NLP Tools for African Languages: Overview

Joaquim Mussandi
Lisbon University
Instituto Superior Técnico
Luanda University
Instituto de Tecnologias de Informação e Comunicação
joaquim.mussandi@tecnico.ulisboa.pt

Andreas Wichert
Lisbon University
Instituto Superior Técnico
andreas.wichert@tecnico.ulisboa.pt

Abstract

In Natural Language Processing (NLP), languages are classified into low-resource languages and high-resource languages, therefore, rich. African languages are grouped among those with few resources, due to little investment and consequently little interest from researchers. To understand this phenomenon, and without wanting to enter into a discussion in the historical-anthropological, sociological forum, or other areas, we carried out a study to identify the languages that have PLN resources, report the points of greatest consensus among experts, identify researchers' motivations, possible opportunities, challenges, possible linguistic patterns, institutions, events and projects that stimulate them, the most used techniques for constructing datasets in vernaculars and the different forms of corpus training. From what we have seen, it seems to us to be a promising field of study.

1 Introduction

Languages are the main communication mechanisms. In this era of digitalization, there is a need for African languages to keep up with this dynamic. Technological linguistic resources are essential for the development of the economic, political, financial, educational, medical and tourist sectors, etc. as they constitute the basis for the development of more advanced research in Artificial Intelligence. Automatic text translation, voice search, sentiment analysis, data analysis and event prediction (Siminyu et al., 2021, 2020) are some of the domains that require contextualized linguistic corpus. However, there is a shortage of funding, documentation and human resources to overcome the challenges in building technological resources in Natural Language Processing (NLP) in languages of African origin (Siminyu et al., 2021; Adda et al., 2016; Ayogu and Abu, 2021), as well as mitigating the possibility of extinction of some languages that are under threat (Sands, 2018). At the same time, there is a need to reduce the difference between the richest languages and those with few resources (Adda et al., 2016).

The proliferation of the Internet in urban and suburban areas of Africa could mean an increase in data in digital format in these languages, which consequently increases their visibility within and outside the continent with the possibility of collecting data through Web tracking (Kandybowicz and Torrence, 2017). The corpus annotated at the token level, identification of orthographic patterns, grammatical classification of texts at the morphological and morphosyntactic level in vernaculars are differentials for the development of other research in this domain (du Toit and Puttkammer, 2021). Like the challenges of languages with greater resources, for corpuses with text-to-speech, actors must be found to record such extracts through simulated scenes (Siminyu et al., 2021) or social communication professionals, teachers, judges, etc. as long as they speak the respective languages.

In this overview we surveyed the languages of African origin that have PLN tools, the institutions that motivate the holding of events and projects to build technological tools in languages of African origin, opportunities and challenges, particularities in corpus construction in these languages. This field of study is promising.

2 Initial considerations: definition of research questions

There are, in the literature, several studies carried out on African languages. Some study them in the linguistic aspect (Matsinhe and Fernando, 2008), others study them in the political aspect (Kanana Erastus and Erastus, 2013) combining linguistic history to identify common ancestors through language (phylogenetic) (Schryver et al., 2015), territories, ethnicities, culture of the people (Pinto and Silva, 2022; Ki-Zerbo, 2010), in an attempt to revitalize languages threatened with extinction (Sands, 2018). NLP specialists motivated to provide linguistic resources for the technology industry are also invited (Kanana Erastus and Erastus, 2013; Loubser and Puttkammer, 2020a; Niekerk et al., 2017; Siminyu et al., 2020).

1. Technologically, is there research carried out on African languages?
2. But, why study African languages from a technological perspective?
3. What NLP tools exist in African languages?

These questions are answered in this article as we read it further.

2.1 African languages as a factor in economic and socio-cultural development

The economic sector is one of the most affected by the lack of technological resources in African languages. In Africa, generally, the languages of the former colonial powers are spoken in large urban centers and local languages are spoken mainly by young people and adults in the interior regions. Classes in national systems are mostly taught in the languages of colonial powers, sometimes providing children with illustrations and examples outside of everyday games, increasing the degree of difficulty in learning. Especially because knowledge of linguistic history helps in the search for the necessary bases for inferences about the cultural history of its speakers (Ki-Zerbo, 2010).

Siminyu et al. state that with an investment in NLP, governments would be the biggest beneficiaries, as they would have factual data for making decisions about investment within the scope of public policies, the analysis of feelings about the needs of local customers, to direct private investors, for example, example (Siminyu et al., 2021). Unfortunately, some governments restrict national languages for use in certain domains considered informal, such as intra-community communication, interpretive roles in local courts, use by politicians at rallies, etc. For Kanana, Tanzania, Ethiopia and Egypt, with Swahili, Amharic and Arabic languages, respectively, and most Arabic-speaking countries, are references in the development of native languages that today serve as national languages used for education, business and commerce (Kanana Erastus and Erastus, 2013).

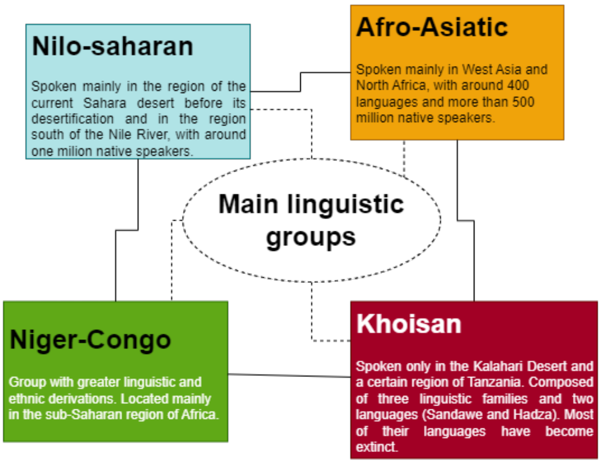
3 Background: work carried out

NLP resources, in particular, can be a means of accelerating the process of digital inclusion. This vision is also shared by Pauw and Schryver who direct their research in seven African languages of Bantu origin, namely Ciluba (Republic of Congo), Gikuyu, Kikamba (Kenya), Sotho and/or Soto and Venda (South Africa), the Nilotic Maa (Kenya) and the Defoid Yoruba (Nigeria) whose spelling is similar (diacritics), having developed applications and technological components under the approach of data-driven methods (Pauw and Gilles-Maurice de Schryver, 2009). Jakobus and Puttkammer have developed some language technologies for four official South African languages, namely Ndebele, Swati, Xhosa and Zulu. These technologies are summarized as lemmatizer, part-of-speech tagger, morphological analyzer for each of the languages. This was possible after building a corpus in each of these languages (du Toit and Puttkammer, 2021). In another research, an implementation of artificial neural networks was used in word embedding to build language technologies in which the data was modeled sequentially to perform all part-of-speech tagging tasks (POS-tagging, grammatical marking of the text), lemmatization, Named Entity Recognition[1] (NER), compound analysis. It is a language model that, for translation, given a sequence of input words, predicts the output, with different lexicons and lengths. This approach was based on ten South African languages[2], having presented good results, the best in the state of art at the height (Loubser and Puttkammer, 2020a).

Two other studies carried out mainly, but not exclusively, in the former territory of the Kongo kingdom, on the Kikongo with a focus on phylogenetics up to half a century. The first done by Gilles-Maurice et. al. directed from two perspectives: the first to present the character-based phylogenetic classification applied to lexical data, which became known as Kikongo Language Clouster (KLC) and, second, to present an exhaustive overview of the field of lexico-statistics of Bantu languages about KLC (Schryver et al., 2015). The second study carried out by Bostoen to examine variation in the expression of tense and aspect (TA) in a universe of 23 varieties of modern and two historical Bantu

[1] NER is a technique used in NLP to categorize and identify key information in a text.

[2] The languages studied are: Ndebele, Afrikaans, Xhosa, Zulu, Swati, Sepedi, Sotho, Swana, Tsonga and Venda.



Nilo-saharan

Spoken mainly in the region of the current Sahara desert before its desertification and in the region south of the Nile River, with around one milion native speakers.

Afro-Asiatic

Spoken mainly in West Asia and North Africa, with around 400 languages and more than 500 million native speakers.

Main linguistic groups

Niger-Congo

Group with greater linguistic and ethnic derivations. Located mainly in the sub-Saharan region of Africa.

Khoisan

Spoken only in the Kalahari Desert and a certain region of Tanzania. Composed of three linguistic families and two languages (Sandawe and Hadza). Most of their languages have become extinct.

Figure 1: Main African linguistic groups

languages, including Kikongo (Dom and Bostoen, 2015).

Yamaguchi and Tanaka-Ishii include two national languages of Angola in their research. These are Umbundu and Kimbundu (Hiroshi Yamaguchi and Tanaka-Ishii, 2012). Outside the domain of NLP, Kikongo is also a regional language (spoken in Angola, DRC, Congo and Gabon) and has been studied by academics for its valorization, for example, in understanding the complexity of its morphology for preliminary exploration of the ordering of affixes verbal (Pinto and Silva, 2022).

Voice searches, automatic text translation, speech-to-text conversion are tasks that receive greater attention from researchers, however, for languages with greater economic power. For African languages, Siminyu et al. performed data collection for open source corpus annotations for varieties of NLP tasks and creating baselines for machine learning tasks. And they state that NLP contributes to the promotion, appreciation and dissemination of linguistic diversity (multilingual), as well as inclusion (Siminyu et al., 2021).

3.1 Understanding African languages: origins

There are over 7 thousand languages spoken in the world. The African continent, like the others, is multilingual. UNESCO recognizes around 2,092 languages and dialects spoken in Africa, in addition to creoles, a mixture of the languages of former settlers and local native languages (Ki-Zerbo, 2010). These languages have various origins, briefly stratified into four main groups, illustrated in Figure 1.

Sub-Saharan Africa, as a result of what happened at the Berlin conference in 1985, became the most plurilingual, pluricultural and pluriethnic part of the continent (Ki-Zerbo, 2010; Pinto and Silva, 2022). Cameroon alone shares around 70 languages with neighboring countries. One of these countries is Nigeria, with which it shares around 45 languages (Kanana Erastus and Erastus, 2013). Zimbabwe and South Africa are the countries in Africa with the most official languages, 16 in the first and 11 in the second with legal support in their constitution, in addition to another 24 languages (Loubser and Puttkammer, 2020b).

3.2 Main Linguistic Events, Projects and Institutions

Since 1953, UNESCO has debated the use of local languages for African education and to rescue sociocultural values prior to the colonial period, the specific educational need. This strategy would be characterized as effective, firstly, by the use of an appropriate teaching medium, contextualized teaching techniques, culturally appropriate curricular content and sufficient financial and material resources. In addition to supporting various initiatives such as *AI4D, International Conference Language Technologies for All (LT4All)* and others (Siminyu et al., 2021). UNESCO celebrates the day of reflection on the languages of minority groups on February 21st. And, from 2022 to 2032, it coordinates the implementation of the international decade of indigenous languages declared by the United Nations in 2019, known as *Los Pinos Declaration on the Decade of Indigenous Languages* (Siminyu et al., 2021).

In 2020, it held the competition called AI4D-Africa Language Dataset Challenge[3] with the objective of creating and curating a dataset on African languages spoken in South Africa, Ghana and Uganda, with quality data to be used in future studies for language models (Siminyu et al., 2021). AI4D has already held seven events, including six competitions and a *hackathon* in African languages, namely:

- In November 2020, *GIZ NLP Agricultural Keyword Spotter for Luganda*;
- In February 2021, *AI4D Swahili News Classification Challenge*, participants were allowed from Tanzania, Kenya, Malawi, Uganda and Rwanda;

[3] Details here: https://zindi.africa/competitions/ai4d-african-language-dataset-challenge.

Language	Mains tools	Sources
Afrikaans	Corpus and Morphological analyser	(Eiselen and Puttkammer, 2014)
Ciluba	Corpus and Morphological analyser	(Pauw and Gilles-Maurice de Schryver, 2009)
Chichewa	Morphological analyser	(Keet, 2016)
Swahili	Machine Translation, Morphological analyser	(Keet, 2016)
Zulu	Machine Translation	(Eiselen and Puttkammer, 2014; Keet, 2016)
Shona (Xhosa)	Morphological analyser	(Eiselen and Puttkammer, 2014; Keet, 2016)
Setswana	Morphological analyser	(Eiselen and Puttkammer, 2014; Keet, 2016)
Ndebele	Morphological analyser, Language Model	(Eiselen and Puttkammer, 2014; Keet, 2016)
Sepedi	Morphological analyser	(Eiselen and Puttkammer, 2014)
Yuruba	Machine Translation	(Siminyu et al., 2021; Ayogu and Abu, 2021)
Swati	Morphological analyser	(Eiselen and Puttkammer, 2014; Keet, 2016)
Venda	Morphological analyser	(Eiselen and Puttkammer, 2014; Keet, 2016)
Sotho	Morphological analyser	(Eiselen and Puttkammer, 2014; Keet, 2016)
Tsonga	Morphological analyser	(Eiselen and Puttkammer, 2014; Keet, 2016)
Luganda	Agricultural Keyword Spotter	(Siminyu et al., 2021)
Chichewa	News Classification	(Siminyu et al., 2021)
Fongbe	Machine Translation	(Siminyu et al., 2021)
Ewe	Machine Translation	(Siminyu et al., 2021)
Tunisian Arabizi	Social Media Sentiment Analysis	(Siminyu et al., 2021)
Wolof	Automatic Speech Recognition	(Siminyu et al., 2021)
Kinyrwanda	CHAT-GPT for Covid-19,	(Forum, 2022)
Kikongo	Clouster	(Schryver et al., 2015)
Umdundu	Corpus	(Hiroshi Yamaguchi and Tanaka-Ishii, 2012)
Kimbundu	Corpus	(Hiroshi Yamaguchi and Tanaka-Ishii, 2012)
Basaa	Machine translation	(Adda et al., 2016)
Myene	Machine translation	(Adda et al., 2016)
Embosi	Machine translation	(Adda et al., 2016)
Emakhuwa	Corpus	(Ali et al., 2021)
Lingala	Corpus	(Sene-Mongaba, 2015)

Table 1: Main NLP toolkits for African languages

- As of March 2021, *AI4D iCompass Social Media Sentiment Analysis for Tunisian Arabizi*;
- In May 2021, *AI4D Yoruba Machine Translation Challenge*;
- In May 2021, *AI4D Takwimu Lab - Machine Translation Challenge: French to Fongbe and Ewe*;
- In May 2021, *AI4D Chichewa News Classification Challenge*;
- Coming May 2021, *AI4D Baamtu Datamation - Automatic Speech Recognition in WOLOF*.

AI4D - Artificial Intelligence for Development is an initiative that involves entities from the academic, business and governmental community whose vision is to leverage AI through high-quality research, responsible innovation and strengthening talent, as well as enabling better political decisions by governments to its people, integrating African languages into digital platforms.

In 2016, Adda et al. carried out a project that became known as **BULB**[4] (*Breaking the Unwritten Language Barrier*), which brought together linguists and computer scientists, with the aim of supporting the documentation of languages with few resources. To achieve this, tools adapted to the documentary needs of linguists were developed, taking advantage of NLP technology and knowledge, especially in *automatic speech recognition* and *machine translation*. Using three Afrikan languages from the Bantu family, from the **Nigerian-Congolese** linguistic group, namely: Basaa, Myene and Embosi. The project was divided into three

[4] Available at:https://www.bulb-project.org/

main stages: in the first stage, data was collected from a large audio corpus (100h per language) at a reasonable cost. To achieve this, standard mobile devices and dedicated software were used – Lig-Aikuma. The recorded data was repeated by a reference speaker to improve signal quality and translated orally into French. In the second stage, automatic transcription of the Bantu languages was carried out at the phoneme level and the French translation at the word level. Recognized Bantu phonemes and French words were automatically aligned. In the third and final stage, tools were developed in close cooperation and discussion between linguists, speech and language technologists, to support linguists in their work, taking into account their needs and technological capabilities (Adda et al., 2016).

Totoeba is a project that involves around 500 languages, with emphasis on those with few resources, whose objective is to create automatic translation tools and models covering the widest languages in the world, in an open way. It has a comprehensive collection of diverse datasets in hundreds of languages with systematic language and *script* annotation. Provides a growing number of pre-trained baseline models for individual language pairs and selected language groups. In line with the main objective is the purpose of encouraging people to develop machine translation in real-world cases into multiple languages (Tiedemann, 2020). With around 500 GB of compressed data for 2,961 language pairs, covering 555 languages. Unlike the annotation made by Agíc and Vulíc (Agić and Vulić, 2019), Tiedemann labeled the stored languages with IDs that associate them with the name of the source *corpus* for the training datasets (Tiedemann, 2020).

The *Department of Arts and Culture, and the Department of Science and Innovation* of the South African Republic, more than two decades later, continue to fund several NLP projects related to African languages, such as HLTs, SADiLar[5] (*South African Center for Digital Language Resources*), SARIR [6] (South African Research Infrastructure Roadmap).

JW300 is a project that involves 343 languages and a total of 1,335,376 articles, around 109 million sentences, 1.48 billion tokens. This data is collected by tracking publications made on the jw.org portal. and cover topics from different areas. The primary language of published information is English. Each published article has an identifier to identify it in any language. These articles were converted to clean text in HTML format with one sentence per line, having aligned more than 50 thousand language pairs with more than 90 thousand parallel sentences per language pair on average. The high-resource languages (English, French, German, Portuguese and Italian) stood out in performance (Agić and Vulić, 2019).

FAIR Forward – Artificial Intelligence for all[7] is a project that is part of the " *Digital Transformation for Sustainable Development* from *German Federal Ministry for Economic Cooperation and Development (BMZ)* and implemented by GIZ. FAIR Forward collaborates closely with the other flagship projects of this initiative, *the global e-learning platform Atingi, the Centers for Digital Transformation in Africa and Asia, the BMZ Digilab, the Data Economy and Data4policy* as well as business initiative projects *Make-IT*. At the African level, Ghana, Rwanda, Kenya, South Africa and Uganda collaborate. While in Asia it has Indonesia and India. With a view to achieving three objectives: *Access to Training Data and AI Technologies for Local Innovation, strengthening local technical know-how on AI and Developing Policy Frameworks for Ethical AI, Data Protection and Privacy*. FAIR Forward, the Mozila Foundation and local partners from Rwanda, Uganda and Kenya, contribute to the development of *open AI training datasets* in the languages Luganda and Kiswahili[8] and Kinyarwanda, which under funding from its local government and partners, developed a national *chatbot* that reports on the status of the Covid-19 pandemic in the local language, Kinyarwanda (Forum, 2022). South Africa has produced documents to standardize its local languages [9], including for the education sector[10]. The measures contained in these documents

[5] **SAiLaR:** is a language training center focusing on all the official languages of South Africa, in the humanities and social sciences, in the linguistic-technological domain. The center is supported by the Department of Science and Innovation for the creation, management and distribution of digital language resources. Available at: `https://sadilar.org/index.php/en/`

[6] **SARIR:** SARIR is an intervention high-level strategic and systemic approach to providing research infrastructure across the public research system, building on existing capabilities and strengths and taking advantage of future needs.

[7] Details at `https://www.bmz-digital.global/en/overview-of-initiatives/fair-forward/`

[8] It is spoken by over 150 million people.

[9] `https://www.gov.za/documents/`

[10] `https://www.gov.za/sites/default/files/gcis_document/201409/langframe0.pdf`

help preserve the cultural significance of each language, consequently, they contribute to the tools and resources that serve its community of language researchers (du Toit and Puttkammer, 2021).

3.3 Corpus in African languages

Artificial Intelligence applications use, for their experiments, preferably huge data sets to improve their performances. In this sector there is a deficit for African languages, in some cases compromising the results of these experiences and in other cases hindering their implementation. Corpus are an important component of the NLP tools necessary for developing your multilingual solutions, such as in the development of programs to perform tasks such as automatic text translation, information extraction, text classification, sentiment analysis, text summarization, among other tasks. (du Toit and Puttkammer, 2021). For its construction, specialists from related areas are needed, such as linguists, translators, computer engineers, as well as native speakers of the language (Siminyu et al., 2020). African languages classified among low-resource languages lack a varied corpus. The few found in the literature are from institutional and individual or independent initiatives with a little more than a dozen African languages, with emphasis on:

- SADiLAR, South African Government repository for datasets in local languages: https://repo.sadilar.org/handle/20.500.12185/1 (du Toit and Puttkammer, 2021);
- Vector words in 157 languages, available at: https://fasttext.cc/docs/en/crawl-vectors.html
- Kinyarwanda dataset https://digitalumuganda.com/dataset/ corpus that serves as the basis for the Government of Rwanda's chatbot functionalities.
- The Alliance of Digital Humanities Organizations (ADHO) held its annual conferences from 2009 to 2019. The data available at https://aflat.org/ serves to catalog its results and make them available to the community of researchers.
- Universal Dependency: coprora developed by a group of independent researchers, available at: https://universaldependencies.org/format.html
- Tatoleba is a project composed of more than 300 languages spoken around the world, including audio of the words, available at: https://tatoeba.org/pt-br/stats/sentences_by_language (Agić and Vulić, 2019).

The collection of data from native speakers of African languages, as well as on other continents, has followed some legal assumptions, regardless of whether the data is in textual format, images, audio or video, whether via web tracking or other alternatives (Siminyu et al., 2020).

3.3.1 Corpus construction techniques

The process of building a corpus follows at least five steps, namely: identification of a primary source of data (data source), definition of the protocol, pre-processing of texts, annotations of the corpus data, part-of- speech tagging[11] (level of morphological decomposition) and lemmatization (du Toit and Puttkammer, 2021; Loubser and Puttkammer, 2020b). However, the form of implementation according to the specificity of each language or vernacular may imply some changes from one researcher to another, as well as depending on the methodology. It has already been used in bag of word, Markov model (Hidden Markov Models) (Loubser and Puttkammer, 2020b), text alignment under heuristics (Tiedemann, 2014) and currently it is recurrently used in artificial neural networks with various configurations (Loubser and Puttkammer, 2020b). However, pre-processing texts in African languages is a relatively difficult task due to the fact that many of these languages use lexicons borrowed from the official languages spoken in the respective countries and in some cases, the lack of orthographic uniformity due to the lack of a writing standard (Siminyu et al., 2021). For these cases, manual text cleaning is used first, after which regular expressions are used to check terms in official languages using linguistic detectors (Siminyu et al., 2021; Agić and Vulić, 2019). However, for Tiedemann, the use of this tool is preceded by cleaning characters and strings that violate Unicode encoding principles using the re-coding forced encoding mode (Tiedemann, 2020).

3.3.2 Main sources of data collection

A data source represents the source that feeds the rest of the process of creating NLP technologi-

[11] Part-Of Speech Tagging is preferable to be done by linguists

cal tools. This is where the main problem lies in building technological tools for solving NLP problems in African languages (du Toit and Puttkammer, 2021; Niekerk et al., 2017). While there tends to be little writing in African languages, there is some media content online in local languages. Some international media powerhouses, such as the BBC and Voice Of America, have versions of their websites aimed at African audiences with content exclusively in African languages (Siminyu et al., 2021), as do a few religious institutions (Tiedemann, 2020). Siminyu et al. present three ways of collecting data, namely: data scraping from online sources, which, through the provision of online data in African languages, performs automatic or manual collection to verify accuracy, in addition to transcriptions of TED talks/films, transcriptions of radio and texts. The translators, with sentences created from scratch based on certain themes, with highly experienced local translators. And later online collection by tracking; and finally audio recordings, with the conversion of texts into audio using recordings of actors (Siminyu et al., 2021), as well as crowdsourcing (Nzeyimana, 2020). In other experiments, a set of recordings was used in a large corpus of speech with around 100 hours per language. To achieve this, standard mobile devices and dedicated software were used – Lig-Aikuma. The recorded data was repeated by reference speakers to improve signal quality and translated orally into French (Adda et al., 2016). Having the internet as the basis for obtaining texts in main and non-main languages, another research was carried out with more than 200 languages from four continents, with greater emphasis on some spoken in Africa (between main and non-main languages), in which, given a document written in several languages, an attempt was made to identify certain parts of the same document in other languages (Hiroshi Yamaguchi and Tanaka-Ishii, 2012). The language identification process is preceded by the segmentation of documents into each of the languages in the document, duly labeled, in a small amount of data, on the one hand, due to the scarcity of data in non-main languages, on the other hand, due to the difficulty of using certain machine learning methods with reduced learning corpuses.

With around 500 GB of compressed data for 2,961 language pairs, covering 555 languages. Unlike the annotation made by Agíc and Vulíc (Agić and Vulić, 2019), Tiedemann labeled the stored languages with IDs that associate them with the name of the source corpus for the training datasets (Tiedemann, 2020). Briefly, the data is collected on the Web in video, text, audio format (Adda et al., 2016; Agić and Vulić, 2019; Tiedemann, 2020), in languages with greater resources and through experienced speakers they are re-recorded and translated into African languages in text format (Adda et al., 2016) . In some cases, actors speaking local languages perform with free themes in environments (Siminyu et al., 2021, 2020).

3.3.3 Machine Learning

To build language models, machine learning has been widely used; Successful examples for language models have generally used supervised learning combined with reinforcement learning. For Agi and Vuli to build linguistic tools for natural language processing without supervision, it is not possible to achieve the minimum quality required (Agić and Vulić, 2019). In either form, artificial neural networks demonstrate excellent results (Loubser and Puttkammer, 2020b; Agić and Vulić, 2019; Tiedemann, 2014). BERT with simple two-layer variations was used for language model training of low-resource languages (Nzeyimana, 2020).

3.4 Analysis and Discussion

NLP resources are used to perform various tasks, generally related to texts and voice. For texts, are used in sentiment analysis, text summarization, automatic translation, document classification, information extraction, document similarity search, keyword extraction, etc (Sefara et al., 2022). As for voice, the tasks are text-to-speech conversion, audio analysis, audio transcription or automatic speech recognition, music information retrieval, audio classification, real-life applications, etc. (Li and Màrquez, 2010; Sefara et al., 2022). NLP provides technology for, for example, people with visual impairments, to obtain information, such as books, in audio form, allowing access to information, which is one of several forms of social inclusion.

The Figure 2 presents the types and quantification (percentage) of NLP tools for African languages, based on Table 1. It is worth adding that all the tools mentioned here are based on a generally annotated linguistic corpus.

Languages of African origin spoken exclusively in Portuguese-speaking African Countries (PALOP) are among those most in need of NLP resources. Angola, for example, has more than

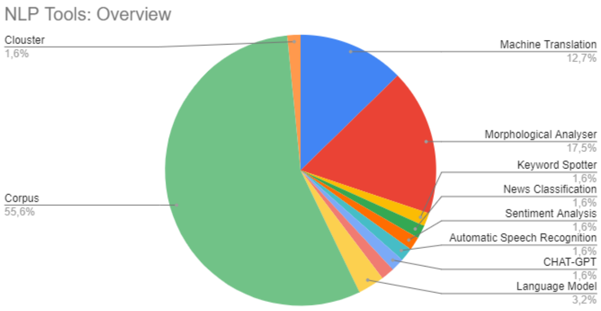


Figure 2: NLP tools overview

eleven languages of African origin, only three (Kimbundu e Umbundu) have unannotated linguistic corpus, in inaccessible repositories (Hiroshi Yamaguchi and Tanaka-Ishii, 2012), and Kikongo, which was included in a phytogenic study carried out in the former territory belonging to the Kongo kingdom (Schryver et al., 2015). The same applies to Mozambique, which has few resources for the Emakhuwa language (Ali et al., 2021).

Artificial Neural Networks (ANN) are widely used as the main tool in building language models for machine translation (Wang et al., 2019), and researchers in African languages have adopted them (Loubser and Puttkammer, 2020b).

On the other that, (Nzeyimana, 2020) argues that language models pre-trained on high-quality monolingual corpora generally present the best performances, especially for morphologically rich languages. This is the case for most African languages.

However, to create the corpus, common principles in NLP are observed, which include cleaning the texts (as plain text), detecting words in the official Western languages (main languages) using the monolingual and multilingual gold standard, annotating the corpus, which are generally done in four different ways: Part-Of-Speech Tagging (POS-Tagging), Name Entity Recognition (NER), compound analysis and lemmatization(Nzeyimana, 2020; Loubser and Puttkammer, 2020b). According to our findings, apart from the aflat (Pauw and Gilles-Maurice de Schryver, 2009) which will have ended in 2019, and (Hiroshi Yamaguchi and Tanaka-Ishii, 2012), the corpus projects in African languages, all others are in progress, according to their own calendar.

Future research should be directed towards building annotated corpus of languages that are not included in Table 1, and/or improving the performance of existing resources in that table, to equate the diversity of annotated and unannotated corpus of South African languages.

4 Conclusions

There has been research carried out in African languages for some time, however the ratio between researchers and languages without technological resources is enormous, although the barrier of scarcity of funding (Siminyu et al., 2021), scarcity of content, the lack of a grammatical standard and spelling rules, especially in languages spoken in more than one country (Sands, 2017), there is a need to study them to build the basis of technological development and reduce the difference with languages with more resources. The Table 1 presents some of the main NLP tools existing in African languages. Development cannot occur where there are linguistic barriers (Kanana Erastus and Erastus, 2013), as languages are present in the daily lives of their speakers and serve as a work, educational tool, when seeking medical appointments, etc. and are also a means of technological inclusion for their speakers (Kandybowicz and Torrence, 2017). In this regard, African languages lagged behind, with few human resources motivated to research in this area. The only reason for the technological delay of African languages is not the economic aspect, but also those related to cultural appreciation, political interest and other aspects (Kanana Erastus and Erastus, 2013). Swuhalli, a language spoken by more than fifty million people in Africa (Kandybowicz and Torrence, 2017; Pauw and Gilles-Maurice de Schryver, 2009), would have much more resources if other governments of countries that have territories inhabited by speakers of this language had the same concern as the government of Rwanda, Uganda and South Africa, which, following these efforts, built a corpus of all its official languages and beyond, some of which are licensed for free use for non-commercial purposes (du Toit and Puttkammer, 2021; Agić and Vulić, 2019) and some cross-language translator tools (Loubser and Puttkammer, 2020b). Kikongo spoken in around four countries, with a universe of around 7 million native speakers, has one or another technological tool without due attention from institutions and researchers in these countries (Schryver et al., 2015). It appears that there is some technological development of a given African language in countries that adopt them as official languages. For example, since 2013, Google Translator has translated from English to Zulu and Facebook has provided an interface in Chichewa, Kiswahili, Zulu and Shona, which can also be found on the Ubuntu (Keet, 2016) operating system. For the linguistic diversity of the continent, these advances, although motivating, are still insignificant. However, South Africa is a model to be taken into account, especially for giving more visibility to its other official languages (Siminyu et al., 2021; Kandybowicz and Torrence, 2017; Loubser and Puttkammer, 2020b; Niekerk et al., 2017; Sands, 2017; Schlunz, 2018).

Acknowledgements

To Projecto de Desenvolvimento de Ciência e Tecnologias (PDCT) – Angola for funding the scholarship and INESC-ID and GAIPS, the research laboratories where I am a guest during my PhD.

References

Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. 2016. Breaking the unwritten language barrier: The bulb project. *Procedia Computer Science*, 81:8–14. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

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. Association for Computational Linguistics.

Felermino D. M. A. Ali, Andrew Caines, and Jaimito L. A. Malawi. 2021. Towards a parallel corpus of portuguese and the bantu language emakhuwa of mozambique. *arXiv:2104.05753v1 [cs.CL] 12 Apr 2021*, 2021.

Ikechukwu Ignatius Ayogu and Onoja Abu. 2021. Automatic Diacritic Recovery with focus on the Quality of the training Corpus for Resource-scarce Languages. In *2020 IEEE 2nd International Conference on Cyberspac (CYBER NIGERIA)*, pages 98–103, Abuja, Nigeria. IEEE.

Sebastian Dom and Koen Bostoen. 2015. Examining variation in the expression of tense/aspect to classify the Kikongo Language Cluster. Africana Linguistica 21, 163-211. *Africana Linguistica*, 21.

Jakobus S. du Toit and Martin J. Puttkammer. 2021. Developing core technologies for resource-scarce nguni languages. *Information*, 12(12).

Roald Eiselen and Martin Puttkammer. 2014. Developing text resources for ten south african languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3698–3703, Reykjavik, Iceland. European Language Resources Association (ELRA).

2022 World Economic Forum. 2022. *Chatbots RESET Framework: Rwanda Artificial Intelligence (AI) Triage Pilot*. World Economic Forum, Rwanda.

Hiroshi Yamaguchi and Kumiko Tanaka-Ishii. 2012. Text Segmentation by Language Using Minimum Description Length.

Fridah Kanana Erastus and Erastus. 2013. Examining african languages as tools for national development: The case of kiswahili. *The Journal of Pan African Studies*, 6:41–68.

Jason Kandybowicz and Harold Torrence. 2017. *Africa's Endangered Languages: Documentary and Theoretical Approaches*. Oxford University Press.

C. Maria Keet. 2016. An assessment of orthographic similarity measures for several African languages. ArXiv:1608.03065 [cs].

Joseph Ki-Zerbo. 2010. Historia geral da africa i: Metodologia e pre-historia da africa.

Hang Li and Lluís Màrquez, editors. 2010. *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Cambridge, MA.

Melinda Loubser and Martin J. Puttkammer. 2020a. Viability of Neural Networks for Core Technologies for Resource-Scarce Languages. *Information*, 11(1):41. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

Melinda Loubser and Martin J. Puttkammer. 2020b. Viability of neural networks for core technologies for resource-scarce languages. *Information*, 11(1).

Sozinho Matsinhe and Mbiavanga Fernando. 2008. A preliminary exploration of verbal affix ordering in Kikongo, a Bantu language of Angola. *Language Matters*, 42(2):332–358. `http://dx.doi.org/10.1007/s40858-017-0164-2`.

Daniel van Niekerk, Charl van Heerden, Marelie Davel, Neil Kleynhans, Oddur Kjartansson, Martin Jansche, and Linne Ha. 2017. Rapid Development of TTS Corpora for Four South African Languages. In *Interspeech 2017*, pages 2178–2182. ISCA.

Antoine Nzeyimana. 2020. Morphological disambiguation from stemming data. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4649–4660, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Guy De Pauw and Gilles-Maurice de Schryver. 2009. African Language Technology: The Data-Driven Perspective. pages 79–96.

Hermenegildo Pinto and Ana Alexandra Silva. 2022. Língua umbundu: caminhos para a sua preservação. *Revista angolana de ciências*, 4(1).

Bonny Sands. 2017. 11The Challenge of Documenting Africa's Least-Known Languages. In *Africa's Endangered Languages: Documentary and Theoretical Approaches*. Oxford University Press.

Bonny Sands. 2018. *Language revitalization in Africa*.

Georg I. Schlunz. 2018. Usability of Text-to-Speech Synthesis to Bridge the Digital Divide in South Africa: Language Practitioner Perspectives. In *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, pages 1–10, Plaine Magnien. IEEE.

Gilles-Maurice de Schryver, Rebecca Grollemund, Simon Branford, and Koen Bostoen. 2015. Introducing a state-of-the-art phylogenetic classification of the Kikongo Language Cluster. *Africana Linguistica*, 21:87–162.

Tshephisho Joseph Sefara, Mahlatse Mbooi, Katlego Mashile, Thompho Rambuda, and Mapitsi Rangata. 2022. A toolkit for text extraction and analysis for natural language processing tasks. In *2022 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, pages 1–6.

Bienvenu Sene-Mongaba. 2015. The making of lingala corpus: An under-resourced language and the internet. *Procedia - Social and Behavioral Sciences*, 198:442–450. Current Work in Corpus Linguistics: Working with Traditionally- conceived Corpora and Beyond. Selected Papers from the 7th International Conference on Corpus Linguistics (CILC2015).

Kathleen Siminyu, Sackey Freshia, Jade Abbott, and Vukosi Marivate. 2020. AI4D – African Language Dataset Challenge. ArXiv:2007.11865 [cs].

Kathleen Siminyu, Godson Kalipe, Davor Orlic, Jade Abbott, Vukosi Marivate, Sackey Freshia, Prateek Sibal, Bhanu Neupane, David I. Adelani, Amelia Taylor, Jamiil Toure ALI, Kevin Degila, Momboladji Balogoun, Thierno Ibrahima DIOP, Davis David, Chayma Fourati, Hatem Haddad, and Malek Naski. 2021. Ai4d – african language program.

Jörg Tiedemann. 2014. Rediscovering Annotation Projection for Cross-Lingual Parser Induction. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1854–1864, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. ArXiv:2010.06354 [cs].

Yuxuan Wang, Wanxiang Che, Jiang Guo, Yijia Liu, and Ting Liu. 2019. Cross-lingual BERT transformation for zero-shot dependency parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5721–5727, Hong Kong, China. Association for Computational Linguistics.