. Low resource neural machine translation: A benchmark for five African languages. *arXiv preprint arXiv:2003.14402*.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Fréjus A A Laleye, Laurent Besacier, Eugène C Ezin, and Cina C Motamed. 2016. First Automatic Fongbe Continuous Speech Recognition System: Development of Acoustic Models and Language Models. In *2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, volume 8, pages 477 – 482, Gdansk, Poland.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Alex Boniface Makulilo. 2012. Privacy and data protection in Africa: a state of the art. *International Data Privacy Law*, 2(3):163–178.

Edgar J. Malatji. 2019. *The impact of social media in conserving African Languages amongst youth in Limpopo Province*. Ph.D. thesis, SCHOOL OF LANGUAGES AND COMMUNICATION STUDIES at the UNIVERSITY OF LIMPOPO.

Laura Martinus and Jade Z Abbott. 2019. A focus on neural machine translation for African languages. *arXiv preprint arXiv:1906.05685*.

Arya D McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020a. The Johns Hopkins University bible corpus: 1600+ tongues for typological exploration. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2884–2892.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020b. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Thipe I Modipa, Febe De Wet, and Marelie H Davel. 2013. Implications of Sepedi/English code switching for asr systems. *Conference Proceedings of the 24th Annual Symposium of the Pattern Recognition Association of South Africa, Johannesburg, South Africa, 3 December 2013*.

Jama Hussein Mohamud, Lloyd Acquaye Thompson, Aissatou Ndoye, and Laurent Besacier. 2021. Fast development of ASR in African languages using self supervised speech representation learning. *CoRR*, abs/2103.08993.

Susanne Mohr. 2018. The changing dynamics of language use and language attitudes in Tanzania. *Language Matters*, 49(3):105–127.

Justin Mott, Ann Bies, Stephanie Strassel, Jordan Kodner, Caitlin Richter, Hongzhi Xu, and Mitch Marcus. 2020. Morphological segmentation for low resource languages. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3996–4002.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Rubungo Andre Niyongabo, Qu Hong, Julia Kreutzer, and Li Huang. 2020. KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5507–5521, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2016. Survey on the use of typological information in natural language processing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1297–1308, Osaka, Japan. The COLING 2016 Organizing Committee.

Alp Öktem. 2021. Swc stt 0.3. Technical Report STT-SWC-0.3, Translators without Borders, `https://github.com/coqui-ai/STT-models`.

Tèmítọ́pẹ́ Olúmúyìwá. 2013. Yorùbá writing: standards and trends. *Journal of Arts and Humanities*, 2(1):40–51.

Iroro Orife, David I Adelani, Timi Fasubaa, Victor Williamson, Wuraola Fisayo Oyewusi, Olamilekan Wahab, and Kola Tubosun. 2020a. Improving Yorùbá diacritic restoration. *arXiv preprint arXiv:2003.10564*.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020b. Masakhane–machine translation for africa. *arXiv preprint arXiv:2003.11529*.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Adama Ouane and Christine Glanz. 2010. *How and why Africa should invest in African languages and multilingual education: an evidence-and practice-based policy advocacy brief*. UNESCO Institute for Lifelong Learning.

Mohamed Outahajala and Paolo Rosso. 2016. Using a small lexicon with crfs confidence measure to improve pos tagging accuracy. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4311–4315.

Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, and Olalekan Akinsande. 2020. Semantic enrichment of Nigerian Pidgin English for contextual sentiment classification. *arXiv preprint arXiv:2003.12450*.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Malin Petzell. 2012. The linguistic situation in Tanzania. *Moderna språk*, 106(1):136–144.

Edoardo Maria Ponti, Helen O'horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *OpenAI*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 African languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1306–1320, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Philip Resnik, Mari Broman Olsen, and Mona Diab. 1999. The bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1):129–153.

Russell G. Schuh and Lawan D. Yalwa. 1993. Hausa. *Journal of the International Phonetic Association*, 23(2):77–82.

Tanja Schultz. 2002. Globalphone: a multilingual speech and text database developed at karlsruhe university. In *Seventh International Conference on Spoken Language Processing*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1351–1361, Online. Association for Computational Linguistics.

Guillaume Segerer. 2008. Closed adjective classes and primary adjectives in African Languages. Working paper or preprint.

Stephanie Strassel and Jennifer Tracey. 2016a. Lorelei language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280.

Stephanie Strassel and Jennifer Tracey. 2016b. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

Ewan Sutherland. 2018. Digital privacy in Africa : cybersecurity, data protection & surveillance. *Data Protection & Surveillance (June 22, 2018)*.

Martha Yifiru Tachbelie, Solomon Teferra Abate, and Tanja Schultz. 2020. Analysis of globalphone and Ethiopian languages speech corpora for multilingual ASR. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4152–4156.

Allahsera Auguste Tapo, Bakary Coulibaly, Sébastien Diarra, Christopher Homan, Julia Kreutzer, Sarah Luger, Arthur Nagashima, Marcos Zampieri, and Michael Leventhal. 2020. Neural machine translation for extremely low-resource African languages: A case study on Bambara. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 23–32, Suzhou, China. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Barbara Trudell. 2005. Language choice, education and community identity. *International Journal of Educational Development*, 25(3):237–251.

Barbara Trudell. 2018. Language and education in Nigeria. *British Council, Nigeria*.

Yarik Turianskyi. 2018. Balancing cyber security and internet freedom in Africa. *SAIIA Occasional Paper No 275,*.

Eddie Williams. 2011. Language policy, politics and development in africa. *Dreams and realities: Developing countries and the English language*, pages 39–56.

Eddie Williams. 2013. *3. Political Perspectives on Language Policies and Development in Africa*, pages 68–87. Multilingual Matters.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

Alp Öktem, Eric DeLuca, Rodrigue Bashizi, Eric Paquin, and Grace Tang. 2021. Congolese Swahili machine translation for humanitarian response.

Appendices

A Effect of Tone in MT

In this experiment on tone, we used the bible for the Yor-En pairs (Adebara et al., 2021), and LDC dataset (Bamanankan Lexicon LDC2016L01.) for the Bam-Eng pairs. Details of the data sizes are available in Table A.1.

Pair	Lang	Sent	Words
Bam-Eng	Bam	$11,154$	$43,786$M
	Eng	$11,154$	$64,571$
Yor-Eng	Yor	$31,086$	$942,663$
	Eng	$31,086$	$822,950$

Table A.1: Number of sentences and words for the training data used for each language pair.

We developed python scripts to remove diacritics from Bambara and Yorùbá no-tone marked settings. In Table A.2, tone significantly affects bleu scores for En-Yor pairs but has marginal effect in the Bam-En pairs. The influence of tones thus needs to be further investigated.

Pair	Tone-Marked	No-Tone Mark
BAM-ENG	1.61	1.61
ENG-BAM	1.07	1.34
ENG-YOR	32.95	11.51
YOR-ENG	38.57	12.76

Table A.2: BLEU scores for tone-marked and no tone mark settings.

B Language Typology Information

In Table B.1, we provide typology information covering tone, vowel harmony and SVC for 116 African languages. The checkmarks indicate the presence of the specified feature in the language. To the best of our ability, this information represents the features in the specified languages and for the specified features. Although we do not claim that this information is complete. This table was created by perusing grammatical descriptions, pedagogical materials, and linguistic research regarding these features in the specified languages.

C The Language Situation in Africa

In Table C.1 we list the status for different languages in Africa. This table was created using information available on ethnologue (Eberhard et al., 2021) for each African country. The national, regional, educational and Indigenous languages are presented as it applies to each country. We present all African countries including those not officially recognized in this list. To the best of our knowledge, this list is a true representation of the status of languages used in Africa.

All African countries, except Ethiopia and Liberia were colonized, with most gaining independence between the 1950s and the 1970s. The colonialist came from different parts of Europe and adopted different language policies which seem to play an important role in the language policies adopted by different African countries today. Although economics, politics, and globalization also play a crucial role. All colonialists interacted derogatorily with Indigenous languages and often referred to them as *vernaculars*. Although, the British colonialists allowed Indigenous languages in their territories if desired. The French, Spanish, and Portuguese on the other hand did not tolerate any Indigenous languages in public. Despite the British's tolerance for Indigenous languages, Indigenous languages were allowed only in early childhood education and Indigenous languages where prohibited after the 4th year in elementary school (Williams, 2013; Ouane and Glanz, 2010).

From Table C.1, it is evident that colonial languages have retained their official status in many African countries till date (Khejeri, 2014). Foreign languages are dominantly used in education, and most official government functions, even in countries where Indigenous languages have official status (Banda, 2009). According to Ouane and Glanz (2010), only 25% of the languages used in secondary education and 5% of the languages in higher education are African. This is despite the known benefits of using Indigenous languages in Education and minority language development (Bühmann and Trudell, 2008; Trudell, 2005; Williams, 2013; Bull, 1955). In cases where policy favours the

	Language	Code	Tone	T.Marked	VH	SVC	Language	Code	Tone	T.Marked	VH	SVC
Afro-Asiatic	Afar	aar					Amharic	amh			✓	
	Amazigh	kab					Coptic	cop				
	Ge'ez	gez					Oromo	gaz	✓		✓	
	Hausa	hau	✓				Somali	som	✓			
	Tachelhit	shi					Tamazight	tzm				
	Tamajaq	ttq					Tamajaq	ttq				
	Wolaytta	wal					Tumbuka	tum				
	Arabic	ara					Arabic Sudanese Spoken	apd				
	Tigré	tig					Tigrinya	tir				
A.	Malagasy	plt										
Niger Congo	Akoose	bss	✓	✓	✓		Akan	aka	✓		✓	✓
	Akoose	bss	✓	✓	✓		Bambara	bam	✓		✓	
	Bassa	bsq	✓	✓	✓		Bemba	bem	✓	✓		
	Chitonga	toi	✓				Chichewa	nya	✓			
	Dagaare	dga	✓		✓	✓	Dagbani	dag	✓			
	Ewondo	ewo	✓	✓			Farefare	gur	✓		✓	
	Fang	fan	✓	✓			Efik	efi	✓	✓		
	Éwé	ewe	✓	✓		✓	Edo	bin	✓	✓		
	Esan	ish	✓				Dangme	ada	✓			
	Fulah	ful					Fulfulde	fuv				
	Limba	lma					Fulfulde	fuv				
	Ga	gaa	✓		✓		Igala	igl	✓	✓	✓	
	Kabiyè	kbp	✓	✓	✓		Kpelle	xpe	✓	✓	✓	
	Kikuyu	kik	✓		✓		Kinyarwanda	kin	✓		✓	
	Igbo	ibo	✓	✓	✓		Mbukushu	mhw				
	Mampruli	maw					Ndonga	ndo				
	Medumba	byv	✓	✓			Mende	men	✓			
	Lunda	lun					Ndebele	nbl				
	Jula	dyu					Kamba	kam				
	Kabiyè	kbp					Isoko	iso				
	Kaonde	kqn					Karaboro	kza				
	Kimbudu	kmb					Fante	aka	✓		✓	
	Kwanyama	kua					Luganda	lug	✓			
	Kongo	kwy	✓	✓	✓		Kwangali	kwn				
	Twi	aka	✓	✓	✓		Chitumbuka	kwn	✓			
	Tswa	tsc					Tshiluba	lua				
	Tsonga	tso					Zama	xuu				
	Limba	lma	✓	✓	✓		Lukpa	dop				
	Pular	fuf					Kissi	kqs	✓			
	Rundi	run					Setswana	tsn	✓			
	Shona	sna	✓		✓		Swahili Congo	swc				
	Swati	ssw	✓				Swahili	swh				
	Swahili	swa					Sepedi	nso				
	Sesotho	sot	✓		✓		Tsonga	tso				
	Themne	tem					Comorian Ngazidja	zdj				
	Urhobo	urh	✓		✓	✓	Tshiluba	lua				
	Venda	ven					Wolof	wol			✓	
	Xhosa	xho	✓				Lingala	lin	✓	✓	✓	
	Yemba	ybb					Yorùbá	yor	✓	✓	✓	✓
	Zande	zne					Mboshi	mdw				
	Zulu	zul	✓				Tiv	tiv	✓			
N.S.	Kanuri	knc	✓	✓	✓		Dinka	dik	✓	✓		
	Kunama	kun					Bari	bfa				
	Luo	luo	✓				Dendi	ddn				
I.E.	Afrikaans	afr					English	eng				
	French	fra					Portuguese	por				
	Spanish	spa					Urdu	urd				
Creole	Kituba	ktu					Juba Arabic	pga				
	Seychelles French Creole	crs					Sango	sag	✓	✓		
	Nigerian Pidgin	pcm					Kabyverdianu	kea				

Table B.1: List of Languages, language codes and typology of the languages presented in this paper across 6 language families: Austronesian (A.), Nilo-Saharan (N.S.), and Indo-European (I.E.). The checkmarks are added to each language to indicate the presence of the corresponding feature.

official use of Indigenous languages, some governments have shown a lack of political will to implement these policies (Williams, 2011). The current linguistic situation thus seem to be one of convenience rather than one from well developed language policies.

Despite a few dissenting voices who argue that the use of several mother tongues will accentuate inter-tribal conflict (Khejeri, 2014), the general consensus is that preserving language diversity through policies that encourage multilingualism are most desirable. Developing a truly multilingual language policy for Africa will certainly be challenging (Ouane and Glanz, 2010), but will be most beneficial even to the progress of NLP on the African continent.

Region	Country	Lang(s)	Ind.	National	Regional	Educational
East Africa	**Burundi**	4	2	run		
	Comoros	7	2	fra, ara	zdj	
	Djibouti	5	2	fra, ara		
	Eritrea	15	9	ara		kun, tig
	Ethiopia	91	87	amh	aar, gaz, som, tir	31 Ind. and 1 foreign
	Kenya	68	61	eng, swa		11 Ind.
	Madagascar	14	12	fra, mlg (higher ed.)		
	Malawi	17	13	eng		tum
	Mauritius	9	2	eng, fra		urd
	Mayotte	4	2	fra		
	Mozambique	44	42	por		
	Reunion	3	1	fra		
	Rwanda	4	2	eng, fra, kin		
	Seychelles	3	1	eng, fra, crs		
	Somalia	13	10	ara, som		eng
	South Sudan	70	59	eng	pga, zne, apd, bfa	8 Ind.
	Tanzania	126	118	swa		eng and swa
	Uganda	44	41	eng, swa		2 Ind. , 1 non-Ind..
	Zambia	46	37	eng	3 Ind.	4 Ind., 1 non-Ind.
	Zimbabwe	22	16	eng		2 Ind, 2 South African
Middle Africa	**Angola**	46	41	por		
	Cameroon	275	271	eng, fra		
	Central Afr. Rep.	75	65	eng, sag		
	Chad	129	123	fra, ara		
	Congo	66	55	fra		
	Dem. Rep. of Congo	214	209	fra		2 Ind.
	Equatorial Guinea	15	12	spa		2 foreign
	Gabon	43	40	fra		
	Sao Tome e Principe	7	3	por		1 foreign
North Africa	**Algeria**	19	14	ara, kab		1 foreign
	Egypt	19	9	ara		
	Libya	9	8	ara		
	Morocco	15	10	ara, tzm		
	Sudan	75	70	ara, eng		
	Western Sahara	4	2	ara		
South Africa	**Botswana**	31	26	eng and tsn		1 foreign
	Eswatini	5	1	eng, ssw		
	Lesotho	5	3	eng and sot		
	Namibia	28	23	eng		3 foreign, 6 Ind.
	South Africa	31	20	eng, afr, nbl, tsn, nso, sot, ssw, tso, ven, xho, zul		1 foreign
West Africa	**Benin**	55	50	fra		
	Burkina Faso	71	66	fra		
	Cape Verde Islands	2	1	kea, por		
	Cote d'ivoire	87	75	fra		1 foreign
	Gambia	11	7	eng		
	Ghana	83	73	eng	gur, maw	5 Ind.
	Guinea	37	35	fra	fuf	3 Ind.
	Guinea-Bissau	23	18	por		
	Liberia	31	27	eng		
	Mali	68	63	fra		5 Ind., 1 foreign
	Mauritania	7	5	ara		
	Niger	23	19	fra		2 Ind., 1 foreign
	Nigeria	522	512	eng	hau, ibo, yor	5 Ind.
	Saint Helena	1	0	eng		
	Senegal	39	31	fra		
	Sierra Leone	24	19	eng	lma, men, tem	
	Togo	44	40	fra		

Table C.1: A statistics of the language use in Africa computed from (Eberhard et al., 2021). For each country, we show the number of languages (lang) reported, the number of Indigenous languages spoken in the country (Ind.), the national languages, the regional languages, and the educational languages.

D Communities

Many communities contribute significantly to the development of NLP for African languages. We list some of them below. Masakhane aims to build an active community geared at creating resources that are truly representative of African culture, facilitating collaborations to develop African NLP and lowering the barriers for NLP participation. They achieve this by having an active slack channel that fosters interaction between stakeholders, organizing workshops, creating easy to use google colab notebooks among several other initiatives. Masakhane so far has over 1,000 members.

Black in AI is an organization that focuses on increasing the presence, inclusion, and visibility of black people in artificial intelligence. They achieve this objective through advocacy, mentorship, and facilitating collaborations. Although BlackinAI encompasses black people beyond the African continent, and they do not specifically restrict their operations to African languages, it is a great community for collaborations.

Deep Learning Indaba is an organisation whose mission is to strengthen machine learning and artificial intelligence in Africa by enabling Africans to be active shapers and owners of AI technologies. Deep Learning Indaba which was inaugurated in 2017 organizes an annual Deep Learning Indaba retreat for teachings and practical sessions on AI. They also provide mentorship programs and grants (the IndabaX) that fund AI gatherings in 26 African countries with plans underway to include more countries. This is in addition to awards for the application of AI to an African problem, for excellence in research in tertiary African institutions, and for services to the machine learning community in Africa- Kambule, Maathai, and Umuntu awards respectively. These programmes aim to build a sustainable pan-African community of AI expertise, create local leadership in AI in every country across the continent, and recognise excellence in research and application of AI technologies, respectively.

Knowledge 4 All Foundation Ltd (K4A) pioneers machine learning methods of pattern analysis, statistical modeling, and computational learning and transforms these into technologies for large scale applications in open education. They organize symposiums, summer schools, workshops, colloquiums, and conferences. They also provide fellowship to develop datasets and strengthen capacities and innovation potential for low resource African languages under the international development program. They have developed resources for Ewe, Fongbe, Yorùbá, Chichewa, Wolof, Kiswahili, Tunisian Arabizi, Twi, and Luganda. They also various competitions to develop or improve methods for NLP of African languages.

Zindi hosts a large community of African data scientists and facilitates collaborations between data scientists and organizations. They provide a place to learn, improve skills and find a job. They also organize competitions for data collection tasks and developing NLP models for various African languages.

ALTI is one of the pioneering NLP communities in Africa. They focus on making computers usable in African languages and develop and grow human talent that take African Languages into the information age. They also provide a hub were NLP enthusiasts can be mentored for NLP work in African languages.

Different organization provide funding for NLP research. Some of these organizations are presented in Table D.1.

Organization	Type
Google	Industry
Microsoft	Industry
The Rockefeller Foundation	Foundation
FAIR Forward	Government
Lacuna Funds	NGO
Knowledge 4 All	Research
IDRC	Research

Table D.1: Some funding Organization for African NLP including Non-Governmental organizations (NGO)s

E Workshops

The ***AfricanNLP*** workshop has run annually alongside ICLR and EACL in 2020 and 2021 respectively. In 2020, 32 papers were presented while in 2021, 40 papers describing different systems were accepted. Currently, papers submitted are non-archival, giving authors the opportunity to submit the papers to other venues. ***BlackInAI*** has organized yearly workshops co-located with Neural Information Processing Systems (NeurIPS) since 2017. Audience is composed of researchers who self-identify as Black and often has many works related to African languages.

F Resources

F.1 Labelled Resources

Majority of labelled corpora is developed as part of the development process of many NLP tasks. This is due to a lack of readily available labelled corpora for many NLP tasks. Labelled corpora has been developed for MT (Adelani et al., 2021a; Nekoto et al., 2020; Tapo et al., 2020; Emezue and Dossou, 2020; Ezeani et al., 2020; Hadgu et al., 2020), classification (Niyongabo et al., 2020; Fourati et al., 2020; Oyewusi et al., 2020), automatic spelling correction (Gezmu et al., 2018), morphological segmentation (Outahajala and Rosso, 2016; Mott et al., 2020), NER (Adelani et al., 2021b; Hedderich et al., 2021), diacritic restoration (Orife et al., 2020a; Asahiah et al., 2017), automatic speech recognition (ASR) (Dossou and Emezue, 2021; Tachbelie et al., 2020), and speech translation (Godard et al., 2018). A summary of labelled corpora can be found in Table F.5.

A few of the labelled corpora are developed by trained linguists and language experts (Strassel and Tracey, 2016a; Adebara et al., 2021) while others are collected by native speakers (Adelani et al., 2021b,a; Nekoto et al., 2020). Furthermore, evaluation is often done using automatic metrics that measure model performance rather than data quality or inter-annotator agreement (Outahajala and Rosso, 2016). Data is also often labelled on the assumption that the data has been proofread (Gezmu et al., 2018), while the procedure for developing the dataset is often not discussed. It is important to mention here that we advocate that trained linguists or language experts, particularly those trained in African languages, be involved in data collection or curation activities for African languages. This is because of the linguistic situation in Africa and the literacy levels in African languages which have been discussed in this paper.

F.2 Unlabelled Corpora

Unlabelled corpora seem to be the bulk of available data for African languages. Most corpora are crawled from the web as part of multilingual corpora development efforts like JW300 (Agić and Vulić, 2019), ParaCrawl (Bañón et al., 2020; Esplà-Gomis et al., 2019), WikiMatrix (Schwenk et al., 2021), OSCAR (Ortiz Suárez et al., 2020), mC4 (Xue et al., 2021), CCAligned (El-Kishky et al., 2020), wikiAnn (Pan et al., 2017). We provide a summary of unlabelled corpora in Table F.5.

F.3 Crosslingual Tools

Pre-trained models like BERT (Devlin et al., 2019), ELMo (Peters et al., 2018), Roberta (Liu et al., 2019), GPT (Radford et al., 2018, 2019; Brown et al., 2020), BART (Lewis et al., 2020) have advanced the state of the art in a wide variety of tasks, suggesting that these models acquire valuable, generalizable linguistic information during the pre-training process. However, training language-specific models is possible for only a few languages which have large amounts of data. A popular alternative has been multilingual language models (MLM) such as mBERT (Devlin et al., 2019), XML-R (Conneau et al., 2020), MT5 (Xue et al., 2021), mBART (Liu et al., 2020) and many others. MLMs are trained on large amounts of unlabelled data from multiple languages so that low resource languages may benefit from shared vocabulary and other linguistic information from high resource languages and other similar languages in the MLM. Very few MLMs have representations for African languages and many of those available are trained with noisy data (Adelani et al., 2021c; Alabi et al., 2020; Kreutzer et al., 2021) which may affect downstream tasks. We provide information about crosslingual tools in Table F.4 and other NLP models in Table tab:modelresources.

F.4 Raw Data

Blog sites, online newspapers, Wikipedia, Jehovah's Witness website are some sources of raw data for African languages. We provide details in Table F.1 and Table F.5.

Country	Site	Language
Ethiopia	Addisadmassnews	amh
Ethiopia	Ethiopian Reporter	amh and eng
Lesotho	Mosotho	sot
Namibia	Republikein	afr
Nigeria	Premiumtimes	hau
Nigeria	Leadership	hau
Nigeria	Hausa Legit	hau
Nigeria	Aminiya	hau
Nigeria	Igbo Radio	ibo
Nigeria	Kaoditaa	ibo
Nigeria	Iroyin Owuro	yor
Somalia	Boramanews	som
Somalia	Caasimada	som
Somalia	Horseedmedia	som
Somalia	Idalenews	eng and som
Somalia	Markacadeey	eng and som
Somalia	Ogaden	eng and som
Somalia	Puntlandpost	eng and som
Somalia	PQarannews	eng and som
Somalia	Shabellemedia	eng and som
Somalia	Simbanews	eng and som
Somalia	Togaherer	eng and som
Somalia	Waagacusub	eng and som
Somaliland	Dhamays news	som
Somaliland	Goobjoog	som
Somaliland	Haatuf	som
Somaliland	Maandeq	som
Somaliland	Qorilugudnews	som
Somaliland	Somalilandpost	eng and som
South Africa	Netwerk24	afr
South Africa	Huisgenoot	afr
South Africa	Dievryburger	afr
South Africa	Isolezwe	zul
Tanzania	Mwananchi	swh
Tanzania	Nipashe	swh
Tanzania	Nipashe-Jumpaili	swh
Uganda	Bukedde	lug
Zimbabwe	Kwayedza	sna
Zimbabwe	Umthunywa	nbl

Table F.1: Newspapers in Indigenous languages of Africa.

G Data Quality

The preliminary evaluation of Flores101 dataset for Yorùbá was done by a native speaker of Yorùbá who is also a linguist. Specifically, 57% of the dataset was randomly selected while keeping track of the word's sentential context and the English source context. We removed all numerals written with digits from the dataset before the random selection. This was to help us focus on lexical items alone. We found **(1)** spelling errors, **(2)** inconsistent spellings, which are instances of different spellings for the same word within the text, **(3)** borrowed words not adapted to the orthographic convention of the target language, without recourse to named entities, and **(4)** incorrect tone marks. Further evaluation will be required to access the quality of the dataset on a semantic and syntactic level. Examples of each of the errors identified is presented in Table G.1.

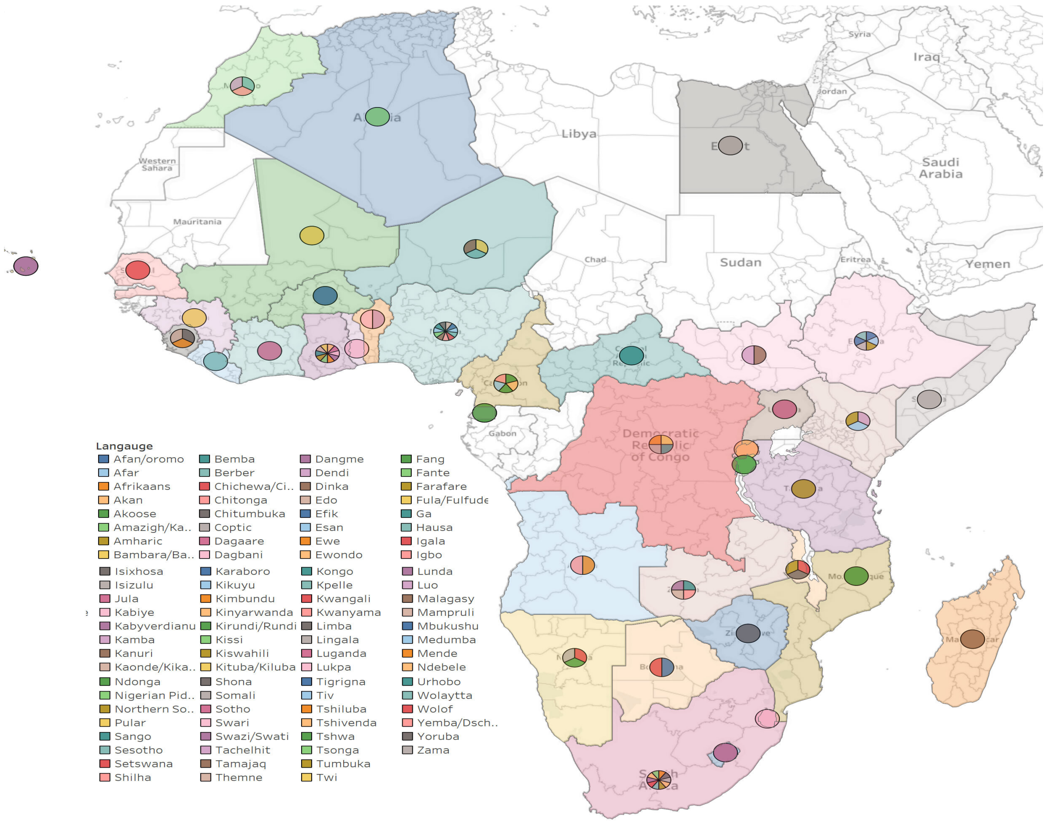

Figure F.1: A high quality (bigger) version of the African languages map provided in this paper.

Num Score		Most Extensive Grammar Description Type	# Languages	
5	long grammar	extensive description of most features of the grammar ≈300+ pages	411	18.9%
4	grammar	a description of most features of the grammar ≈150 pages	243	11.1%
3	grammar sketch	a less extensive description of many features of the grammar ≈50 pages	562	25.9%
2	specific feature	a description of some features of the grammar (i.e noun class system, verb morphology, etc)	157	7.2%
2	phonology	a description of the sound inventory using minimal pairs	82	3.7%
2	dictionary	≈75+ pages	53	2.4%
2	text	text material	13	0.5%
1	wordlist	≈100 − 200 words	13	0.5%
0	minimal	a small number of morphemes	124	5.7%
0	overview	document with meta-information about the language (i.e where spoken, non-intelligibility to other languages etc)	48	2.2%
		Total:	2,169	

Table F.2: Available linguistic resources for African languages. Adapted from (Güldemann, 2018).

Model	Language(s)	URL
Word embeddings	Twi-Yorùbá	https://github.com/ajesujoba/YorubaTwi-Embedding
Okwugbe (ASR)	Igbo-Fon	https://github.com/bonaventuredossou/fonasr
Automatic Diacritic Restoration	Yorùbá	https://github.com/Niger-Volta-LTI/yoruba-adr
FFR v.1.1 model	Fon-French	https://github.com/bonaventuredossou/ffr-v1/blob/master/model_train_test/fon_fr.py
Masakhane MT	30 languages	https://github.com/masakhane-io/masakhane-mt
AfriBERT	Afrikaans	https://github.com/sello-ralethe/AfriBERT

Table F.3: A list of available models.

Language Model	African Languages Represented
MT5	afr, nya, mlg hau, ibo, sna, som, sot / nso, swa, xho, yor, zul
MBERT	afr, swa, yor
XLM-R	afr, amh, hau, gaz, som, swa, xho.

Table F.4: Language models with African languages represented.

	Name	Language(s)	Task.	References
Labelled	KINNEWS and KIRNEWS Corpus	kin, run	POS, NER, Parsing	(Niyongabo et al., 2020)
	Masakhane NER	amh, hau, ibo, kin, lug, luo, pcm, swa, wol, yor	NER	(Adelani et al., 2021b)
	Nigerian Pidgin Tweets	pcm	Sentiment	(Ahia and Ogueji, 2020)
	Swahili News Classification	swa	Classification	
	Amharic News classification	amh	Classification	(Azime and Mohammed, 2021)
	A study on African Language	hau, yor	NER, TC	(Hedderich et al., 2020)
	YorùbáTwi-Embedding	aka, yor	NER, embedding	(Alabi et al., 2020)
	XL Sum	40 languages including: amh hau ibo, kin, gaz, pcm, som, swa, yor	Summarization	(Hasan et al., 2021)
	WikiAnn	282 languages including aar, afr, amh, bam, ewe, Fula*, hau, ibo, kab, kon, kik, kua, kau, lin, mlg, ndo, nso, gaz, run, kin, sna, som, sot, ssw, swa, tsn, tso, wol, xho, yor, zul	NER	(Pan et al., 2017)
	DARPA LORELEI	aka, amh, gaz, som, tir	NER, SemAnal	(Strassel and Tracey, 2016b)
	Automatic Diacritic Restoration	yor	ADR	(Orife et al., 2020a)
Unlabelled	mC4	10 languages including: afr, nya, hau, ibo, sna, som, Sotho*, swa,xho, yor, zul	LM	(Xue et al., 2021)
	Swahili Language Modeling	swa	LM	
	The John Hopkins University Bible Corpus	1600+ (including 313 Niger-Congo), 67 Afro-Asiatic, and 52 Nilo-Saharan	LM	(McCarthy et al., 2020b)
	Monolingual xho corpus	swa	LM	
	OSCAR	166 languages including: afr, swa, yor	LM	(Ortiz Suárez et al., 2020)
Raw Data	Wikipedia	37 African languages	LM	
	Alaroye	yor	LM	
	VOA	gaz, amh, bam, hau, kin, lin, nbl, sna, som, swa, tir, Zimbabwe	LM	
	Jehovah's witness	More than 100 African languages	MT	(Agić and Vulić, 2019)
	BBC News	gaz, amh, hau, ibo, run, pcm, som, swa, tir, yor	LM	
Parallel	Tanzil	kab, amh, hau, som swa	MT	(Tiedemann, 2012)
	Amharic Evaluation Dataset	amh-eng	MT	(Hadgu et al., 2020)
	Parallel Corpora for Ethiopian Languages	eng-amh, tir, gaz, wal, gez	MT	(Abate et al., 2018)
	English-Luganda Parallel Corpora	eng-lug	MT	
	Back-translated Swahili-French 1M sentence parallel data	swa-fra	MT	(Öktem et al., 2021)
	Gamayun Mini kit 5k	swa-eng	MT	
	Gamayun Mini kit 5k	kau-eng	MT	
	English-Akuapem Twi parallel corpus	eng-aka	MT	
		afr, amh, ful, hau, ibo, kea, kam, luo, nso, nyj, gaz, som	MT	(Guzmán et al., 2019a)
	FLORES-101	swa, wol, xho, yor, zul		(Guzmán et al., 2019a)
	Xhosa-English	xho-eng	MT	(Tiedemann, 2012)
	Bamanankan Lexicon	bam-eng	MT	
	Autshumato	eng-tsn	MT	
	Masakhane MT	efi, afr, amh, bin, ddn, fon, hau, ibo, ish, iso, kam, kik, kmb, lin, lua, luo, nbl, nso, nya, pcm, sna, sot, swa, tir, tiv, tsn, aka, urh, ven, xho, yor, zul, swc	MT	(Orife et al., 2020b)
	Bambara Dataset	bam, eng and fra	MT	(Tapo et al., 2020)
	AFRONMT	eng, swa, amh, tir, gaz, som	MT	(Lakew et al., 2020)
	AFROMT	afr, xho, zul, run, sot, swa, bem, lin	MT	(Reid et al., 2021)
	CCAligned	137 languages including yor, afr, aka, amh, Fulfulde, ibo, som, swa, wol, yor, zul	MT	(El-Kishky et al., 2020)
	IgboNLP	ibo-eng	MT	(Ezeani et al., 2020)
	MENYO-20k	yor-eng	MT	(Adelani et al., 2021a)
	Extended Amharic-English bilingual corpus	amh, eng	MT	(Gezmu et al., 2021)
	WikiMatrix	85 languages including swa	MT	(Schwenk et al., 2021)
	Lorelei	aka	MT	(Schwenk et al., 2021)
	Paracrawl	39 languages including som and swa	MT	(Bañón et al., 2020)
	Parallel Bible Corpus	100 languages including afr, amh, Coptic, din, ewe, kab, dop, som, swa, shi, ttq, wal, wol, xho, xuu, zul	MT	C&S
	FFR v1.1	fon-fra	MT	(Emezue and Dossou, 2020)
Speech	SPCS Speech Corpus	eng, nso	Speech	(Modipa et al., 2013)
	TTS data for four South African languages	afr, sot, tsn and xho	Speech	
	Mboshi French Parallel Corpus	mdw, fra	Speech	(Godard et al., 2018)
	IWSLT Low Resource Shared Task	swh-eng, swc-fra	Speech	(Anastasopoulos et al., 2021)
	Tico-19	swc	Speech	(Anastasopoulos et al., 2020)
	GlobalPhone	hau, Swahili	Speech	(Schultz, 2002)
	The NCHLT Speech Corpus of the South African languages	afr, tso, tsn, sot, nso, zul, ven, ssw, xho, nbl	Speech	(Davel et al., 2014)
	ALFFA	amh, swh, hau, wol	Speech	(Gauthier et al., 2016)
	Fon ASR	fon	Speech	(Laleye et al., 2016)
	Swahili audio mini-kit	swh	Speech	
	Swahili (Congo) STT v0.3.0	swc	Speech	(Öktem, 2021)
	AIMS	hau, lug, kab, kin	Speech	(Mohamud et al., 2021)
	IARPA Corpus	amh, ibo, luo, amh, ibo, luo, swh, zul	Speech	(Cui et al., 2013)

Table F.5: List of available data resources. TC=Topic Classification. C&S=(Christodouloupoulos and Steedman, 2015).

Lang	Lang	Lang	Lang	Lang	Lang	Lang
Kásim,0	isekiri , 0	ndonga , 0	matuumbi, 0	bété , 0	bini, 0	babole , 0
obolo, 0	ghulfan, 0	masakin , 0	alagwa , 0	tem , 0	miisiirii , 0	gokana , 0
baga sitemu , 0	vagla , 0	mundani , 0	mbole , 0	kom , 0	ndut, 0	gurenne , 0
hemba, 0	gbeya bossangoa , 0	seychelles creole , 0	grebo , 0	guere , 0	majang , 0	waama , 0
bujeba , 0	ewondo , 0	mankanya , 0	emai , 0	moro , 0	lamé , 0	shatt , 0
kohumono , 0	tetela , 0	baka , 0	qafar , 0	wan , 0	talinga , 0	soninke , 0
gbaya kara , 0	yaka , 0	bororo , 0	vili , 0	tennet , 0	palor , 0	buduma , 0
balanta , 0	bai , 0	mandinka , 0	mango , 0	iraqw , 0	ajagbe , 0	bafut , 0
nubi , 0	migama , 0	burunge, 0	bobo madaré , 0	lobi , 0	yamba , 0	tera , 0
manjaku , 0	tommo so , 0	otoro , 0	shuri , 0	dyula , 0	tenyer , 0	koyraboro senni , 0
comorian , 0	duma , 0	mamvu , 0	hamer , 0	kasem , 0	mara , 0	temne , 0
bankon , 0	kisi , 0	sama , 0	yeyi , 0	tuki , 0	kxoe , 0	guduf , 0
kwangali , 0	supyire , 0	dangaléat , 0	mofu-gudur , 0	mokilko , 0	tigré , 0	ful , 0
bandi , 0	herero , 0	!xun , 0	bangime , 0	tuareg , 0	mbe' , 0	mayogo , 0
ko , 0	sena , 0	chumburung , 0	bafia , 0	bori , 0	kilba , 0	avokaya , 0
ejagham , 0	londo , 0	avatime , 0	sisaala , 0	ma'di , 0	bakundu , 0	nyimang , 0
darma , 0	tunen , 0	wolaytta , 0	mbodomo , 0	mupun , 0	kenyan S.L , 0	gamo , 0
ciluba , 0	turkana , 0	sungor , 0	uma , 0	degema , 0	akwa , 0	aghem , 0
kpelle , 0	päri , 0	tamashek , 0	aizi , 0	katcha , 0	ijo , 0	baale , 0
hunde , 0	samba leko , 0	ngizim , 0	príncipense , 0	nupe , 0	tumak , 0	ncàm , 0
me'en , 0	duala , 0	ghotuo , 0	ik , 0	mwera , 0	kanakuru , 0	nsenga , 0
kera , 0	seme , 0	bidiya , 0	birri , 0	fongbe , 0	jukun , 0	burum , 0
bobangi , 0	ekoti , 0	midob , 0	mbugu , 0	aja , 0	sukumam , 0	tama , 0
hadza , 0	ugandan S.L , 0	bushoong , 0	maninka , 0	efik , 0	kotoko , 0	kukú , 0
kunama , 0	rundi , 0	muher , 0	mauka , 0	lua , 0	moru , 0	avikam , 0
daba , 0	mundang , 0	dongo , 0	beembe , 0	mankon , 0	toro so , 0	krongo , 0
bamun , 0	tiv , 0	wobe , 0	miya , 0	diola-fogny , 0	mbili , 0	basaá , 0
kuanyama , 0	sebei , 0	karimojong , 0	orig , 0	budu , 0	sandawe , 0	yakoma , 0
laal , 0	pero , 0	//ani , 0	awngi , 0	kete , 0	daju , 0	lebeo , 0
leko , 0	mambwe , 0	lango , 0	hdi , 0	shinassha , 0	songe , 0	mpongwe , 0
bimoba , 0	ogbronuagum , 0	bayso , 0	kinga , 0	acholi , 0	bilin , 0	chaga , 0
nara , 0	dizi , 0	nyanga, 0	jeli , 0	hehe , 0	pokot , 0	burji , 0
enya , 0	mano , 0	nharo , 0	baule , 0	maasai , 0	mondunga , 0	kagoma , 0
ngbandi , 0	lendu , 0	tirmaga , 0	leti , 0	nande , 0	runyankore , 0	shambala , 0
fyem , 0	yemsa , 0	lafofa , 0	ingessana , 0	nandi , 0	lele , 0	senadi , 0
mituku , 0	gula iro , 0	fur , 0	kirma , 0	fe'fe' , 0	gula , 0	niuafo'ou , 0
malgwa , 0	ebira , 0	berber , 0	jul'hoan , 0	mono , 0	ama , 0	ngambay , 0
bura-pabir , 0	gusii , 0	bolia , 0	buli , 0	sangu , 0	ika , 0	shabo , 0
kele , 0	kullo , 0	nkem , 0	gan , 0	beria , 0	nkonya , 0	langi , 0
izi , 0	makonde , 0	bariba , 0	babungo , 0	kposo , 0	giziga , 0	oku , 0
mongo , 0	!xóõ , 0	jomang , 0	kenga , 0	temein , 0	Kami , 0	gorowa , 0
ding , 0	kalanga , 0	coptic , 0	urhobo , 0	gumuz , 0	gunu , 0	bukusu , 0
dagaare , 0	uldeme , 0	gworok , 0	afar , 0	bakueri , 0	bana , 0	karó , 0
tampulma , 0	mende , 0	lunda , 0	haya , 0	nkore-kiga , 0	guinea bissau c. , 0	amele , 0
neyo , 0	bira , 0	fulfulde , 0	kanyok , 0	bahnar , 0	miri , 0	nyiha , 0
bodo , 0	lelemi , 0	logoti , 0	mbalanhu , 0	munzombo , 0	kenyang , 0	dabida , 0
bozo, 0	karanga , 0	bisa , 0	konyagi , 0	tashlhiyt , 0	ndebele , 0	dullay , 0
mbosi , 0	goemai , 0	murle , 0	=lhoan , 0	teso , 0	ngbaka , 0	kefa , 0
ndogo , 0	ronga , 0	tonga , 0	kresh , 0	gungbe , 0	bubi , 0	koranko , 0
konni , 0	guro , 0	mambila , 0	mündü , 0	da'a , 0	nuer , 0	runyoro-rutooro , 0
maale , 0	dhaasanac , 0	angas , 0	harari , 0	bagiro , 0	bade , 0	ngoni , 0
ibibio , 0	pa'a , 0	mooré , 0	lozi , 0	toussian , 0	nzakara , 0	rimi , 0
zayse , 0	gimira , 0	birom , 0	leggbó , 0	benga , 0	lagwan , 0	margi , 0
pangwa , 0	zande , 0	isoko , 0	mampruli , 0	kpan , 0	masalit , 0	konkomba , 0
gola , 0	beng , 0	maba , 0	nyangi , 0	ngemba , 0	saho, 0	suku , 0
musgu , 0	adioukrou , 0	/xam , 0	tikar , 0	broken , 0	yana , 0	mada , 0
nuni , 0	binga , 0	kagulu , 0	ndumu , 0	holoholo , 0	jur mödö , 0	mumuye , 0
nyamwezi , 0	shilluk , 0	ron , 0	dime , 0	ngombe , 0	buma , 0	dahalo , 0
dhivehi , 0	kosop , 0	defaka , 0	bongo , 0	luwo , 0	lugbara , 0	koyra chiini , 0
kituba , 0	dii , 0	abidji , 0	boko , 0	komo , 0	lamnso' , 0	klao , 0
kadugli , 0	kabiyé , 0	nyambo , 0	mbum , 0	bole , 0	linda , 0	ila , 0
ntomba , 0	lese , 0	luvale , 0	lyele , 0	busa , 0	doko , 0	igede , 0
aka , 0	nateni , 0	idoma , 0	kara , 0	n'ko , 0	khoekhoe , 0	rendille , 0
katla , 0	tabwa , 0	korana , 0	koh , 0	pogoro , 0	didinga , 0	luri , 0
vata , 0	podoko , 0	yulu , 0	tangale , 0	lamang , 0	engenni , 0	dadjriwalé , 0
berta , 0	tsogo , 0	dagbani , 0	bulu , 0	kiluba , 0	tarok , 0	datooga , 0
bari , 0	mungaka , 0	ega , 0	ifumu , 0	mahican , 0	gude , 0	runga , 0
sare , 0	masa , 0	yansi , 0	mbay , 0	wéménugbé , 0	sengele , 0	kela , 0
anyi , 0	fulani , 0	mursi , 0	soddo , 0	diola-kasa , 0	jamsay , 0	koorete , 0
ga'anda , 0	arbore , 0	anywa , 0	loma , 0	fiote , 0	dyimini , 0	alladian , 0
bena-lulua , 0	mbere , 0	doyayo , 0	gidar , 0	etsako , 0	ngiti , 0	ogbia , 0
subiya , 0	mba , 0	chai , 0	tupuri , 0	kanembu , 0	tima , 0	koromfe , 0
godié , 0	nanerge , 0	mambai , 0	koegu , 0	lucazi , 0	adamorobe S.L , 0	anufo , 0
sotho , 0	dong , 0	aari , 0	kemant , 0	kanuri , 0	sidaama , 0	donno so , 0
bemba , 0	deti , 0	lamba , 0	angolar , 0	gade , 0	gunya , 0	barambu , 0
bagirmi , 0	kamba , 0	mbara , 0	vai , 0	makaa , 0	gwari , 0	nafaanra , 0
nigerian pidgin , 0	rashad , 0	mangbetu , 0	somali , 1	igbo , 1	bambara, 1	venda , 1
tumbuka , 1	twi, 1	sango, 1	kikuyu, 1	kirundi, 1	ndonga, 1	lingala, 1
sesotho , 1	chichewa , 1	dinka , 1	malagasy , 1	ewe , 1	kinyarwanda , 1	kabiye , 1
kongo , 1	northern sotho , 1	kabyle , 1	oromo , 1	akan , 1	tsonga , 1	luganda , 1
amharic , 2	hausa , 2	xhosa , 2	swahili , 2	zulu , 2	tswana , 2	wolof , 2
tigrinya , 2	Yorùbá , 2	afrikaans , 3				

Table F.6: Language diversity index. Adapted from Joshi et al. (2020).

Output	Sentence
Spelling errors	
Yorùbá Source	"Mo **dúpé** lọ́wọ́ àdọ tó gbórùkù ti ẹlẹ́wọ̀n bíi témi" ...
English Source	"Thanks for those who supported a convict like me",...
Yorùbá Target	**Ìflrúnú** hàn bẹ̀rẹ̀ ní ago mọ́kànlá (UTC+1) ní Whitehall ní wájú ẹnu ọ̀nà ilé iṣẹ́ ọlọ́pàá sí òpópónà Downing, ilé árẹ orílẹ̀ ‘ed‘e.
English Source	The protest started around 11:00 local time (UTC+1) on Whitehall opposite the police-guarded entrance to Downing Street, the Prime Minister’s official residence.
Inconsistent spellings	
Yorùbá Target	Fidali, omo odun-28 ti darapọ̀ mọ́ ẹgbẹ́ **agbáboolu** Basilona ...
English Source	28-year-old Vidal had joined Barça ...
Yorùbá Target	**Agbábọ̀lù** Tòní ní Alex Overchkin ti Washington Capitals.
English Source	Today’s Player of the Day is Alex Ovechkin of the Washington Capitals.
Yorùbá Target	Kóṣẹ́lòmín tó ṣeré jù tàbí jẹ góòlù jù fún ikọ̀ **Agbábọ̀ọ̀lù** ju Bobek
English Source	No one else has ever made more appearances or scored more goals for the club than Bobek.
Borrowed words not adapted to orthographic conventions of target language	
Yorùbá Target	Àwọn kan gbàgbọ́ pẹ̀lú john Grant, pé àti **funding crunch** àti sísún ní ẹ̀kọ́ ìmòye ètó orí amóhùnmáwòrán dási láti parí eré náà.
English Source	It is believed by some, including John Grant, that both the funding crunch and a shift in the philosophy of educational television programming contributed to ending the series.
Yoruba Target	Àwọn onímọ̀ sáyẹ́nsị̀ ma n pè ní **“stimulated emission of radiation"** torí àwọn átọ́mṣokù ma ń fura sí iná tó ràn èyí ṣokùn fa kí fotoni ina maa jáde, iná dẹ̀ jẹ́ irúfẹ́ redieṣọ́ni.
English Source	Scientists call this process "stimulated emission of radiation" because the atoms are stimulated by the bright light, causing the emission of a photon of light, and light is a type of radiation.
Incorrect tone marking	
Yorùbá Target	**Awon iIwe naa fihan ile ifowopamo merinla to ran awon onisowo olola pa ilopo bilioni owo Amerika mo lati le sa fun owo ori ati awon ofin miin.**
English Source	The documents showed fourteen banks helped wealthy clients hide billions of US dollars of wealth to avoid taxes and other regulations.

Table G.1: Some errors from flores101 for Yorùbá. We indicate the errors with bold type fonts.