Learnings From Developing Translation Technology for Lingala and other Low-Resource Languages

Kirthi Shankar
kirthis2@illinois.edu

Abstract

In this paper, we investigate methodologies and processes required to build NLP tools for communities that speak low-resource languages, with a focus on the Congolese community in Urbana-Champaign, who speak primarily in Lingala, French, and English. A member of the community expressed interest in a Lingala-French + Lingala-English translation application, so we summarize research that contributes to this goal and identify ways to improve upon existing translation tools for Lingala. We explore ethical methods for expanding the existing parallel data for Lingala machine translation tasks, and how to ensure that the technology properly assesses code-switching that occurs in natural speech used by the community. We then lay out next steps for building the requested tool, such as adding code-switched data during the pretraining stage to expand the domain of translation past religion, working with the community to grow current datasets, supporting French-Lingala translation, and incorporating the final results into a mobile application. This research can also apply to other low-resource languages.

1 Introduction

Language technologies such as translation services, spell-checking, and online educational games have the potential to greatly serve the goals of minority communities that speak low-resource languages, meaning that there is little available data in these languages that can be used for natural language processing(NLP) tasks.

One such goal is accessibility. For example, Masakhane, a grassroots NLP organization in Africa, highlights that most important news in Africa is delivered in French, English, or Arabic, making it inaccessible to people who only speak African languages (Adelani et al., 2022). They also share that most educational material is not administered in African languages, preventing accessibility and excluding African languages from scientific domains. These are two of their incentives for developing better translation technology for African languages (Masakhane).

Another goal is language maintenance, which in this case refers to a minority language community keeping their language practices intact amidst pressure from the more dominant language to change their language or use it less. For example, The Yukon Native Language Center made efforts to provide language education by creating online learning material for eight indigenous languages, which maintains the languages' usage by helping the younger generation continue to use them (Littell et al., 2018).

However, producing machine-learning-based language technology for minority language communities presents challenges such as lack of data and potential community exploitation. There are also considerations about the different ways in which language is used by a community, such as code-switching and dialectal differences, bringing up the question of how we can accurately represent the diversity of a low-resource language in machine learning tasks.

In this paper, we cover how one would go about producing language technologies for a language community that uses a low-resource language. In particular, we focus on the Congolese community in Urbana-Champaign, who speak a mix of the low-resource language Lingala, French, English, and Swahili, and their request for improved translation technology. The resulting ideas are also relevant for other low-resource languages.

We cover general examples of NLP technologies that can be used to support minority language communities, prior machine translation work with Lingala, data collection procedures for low-resource languages, and the consideration of code-switching in machine translation.

Ultimately, this paper provides a starting point

Technology	Amount of Data
Keyboard	Low
Orthography conversion	Low
Text-to-Speech	Low
Optical Character Recognition	Low
Predictive text	High
Spell-checking	High
Machine Translation	High
Speech-to-Text	High
Audio Segmentation	High
Education	High

Table 1: Language technologies that can be used for language maintenance in minority language communities and their corresponding data requirements (Littell et al., 2018).

for future researchers working with a community that speaks a low-resource language, supporting future exploration of this area.

2 Language Technology in Support of Minority Language Community Goals

A paper by Littell et al. summarizes the various ways in which languages technologies can be used to support indigenous communities, as well as the corresponding amount of data that would be required to produce these technologies (Littell et al., 2018). The same tools can also be used by other minority language communities. Table 1 provides a succinct list containing this information, and the following sections provide details about prior work accomplished using some key technologies listed.

2.1 Educational Material

One subset of language technology, online educational content, can make language learning much more accessible, contributing to the efforts of minority language communities to maintain language use within the younger generation. Apps like Duolingo invite millions of users to engage with language learning in an accessible way, including the indigenous languages of Navajo and Guarani, for which the company partnered with community members to create. This partnership contributed to the communities' goal to prevent these endangered languages from going extinct (Alcántara, 2018).

Another effort toward language education was made by Danielle Boyer, a member of the Sault Ste Marie Tribe of Chippewa Indians, who developed an ethically-created AI robot called the SkoBot that helps children learn the endangered language Anishinaabemowin, aiding in language revitalization while working directly with her community to build the product. This project was entirely community-contained and freely distributed to children to provide access to the learning materials provided, which directly supports the community (Connection, 2024).

With the right amount of data, simple NLP tools can be used to develop grammatical exercises for languages.

For example, researchers created a method to generate multiple choice questions(MCQ) testing grammatical knowledge of the indigenous languages Maya, Guarani, and Bribri (Chiruzzo et al., 2024). The questions involved transforming a source string by correctly making the specified grammatical change (ex. a new conjugation or tense). The researchers used Prefer Observed Edit Trees from spaCy to generate transformations between each source and target sentence for each possible transformation between them. For Maya, they took a subset of 1,400 phrases from a corpus of 13,873 to be grammatically annotated by experts for the sake of this task. For Guarani, they used data from the Jojajovai corpus, Mozilla Common Voice corpus, and a generator for Guarani to Spanish pairs. Once the transformations and annotations were complete, they had a dataset of size 621. For BriBri, they used sentences from existing textbooks, a grammar book, and a treebank, gaining 1,001 sentences after applying new conjugations (Chiruzzo et al., 2024).

Another group of researchers used POS-tagging to create a game called Cipher: Faoi Gheasa based on identifying errors in Irish text with the goal to make Irish language learning more enjoyable (Ward et al., 2022). The game gave students practice in reading texts to identify errors and writing new words to change the ending of the stories. This game added an additional level of cultural education as some of the texts the students worked with were culture-specific stories. However, this research may be challenging to apply in a situation with less resources, as the researchers had support from the school system to test their product in an actual setting (Ward et al., 2022). POS-tagging also requires annotated data in the low resource language.

Overall, educational materials can be made using simple NLP technologies (ex. POS-tagging) to LLMs (ex. Duolingo). They can also be made without NLP via online lessons created by educa-

tors. However, the simpler NLP examples above required data to be annotated for grammatical structures or parts of speech, which involves hiring experts or volunteers if those datasets do not already exist for the low-resource language (5.3).

2.2 Modern Communication

One way to maintain minority languages is to incorporate their usage into speakers' everyday lives. Thus, technology that supports the usage of the language in online settings can contribute to language maintenance (Littell et al., 2018; Levin, 2009). One example of this is creating a keyboard that supports typing in that language's orthography, which allows digital communication in the language. For example, researchers Eddie Santos and Atticus Harrington developed a keyboard that allows typing using Plains Cree syllabics to promote online usage, using multiple layers to account for all variations to consonants(given that the language is an abugida). This project did not require a lot of data, as the researchers mainly needed character co-occurrence frequencies to determine which keys to place each symbol on (Santos and Harrigan, 2020).

Spell-checking and predictive text contribute to the same goal by increasing ease of communication, but these have been explored less for low-resource languages, and require more data (Littell et al., 2018).

2.3 Translation

Machine translation, typically between a low-resource language and the dominant language in the region, is another language technology that supports minority language communities by increasing accessibility of legal, medical, educational, and other services for members of the community who may not speak the dominant language(Adelani et al., 2022).

There have been many efforts to produce machine translation tools for low-resource languages. We will highlight the work of the Masakhane grassroots organization in Africa, which produces NLP research focused on African languages. The organization seeks to decolonize research done on African languages via community ownership and collaboration (Masakhane). In a recent publication by Masakhane, researchers developed new parallel datasets for sixteen low-resource African languages based on news articles, expanding the scope of existing data for these African languages past religious content (Adelani et al., 2022).

The researchers hired a team of translators and native speakers to translate and validate up to eight thousand sentences per language. They tested baseline Transformer sequence-to-sequence models for each language, as well as the pretrained models Byt5, mBART50, and M2M-100, a multilingual model. They tried three fine-tuning methods related to domain adaptation: 1) training on pre-existing religious corpora and fine-tuning on the news corpora, 2) training on the religious corpora and part of the news corpora, then fine-tuning on the rest of the news corpora, and 3) fine-tuning on both the religious and news corpora. They found that the M2M-100 model and the second domain adaptation method performed the best. They also found that a higher amount of initial religious text gives the model more information for training, which was beneficial(Adelani et al., 2022).

This research required hiring individuals to help generate the new corpora(Adelani et al., 2022). In the absence of this resource, researchers may use smaller corpora that already exist, such as Bible translations or JW300. However, it is important to identify whether religious domains are appropriate for all languages, depending on a community's historical background(5.1).

3 About the Congolese Community

3.1 Urbana-Champaign Congolese Community

One of the minority communities in the Urbana-Champaign area of Illinois is the Congolese community. The community is tight-knit and often meets in community, religious, and cultural events. They primarily speak in Lingala, French, English, and Swahili, often code-switching between these four languages, with Swahili being in the minority. Sociolinguistic research on the community by expert Williams Asamoah Frimpong indicates that in social settings, such as community or religious events, individuals primarily use Lingala or a mix between French and Lingala, while in non-community-centered settings, such as education, they use English. So, when engaging with individuals who are not community members, or people they do not have close ties with, they use English.

Lingala, English, and French are important in the community for different reasons. Lingala is used in social settings and at home to keep the language alive and maintain cultural practices, and has the highest rank in value among the three lan-

guages. French is the primary language used for all services. According to Frimpong's research, although individuals self-report proficiency in all the languages, English has the lowest self-reported proficiency, but second highest value after Lingala.

A member of the community shared that a mobile application for translation technology would be a valuable tool for the community, specifically translation between French and Lingala as well as English and Lingala. The primary need stated was that new members of the community might require support in adapting to the area via translation services. A mobile app that performs automatic speech translation would be the most helpful, with an online site being a reasonable starting point. Since there are existing technologies that perform this type of translation, including Google Translate, we explore the current state of these technologies and how they can be improved on to have a higher accuracy that better serves the needs of the community.

3.2 Lingala Properties

Lingala is a Bantu language spoken in the Democratic Republic of the Congo, the Republic of the Congo, and in diaspora communities, such as the Congolese community in Urbana-Champaign. It is a tonal language with an agglutinative verbal structure. The language also distinguishes nouns based on animacy and inanimacy, with verbal agreement based on these factors.

NLP technologies based on languages such as English treat individual words as a single unit of meaning, which can be troublesome for agglutinative languages in which one word can contain the combined meanings of multiple morphemes. Thus, for tasks such as transfer learning, it would be beneficial to base technology for Lingala off of technology for high-resource agglutinative languages.

Lingala is also a low-resource language, meaning it does not have a lot of existing data that can be used for training NLP tools. Current parallel data sources include the translated Bible, JW300, and Tatoeba, although Tatoeba is not as structured as the other two sources. For monolingual data, researchers have utilized the CC-100 dataset (Nguefack et al., 2025).

4 Prior Work in Machine Translation for Lingala

We will summarize previous efforts made to produce translation technology for Lingala.

4.1 Existing Publicly-Available Translation Tools

Current existing translation tools for Lingala that are publicly accessible (meaning they don't require programming knowledge to access) include Google Translate, Glosbe Translator, dic.lingala.be, Online Translation Pro, and a mobile app called "English to Lingala Translator". A notable oddity in the mobile app is that translating the phrase "Easily translate from English to Lingala" produces the translation "Traduire facilement de Anglais a Lingala", which is French. Although French-Lingala code-switching occurs often in the community, the application is explicitly set to translate to Lingala, indicating a lack of sufficient or clean data in Lingala that was used for this translation task.

Additionally, a key problem reported by Lingala speakers is that currently existing software does not always provide accurate translations, and thus better technology must be developed to satisfy this need.

4.2 LiSTra

LiSTra is an English to Lingala Automatic Speech Translation(AST) project by Kabongo Kabenamualu et al (Kabongo Kabenamualu et al., 2022). To create their dataset, the researchers utilized pre-existing English audio recordings of the Bible, and the Lingala translation of the Bible. They used WebMAUSBasic to split the English audio into verses and then paired these with the corresponding verses in Lingala.

They tested two methods of AST: 1) first convert the English audio to English text, then translate the English text into Lingala and 2) create a single model that takes in English audio and outputs Lingala text.

For the first method, they used a pretrained Sirelo model for the speech-to-text task, and then input the resulting English text into a standard transformer-based architecture for machine translation in order to get the translation in Lingala.

For the second method, end-2-end, they used a transformer-based model with one encoder and two decoders, and an interactive attention sub-layer allowing information sharing between the audio

transcription and text translation tasks.

Ultimately, their end-2-end model showed the best performance, with a BLEU score for Lingala of 28.52 and a BLEU score for English of 84.90. The two-stage model had the same BLEU score for English but a lower BLEU score for Lingala at 13.92.

This research established a baseline for automatic speech translation between English and Lingala, which is important to our work as the community requested a tool that can also handle audio inputs. The researchers mentioned that they could not get better results than previous translation work by Masakhane, attributing this to their translation decoder lacking adequate pretraining, which presents an area for future improvement (Kabongo Kabenamualu et al., 2022).

4.3 AfroMT

AfroMT is a machine learning baseline for translating between English and eight African languages, including Lingala (Reid et al., 2021).

For datasets, the researchers used JW300. They created a model, AfroBART, that uses a Transformer-base architecture allocating six layers for the encoder and decoder.

The researchers found that their methods of preprocessing the data improved the accuracy of the machine translation results. They removed parallel texts where either the source or target were missing and very short sentences that were just numbers/symbols. To tokenize data across multiple sources, the researchers used the Moses toolkit's detokenizer. Additionally, they removed sentences with too much overlap to reduce data leakage.

They also used the following data augmentation strategies to account for the lack of monolingual data that would help in pretraining tasks: 1) Generating synthetic code-switched data by replacing words in English with its corresponding translation in Lingala (or the other African languages they worked with) to introduce more African words to the model, and 2) using a pretrained machine learning model to generate new monolingual data in the African languages and introduce new domains of monolingual data.

The researchers pretrained their model on the eight languages they tested as well as English, French, and Dutch, using the monolingual corpora they generated as well as data from CC100, making sure the data was sampled in a balanced way.

AfroBART ultimately produced a BLEU score of 29.46 for translating between English and Lingala (Reid et al., 2021).

4.4 AfroMT: Later Experiments

Later work by Nguefack et al. builds on AfroMT by using AfroBART as a baseline and exploring new pretraining strategies in an attempt to improve the scores (Nguefack et al., 2025).

For their first experiment, they pre-trained only using monolingual data in Lingala from AfroMT, using the architecture of mBART and fairseq for denoising. This had the lower BLEU score of 25.34, which the authors attributed to the data being pretrained on only one language instead of multiple.

Based on this, they altered their experiment, adding monolingual data from other minority African languages and removing the monolingual data for English. This increased the first experiment's BLEU score by 2 points, achieving a BLEU score of 27.38.

Their following experiments did not result in a BLEU score higher than 27.38. They next attempted to include monolingual English data, which decreased their BLEU score. They also tried using only the English-Lingala parallel and monolingual data, which did not improve their BLEU score.

Ultimately, their experiments were not able to improve on AfroMT's BLEU scores, but they did identify that out of their experiments, keeping monolingual data from other African languages was most beneficial (Nguefack et al., 2025).

4.5 Challenges and Considerations

LiSTra, AfroMT, and the subsequent AfroMT experiments stated the challenge of data being limited to the Bible and JW300, which forced the translation task into a purely religious domain. Although the JW300 collection does feature some non-religious content, that content ultimately is linked with religious contexts or ideas, limiting the scope of the data. AfroMT provided the idea of iteratively using pretrained machine learning models to generate monolingual data of Lingala in different domains (Reid et al., 2021), however, the accuracy of this approach may decline the further the domain gets from the original domain of the parallel texts.

Thus, the next area we explore is general considerations when generating new datasets for a low-resource language, with a focus on community interactions.

5 Considerations when Expanding Datasets for Low-Resource Languages

5.1 Domain Limitations

As we saw, the majority of pre-existing parallel corpora available have purely religious domains. While this provides a great baseline for creating translation technology, it narrows the translation power to one domain, reducing the technology's scope of usage. For the Congolese community specifically, our contact cited a main reason for wanting this tool to be ease of communication in a work environment, emphasizing the necessity of translation in multiple domains.

Although this may not apply for the Congolese community in Urbana-Champaign since the community is majority-Christian, the use of religious corpora may not always be an appropriate option for other minority language communities due to the colonial histories behind translation. For example, certain indigenous communities have been subject to translation between their language and English as a form of oppression via forced conversion, and thus do not approve of using religious texts as a mode of building translation technology. Thus, before building translation technology, it is important to consider the historical contexts of the domain being used and the act of translation itself, for the community one is working with before jumping into the task (Mager et al., 2023).

5.2 Community Ownership and Permissions

When working on dataset generation with a community, it is important to involve the community in every aspect of the project. The initial stage of data collection and annotation as well as the decision of how the model or results will be used are aspects that the community must have a say in. Additionally, if the corpus containing the data is private, the community that contributed to the data should not be excluded from accessing it. The community must also have a say in any commercial uses of the product (Mager et al., 2023).

In research done with indigenous communities by Mager et al., members of the community were surveyed in order to understand their opinions about developing machine translation for their indigenous language. Some concerns mentioned by the community members were that translation quality may not be good, language diversity may be lost through attempted standardization, and that translations could have an adversarial impact on their culture and religion by revealing sacred rituals or information that belongs within the community (Mager et al., 2023). Thus, it is important to investigate potential community concerns behind translation before producing such technology.

Additionally, some communities will want to see official permission from a government organization before their language is studied, while others do not need this permission(Mager et al., 2023). Thus, it is necessary to understand what permissions are required by the community itself, and to also get Institutional Review Board approval before involving community members in research.

5.3 Expert Involvement

Another important aspect of expanding corpora for NLP use is that data creation and annotation often requires the support of experts, such as native speakers or linguists. The development of the Maya-Spanish parallel corpus utilized in the grammar MCQ project mentioned in section 2.1 required support from many individuals, such as Maya speakers, linguists, NLP researchers, and volunteers (Chiruzzo et al., 2024). Masakhane's sixteen-language translation project required a team of translators, and the time spent on the project depended on the schedules of these translators (Adelani et al., 2022). Hiring individuals to help create new datasets requires funding, which is covered in the next section.

5.4 Research Funding

In order to obtain funding for dataset expansion, one must look to organizations with needs or interests that match the desired outcome of the research being performed. Research is typically funded by a government agency, university, or business (Droegemeier, 2023), which we can see in the examples of past language technologies produced in support of minority language communities.

For example, the grammar MCQ project mentioned in section 2.1 used a Maya-Spanish dataset that was developed with the support of large companies and organizations, starting with Duolingo, then The Secretariat of Culture and the Arts of Yucatan, then The Geospatial Information Sciences Research Center (Chiruzzo et al., 2024). The educational game Cipher: Faoi Gheasa, required support from the school system to administer tests so that the researchers could see how students interacted with the game (Ward et al., 2022). Another set of projects by Carnegie Mellon University (AVENUE

and LETRAS) to translate Mapudungun and Iñupiaq were supported by Alaska Native Language Center at the University of Alaska at Fairbanks, the Universidad de la Frontera, and the Chilean Ministry of Education (Levin, 2009).

For the Maya-Spanish dataset, Duolingo initially intended to add Maya to their curriculum, but halted the project when they identified that the language's properties did not easily mesh with their teaching structure. However, The Geospatial Information Sciences Research Center maintained support for the project due to their desire to make technologies using Maya (Chiruzzo et al., 2024). This illustrates the importance of the intention of the research project: will it produce an outcome meant to be commercialized, or is it for community usage? Although there is certainly an overlap, typically, private institutions like businesses will fund the first type of research, and government organizations or universities are the ones that will fund the second type of research (Droegemeier, 2023).

5.5 Avoiding Extractive Research

There is a third intention for performing this kind of research: pure curiosity (Droegemeier, 2023). However, when involving a community in developing data that aids one's research, it is important to consider what the community desires out of their direct or indirect participation in order to avoid extractive research, where a community contributes to a researcher's knowledge, but does not gain anything from their contribution.

In his paper, "Must NLP be Extractive?", researcher Steven Bird recounts his experiences working with the Mok clan in Kabulwarnamyo, an Aboriginal settlement in Australia, who mostly rely on oral traditions (Bird, 2024). Bird argues that languages that are primarily oral and have very localized traditions do not need to be extracted into a text-based format so that they can become a part of the wider database of parallel texts used for machine translation. In order to create parallel text that involves such a unique language, relevant cultural aspects of that language are bound to be stripped off, taking ownership away from the producers of the language (Bird, 2024). This emphasizes the importance of understanding the desires of a community when considering the outcome of researching their language: one result of creating a parallel corpus could be a product that has no cultural relevance to the community, and is therefore not useful to them.

Bird collected data through recordings during guided tours. The following are practices he performed to ensure that his research was mutually beneficial, ethical, and non-exploitative:

- He stated his full intentions for his research before each tour.
- All participants gave their consent to be recorded.
- Participants were compensated for their time and not pressured to join.
- The overall goal of the research was "knowledge of the Country, not valorising or commodifying the language" (Bird, 2024).
- If there was audio from non-consenting individuals that was recorded, it was not used.
- He avoided asking participants to repeat tasks in order to not bore them.
- He avoided exposing a lack of knowledge about certain locations during the tour by not asking specific questions that the participants might not know the answers to. This was to make the process feel less like a test and more like a conversation for the participants.

Overall, the research has to be transparent, optional, compensated, comfortable for participants, and have the goal of serving the community instead of framing their needs in terms of the values of outsiders (Bird, 2024).

6 Code-Switching in Machine Translation Tasks

Code-switching is when more than one language is used in the same utterance. In code-switching, a matrix language is the language from which code-switched text derives its syntactic structure, and an embedded language is the language whose vocabulary is inserted into the matrix language.

As mentioned in section 5.1, one concern when producing translation services for a low-resource language is that machine translation may only work with a standardized version of the language that does not represent the way it is actually used by individuals in the community. This is relevant to the Congolese community in Urbana-Champaign because they often code-switch between Lingala, French, English, and occasionally Swahili. Thus,

it would be beneficial for machine translation tools to be able to adapt to code-switched language.

We will look at prior research in NLP tasks focusing on code-switched data to see how they can be applied to Lingala, as well as the potential for developing a code-switching corpus for Lingala.

6.1 Generating Synthetic Code-Switched Data

With the lack of parallel data for low resource languages, one can imagine that there is also a lack of parallel data that includes natural code-switching. This is also the case for high-resource languages. Thus, researchers have searched for ways to create synthetic code-switched data that can be used to make models more robust when encountering natural code-switching.

6.1.1 Replacement-based Generation

One group of researchers identified a method for creating synthetic English-Spanish and English-French code-switched data for translation tasks that surpassed the performance of multilingual models acting on code-switched text and performed similarly to machine translators on monolingual data (Xu and Yvon, 2021). They used fastalign to find word alignments in parallel texts, then statistically identified bilingual phrase pairs. Then, within the matrix language, they replaced these phrases with their counterparts to generate synthetic data. The researchers progressively replaced more and more phrases and identified that the more phrases that were replaced, the more accurate the translation task was (irrespective of whether the matrix language was the source language or not). For their machine translation model, they used a fairseq transformer model. They ended up using 13 million sentences for English and Spanish code-switching, and 33 million sentences for English and French code-switching (Xu and Yvon, 2021).

It is important to note that their model trained on synthetic code-switching data performed better in code-switching translation tasks than multilingual models trained on monolingual data do, which indicates that naturally-occurring code-switching will be better translated if the model used has been explicitly trained on real or synthetic code-switched data. Although the amount of data used exceeds what is currently possible for Lingala, the method of generating code-switched data itself is feasible with the data that currently exists.

6.1.2 Translation-based Generation

Another group of researchers created a translation-based method for generating synthetic corpora for code-switching (Tarunesh et al., 2021). To create synthetic code-switched data with Hindi as the matrix language and English as the embedded language, they created a machine learning model that translates from Hindi to Hindi-English code-switched text, using this model as a synthetic code-switched text generator.

The code-switched text they used was derived from movie conversations and a Hindi Treebank. They hired professional annotators to transcribe this data and transform each sentence into multiple code-switched sentences by translating different clauses into English. Additionally, they generated synthetic code-switched data using two different methods: 1) replacing Hindi words with their corresponding English translations based on the frequency this word is replaced in natural code-switching, and 2) another clause replacement method that is similar to that used by the previous researchers.

Based on the BERTScore evaluation metric for code-switching in NLP tasks, the researchers' supervised and unsupervised translation techniques for generating synthetic code-switching performed much better than their replacement-based generation methods. However, a BERT-based classifier for synthetic vs. real code-switching was able to distinguish between the translation-derived code-switched data and real code-switched data with high accuracy, indicating that this method of generating synthetic code-switching does not mimic natural code-switched text well.

Unlike the replacement-based methods, this translation-based method required annotators and evaluators for the generated results. It also relied on vast amounts of monolingual data, which may not be available for low-resource languages (Tarunesh et al., 2021).

6.2 Comparing Natural Versus Synthetic Code-Switched Data

Past research on Spanglish analyzes how the models mBERT, XLM-R-base, and XLM-R-large handle various NLP tasks for real and synthetic code-switched data in comparison to monolingual data (Laureano De Leon et al., 2024). For data collection, they collected tweets from X and asked someone fluent in Spanish and English to vali-

date that they are true examples of code-switching. For the researchers' syntax experiments, tweets were rewritten by bilingual individuals to have proper grammar(punctuation, capitalization, and shorthand correction i.e "u" to "you").

They also tested some of their experiments on synthetic code-switched data that the researchers produced themselves to see if synthetic data would also be a viable option. To create synthetic data, they first took the English and Spanish translations of the dataset they generated from X. Then, they applied two replacement methods: 1) replacing random tokens with their translation in the other language, and 2) using spaCy to identify noun phrases in the original language and translate them into the other language.

The researchers identified that XLM-R-large performed code-switching identification and language identification tasks well. They also found that models are able to identify syntactic structures and semantic meaning similarly as they would for monolingual data in the real code-switched data, but not in the synthetic data (Laureano De Leon et al., 2024). Although this could indicate that synthetic code-switched data is not effective in NLP tasks, the prior researchers used more advanced generation methods that had promising results for the specific task of machine translation, so developing synthetic code-switched data should still be explored.

6.3 Datasets for Code-Switching using Lingala

A lot of natural code-switched text comes from social media posts, as social media provides an informal setting where code-switching is free to take place. In order to use this data, it needs to first be appropriately cleaned (removing hashtags and emojis, standardizing punctuation, etc.). Then it must also be translated into one or both of the matrix and embedded languages.

A dataset containing millions of tweets from Africa from 2017 to 2023 has such examples of naturally occurring code-switched text for various languages spoken in Africa (Dunn, 2020). Although the tweets are not translated, they are cleaned and classified into languages using three different language models, the third being the most accurate. To create a corpus of tweets that use our languages of interest, Lingala, French, English, and Swahili, we use Python to filter the dataset, first removing tweets that have not been classified as any of these four languages, then keeping all the tweets classified by the third classifier as Lingala, which totals to 10,014 tweets.

Taking a closer look at these tweets, they contain code-switching between French, Lingala, English, and Swahili. The amount of code-switching in each tweet varies. For example, some tweets only seem to be classified as Lingala because of the presence of proper nouns, which technically does not count as code-switching(ex. "congolese residents in south africa march to press katumbi s return"). However, there are examples of true code-switching as well, such as the following tweet containing Lingala and French: "eza mawa soki tozoloba salaire et taux nul part qu en rdcongo".

To make this a viable code-switching corpus, we would need to perform the following steps:

- Validate the classifications identified by the model as well as whether the sentences contain true code-switching.
- Decide if grammatical alterations need to be made to make the text mimic conversational Lingala, and perform this "clean up" if required.
- Translate the code-switched text into corresponding Lingala, French, and English translations.

This would require hiring experts in the three languages, such as native speakers.

Alternatively, the same way monolingual and synthetic code-switching data is used in pretraining for AfroBART (Reid et al., 2021), so could this data for the sake of domain adaptation.

7 Conclusion

Working toward the goal of developing language technology for the Congolese community in Urbana-Champaign, we explored various ways in which NLP technologies can be used to empower minority language communities, prior work in developing translation technology for Lingala, and how this work has been challenged by a lack of Lingala-English and Lingala-French parallel data. To address this, we explored best practices for ethically and non-extractively working with a community to build on existing corpora for their language. Lastly, we wanted to understand how to make sure the natural code-switching that the Congolese community uses is properly translated by

machine learning models, so we looked at how researchers used synthetic and real code-switched data to prime machine translation models for code-switched data. Although this survey primarily focused on Lingala, the research on translation tools, corpus development, and code-switching incorporation can be applied to any low-resource language.

7.1 Next Steps

7.1.1 Improving Existing Models

In order to improve existing translation models for Lingala, one approach is to work with the community to develop more parallel corpora that can be used to expand the domain of translation tasks, as Masakhane did for sixteen African languages by building a new set of news corpora (Adelani et al., 2022).

Another option is to take advantage of existing data from social media, such as the corpus mentioned in section 6.3, which contains code-switched data in Lingala, French, English, and Swahili. As mentioned in the AfroMT projects, the addition of monolingual African data in pretraining stages improves model accuracy (Reid et al., 2021; Nguefack et al., 2025). Adding naturally occurring code-switched data to the model's pretraining could potentially expand its domain, prime it for code-switched data, and increase its accuracy.

Additionally, current translation tasks have been focused on translation from English to Lingala, but the request was for bidirectional translation between Lingala and English, as well as Lingala and French. Thus, research would have to be expanded to account for these three new translation directions.

7.1.2 Mobile Application

It is important that the community directly benefits from the developed model. Thus, the resulting machine learning model should be integrated into a mobile application that takes text or audio input and provides the corresponding translation to make the product accessible to the community.

8 Acknowledgements

Thank you to Dr. Jonathan Dunn and Williams Asamoah Frimpong for their guidance throughout this project. Thank you to the member of the Congolese community who provided feedback for this project about community needs, whose name is ommitted for privacy.

References

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen H. Muhammad, Guyo D. Jarso, Oreen Yousuf, and 26 others. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Amanda Alcántara. 2018. Popular language learning platform adds navajo and hawaiian languages to curriculum.

Steven Bird. 2024. Must NLP be extractive? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14915–14929, Bangkok, Thailand. Association for Computational Linguistics.

Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2024. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 224–235, Mexico City, Mexico. Association for Computational Linguistics.

Steam Connection. 2024. Meet the skobot.

Kelvin K. Droegemeier. 2023. *Demystifying the Academic Research Enterprise: Becoming a Successful Scholar in a Complex and Competitive Environment*. The MIT Press.

Jonathan Dunn. 2020. Mapping languages: the corpus of global language use. *Language Resources and Evaluation*, 54(4):999–1018.

Salomon Kabongo Kabenamualu, Vukosi Marivate, and Herman Kamper. 2022. LiSTra automatic speech translation: English to Lingala case study. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 63–67, Marseille, France. European Language Resources Association.

Frances Adriana Laureano De Leon, Harish Tayyar Madabushi, and Mark Lee. 2024. Code-mixed probes show how pre-trained models generalise on code-switched text. In *Proceedings of the 2024 Joint*

International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 3457–3468, Torino, Italia. ELRA and ICCL.

Lori Levin. 2009. Adaptable, community-controlled, language technologies for language maintenance. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation*, Barcelona, Spain. European Association for Machine Translation.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada. Association for Computational Linguistics.

Masakhane. Our mission.

Idriss Nguepi Nguefack, Mara Finkelstein, and Toadoum Sari Sakayo. 2025. Pretraining strategies using monolingual and parallel data for low-resource machine translation. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 31–38, Vienna, Austria. Association for Computational Linguistics.

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.

Eddie Antonio Santos and Atticus Harrigan. 2020. Design and evaluation of a smartphone keyboard for Plains Cree syllabics. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 88–96, Marseille, France. European Language Resources association.

Ishan Tarunesh, Syamantak Kumar, and Preethi Jyothi. 2021. From machine translation to code-switching: Generating high-quality code-switched text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3154–3169, Online. Association for Computational Linguistics.

Monica Ward, Liang Xu, and Elaine Uí Dhonnchadha. 2022. How NLP can strengthen digital game based language learning resources for less resourced languages. In *Proceedings of the 9th Workshop on Games and Natural Language Processing within the 13th Language Resources and Evaluation Conference*, pages 40–48, Marseille, France. European Language Resources Association.

Jitao Xu and François Yvon. 2021. Can you traducir this? machine translation for code-switched input. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 84–94, Online. Association for Computational Linguistics.