NAMED ENTITY RECOGNITION FOR AFRICAN LANGUAGES: A FOCUS ON IGBO LANGUAGE

Chiamaka Ijeoma Chukwuneke, BSc (Hons), MSc
School of Computing and Communications
Lancaster University

Supervisors:
Professor Paul Rayson
Dr Ignatius Ezeani
Dr Mahmoud El-Haj

A thesis submitted for the degree of
Doctor of Philosophy

June, 2025

Declaration

I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university. This thesis does not exceed the maximum permitted word length of 80,000 words including appendices and footnotes, but excluding the bibliography. A rough estimate of the word count is: 26,248

Chiamaka Ijeoma Chukwuneke

Named Entity Recognition for African Languages: a focus on Igbo

Chiamaka Chukwuneke B.Sc. (Hons), MSc.
School of Computing and Communications, Lancaster University

A thesis submitted for the degree of Doctor of Philosophy. December, 2024.

Abstract

Named Entity Recognition (NER) is a crucial task for many downstream NLP applications, including text summarization, document indexing, question answering, classification, and machine translation. Analysis of research reveals that 95% NLP efforts are concentrated on English and a few other languages like Japanese, German, and French, even though there are over 7,000 languages globally. Around 90% of African languages are considered under-resourced in NLP highlighting the gap in resources for African languages

The work presented in this thesis significantly advances Named Entity Recognition (NER) for low-resource languages, particularly African languages like Igbo, which, despite having millions of speakers, has remained largely underrepresented in NLP research. Focusing on Igbo, this research addresses a critical gap where foundational tools and resources, such as IgboNER, have been unavailable, thus limiting the language's integration into broader computational applications. Prior to this work, the Igbo language lacked dedicated NER resources and a specialised language model essential for accurate information extraction and analysis, which has kept Igbo on the periphery of digital advancements in NLP.

To address this gap, we developed IgboBERT, the first transformer-based language model pre-trained from scratch on the Igbo language, to serve as a baseline model. We created a parallel English-Igbo corpus and utilized spaCy, an existing NER tool for the high-resource English language, to tag the English sentences. These tags were then transferred to Igbo using a projection method, aided by our semi-automatically created mapping dictionary to facilitate the tag transfer process. Additionally, we designed a framework for the creation of the IgboNER dataset, which can be extended to other low-resource languages.

We fine-tuned IgboBERT and several state-of-the-art models, including mBERT, XLM-R, and DistilBERT, for the downstream IgboNER task using transfer learning. Our evaluation across various data sizes indicated that while large transformer models significantly benefited the IgboNER task, fine-tuning a transformer model built from scratch with relatively little Igbo text data also produced commendable results. This work substantially contributes to IgboNLP and the broader African and low-resource NLP landscape.

Publications

The following publications have significantly influenced and shaped the development of this thesis:

- **Chukwuneke, C. I.**, Rayson, P., Ezeani, I., El-Haj, M., Asogwa, D. C., Okpalla, C. L., & Mbonu, C. E. (2023).
  IgboNER 2.0: Expanding Named Entity Recognition Datasets via Projection.
  In 4th Workshop on African Natural Language Processing.
  `https://openreview.net/forum?id=tHUS9-vmUfC`
  I contributed to all aspects of this work, including data collection, analysis, and paper writing. The details of this work will be discussed in Chapter 7

- **Chukwuneke, C. I.**, Ezeani, I., Rayson, P., & El-Haj, M. (2022).
  IgboBERT Models: Building and Training Transformer Models for the Igbo Language. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 5114-5122).
  `https://aclanthology.org/2022.lrec-1.547/`
  I contributed to all aspects of this work, including data collection, analysis, and paper writing. The details of this work will be discussed in Chapter 5

- Adelani, D. I., Neubig, G., Ruder, S., Ezeani, I., **Chukwuneke, C. I.**, … Klakow, D., Gwadabe, T., ... & Adewumi, T.(2022)
  MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 4488–4508, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
  `https://aclanthology.org/2022.emnlp-main.298/`.
  I contributed to the dataset annotation for the Igbo language. The details of this work will be discussed in Chapter 6

- Adelani, D. I., Abbott, J., Neubig, G.,Ezeani, I., Rayson, P., **Chukwuneke, C. I.**, Kreutzer, J., Lignos, C., ... & Osei, S. (2021).
  MasakhaNER: Named entity recognition for African languages. Transactions of the Association for Computational Linguistics, 9, 1116-1131.
  `https://aclanthology.org/2021.tacl-1.66/`
  I contributed to the dataset annotation for the Igbo language. The details of this work will be discussed in Chapter 4

Publications Beyond this Thesis

Listed here are other research publications I collaborated on during this PhD and are not included in this thesis. All are contributions to African NLP and NLP at large and the the topics are focused on Question answering, Stopwords, COMET, information retrieval, and Summarization. The summary of this contribution is in Chapter 8.

- Odunayo, O., Gwadabe, T. R., Rivera, C. E., Clark, J. H., Ruder, S., **Chukwuneke C.I.**, Adelani, D. I., ...& Dossou, B. FP. (2023)
Cross-lingual Open-Retrieval Question Answering for African Languages.
`https://aclanthology.org/2023.findings-emnlp.997/`. I contributed to this work by assisting in the paper writing and serving as one of the Igbo language annotators

- Tinner F.,Adelani D. I., Emezue, Chris, **Chukwuneke C.I.**, ...& Mbonu, C. E., Aziza, M., Müge K., Duygu, A. (2023)
Findings of the 1st Shared Task on Multi-lingual Multi-task Information Retrieval at MRL 2023
`https://aclanthology.org/2023.mrl-1.24` I contributed to both subtasks by creating questions, labeling the answers, and annotating named entities for the Igbo language.

- Wang, J., Adelani, D. I., Agrawal, A., Rei, R., Briakou, E., **Chukwuneke, C. I.**, Obiefuna, N., Ogbu, O. R., Mbonu, C. E., ...& Masiak, M. (2023)
AfriMTE and AfriCOMET: Empowering COMET to Embrace Under-resourced African Languages.
`https://arxiv.org/abs/2311.09828` I contributed to the dataset annotation for the Igbo language.

- Adelani, D. I., , Masiak M., Azime, I. A., Alabi, J., **Chukwuneke, C. I.**, Tonja, A. L.,...& Mwase, C.(2023)
MasakhaNEWS: News Topic Classification for African languages.
`https://aclanthology.org/2023.ijcnlp-main.10/`I contributed to the dataset annotation for the Igbo language.

- Mbonu, C. E., **Chukwuneke, C.I.**, Paul, R.U., Ezeani, I., Onyenwe, I. (2021)
IGBOSUM1500 - INTRODUCING THE IGBO TEXT SUMMARIZATION DATASET.
`https://openreview.net/forum?id=rMUccG4LZq` I assisted with the paper writing.

Acknowledgements

I am grateful to God Almighty, whose grace was sufficient for me throughout this PhD journey. To all the people He used to make this a reality in some way or another, I want to use this opportunity to appreciate you, though this space is not enough to name you all. Please, know that I am deeply grateful.

Professor Paul Rayson, my supervisor, you are one in a million. I am so grateful for your ideas, suggestions, guidance, concern and support during this journey. Yes, we came through despite the challenges.

To the other members of my supervisory team: Dr. Mo El-Haj, thank you for your insightful feedback; and Dr Ignatius Ezeani, thank you for your constant encouragement. You always got my back when I felt overwhelmed, calmly urging me to keep going. I also thank my examiners, Professor Eric Atwell (external) and Dr. Alistair Baron (internal) for their constructive and professional approach, which made the viva a positive and enriching experience.

I am grateful to Professor Ikechukwu Onyenwe, who introduced me to IgboNLP. My immense gratitude to the Masakhane community, for giving me the opportunity to carry out the first NLP task at the inception of this PhD. To my friends and colleagues in the Lancaster University NLP group and UCREL, thank you for your support.

To my husband *(di dọkịta)*, my hero and support system, thank you for your patience and understanding. I would not have embarked on this journey without your unwavering support. To the amazing gifts from God, our children: Ikechukwu, Chizitelu, and Ifechukwu, thank you for the many hours you allowed me to spend working on my computer, even when it meant not giving you the full attention you deserved. Obinna, my amazing brother and Dr. Sam Amaefule, you will never lack help. I will always be grateful to my parents, Nwora and Adaora, for your prayers and encouraging words, may God continue to bless and reward you abundantly.

Finally, I am grateful to Nnamdi Azikiwe University, Nigeria, for making this PhD journey a reality.

Contents

List of Figures

List of Tables

Section A
Introduction and Related Works

Chapter 1

Introduction

1.1 Introduction

Natural Language Processing (NLP) has significantly influenced technology development by transforming human-machine interaction and enhancing machine understanding of human language (Jurafsky, 2020). With the evolution of NLP methods and algorithms, intelligent and smart devices capable of comprehending, interpreting, and generating natural language have emerged. The outcomes of NLP research have led to the development of various tools widely used today. For instance, *Google Translate*[1] and *DeepL*[2] facilitate the translation of texts from one language to another. *ChatGPT*[3] is significantly better than previous NLP approaches to achieve some level of machine understanding of natural language instructions and providing detailed responses, and *displaCy*[4] allow users to highlight named entities and their labels in a text directly within a web browser. Virtual assistants like *Amazon Alexa*, *Google Assistant*, *Apple's Siri* can comprehend voice commands in natural language and execute tasks as instructed by the user. These tools showcase the practical applications of NLP research in enhancing communication, information retrieval, and task automation. These advancements have reshaped various sectors, including customer service, healthcare, hospitality, finance, education, and more (C. Park, Jeong, and J. Kim, 2023). This shift is possible through the availability of NLP applications and tools. For example, tools like Alexa and Siri can identify and categorize specific types of information within spoken language, such as names, locations, and dates, with the help of one key component called named entity recognition (NER). NER, a fundamental task in NLP, involves identifying and classifying named entities like individuals, organizations, locations, dates, and others in unstructured text data. Its significance in NLP lies in serving as a bridge between unstructured text and structured data, enabling machines to extract organized information from text

[1] https://translate.google.co.uk/
[2] https://www.deepl.com/translator
[3] https://chat.openai.com/
[4] https://demos.explosion.ai/displacy-ent

for advanced information retrieval, knowledge extraction, and text comprehension (Pakhale, 2023). NER indeed plays a pivotal role in harnessing the capabilities of NLP technologies and fostering innovation across various industries. In healthcare, for instance, NER facilitates the extraction of vital medical information from patient records, aiding healthcare providers in delivering more personalized and efficient care. Additionally, it enables resume filtering by identifying specific skill sets as entities, streamlining the recruitment process for employers. In virtual assistants and chatbots used for customer support, NER helps identify and understand the type of requests made by users, enhancing the overall user experience. Moreover, NER assists in establishing relationships between textual data and entities, as observed in search engines, thereby improving search relevance and accuracy. Furthermore, it plays a crucial role in creating summaries of articles, research papers, and blogs, enabling users to grasp the key points of lengthy texts quickly. As NLP techniques continue to advance, NER tasks evolve and expand across diverse applications, further enhancing the efficiency and effectiveness of NLP technologies.

Unfortunately, 95% of NLP research is focused on English and a few other languages like Japanese, German, and French out of above 7,000 languages (David M. Eberhard, Gary F. Simons, and (eds.), 2023) spoken in the world. For example, the only known public tool that supports African languages is Google Translate[5] and some of the languages are Igbo, Yoruba, Hausa, Somali, Zulu. This shows that if not included in the research, many languages will continue to be digitally disadvantaged and left behind as technology advances which can lead to language extinction. Hence, there exists a significant imperative for NLP research across languages worldwide. This thesis is motivated by the goal of advancing NER in African languages, with a particular emphasis on the Igbo language. This work is a contribution to the community of Igbo NLP researchers (IgboNLP) in specific and to the wider community of African NLP (AfricaNLP) researchers at large. Igbo is an institutional[6] language, the fourth most spoken as a first language (L1) by a substantial population estimated to be 12.9% [7] of the Nigerian population of 227,120,344 [8]. It is also an official minority language in Equatorial Guinea and Cameroon in West Africa. Despite the number of speakers, Igbo is digitally disadvantaged because it lacks available tools and resources for performing a wide variety of natural language computer interaction and NLP tasks.

[5]https://africa.googleblog.com/2013/12/google-translate-now-in-80-languages.html

[6]A language utilized in offices and workplaces, educational institutions, mass media, and government administration.

[7]https://www.statista.com/statistics/1268798/main-languages-spoken-at-home-in-nigeria/

[8]https://www.worldometers.info/world-population/nigeria-population/

1.2 Motivation

In an era where technology is reshaping our world, the availability of NLP resources and tools plays a pivotal role in revolutionizing how we interact with computers and electronic devices. Beyond this transformative shift, a pressing need exists to bridge the digital divide by ensuring equitable access to these tools across a wide range of languages. Further motivation will be discussed under the following subsections- "Why African NLP?", "Why Igbo?" and "Why IgboNER".

1.2.1 Why African NLP?

NLP is important for African languages to enhance easy accessibility of the speakers with their language online promoting inclusivity in technology. Cultural heritage will also be preserved and maintained for future generations. At the commencement of this doctoral journey, I became a member of the Masakhane Community[9]. The goal of Masakhane is "for Africans to shape and own these technological advances towards human dignity, well-being, and equity, through inclusive community building, open participatory research, and multidisciplinary". Through this involvement, I gained profound insights into the stark underrepresentation of African languages in technological spheres and was motivated to actively participate in collaborative endeavors to advance this community's goal.

Table 1.1 gives a statistics overview of author affiliations at five major conferences in 2018 (ACL, NAACL, EMNLP, COLING, and CoNLL) by Caines (2019) revealed a notable absence of African representation in the dataset, indicating a gap in geographic diversity within NLP. In 2023, the Proceedings of the 61st Conference of the Association for Computational Linguistics (ACL 2023)[10] documented a record-breaking 4864 submissions from 13,658 authors, which 4490 reviewers reviewed. Figure 1.1 illustrates the comprehensive distribution of affiliations among both authors and reviewers, as indicated in their START profiles. We cannot overemphasize that Africa, the fastest-growing continent with a growth rate of 2.55%[11], the third most densely populated continent in the world with a population of 1.48 billion, representing 17.89% of the world's population[12] is geographically underrepresented in NLP research as seen in Table 1.1 and Figure 1.1. This answers why research in African NLP is a key focus of this thesis.

[9] a grassroots organization whose mission is to strengthen and spur NLP research in African languages, for Africans, by Africans

[10] https://aclanthology.org/2023.acl-long.0.pdf

[11] https://www.worldatlas.com/articles/continents-by-population-density.html

[12] https://www.worldometers.info/world-population/africa-population/

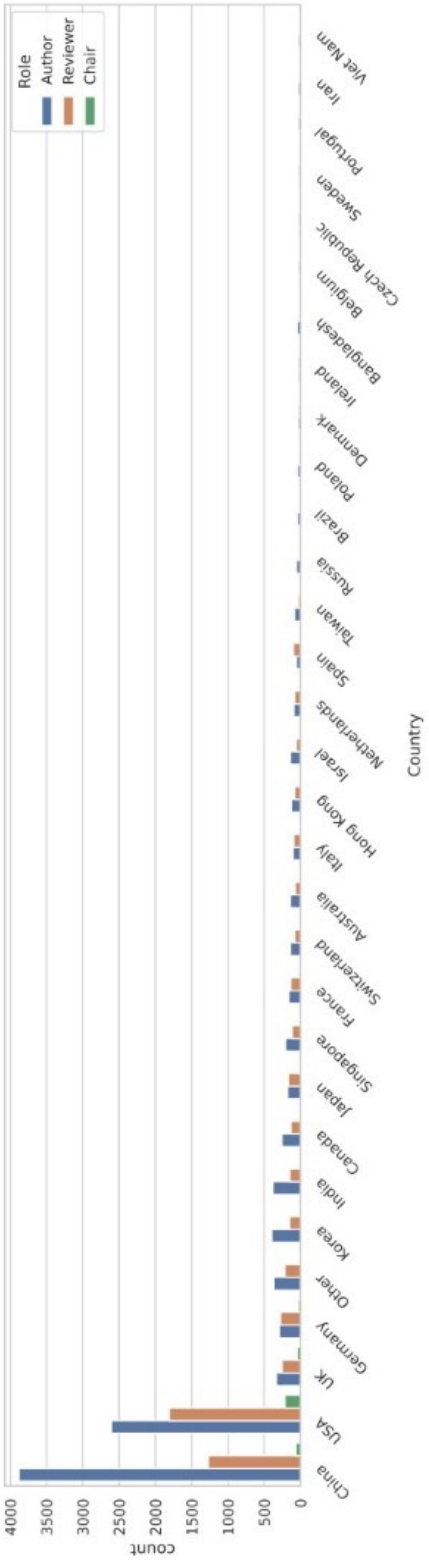


Figure 1.1: The ACL author and reviewer profiles include a listing of the top 30 countries.

Region	No. of author affiliations
North America	1114
Asia	826
Europe	641
Middle East	54
Oceania	44
South America	11
Africa	5

Table 1.1: Count of author affiliations by region for NLP Conferences in 2018.

1.2.2 Why Igbo?

NLP primarily centers around English, resulting in a lack of multilingual perspectives and inadequate representation for most languages globally. Sebastian Ruder (2020) in Why You Should Do NLP Beyond English presented arguments advocating for exploring languages beyond English, articulating his reasons across societal, linguistic, machine learning, cultural and normative, and cognitive perspectives. To be part of the development of African NLP, Igbo, which is the native language of the author of this thesis was considered to be the focus language in this research. The African language, Igbo, is spoken in the following seven states of Nigeria: Abia, Anambra, Ebonyi, Enugu, and Imo, as well as parts of Delta and Rivers states(Nwagu, 2023). This ethnic group, Igbo, is culturally rich and renowned for its entrepreneurial spirit, with business deeply embedded in its cultural fabric. Igbo people are found across Nigeria and in various parts of the world. They stand out as the most intellectually and commercially adept tribe in Nigeria (Nwagu, 2023). Two of the largest international markets in Nigeria: Onitsha Main Market and Ariaria International Market are located in the eastern part of Nigeria [13]. This brings in a lot of people from various parts of the world. Their ability to thrive in business and academia, often making the most of limited resources, sets them apart (Nwagu, 2023). Additionally, the cultural heritage of Igbo land which is a tourist attraction will be preserved thus presenting opportunities in the tourism and hospitality sector which will bring people from other parts of the world. With all these, Igbo should not be left out in the NLP world and global village at large.

1.2.3 Why IgboNER?

According to Ikechukwu Ekene Onyenwe, 2017, the Igbo language is one of the African languages with zero available NLP tools as of May 2013 when he started his research and if there were any, they are not easily found online. At the inception of

[13]https://nigerianinformer.com/largest-markets-in-nigeria/

this PhD journey in 2019, a survey of works on IgboNLP resources was conducted to assess existing tools and identify gaps to address. Below are the existing works found from the study:

1. Uchechukwu (2005) focuses on the technological means of representing Igbo language by having an appropriate computer keyboard.

2. Ikechukwu E Onyenwe, Uchechukwu, and Hepple (2014) incorporated additional language internal features to create a new part-of-speech tagging scheme based on the EAGLES tagset guidelines for annotation tasks for developing a POS-tagged Igbo corpus.

3. Ikechukwu E Onyenwe, Hepple, Chinedu, et al. (2018) discussed the method taken to create the following initial set of resources for Igbo: an electronic text corpus, a part-of-speech (POS) tagset, and a POS-tagged subcorpus, problems and the solutions in the process.

4. Ezeani, Hepple, and I. Onyenwe (2016) explored various word-level diacritic restoration techniques, primarily based on n-grams, to restore diacritics in Igbo texts. This investigation utilized an Igbo bible corpus as the primary dataset.

5. Ikechukwu E Onyenwe, Hepple, Chinedu, et al. (2019) describes the POS tagging experiments using the Igbo POS corpus as a benchmark and also identified the best-performing Machine Learning(ML) method for the limited Igbo POS dataset.

6. I. Onyenwe et al. (2015) presented a novel way to improve a part-of-speech (POS) tagged corpus for the African language Igbo in a semi-automated manner using transformation-based learning (TBL) to identify candidates for correction and to propose possible tag corrections. I. E. M. H. I. Onyenwe and Enemuo (2018) used various existing English embeddings to create Igbo word embeddings using transfer learning. Ezeani, Hepple, and I. Onyenwe (2017) provides a more standardized method for Diacritic restoration in Igbo language using machine learning algorithms.

The study not only uncovered that little has been done in IgboNLP but also a gap in the text-processing steps of the NLP pipeline for enhancing accuracy and efficiency by organising data through entity recognition and classification. This gap is the absence of a Named Entity Recognition (NER) system for the Igbo language. Given the pivotal role of NER in identifying and categorizing named entities within text, addressing this gap becomes imperative for enhancing Igbo language processing capabilities across a wide range of applications.
Furthermore, the challenges inherent within the Igbo language listed below emphasize the necessity for the creation of linguistic resources, such as a NER system tailored specifically to Igbo. Chapter 5 section 5.4 describes the creation of

a mapping dictionary which helps to address the challenge of orthography variation and multi-word expression.

1. **Orthographic Variation**:
   The lack of standardized orthography in Igbo language usage often leads to the combination of various orthographic conventions when writing certain Igbo words in texts. This variance in orthography can contribute to an increase in unseen entities and out-of-vocabulary words for NER systems. For example, the word "Lagos"- a geographical location in Nigeria can be spelled differently as *"Legọs"* or *"Legos"* or *"Lagos"*

2. **Multi-word Expressions**: Multi-word expressions in the Igbo language refer to phrases or combinations of words that convey a specific meaning as a unit, rather than as individual words. These expressions may include idiomatic phrases, compound nouns, time expressions, and more. Some examples of multi-word expressions in Igbo include:

   (a) "*Elekere anọ nke ehihie*" - Four o'clock in the afternoon.

   (b) "*Ụlọ akwụkwọ*" - School.

   (c) "*Ndị* uwe ojii" - Police.

   These multi-word expressions reflect the richness and complexity of the Igbo language, encompassing various aspects of life, culture, and experience. They highlight the importance of considering context and semantics when processing Igbo text data for natural language processing tasks. Therefore, there is a need to design a NER system for Igbo to account for these multi-word expressions to accurately identify and classify named entities in text data.

3. **Diacritics**: Diacritics in the Igbo language play a crucial role in disambiguating words spelled similarly but with different meanings and pronunciations. Speakers and writers can distinguish between homographs and convey the intended meaning more accurately by adding diacritics to certain letters in a word. Here are some examples:

   (a) "Èkè" - "Market day" or "Éké" - "Python": In this example, without the diacritic, the word "Eke" would be ambiguous.

By developing a NER system trained on a corpus written with various Igbo orthography, the effectiveness and accuracy of various NLP applications such as machine translation, summarization, and question-answering systems will significantly improve ensuring entities in Igbo text are correctly identified and processed. Hence, improving Igbo language technology and facilitating its advancement.

1.3 Research Questions

Natural Language Processing (NLP) has tremendously advanced with the persistent improvement of methods and models used in the field. NLP has transitioned from rule-based systems, which depended on linguistic rules that were manually created by experts, to statistical models that leverage large amounts of data. Recently, deep learning and neural network architectures have become state-of-the-art (SOTA) and are known to require much larger amounts of data to learn effectively, which is a challenge for under-resourced languages that lack annotated corpora, dictionaries, and linguistic tools. This research answers the following research questions (RQ) for a digitally disadvantaged language, Igbo.

RQ1 *What efficient methods can be used to create NER annotated datasets for African languages e.g. Igbo and other digitally disadvantaged languages?*

RQ2 *How can we leverage existing high-resource language models in NER to support the advancement of African languages, using Igbo as a case study?*

RQ3 *How might the approaches used in Igbo NER be adapted and applied to other languages?*

RQ4 *What effect does the size of the typically small datasets have on large NER models?*

RQ5 *How can the resources developed for Igbo NER, such as datasets and annotation methodologies, be shared or modified for use in other African language NLP tasks?*

1.4 Thesis Contributions

The need to create NLP resources for disadvantaged languages cannot be overstated. Therefore, at the end of this thesis, our goal is to contribute the following to support African NLP and the entire NLP community at large.

1. The development of the first transformer-based language model, IgboBERT[14], trained on Igbo language data from scratch. This approach ensured that the model was tailored specifically to the linguistic features and characteristics of Igbo. By training from scratch, IgboBERT can effectively learn the intricate linguistic patterns and structures unique to the Igbo language, enhancing the model's ability to comprehend and generate text in Igbo with about 97% accuracy, as measured by the percentage of exactly matched entities.

[14]https://huggingface.co/chymaks/IgboBERT-NER-finetuned-Final-Version

2. The creation of the IgboNER dataset[15] using the projection technique. This technique is used to expand the dataset creation for Igbo and can be extended to other languages.

3. The creation of a mapping dictionary for IgboNER. The mapping dictionary will contain a list of English entities, their Igbo translations, and their tags. This would enhance NLP systems' accuracy, efficiency, and interpretability across various applications and domains.

4. The development of a framework for the creation of NER resources for different languages.

5. The creation of IgboNER visualisation tool[16] to aid users in comprehending and analyzing named entity recognition results in Igbo text.

6. Collaborated in developing the largest human-annotated NER dataset for African languages (David Ifeoluwa Adelani, J. Abbott, et al. (2021) and Adelani et al. (2022)), demonstrating the significance of selecting the optimal transfer language for diverse African linguistic groups.

While IgboBERT is designed for the Igbo language, its underlying transformer-based architecture and pre-training methodology can be adapted beyond Igbo, contributing to advancements in multilingual NLP research and the development of NLP tools for resource-limited languages. Like Igbo, many languages have limited linguistic resources and NLP tools. The development and effectiveness of IgboBERT will inspire similar initiatives for other resource-limited languages. Researchers and practitioners can leverage similar approaches to develop language-specific models tailored to their respective languages and extend the same technique to create dataset for their languages. Chapter 4 outlines the development process and design of IgboBERT.

1.5 Structure of Thesis

The thesis is structured as follows;

Section A - Introduction and Related Works

Chapter 1 Introduction which introduces the research, motivations, objectives, research questions, and contributions of the work.
Chapter 2 Related Work presents a summary of Igbo people, their orthography, and IgboNLP. We also review research on named entity recognition (NER), uses, approaches to NER (methods), NER models and Datasets, digitally disadvantaged

[15] https://github.com/Chiamakac/IgboNER-Models/tree/main/Igbo$_{d}ataset$
[16] https://igbo-demo.streamlit.app/

NER scenarios, NER Evaluation Metrics, and IgboNLP.

Section B- NER for African Languages

Chapter 3 A Framework for Named Entity Recognition describes the systematic approach and methodologies used for the creation of NER tools and models in this work. This involves various phases, from data collection and annotation to model evaluation. The approach employed in this work can be adapted to a wide range of other languages.

Section C- IgboNER

Chapter 4 Transformer Models for the Igbo Language presents a standard Igbo named entity recognition (IgboNER) dataset and the IgboBERT language model, which was pretrained from scratch. Additionally, it includes the fine-tuning of IgboBERT and other state-of-the-art transformer models, which were pre-trained on non-Igbo languages, for the downstream IgboNER task.
Chapter 5: Expanding Named Entity Recognition Datasets Via Projection details the generation of additional IgboNER datasets by leveraging an existing English Named Entity Recognition (NER) tool. The process involves the application of a cross-language projection technique to semi-automatically create a mapping dictionary from a parallel English-Igbo corpus.
Chapter 6: Named Entity Recognition for African languages outlines an effort to alleviate the scarcity of representation for the African continent in NLP research, this initiative focuses on developing the first extensive, publicly accessible, and high-quality dataset tailored for named entity recognition (NER) across ten African languages. Igbo language is one of the languages and I contributed to the annotation task for Igbo.
Chapter 7: Africa-Centric Transfer Learning for Named Entity Recognition presents the creation of a named entity recognition (NER) dataset encompassing 20 diverse African languages, this work provides strong baseline outcomes through fine-tuning of multilingual pre-trained language models on both in-language NER and multilingual datasets. Additionally, we investigate cross-lingual transfer within an Africa-centric context, showing the significance of selecting the optimal transfer language in both zero-shot and few-shot scenarios.

Section D- Contributions beyond NER and Conclusion

Chapter 8: Conclusion provides a summary of the thesis and outlines directions for future work. Also, gives a summary of contributions beyond NER in the course of this work.

Chapter 2

Related Work

2.1 Overview of Igbo language

Igbo is among the 10 most spoken native languages in Africa. It is one of Nigeria's three (3) major official languages and an official minority language in Equatorial Guinea and Cameroon in West Africa. The users of Igbo as their primary language are estimated to be around 31 million people(David M Eberhard, Gary F Simons, and Fennig, 2024). Igbo language belongs to the Benue-Congo group of the Niger-Congo family. It is the native language of the ethnic group found in Southeastern Nigeria known as Igboland and called Igbo people. Surrounded by various closely related Niger-Congo languages, this linguistic environment includes Edoid to the west, Defoid spanning the west and northwest, Idomoid in the north, Lower Cross in the east and south, and Ijo in the southern regions. Igboland consists of the states of Abia, Anambra, Ebonyi, Enugu, and Imo. Igbo is also spoken in the northeast of the Delta state and the southeast of the Rivers state, Nigeria. Igbo is written in Latin script and has over 30 dialects. It is an agglutinative language, a single stem can yield many word forms by the addition of affixes that extend its original meaning (Ikechukwu E Onyenwe and Hepple, 2016). Igbo is a tonal language and is written with diacritics.

2.1.1 Writing System

2.1.1.1 NSIBIDI

Before the colonization of Africa, Nsibidi alternatively referred to as nsibiri or nchibiddi was the native name for the system of writing used in the southern region of Nigeria. The writing system belongs to an exclusive secret society known as the Nsibidi society, where men regularly undergo an initiation process after a preparatory phase (Macgregor, 1909). This constitutes a form of picture-writing with a significant history, as some of the signs already exhibit a degree of what is considered acceptable by society in general.

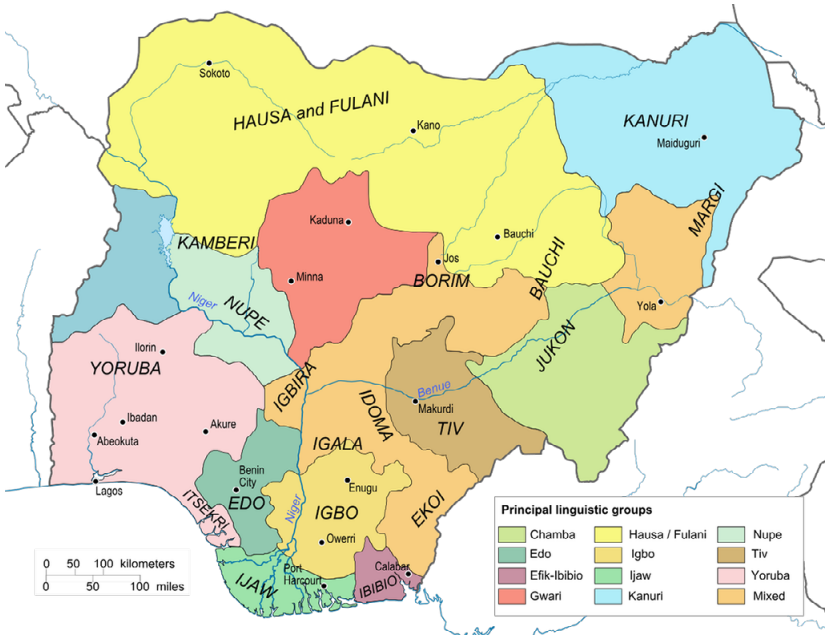


Figure 2.1: Map of Nigeria showing the location of Igboland. [1]

.

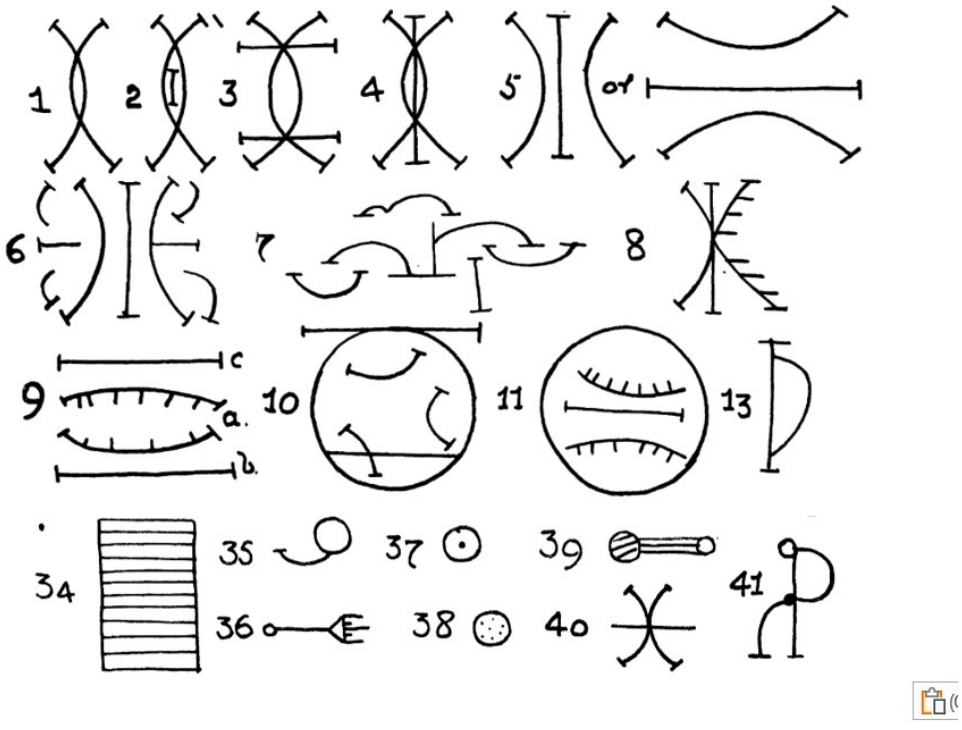


Figure 2.2: Nsibidi signs (Macgregor, 1909). Some of the Nsibidi writing system symbols. The descriptions are below.

- **Symbol 1 and 2**: Married love (2, with pillow).
- **Symbol 3**: Married love with pillows for head and feet- a sign of wealth.
- **Symbol 4**: Married love with pillow.
- **Symbol 5:** Quarrel between husband and wife. This is indicated by the pillow being between them.
- **Symbol 6**: Violent quarrel between husband and wife.
- **Symbol 7**: One who causes a disturbance between husband and wife.
- **Symbol 8**: A woman with six children and her husband; a pillow is between them.
- **Symbol 9**: Two wives with their children (a), of one man (b), with the roof-tree of the house in which they live (c). The tree is put for the whole house.
- **Symbol 10**: A house (a) in which are three women and a man. The dots have no meaning.
- **Symbol 11**: Two women with many children in the house with their husband.
- **Symbol 1**: A woman with child.
- **Symbol 34**: A native mat, used as a bed.
- **Symbol 35**: A gourd for a drinking cup.
- **Symbol 36**: Native comb.
- **Symbol 37**: Toilet soap.
- **Symbol 38**: Basin and water.
- **Symbol 39**: Calabash with 400 chittims inside it. A chittim is a copper wire worth one-twentieth of a rod. Such calabashes have hinges of three strings.
- **Symbol 40**: Slaves.
- **Symbol 41**: Fire.

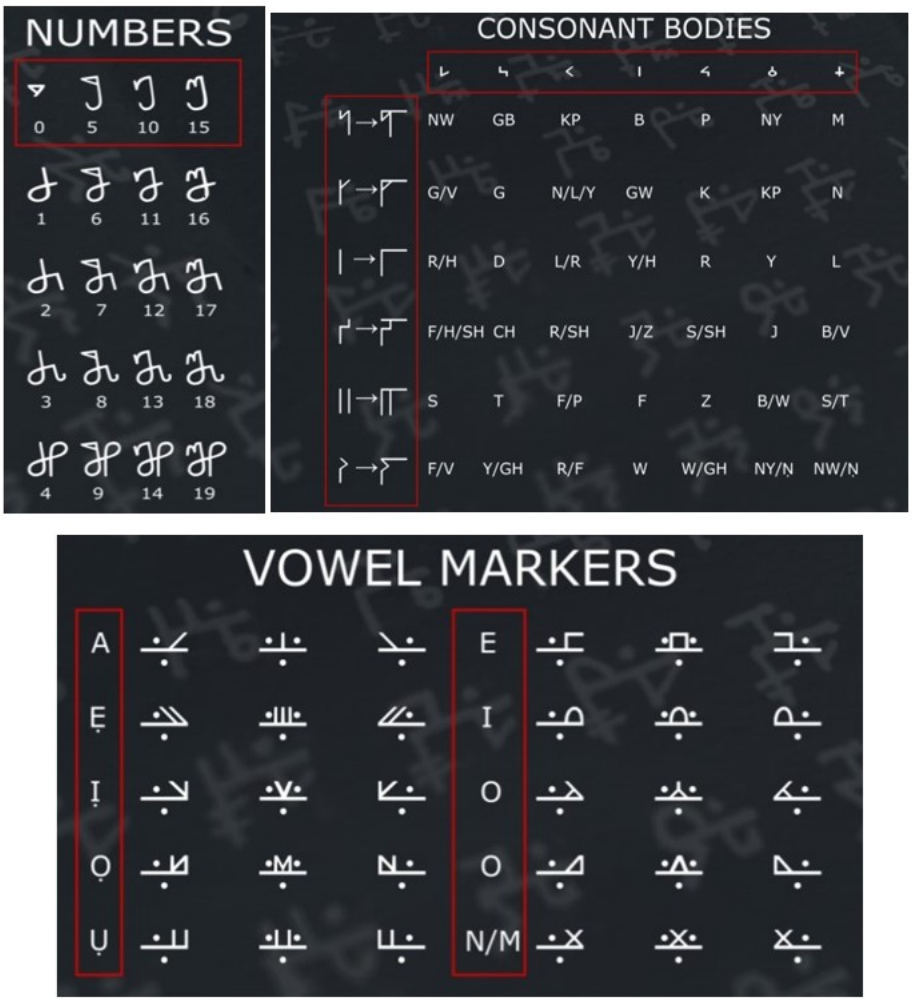


Figure 2.3: Ńdébé number system, consonants, and vowel markers[3]

2.1.1.2 NDÉBÉ

The Ńdébé[2] script is a modern writing system that merges ancient Ìgbò designs with contemporary practicality. It was invented by Lotanna Igwe-Odunze in 2009. Ńdébé uses pre-combined characters to represent all possible Ìgbò sounds as syllables. The script comprises 6 stems, 7 radicals, and 27 diacritic vowel markers. Each syllable is depicted by a character composed of three parts: stem, radical, and diacritic. The Ndebe number system operates on a base-20 (sub-base 5) and features numeral symbols from 0 to 19. The complete syllabary includes numbers, symbols, punctuation, and all syllables categorized by vowel and tone. Figure 2.3 shows Ńdébé numbers, consonants, and vowel markers.

[2]https://ndebe.org/

2.1.1.3 Igbo orthography

The colonial era drastically reduced this oldest form of African writing system, Nsibidi. Historically, a lot of controversy was associated with the adoption of an orthography for the Igbo language. This lasted about three decades (1929-1961) and greatly affected the language, resulting in Igbo not having a long written tradition compared to languages like Arabic, English, and French (Agbo, 2013). The first Igbo orthography, *Lepsuis* was named after the German philologist Karl Richard Lepsuis who published the first Standard Alphabet for African languages in the 1854 (Oraka, 1983). The Lepsuis Standard Alphabet has 34 alphabets consisting of the following:

- 6 vowels - *a, e, i, o, u, ọ* ;
- 9 digraphs consonants - *gb, gh, gw, kp, kw, nw, ny, ds, ts*;
- 19 monographs consonants -*b, d, f, g, h, k, l, m, n, p, r, s, t, v, w, y, z, ṅ, s.*

It made use of diacritics. This orthography was adopted by the Church Mission Society (C.M.S) and has used it *Lepsuis* in all their Igbo publications including the most prominent of all: *Bible Nso*(Holy Bible). In 1861, a Christian missionary, J.F. Schon adopted the Lepsius orthography in his book *Oku Ibo: Grammatical Elements of the Ibo language* thus the earliest dated written form of Igbo language. In 1870, a catechist, F.W. Smart wrote a book *An Ibo primer* and was published by the Church Missionary Society (C.M.S). As recorded by Oraka (1983), approximately 50 books have been written and published using the Lepsius orthography by 1883.

In 1927, *Practical Orthography of African Languages* (*Africa* Orthography) was introduced by the International Institute of African Languages and Cultures (IIALC). The *Africa* Orthography was adopted by the colonial administration of Eastern Nigeria in 1929 (James, 1928). The *Africa* orthography has 36 alphabets and is made away with the diacritical marks. The alphabet consists of the following:

- 8 vowels - *a, ɛ, i, e, o, ɔ , u, ɵ*
- 28 consonants - *b, c, d, f, g, gb, gh, h, j, k, kp, l, m, n, ŋ, ny, p, r, s, sh, t, v, w, y, z, gw, kw, nw* of which 9 are digraphs.

The Roman Catholic Missions also adopted this new orthography but the Church Mission Society has produced a lot of publications using *Lepsuis* rejected the *Africa* Orthography (Oraka, 1983). Therefore, the *Lepsuis* and *Africa* orthography existed together. The three decades of disagreement between the Church Mission Society and the Roman Catholic Mission who are the major stakeholders in education brought the development of Igbo language to a halt. No teaching of the Igbo language in schools, no publications in the language at

that time and Igbo was not part of the African languages written in the Senior Certificate Examinations introduced by Cambridge University in 1935. This led to the official standardization of Igbo orthography by the Ọnwụ committee in 1961. The Ọnwụ committee set up by the Eastern government of Nigeria published the official Igbo orthography known as *Ọnwụ* orthography in 1961. This orthography was named after Mr. S.E. Ọnwụ who headed the committee. The use of diacritics was re-introduced in this orthography and it has 36 characters. The alphabets consist of:

- 8 vowels - *a, e, o, ọ, u, ụ, i, ị*
- 28 consonants - *b, gb, ch, d, f, g, gh, gw, h, j, k, kw, kp, l, m, n, nw, ny, ṅ, p, r, s, sh, t, v, w, y, z* of which 9 are digraphs.

Agreement on a standardized orthography for the Igbo language remains difficult, resulting in the writing of Igbo texts with a combination of orthographies. Some location and person names are written with African orthography which is the effect of the post-colonial government of Nigeria. For example, *"Ọka-Etiti"* a town in Anambra state of Nigeria is most times written with African orthography as *"Awka-Etiti"*. The day of the week *"Thursday"* is written in texts in these forms *"Tọọzude"*, *"Thursday"*, *"Tọzde"*, *"Tọọzdee"* etc. These variations in writing Igbo words could pose a problem for IgboNER. Table 2.1 shows some Igbo words written in these different orthographies.

Lepsius	**Africa (Anglicized)**	**Ọnwụ**
Nàịjirịà	Naijiria	Nigeria
Ìgbò	Ibo	Igbo
Legọs	Lagos	Legos

Table 2.1: Igbo Orthography

2.1.1.4 Diacritics

"Igbo" is a tonal language that is written using diacritical marks to ensure accurate pronunciation and disambiguation of Igbo texts. It is quite fascinating how some words in Igbo have the same spelling but with different meanings and pronunciations, making it challenging to read and understand if not written with diacritics. Unfortunately, these diacritics are frequently missing from the electronic texts we aim to process and use for various tasks in NLP (Ezeani, Hepple, and I. Onyenwe, 2017). This is a challenge as the meaning of such words will be wrongly interpreted thereby affecting the performance and reliability of systems trained with them. Igbo diacritics can be tonal or orthographic. Marks denoting tonal variations are predominantly located above vowels, indicating whether they

are pronounced with a high (◌́), low (◌̀), or mid-tone (◌̄). The orthographic diacritics are the dots (.) placed under some alphabets. Table 2.2 displays the alphabets written with Ọnwụ orthography and their tonal diacritics. The absence

Alphabet	Diacritics
a	à á ā
e	è é ē
i	ì í ī
ị	ị̀ ị́ ị̄
o	ò ó ō
ọ	ọ̀ ọ́ ọ̄
u	ú ù ū
ụ	ụ̀ ụ́ ụ
m	m̀ ḿ m̄
n	ǹ ń n̄

Table 2.2: Igbo Diacritics

of diacritics significantly hinders language processing tasks such as NER. This is explained in the example sentence

1. *Ngozi gosiri m oke nke anyi. Akwa ya na-acha uhie uhie.*

2. *Ngozi gosiri m òkè nke anyi. Ákwà ya na-acha uhie uhie.*

Sentence number 1 can be wrongly read as seen below because of the absence of diacritics:

1. *Ngozi showed me our rat.*

2. *The egg is red.*

While the diacritics, it reads:

1. *Ngozi showed me our share.*

2. *The cloth is red.*

Sentence number 2 written with diacritics helps to disambiguate the sentences giving the correct meaning of words in Igbo. ambiguity in the Igbo language can result in confusion and wrong classification of entities. The work of Ezeani (2019) to restore diacritics in electronic Igbo texts is a positive step towards addressing this challenge.

2.1.2 IgboNLP

NLP research in Igbo (IgboNLP) started with the Ph.D. work of I.E. Onyenwe in 2013 (Ikechukwu Ekene Onyenwe, 2017). The work produced:

1. An automatic POS tagger.

2. EAGLES tagset guidelines were adapted to incorporate Igbo language and used to develop an annotation scheme (tagset) for Igbo, an automated approach utilizing morphological reconstruction to assign suitable tags to all morphologically inflected words that were incorrectly tagged in the corpus.

3. A POS-tagged Igbo corpus

4. A method that leveraged the morphological characteristics of Igbo to address the inadequate handling of unknown words by current taggers.

Other NLP resources created since the inception of IgboNLP include -

1. Diacritic Restoration by (Ezeani, Hepple, and I. Onyenwe, 2016).

2. Machine Translation (Nekoto et al., 2020),

3. MasakhaNER (David Ifeoluwa Adelani, J. Abbott, et al., 2021). This work focused on the creation of an NER dataset for 10 African languages which Igbo is one of the languages. The author of this thesis collaborated on this work by volunteering to be one of the Igbo annotators.

4. Nkowa Okwu[4] (Igbo-English dictionary)

Igbo language is termed "low-resourced" because of the availability of only a few NLP resources.

Named entity recognition (NER), is a subtask of NLP that recognizes entities that are present in a text and classifies them into predefined categories such as person names, dates, organizations, locations, time expressions, quantities, monetary values, Nationalities or religious or political groups, Geopolitical entity, product, event, work of art, language, percent, cardinal, ordinal, etc. NER plays a pivotal role in numerous domains and stands as a fundamental task that underlies the development of various NLP applications including information retrieval, Recommender systems, Robotic Process Automation (RPA), Resume Filtering, Electronic Health Record (EHR) Entity Recognition, etc.

[4]https://nkowaokwu.com/

2.1.3 NER Techniques

Approaches like rule-based approaches, learning-based approaches, and hybrid approaches have been applied to NER tasks since their inception (Goyal, Gupta, and Kumar, 2018) and also the deep-learning-based approach (Keraghel, Morbieu, and Nadif, 2024).

1. **Rule-based** - This involves the application of a set of rules conscientiously created by humans to solve problems which could be extraction of information, classification or finding specific structures, etc. These rules extract patterns and are considered efficient as experts who have domain knowledge and understand the syntactic and linguistic features of a domain create these rules. A rule-based system also applies a list of dictionaries. These rules can be easy to understand and interpret. They can be modified, or updated based on new information or changing requirements. However, its disadvantages are that it is expensive, domain-specific, non-portable, time-consuming, and cumbersome. Mengliev et al. (2023) designed two Named Entity Recognition (NER) algorithms, both utilizing rule-based mechanisms and gazetteers. The first algorithm predominantly depends on morphological analysis of word forms and proves highly effective when the gazetteer contains the relevant words. It accurately identified all geographic objects not present in the dictionary while identifying only 24% of those not contained in the dictionary. The second algorithm, incorporating both morphological and syntactic analyses, demonstrated substantial enhancements. It accurately identified 68% of geographic objects not present in the dictionary. Alfred et al. (2013) constructed rules based on the context of POS-tagging to determine the part-of-speech tag for a given word. If the word is identified as a proper noun, a specific rule is then applied to ascertain whether it qualifies as an entity. The study by K. F. Shaalan and Raza (2009) on Named Entity Recognition for Arabic (NERA) also embraced a rule-based methodology to address the distinctive features and intricacies of the Arabic language. This involves leveraging a Whitelist that functions as a dictionary of names, along with grammar expressed through regular expressions, responsible for identifying named entities. This framework incorporates a filtration mechanism to facilitate revision capabilities within the system.

2. **Learning-based** - This is the application of Machine learning (ML) algorithms. ML is a subset of Artificial Intelligence (AI) that equips computers with the capacity to learn from data, enabling self-improvement. Usually, a classifier is trained to acquire knowledge from data. Machine learning-based frameworks often provide more flexibility than rule-based techniques (Haq et al., 2023). It is grouped into these two types:

    - **Supervised Learning** - This is based on the use of labeled training data with the correct tag. The labeled data is used to

train an algorithm that learns from the labeled data. The trained algorithm is then used to predict or classify new data accurately based on past data. The target output is already known and the output of the algorithm is compared with the correct target output. The features or properties learned by the algorithm from the training data are very important as it is used to generate a model that recognizes and classifies data with similar patterns and relationships in unlabelled data (Goyal, Gupta, and Kumar, 2018). Supervised machine learning is effective in addressing a wide array of real-world computational challenges; however, it relies on a labeled dataset, which is a limitation. Pande, Kanna, Qureshi, et al. (2022) developed a Hidden Markov Model (HMM) integrated with named entities to calculate the ranking probability state of work entities for the categorization of variables in the network. The HMMNE model proposed attains a higher precision value of 99% for locations and 98% for names and organizations. In Azarine, Bijaksana, and Asror (2019) study, named entities such as Person, Location, and Organization in tweets were identified. In this research, each word in the previously labeled NER training data was annotated with POS tags. The labeled training data was then processed using the initial probability, emission probability, and transition probability to determine the optimal tag value through the Hidden Markov Model algorithm. The output of the Named Entity system is determined by the highest probability results. Their experimental findings suggest that the addition of POS tags is the most effective feature for NER modeling using the Hidden Markov Model, resulting in an increased F1 score of 3.65% and an overall F1 score of 64.06%. The MarathiNER system, known as Mner-CRF (N. Patil, A. Patil, and Pawar, 2020), employs a feature function that considers various parameters, including the sentence, current word position, and labels of the current and previous words. These parameters are utilized from the training dataset to predict the most suitable NE tag for each word in the sentence based on learned patterns. The system classifies words into twelve different NE classes, including person, location, organization, miscellaneous, amount, number, measure, date, time, weekday, month, and year. To train and test the system, a Marathi news text corpus from the FIRE 2010[5] dataset was used, comprising 27,177 sentences and 63,236 unique words, with manual annotations for 40 different tags. Using the Conditional Random Fields (CRF) algorithm, Mner-CRF achieved precision, recall, and F1-measure scores of 82.33%, 70.68%, and 75.51%, respectively.

[5]https://www.isical.ac.in/ fire/2010/

- **Unsupervised Learning** - The unsupervised machine learning algorithm functions without relying on labeled datasets. Instead, it independently examines data to uncover underlying hidden patterns and relationships within unlabeled datasets. Its goal is to organize unstructured data based on similarities, patterns, and differences without the need for prior training. Unsupervised learning is grouped into the clustering (Nadeau and Sekine, 2007) and association rules-based approach (Jain, D. Yadav, and Tayal, 2014). The study by Iovine et al. (2022) introduced CycleNER, an unsupervised NER method that uses cycle-consistency training to learn an effective mapping between sentences and entities through two cycles by training sentence-to-entity (S2E) and entity-to-sentence(E2S). CycleNER performed the NER task on a set of sentences without entity labels nor an independent set of entity examples. To enable unsupervised cycle-consistency training, the output of one function was used as the input for the other (e.g., S2E $\rightarrow$ E2S), aligning the representation spaces of both functions. Evaluation results showed that CycleNER achieves competitive performance compared to supervised approaches. CycleNER attains 73% of SOTA in CoNLL03 dataset. S. Zhang and Elhadad, 2013 utilized an unsupervised approach to extracting named entities from biomedical text. They presented a stepwise approach to address the challenges of entity boundary detection and entity type classification without the use of handcrafted rules, heuristics, or annotated data. A noun phrase chunker, followed by a filter based on inverse document frequency, is used to extract candidate entities from free text. The classification of these candidate entities into target categories is performed by applying principles of distributional semantics. Evaluations on on two popular biomedical datasets: i2b2 (clinical notes) and GENIA (biological literature) corpora shows that their system, especially the entity classification step, yields competitive results demonstrating the effectiveness and generalizability of their methods.

3. **Hybrid approaches** - This combines different NER techniques to leverage their strengths enhancing performance and adaptability beyond individual NER techniques. Bharathi et al. (2024) developed a NER system tailored for aviation entities by integrating rule-based and supervised methods. They generated data with silver labels [6] for seven entities related to aviation using RegEx combined with a pre-trained SpaCy (Honnibal and Montani, 2017) model. This silver-labeled data was utilized to train a custom SpaCy NER model. Through experimental testing with various hyperparameters and

[6] high-quality entity- annotated training data

features, they compared the performance of their model with a baseline pre-trained SpaCy model. Their model achieved an F1 score of 93% on both the validation and test sets, surpassing the baseline model's performance. The study by Haider et al. (2023) employed a combination of neural and heuristic methods to identify food or recipe names. They fine-tuned spaCy NER to detect food names and implemented a heuristic-based filtering method to boost precision in recognizing food entities for downstream tasks. To streamline dataset labeling, they introduced a template-driven approach for automatically annotating datasets with labeled food entities, eliminating the need for manual labeling. Their system achieved an impressive F1 accuracy score of 0.97 on a dataset compiled from multiple publicly available resources.

4. **Deep Learning approaches** - Deep learning, a subset of machine learning, has emerged as the leading approach in artificial intelligence due to its superior performance compared to previous methods (J. Li et al., 2020). It mimics the human brain's data processing mechanisms through neural networks, which analyze vast amounts of information, or training data, to make decisions. By repeatedly performing tasks with this data, neural networks improve their accuracy over time. The introduction of the Transformer (Vaswani et al., 2017), a neural network pre-trained on extensive text corpora, has revolutionized the field of natural language processing (NLP) by effectively addressing tasks involving sequence-to-sequence transformations and managing long-range dependencies. This advancement has had a significant impact on NER and other NLP tasks. Deep learning excels in handling unstructured data and extracting features automatically, enabling it to analyze large datasets thoroughly and uncover novel insights. However, a significant drawback is its reliance on large labeled datasets for optimal performance. Feng (2023) research focused on medical named entity recognition and introduced a model called DWI-Pos. This model integrates the position information of entity words and POS features, utilizing a Dynamic Windows Interception mechanism for accurate named entity recognition. The dataset used in the study was derived from the sub-tasks of CCKS2019[7], which includes six entity types. Comparative experiments were conducted between the DWI-Pos[8], BERT-CRF(S. Hu et al., 2022), and LSTM-CRF(Lample et al., 2016) models. DWI-Pos achieved an F1 value of 0.95, outperforming the ELMo-ET-CRF(Wan et al., 2020) model by 0.09 in terms of F1 value. Y. Hu et al. (2024) conducted a study to assess ChatGPT's zero-shot capability in clinical NER tasks. In a similar zero-shot scenario, ChatGPT's performance was compared to that of GPT-3 and a baseline model BioClinicalBERT, which was fine-tuned on synthetic MTSamples[9] and VAERS(Du et al., 2021)

[7]China Conference on Knowledge Graph and Semantic Computing (CCKS)
[8]Dynamic Windows Interception mechanism-position information
[9]a set of 163 artificially generated patient discharge summaries

datasets. The results indicated that ChatGPT outperformed GPT-3 in NER tasks. Furthermore, a clinical task-specific prompt framework was developed, incorporating annotation guidelines, error analysis-based instructions, and annotated samples via few-shot learning. Evaluation of this framework on two clinical NER tasks demonstrated that the GPT-4 model with prompts achieved performance close to that of the state-of-the-art BioClinicalBERT (Alsentzer et al., 2019) model.

2.1.4 Tools used for NER

Here, we outline various tools employed by researchers and developers within the NLP community for NER.

1. **NLTK (Natural Language Toolkit):** NLTK (Bird and Loper, 2004) is a free, open-source, community-driven project-leading platform for developing Python programs geared towards handling human language data. Its toolkit encompasses diverse functionalities, including tokenization, tagging, parsing, stemming, named entity recognition, and wrappers for industrial-strength NLP libraries. NLTK is accessible to users on various operating systems including Windows, Mac OS X, and Linux, and is widely used for teaching and research. NLTK supports the following languages: English, Spanish, French, German, Italian, Dutch, Portuguese, Russian, Chinese, Japanese, Arabic, Hindi, Turkish, Swedish, Norwegian, Danish, Finnish, Greek, Polish, Czech.

2. **spaCy:** spaCy (Honnibal and Montani, 2017) is an open-source library designed for advanced natural language processing tasks in Python, offering a range of pre-trained pipelines. It supports tokenization, training for over 70 languages, and customization of models using frameworks like PyTorch and TensorFlow. It promotes multitask learning with transformers such as BERT. Key components include a named entity recognizer, part-of-speech tagging, dependency parsing, sentence segmentation, text classification, lemmatization, morphological analysis, and entity linking. It also provides built-in visualizers for syntax and named entity recognition. Yoruba, English, Spanish, French, German, Italian, Dutch, Portuguese, Russian, Chinese, Japanese, Greek, Danish, Norwegian, Swedish, and Polish are just a few of the languages that spaCY supports[10]. It also has multi-language support, which enables it to function in a variety of languages to meet a range of linguistic needs.

3. **Stanford NER:** Stanford[11] NER is a Java library offering a range of tools for named entity recognition, with various options for defining feature

[10]https://spacy.io/usage/models/
[11]https://nlp.stanford.edu/software/CRF-NER.shtml

extractors. It includes pre-trained models utilizing an advanced statistical learning algorithm. Stanford NER provides several models for extracting named entities, including:

(a) A 3-class model for recognizing locations, persons, and organizations.

(b) A 4-class model for recognizing locations, persons, organizations, and miscellaneous entities.

(c) A 7-class model for recognizing locations, persons, organizations, times, money, percentages, and dates.

Stanford NER supports the recognition of named entities in English, German, and Arabic texts. However, it can be extended to support other languages through custom training with labeled data.

4. **Flair:** Flair (Akbik, Bergmann, et al., 2019) is a Python framework designed for cutting-edge natural language processing tasks. It provides pre-trained models for various tasks including named entity recognition, sentiment analysis, and part-of-speech tagging (PoS), and offers specialized support for biomedical data, sense disambiguation, and classification. Flair includes a wide range of word embeddings such as GloVe, BERT, ELMo, and Character Embeddings. It enables users to train custom models easily, supports multiple languages including but not limited to English, Spanish, and German, and continually expands its language support.

5. **GATE (General Architecture for Engineering):** GATE (Cunningham, Wilks, and Gaizauskas, 1996) is open-source free software. GATE includes a desktop client for developers, a workflow-based web application, a Java library, an architecture, and a process for the creation of robust and maintainable services. One of its key components is ANNIE (A Nearly-New IE system), an Information Extraction (IE) pipeline that comes built-in with GATE. ANNIE includes a named entity recognition module capable of identifying basic entity types such as Person, Location, Organization, Money amounts, and expressions for Time and Date. ANNIE does not inherently support specific languages; its language support depends on the linguistic resources and modules integrated into the GATE pipeline.

6. **Apache OpenNLP:** Apache OpenNLP (Kwartler, 2017) is a Java-based open-source library used for processing natural language text. It offers a wide range of services including tokenization, sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, coreference resolution, language detection, and more. For named entity recognition tasks, OpenNLP utilizes predefined models like en-ner-date.bin, en-ner-location.bin, en-ner-organization.bin, en-ner-person.bin, and en-ner-time.bin. The following is a list of some languages supported by Apache

OpenNLP: English, French, German, Spanish, Dutch, Italian, Portuguese, Russian, Chinese, Japanese, etc.

2.1.5 NER in Low-resourced Settings

In low-resource settings, NER tasks face a significant challenge due to the scarcity of labeled and unlabeled data (Sebastian Ruder, Søgaard, and Vulić, 2019). While considerable strides have been made in NER for well-resourced languages like English, which benefit from ample gold-annotated training data, the same cannot be said for low-resourced languages. With the advent of data-intensive deep learning approaches, addressing this challenge becomes particularly daunting for low-resourced languages. Below, we delve into several studies that outline various methods to address the challenges posed by limited or nonexistent training data, categorizing them into works focused on African Languages and those focused on non-African languages.

2.1.5.1 Works in Non-African Languages

Pandey and Nathani (2024) evaluated different approaches and methodologies utilized in Named Entity Recognition (NER) for Indian languages, assessing the merits and drawbacks of each method. Additionally, they explored recent advancements in transfer learning and multilingual models, highlighting their potential to enhance NER performance across Indian languages.

Shrestha (2024) explores using NERNepal, a pre-trained BERT-based model to retrieve named entities from texts of a low-resourced language, Nepali. The model's performance was evaluated on two datasets, highlighting unique linguistic challenges specific to Nepali. The research established that the NERNepal model performs better on the Nepali_NER dataset compared to the EverestNER dataset.

Puccetti et al. (2023) studied the application of Named Entity Recognition (NER) to detect the technologies referenced in the titles, abstracts, claims, and descriptions of the state-of-the-art patents. The study compares the effectiveness of three NER methods: gazetteer-based, rule-based, and deep learning-based (such as BERT), assessing their precision, recall, and computational efficiency.

Das et al. (2022) introduced CONTAINER, a contrastive learning technique for Few-Shot NER, which aims to minimize the distance between token embeddings of similar entities while increasing it for dissimilar ones. This approach effectively mitigates overfitting concerns stemming from training domains. CONTAINER demonstrates superior performance compared to previous methods, achieving an improvement of 3% to 13% in absolute F1 points. effectiveness.

Zevallos et al. (2022) created the first comprehensive combined corpus for deep

learning in Quechua, an indigenous South American low-resource language. Additionally, they introduced QuBERT, a publicly available pre-trained BERT model tailored specifically for Quechua, encompassing not just the southern dialect but also other Quechua variants. Evaluation of their corpus and BERT model yielded F1 scores ranging from 71% to 74% for NER tasks and 84% to 87% for POS tasks.

Tsygankova et al. (2021) research demonstrates that utilizing annotations from non-native speakers can serve as an alternative to cross-lingual methods for developing low-resource NER systems. An annotation experiment comparing the performance of non-speaker (NS) annotators with that of fluent speakers (FS) was conducted in Indonesian, Russian, and Hindi.

2.1.5.2 Works in African Languages

Michael A. Hedderich, Lange, and Klakow (2021) study created ANEA an open-source NER tool to obtain large amounts of training data based on distant supervision. Evaluation on the following low-resource language datasets: Estonian (Tkachenko, Petmanson, and Laur, 2013), West Frisian (Pan et al., 2017), Yoruba (Jesujoba Alabi et al., 2020), and manually annotated news articles for Spanish showed an improvement in the F1-score by an average of 18 points.

Michael A Hedderich, Lange, et al. (2021) survey detailed strategies for generating extra labeled data, such as data augmentation, distant supervision, and transfer learning, in low-resource environments. They elucidated the distinctions among these methods and emphasized the importance of comprehending their requirements to select an appropriate technique for a particular low-resource scenario.

David Ifeoluwa Adelani, Michael A Hedderich, et al. (2020) employed label-noise handling techniques and utilized two sources of distant supervision—rules and a list of entities—for NER tasks in Hausa and Yoruba. Experimental results demonstrated that these strategies can effectively enhance classifier performance in practical low-resource scenarios, potentially doubling the model's

Shruti Rijhwani et al. (2020) utilized entity-linking techniques to extract data from well-resourced languages and extensive English knowledge bases like Wikipedia. This extracted information was then incorporated into the CNN-LSTM-CRF NER model (Ma and Hovy, 2016) using a meticulously crafted feature set. Through experiments conducted across four low-resource languages—Kinyarwanda, Oromo, Sinhala, and Tigrinya—they showcased the efficacy of soft gazetteer features, resulting in an average enhancement of 4 F1 points compared to the baseline model.

Cai et al. (2023) introduced a novel Graph Propagated Data Augmentation (GPDA) framework tailored for low-resource NER scenarios. By leveraging graph

propagation alongside natural text, GPDA significantly outperformed prior data augmentation techniques across various low-resource NER datasets, as evidenced by experimental results.

A comprehensive dataset for Named Entity Recognition (NER) in ten African languages, including Igbo, was a collaborative effort led by David Ifeoluwa Adelani, J. Abbott, et al. (2021). I was a member of the Igbo language annotation team. In this work, we trained and evaluated multiple NER models for each of the ten languages, shedding light on transfer learning across languages and establishing robust baseline performance benchmarks. The entire study is discussed fully in Chapter 4 of this thesis.

We expanded our research efforts to develop the most extensive human-annotated NER dataset encompassing 20 African languages (Adelani et al., 2022). Our investigation into the performance of cutting-edge cross-lingual transfer techniques within an African-centric context revealed a significant improvement in zero-shot F1 scores. Specifically, selecting the optimal transfer language enhanced F1 scores by an average of 14 points across all 20 languages, surpassing the performance achieved by using English alone. This underscores the importance of establishing benchmark datasets and models that encompass a wide range of typologically diverse African languages. The entire study is discussed fully in Chapter 6 of this thesis.

2.1.6 NER Datasets

A NER dataset comprises text documents annotated to identify and categorize named entities like persons, organizations, places, dates, and other pertinent entities. These datasets serve the purpose of building, enhancing, testing, and assessing NER models and techniques. They come in diverse sizes and formats, ranging from domain-specific to multilingual, and are crucial in advancing NER research and applications. This section outlines NER benchmark datasets extensively utilized by researchers and professionals for training and assessing NER models and methodologies.

1. **WikiNER**- Nothman et al. (2013) introduced the WikiNER Dataset, which comprises 7,200 Wikipedia articles manually labeled across nine languages: English, German, French, Polish, Italian, Spanish, Dutch, Portuguese, and Russian, and is useful for cross-lingual NER tasks.

2. **GENIA**- The GENIA (J.-D. Kim et al., 2003) dataset is a corpus, frequently utilized in biomedical natural language processing (NLP) endeavors. It is a collection of PubMed abstracts annotated with biomedical entities like genes, proteins, cell types, and cell lines. These annotations offer crucial labeled data for training and assessing NER models within the biomedical field.

3. **Message Understanding Conference (MUC) Datasets** - The Message Understanding Conference (MUC) (Grishman and B. M. Sundheim, 1996) dataset was introduced during the 1990s as a component of the MUC conference series. It comprises annotated English text documents containing named entities. The data originates from military and news reports, and the dataset is available in seven versions, as indicated in Table 2.3.

Dataset Name	Year	Language	Source Type
MUC 1	1987	English	Military reports
MUC 2	1989	English	Military reports
MUC 3	1991	English	Reports from News
MUC 4	1992	English	Reports from News
MUC 5	1993	English, Japanese	Reports from News
MUC 6	1995	English	Reports from News
MUC 7	1997	English	Reports from News

Table 2.3: MUC 1 - MUC 7 dataset series

1. **Automatic Content Extraction (ACE)** - ACE (Doddington et al., 2004) program aims to extract information from audio and image sources alongside textual data with a focus on text extraction. This endeavor entails meticulous task definition, comprehensive data collection, annotation for training, development, and evaluation purposes, and bolstering the research with evaluation tools and workshops. The pilot study for this program commenced in 1999. Table 2.4 provides the details of the ACE corpora and tasks [12].

2. **Conference on Computational Natural Language Learning (CoNLL) Corpus** - CoNLL NER datasets (Tjong Kim Sang (2002); Tjong Kim Sang and De Meulder (2003)) emanated from the Conference on Computational Natural Language Learning (CoNLL) shared task: language-independent named entity recognition. All these entities are annotated from newswire articles. Table 2.6 gives the details of the datasets.

3. **OntoNotes Corpus**- The OntoNotes (Weischedel et al., 2017) dataset is a widely used multilingual annotated corpus in NLP. It comprises text documents from various genres annotated with linguistic information like named entities, syntactic parses, and coreference chains. Table 2.6 describes the various releases of this dataset. OntoNotes is one of the largest benchmark datasets widely used to assess NER tasks.

[12]https://www.ldc.upenn.edu/collaborations/past-projects/ace/annotation-tasks-and-specifications

Dataset Name	Task	Language
ACE Phase 1 and ACE Pilot	Entities	English
ACE Phase 2	Entities and Relations	English
ACE 2003	Entities and Relations Entities	Chinese Arabic
ACE 2004	Entities and Relations	English, Chinese, Arabic
2005 ACE	Event annotation	Arabic, English, Chinese
ACE 2007	A pilot evaluation for EDR [13] and TERN [14]	Spanish
ACE 2008	Local EDR, RDR and pilot task for Global EDR, RDR [15]	English, Arabic

Table 2.4: ACE Annotation Tasks and Specifications

Dataset Name	Year	Language	Source Type
CoNLL'02	2002	Dutch, Spanish	Newswire Articles
CoNLL'03	2003	English, German	Newswire Articles

Table 2.5: CoNLL datasets

4. **Workshop on Noisy User-generated Text (WUNUT) Dataset**- The benchmark dataset for the WNUT2017 Shared Task on Novel and Emerging Entity Recognition comprises 1,008 development and 1,287 test documents, totaling nearly 2,000 entity mentions (Derczynski et al., 2017). These documents are sourced from social media platforms like Twitter and are specifically selected to contain predominantly rare and emerging entities such as persons, locations, corporations, products, creative works, and groups.

5. **Journal of Natural Language Processing for Biomedicine and its Applications (JNLPBA)**- The JNLPBA shared task focuses on bio-entity recognition, utilizing an expanded edition of the GENIA version 3 named entity corpus extracted from MEDLINE abstracts (Collier et al., 2004).

2.1.7 NER Datasets with African Languages

The growing presence of the Machine Learning (ML) Community in Africa has spurred collaborative endeavors to develop Natural Language Processing (NLP) resources such as corpora, datasets, and tools for numerous low-resourced African languages. Although NER datasets are available for several languages, only a handful of African languages have dedicated NER datasets. Before this Ph.D.

Dataset Name	Year	Language	Source Type
OntoNotes 1.0	2007	English, Mandarin Chinese	Newswire Articles
OntoNotes 2.0	2008	English, Mandarin Chinese	Broadcast News
OntoNotes 3.0	2009	English, Chinese, Arabic	Broadcast Conversation Newswire Articles
OntoNotes 4.0	2011	English, Chinese, Arabic	Reports from News
OntoNotes 5.0	2013	English, Chinese, Arabic	Reports from News

Table 2.6: OntoNotes dataset versions

study, the WikiAnn dataset was the sole resource containing tagged name mentions for the Igbo language, with only 968 entries (David Ifeoluwa Adelani, J. Abbott, et al., 2021). Table 2.7 summarizes NER datasets that contain Igbo-tagged entities. Table 2.8 summarizes the NER datasets without Igbo-tagged entities. In this thesis, we created an IgboNER dataset for the Igbo language, MasakhaNER (details in Chapter 4) for ten languages, and we also extended it to 20 languages (details in Chapter 6) to contribute to AfricaNLP and the NLP community at large.

2.1.8 NER Label Sets

NER label sets comprise predetermined classes of named entities that a NER model is trained to recognize and categorize within text data. These sets encompass diverse entity types like person, organization, location, date, concept, product, event, technology, module, physiology, a unit, feature, medical condition, animals, cell feature, work-of-art, input device, and others (Zhou et al., 2023), tailored to the particular task and domain. In this thesis, we labeled entities belonging to four categories: person names (PER), locations (LOC), organizations (ORG), and dates and times (DATE). Our entity labeling set was influenced by the English CoNLL-2003 Corpus (Tjong Kim Sang, 2002), with the MISC tag being substituted with the DATE tag in alignment with our MasakaNER (David Ifeoluwa Adelani, J. Abbott, et al., 2021) dataset.

Dataset	Languages	Tokens or Sentences	Data source	Availability
WikiAnn corpus (Pan et al., 2017)	Covers 282 languages including Igbo.	10 million English pages labeled with 968 Igbo-tagged tokens. Automatically annotated	Wikipedia	public
Masakha-NER (David Ifeoluwa Adelani, J. Abbott, et al., 2021)	Amharic, Hausa, Igbo, Kinyarwanda, Luganda, Luo, Nigerian Pidgin, Swahili, Wolof, Yorùbá	Less than 4k sentences each. Manually annotated	News	public
MasakhaNER (Adelani et al., 2022)	Covers 20 languages including Igbo	Between 8K–11K sentences per language. Manually annotated	News	public

Table 2.7: African NER Datasets containing Igbo language

2.1.9 NER Annotation schemes

NER annotation schemes serve as frameworks for annotating named entities within text data. These schemes offer a structured approach to marking named entities in text, aiding in identifying their type and position within sentences (Keraghel, Morbieu, and Nadif, 2024). By providing a systematic method for annotation, these schemes facilitate the training and assessment of NER models. They also promote uniform labeling across annotators and datasets, which is essential for the reliable development and implementation of NER systems. In annotation schemes, the initial token of an entity is identified as "B" (Beginning), the tokens within the entity are designated as "I" (Inside), and the final token is indicated as "E" (End). Tokens that do not belong to an entity are labeled as "O" (Outside). Frequently used annotation schemes include BIO, IO, IOE, IOBES, IE, and BIES.

1. **BIO/IOB**- This schema is adopted by CoNLL (Tjong Kim Sang and Buchholz, 2000) and is the most commonly used for NER tasks. Each token in a sequence is labeled with one of three tags: "B" indicates that the token is the beginning of a named entity. "I" indicates that the token is inside a named entity. "O" indicates that the token is outside of any named entity (non-entity words).

2. **IOE**- In this scheme, named entities are marked by "I" for tokens inside the entity and "E" for tokens marking the end of the entity, instead of using "B"

Dataset	Languages	Tokens or Sentences	Data source	Availability
GV-Yorùbá-NER (Jesujoba Alabi et al., 2020)	Yoruba	1,101 sentences (26,240 tokens) Manually annotated	Global Voices news	public
SADiLaR (Eiselen, 2016)	Afrikaans, isiNdebele, isiXhosa, isiZulu, Sesotho sa Leboa, Sesotho, Setswana, SiSwati, Tshivenda, Xitsonga	15,000 tokens. Manually annotated	Government data	not public
Hausa (Michael A Hedderich, David Adelani, et al., 2020)	Hausa	Less than 2k sentences. Manually annotated	Voice of America Hausa (News)	public
LORELEI language packs (Strassel and Tracey, 2016)	23 languages plus Yorùbá, Hausa, Amharic, Somali, Swahili, Wolof, Kinyarwanda, Zulu	75k words per languages labeled for Simple Named Entity & 25k words for Full Entity. Manually annotated	News, blogs, discussion forums, microblogs	not public
ANEC (Jibril and A. Cüneyd Tantuğ, 2023)	Amharic	8,070 sentences, which has 182,691 tokens. Manually annotated	News	public
Tigrinya-NER (Yohannes and Amagasa, 2022)	Tigrinya	69,309 entity tags containing 3625 sentences. Manually annotated	News and some freely available e-books, including the Bible	public

Table 2.8: African NER Datasets not containing Igbo language

to indicate the beginning as in the IOB scheme.

3. **IO**- In this tagging scheme, each token representing a named entity is tagged as "I", while all other tokens are tagged as "O".

4. **IOBES**- This scheme offers detailed boundary information for named entities. It employs four tags: "B" for the start of an entity, "I" for tokens within the entity, "E" for the end of an entity, "S" for single-token entities, and "O" for non-entity words outside named entities.

5. **IE**- This annotation scheme functions similarly to IOE but differs in labeling the end of non-entity words with the tag "E-O" and the remainder of the non-entity words as "I-O".

6. **BIES**- This scheme is similar to IOBES but employs tags like "B-O" to mark the beginning of non-entity words, "I-O" for inside tags within non-entity words, "E-O" to signify the end of non-entity words, and "S-O" for individual non-entity tokens situated between two entities.

2.1.10 NER Evaluation Metrics

NER evaluation metrics are essential for assessing the performance of NER models by comparing their predicted named entities with the ground truth annotations in a dataset. These metrics measure the accuracy of identifying named entities regarding both entity types and their boundaries. Commonly used evaluation metrics in the literature include MUC, CoNLL, and SemEval metrics, which are named after the conferences where they were introduced (Jibril and A. Cüneyd Tantuğ, 2023). These metrics provide standardized measures to evaluate the effectiveness of NER models across different datasets and tasks.

1. **MUC Metric**- According to (Grishman and B. M. Sundheim, 1996) these metrics involve comparing the system's answers against the human-created (gold) annotation and categorizing different types of errors. The recall-precision evaluation metrics expanded for application to MUC originated from the field of Information Retrieval (IR). MUC metrics consider scenarios like:

   (a) **Correct (COR)**: If the system answers and the gold annotation are deemed to be equivalent.

   (b) **Incorrect (INC)**: If the gold annotation and system answers do not match.

   (c) **Partial (PAR)**: If the system answers and the gold annotation are judged to be a near match.

   (d) **Missing (MIS)**: A golden annotation is not captured by a system.

   (e) **Spurious (SPU)**: System produces a response, which does not exist in the golden annotation.

$$Recall = \frac{Correct + (0.5 * Partial)}{Possible} \tag{2.1}$$

$$Precision = \frac{Correct + (0.5 * Partial)}{Actual} \tag{2.2}$$

$$F1 = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{2.3}$$

Possible is the sum of the correct, partial, incorrect, and missing.
Actual is the sum of the correct, partial, incorrect, and spurious.
Recall (REC) is the percentage of possible system answers that are correct.
Precision (PRE) is the percentage of actual system answers given which were correct. ***F1-score*** is the harmonic mean of precision and recall.

2. **CoNLL Metric**- The Language-Independent Named Entity Recognition task, as introduced at CoNLL-2003 (Tjong Kim Sang and De Meulder, 2003) evaluates the task's performance using the $F\beta = 1$ rate, with $\beta = 1$ (Rijsbergen, 1979). Precision represents the proportion of correct named entity results out of all results retrieved by the system that is correct. Recall is the proportion of named entities in the corpus found by the system. A named entity is correct only if it is an exact match of the respective entity in the data file.

$$Precision = \frac{TP}{TP + FP} \tag{2.4}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.5}$$

$$F_\beta = \frac{(\beta^2 + 1) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \tag{2.6}$$

where TP is the number of True Positives, FP is the number of False Positives, and FN is the number of False Negatives.

3. **International Workshop on Semantic Evaluation (SemEval) Metric**- This evaluation metrics aim to assess whether a system can accurately identify the precise span of an entity, irrespective of its type, and determine if it can correctly assign the entity type, regardless of its boundaries (Segura-Bedmar, Martínez, and Herrero-Zazo, 2013). The SemEval evaluation script will generate four sets of scores as output:

   (a) **Strict** evaluation (exact-boundary and type matching).

   (b) **Exact** boundary matching (regardless of the type).

   (c) **Partial** boundary matching (regardless of the type).

   (d) **Type** matching (some overlap between the tagged entity and the gold entity is required).

Evaluation results are reported using the standard precision/recall/f-score metrics as the scoring categories proposed by MUC in different ways (Jibril and A. Cüneyd Tantuğ, 2022).

Exact Match (i.e., strict and exact):

$$Precision = \frac{COR}{ACT} = \frac{TP}{TP + FP} \tag{2.7}$$

$$Recall = \frac{COR}{POS} = \frac{TP}{TP + FN} \tag{2.8}$$

Partial Match (i.e., Partial and Type):

$$Precision = \frac{COR + (0.5 * PAR)}{ACT} = \frac{TP}{TP + FP} \tag{2.9}$$

$$Recall = \frac{COR + (0.5 * PAR)}{POS} = \frac{COR}{ACT} = \frac{TP}{TP + FP} \tag{2.10}$$

Possible and Actual scores are calculated as:

$$Possible(POS) = COR + INC + PAR + MIS = TP + FN \tag{2.11}$$

$$Actual(ACT) = COR + INC + PAR + SPU = TP + FP \tag{2.12}$$

2.2 Chapter Summary

This chapter presented an overview of the focus language, Igbo, its writing system before colonization using symbols, its orthography, and controversies surrounding the adoption of an orthography. The effect of diacritics was explained. NLP research in Igbo and the available resources for the Igbo language were discussed under IgboNLP.

The meaning of NER was explained, techniques applied to NER tasks, tools used, NER benchmark datasets extensively used by researchers, evaluation metrics, datasets with African languages before this thesis, annotation schemes, and NER label sets. We also presented some studies on NER in low-resourced settings.

Section B
IgboNER

Chapter 3

A Framework for Named Entity Recognition

3.1 Introduction

This chapter outlines our approach to developing fundamental Natural Language Processing (NLP) resources for Igbo, a low-resource language, to accommodate its distinctive linguistic characteristics. We present a methodology for creating three key components: a dataset, a Named Entity Recognition (NER) model, and a visualizer specifically designed for Igbo[1]. Our methodology is crafted with scalability and generalization in mind, allowing for expanding the NER framework to encompass other languages and domains. This inclusive approach will empower researchers engaged in low-resource language studies to devise NLP solutions that address the unique requirements of diverse linguistic communities.

3.2 Comprehensive Framework

This section introduces the architectural framework for NER, as illustrated in Figure 3.1. The framework is designed to provide a systematic approach to developing various NLP resources for the Igbo language. This adaptable framework can be applied to any language by replacing "Igbo" with the desired language. It comprises three phases:

- ***Phase I: Data Collection***- This shows that written data was crawled from the Web and also collected locally[2]. The machine-translated parallel data (Igbo and English) was adapted from the work by Ezeani, Rayson, et al., 2020b. This data collection process is employed due to the absence of digitally, culturally relevant content that adequately reflects the subtle language distinctions present in the Igbo language. In this phase, we

[1]https://igbo-demo.streamlit.app/

[2]locally in this context means data from Igbo novels which are not in a digital format

conducted manual observation and correction of the data to ensure that all English sentences had their corresponding Igbo translations. Each sentence was verified and its translation was correctly aligned, ensuring accuracy and consistency throughout the dataset. Section 5.3 of Chapter 5 describes the data collection.

- ***Phase II: Dataset Annotation***- The projection technique is applied here and the process is split into two: the manual (left) and the automatic (right). A semi-automatic process is used to create a mapping dictionary to transfer tags to Igbo sentences. The use of a mapping dictionary handles the challenge of orthographic variations and multi-word expressions in the Igbo language. The projection approach offers an efficient alternative to the human-intensive task of dataset creation, particularly beneficial for languages with scarce digital text resources. In this phase, the English sentences were tagged using spaCy. A mapping dictionary was then created, facilitating the transfer of tags to the corresponding Igbo sentences and ensuring accurate tagging across both languages. This process is explained in Section 5.4 of Chapter 5 of this work.

- ***Phase III: Model development and Evaluation***- This phase is about training an IgboNER from scratch. In this phase, we fine-tuned several state-of-the-art models, such as mBERT and DistillBERT, using the IgboNER dataset and conducted evaluations to assess their performance. This process is explained in Chapter 4 of this work.

3.3 Manual Annotation Process

In the course of this PhD, I collaborated in the following studies: David Ifeoluwa Adelani, J. Abbott, et al. (2021) and Adelani et al. (2022) described in Chapter 6 and Chapter7 of this thesis and was the lead annotator. As the lead annotator, I acted as the intermediary between the Igbo annotators and the language coordinator, giving updates on the annotation progress. I supported other annotators, clarified annotation rules, and resolved ambiguities in labelling. We used Elisa(Y. Lin et al., 2018), an annotation tool that allows trained human annotators to highlight entities within the text. The project coordinator provided predefined annotation guidelines [3] to ensure accuracy and reduce ambiguity. A training session was held for the annotators to explain the task, how to use the Elisa tool and the guidelines. Each sentence was annotated by two different annotators to ensure the quality, consistency, and reliability of the labels. Elisa supports adjudication and provides the interface for disagreements between annotators to be reviewed. This exercise was carried out with the language coordinator and

[3]https://github.com/masakhane-io/masakhane-ner/blob/main/MasakhaNER%20Annotation%20Guideline.pdf

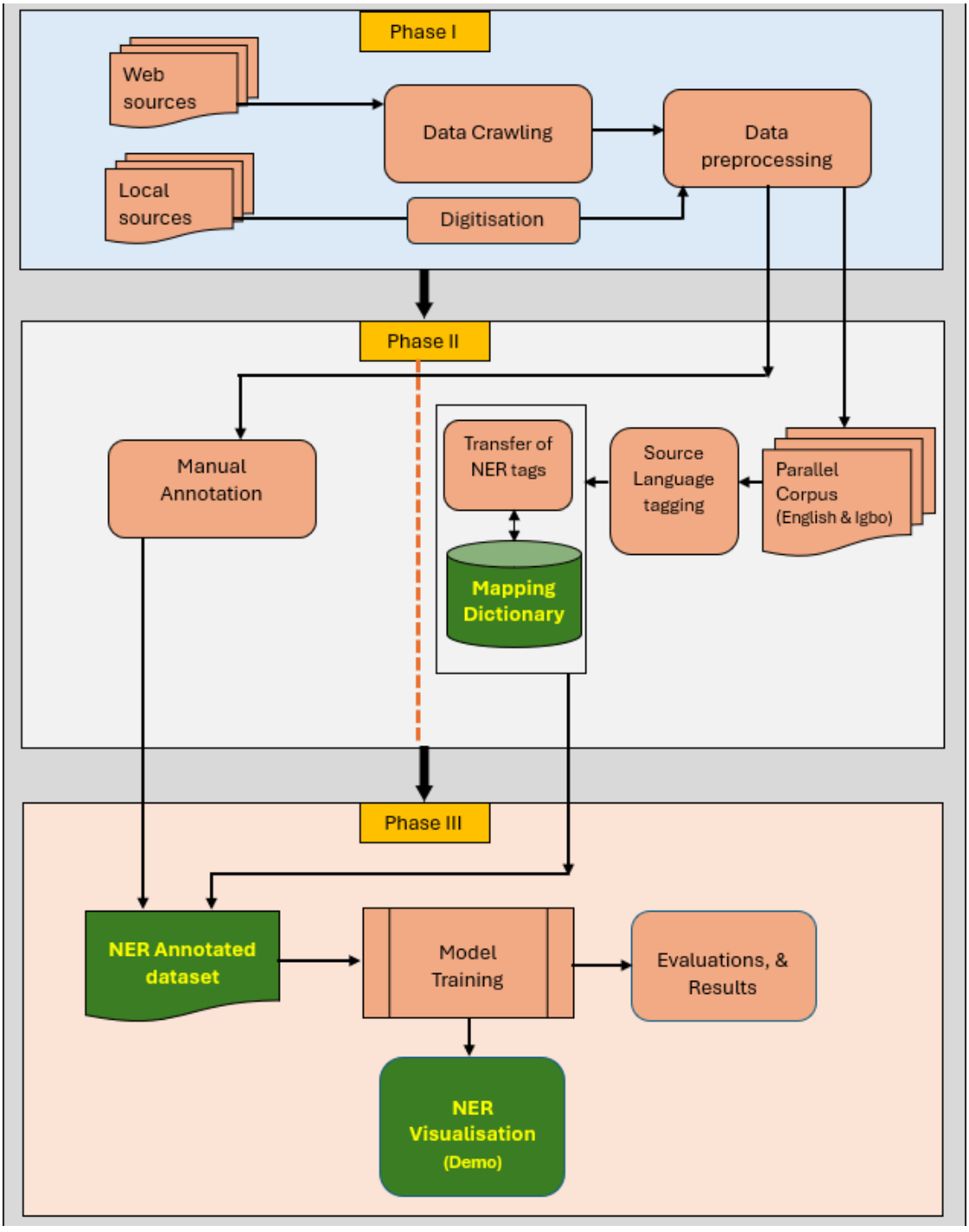


Figure 3.1: Named Entity Recognition Framework.

the annotators to produce the final validated labels. The finalised labels are then exported by the coordinator for further quality checks before they are used for model training.

3.4 Chapter Summary

This chapter outlines our approach to building NLP tools such as NER datasets, models, and mapping dictionaries for languages with limited resources. The Elisa tool for manual annotation and the process of annotation is also discussed. The IgboNER model and the mapping dictionary are novel as they are the first to be created for the Igbo language. These efforts aim to increase the accessibility of language technology to diverse communities. This chapter addressed RQ3.

Chapter 4

Transformer Models for the Igbo Language

4.1 Introduction

This chapter is about the NER model and dataset creation for Igbo language. This chapter aimed to train a baseline IgboNER model from scratch and create a dataset for the Igbo language. This Chapter is derived from the published paper titled " IgboBERT Models: Building and Training Transformer Models for the Igbo Language" (C. Chukwuneke et al., 2022)

Information is stored digitally in many languages. To digitally interact in these languages, localization of computer interfaces and tools is vital, leading to the need for NLP research to build these tools. Some of the contributing factors to the lack of research in various countries include very few available language resources and computing capacity to handle such research. This limits the development and creation of tools and resources for performing a wide variety of NLP tasks such as named entity recognition (NER), machine translation (MT), information retrieval, etc.

This work is an addition to the efforts made by (Ikechukwu E Onyenwe, Uchechukwu, and Hepple, 2014; Ezeani, Hepple, and I. Onyenwe, 2016; Ezeani, Rayson, et al., 2020a; Ikechukwu E Onyenwe, Hepple, Chinedu, et al., 2018) and also presented the Igbo part of the MasakhaNER dataset to support Igbo NLP in particular but also to contribute to the African NLP efforts. Secondly, we are interested in finding out to what extent a small language model pre-trained from scratch with the target language compares to existing multilingual transformer models like mBERT and XLM-RoBERTa.

4.2 Related Work

4.2.1 Low Resource Named Entity Recognition

Significant research has been in other languages, mainly English, has been performed on NER since the inception of the task at the MUC-6 conference in 1996 ((Lample et al., 2016); (David Ifeoluwa Adelani, J. Abbott, et al., 2021); (Ratinov and Roth, 2009)). One major problem facing NER tasks in low-resourced scenarios is the availability of labeled data (Sebastian Ruder, Søgaard, and Vulić, 2019). Manually labeling large corpora is task-intensive, time-consuming, and expensive. With the recent more data-hungry deep learning approach, it has become a bigger challenge to work in this area. Here, we will focus on recent work on NER for low-resourced languages. David Ifeoluwa Adelani, J. Abbott, et al., 2021 created high-quality data sets of less than 4k sentences each for 10 African languages by manual annotation. Transfer learning and gazetteer approaches were applied in building a model that can recognize named entities for 10 African languages. The model was evaluated on multiple state-of-the-art NER models and showed improvement. ANEA, a tool to automatically annotate named entities based on distant supervision to obtain large amount of training data was presented by (Michael A. Hedderich, Lange, and Klakow, 2021). ANEA allows users to add their expertise by allowing a tuning step to improve the automatically annotated data. Evaluations on 16 entity types in the following different languages (Spanish, Yoruba, Estonian, and West Frisian) showed an improvement on 14 entity types with an F1-score of average.

Tsygankova et al., 2021's study proved that using non-native speakers annotation is an alternative to cross-lingual methods for building low-resource NER. One of the reasons for its success is the ability of human non-speaker annotators to make inferences over common sense world knowledge, unlike an automatic system. Michael A Hedderich, David Adelani, et al., 2020's work on NER and topic classification showed that data sizes affect the performance of models. Transfer learning and distant supervision on multilingual transformer models were evaluated on three African languages: Hausa, isiXhosa, and Yorùbá, each with different amounts of available resources. This study achieved the same performance as baselines with little data but not for all the cases.

4.3 Language Resources

4.3.1 Data Collection

In this work, we used the MasakhaNER dataset created by the Masakhane Community (David Ifeoluwa Adelani, J. Abbott, et al., 2021). The data was

obtained from BBC Igbo news[1] and is 3,190 sentences containing 61,668 tokens. Additionally, 8,000 Igbo sentences from Lacuna project[2] in the Masakhane community were also used. The contents are from Igbo-Radio and Kaoditaa[3].

We also used 383,449 raw monolingual Igbo sentences from the study by (Ezeani, Rayson, et al., 2020a). A large section of the data was collected from the Jehova's Witness Igbo[4] and the contents include the Bible, more contemporary contents (books and magazines e.g. Teta! (Awake!), UloNche! (WatchTower)). Also collected are contents from BBC-Igbo[5], igbo-radio[6] as well as Igbo literary works (Eze Goes To School[7] and Mmadu Ka A Na-Aria by Chuma Okeke). The table 4.1 shows the statistics of the raw data used in this work.

Source	**Sentences**	**Tokens**	**Orthography**
eze-goes-to-school	1272	25413	Ọnwụ
mmadu-ka-a-na-aria	2023	39731	Ọnwụ
bbc-igbo	34056	566804	Africa, Ọnwụ
igbo-radio	7251	202623	Lepsuis, Africa, Ọnwụ
jw-ot-igbo	32251	712349	Lepsuis, Ọnwụ
jw-nt-igbo	10334	253806	Lepsuis, Ọnwụ
jw-books	142753	1879755	Lepsuis, Ọnwụ
jw-teta	14097	196818	Lepsuis, Ọnwụ
jw-ulo-nche	27760	392412	Lepsuis, Ọnwụ
jw-ulo-nche-naamu	113772	1465663	Lepsuis, Ọnwụ
kaoditaa	5880	22557	Lepsuis, Africa, Ọnwụ
Total	391449	5757931	

Table 4.1: Data Sources and Counts

4.3.2 Annotation

We used the BIO (Beginning, Inside, Outside) tagging scheme to label the entities. The entity tags correspond to this list: "O", "B-PER", "I-PER", "B-ORG", "I-ORG", "B-LOC", "I-LOC", "B-DATE", "I-DATE" where O denotes non-entity words, B-PER/I-PER denotes the beginning of/is inside a person entity, B-ORG/I-ORG denotes the beginning of/is inside an organization entity, B-LOC/I-LOC denotes the beginning of/is inside a location entity, and B-DATE/I-DATE denotes the beginning of/is inside a date entity. Throughout, 'B' indicates the beginning

[1]https://www.bbc.com/igbo
[2]`https://github.com/Chiamakac/lacuna_pos_ner/tree/main/language_corpus/ibo`
[3]https://kaoditaa.com/
[4]https://www.jw.org/ig/
[5]https://www.bbc.com/igbo
[6]https://igboradio.com/
[7]https://bit.ly/2vdGvKN

of a tag, 'I' indicates inside of a tag, and 'O' indicates outside i.e. the token belongs to no tag. The annotation of the IgboNER high-quality was performed using the ELISA tool (Y. Lin et al., 2018) by Igbo native speakers from the Masakhane community[8] of which the thesis author is a member. The ELISA tool was used because it provides an interface for annotators to correct their mistakes, making it easy to achieve a high inter-annotator agreement, and also provides an entity-level F1 score. Training was given to the annotators to ensure high-quality annotation. The guideline for the annotation and a video explaining the guideline was shared with the annotators. Then a virtual meeting was held for questions and discussion to ensure proper understanding of the guidelines. Fleiss' Kappa (Fleiss, 1971) was used to calculate the inter-annotator agreement and it considers each span that an annotator proposed as an entity. The data set has an inter-annotator agreement of 0.995 and 0.9830 at the token and entity level respectively. The annotators annotated four entity tags/types: personal name entity (PER), location entity (LOC), organization (ORG), and date & time (DATE) using the MUC-6 annotation guide[9]. The annotated entities were based on the state-of-the-art English CoNLL2003 Corpus (Tjong Kim Sang, 2002) but the Miscellaneous (MISC) tag was replaced with the DATE tag in the MasakhaNER data set following previous work (Jesujoba Alabi et al., 2020). Figure 4.1 below shows the distribution of the entities annotated.

The major issues faced when annotating the Igbo language that we discovered during the process are:

- Orthography: Igbo text corpora are written with a combination of Lepsuis, Africa, and Ọnwụ (see section 2.1.1.3).

- Ambiguity: Some Igbo words are relatively ambiguous. For instance, some person names have other meanings, e.g. "`Eze`" can be the name of a person (proper noun), a part of the human body for chewing (plural noun), and also it can be a male ruler of an independent state (noun).

4.3.3 Dataset Splits

The data set is split into three parts named: train, development (dev), and test originally and they correspond to the train, validation, and test splits (David Ifeoluwa Adelani, J. Abbott, et al., 2021). This was used in fine-tuning all the models in this work. Table 4.2 shows a summary of the data set splits.

[8] https://www.masakhane.io
[9] https://cs.nyu.edu/∼*grishman*/muc6.html

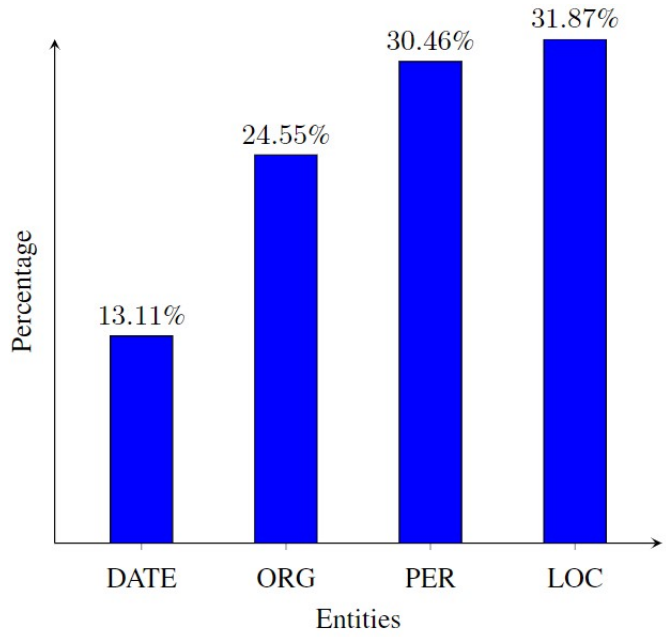


Figure 4.1: Annotated entity distribution. This shows the percentage distribution of the entities: person (PER), location (LOC), organization (ORG), and date (DATE).

Data set	Number of sentences	Number of tokens
Training set	2233	42719
Development set	319	6304
Test set	638	12645

Table 4.2: Summary of dataset splits

4.4 Experimental Setup

In this section, we describe the baseline model we pre-trained for the Igbo language and some state-of-the-art transformer models we fine-tuned to the downstream Igbo NER task.

4.4.1 Baseline Model

The first experiment was the training of an Igbo language model (IgboBERT) from scratch using transformers and tokenizers to have a baseline model for Igbo language NER[10]. The model was trained with the raw data described in Table 4.1 with a masked language modeling (MLM) objective. We trained a byte-level Byte-pair encoding tokenizer (the same as GPT-2) of size 52,000, with the same special

[10]https://huggingface.co/blog/how-to-train

tokens as RoBERTa[11]. Our tokenizer is optimized for Igbo by encoding native words and diacritics in Igbo language characters. Byte-level Byte-pair encoding tokenizer was chosen because it starts building its vocabulary from an alphabet of single bytes, so all words will be broken down into tokens to eliminate unknown (<unk>) tokens. The small model consists of 6 layers with 768 hidden size, 12 attention heads, and 84M parameters, the same number of layers and heads as DistilBERT. IgboBERT was trained at a learning rate of 1e-4 for 5 epochs and a batch size of 16. We carried out only 5 epoch training because of limited compute resources such as GPU at the time of the experiment. We then fine-tuned the IgboBERT model on the IgboNER downstream task using our MasakhaNER dataset.

4.4.2 Fine-tuned Models

The following state-of-the-art transformer models pre-trained on raw texts only were fine-tuned to a downstream IgboNER task using the MasakhaNER dataset. We added a linear classification layer to the pre-trained transformer models to predict entity types. 20 epoch training with a batch size of 8 at a learning rate of 2e-5 and 1e-4 was run.

- Multilingual BERT (mBERT): mBERT (Devlin et al., 2019) is a transformer model pre-trained with a large corpus of multilingual data from Wikipedia on 104 languages including only two African languages: Swahili and Yorùbá. This model was trained with two objectives: masked language modeling (MLM) and Next sentence prediction (NSP). We use the mBERT-base cased model with 12-layer Transformer blocks consisting of 768-hidden size and 110M parameters

- XLM-RoBERTa (XLM-R): XLM-R (Conneau et al., 2020) is a multilingual version of RoBERTa pre-trained on 2.5TB of filtered CommonCrawl data containing 100 languages including three African languages: Amharic, Hausa, and Swahili. We use the XLM-R base model consisting of 12 layers, with a hidden size of 768 and 270M parameters.

- DistilBERT (Victor Sanh et al., 2019): a smaller and faster version of BERT, which was pre-trained on the same corpus as BERT. We use the DistilBERT base model uncased with 6-layer Transformer blocks consisting of 768-hidden size and 66M parameters

4.5 Results

Table 4.3 and 4.4 show the fine-tuned results of the mBERT, XL-MR, DistilBERT, and IgboBERT after 20 epochs at learning rates 1e-4 and 2e-5, respectively using

[11]https://huggingface.co/docs/transformers/model_doc/roberta

the IgboNER dataset created in this work. From the results, mBERT with an F1 score of 89.02 and accuracy of 98.05%, consistently outperformed the other models across all criteria, with its highest performance at a lower learning rate (2e-5). XLM-R was next in performance with an F1 score of 87.93 and an accuracy of 97.74%, then DistilBERT with an F1 score of 80.37 and 96.67% accuracy. IgboBERT performed poorly consistently, with its highest performance at a higher learning rate (1e-4), producing an F1 score of 77.94% and 95.61% accuracy. Our IgboBERT was outperformed by mBERT, XLM-R, and DistilBERT, as shown in the tables, which is likely because of the limited quantity of its training data. We conclude that the performance of IgboBERT is reasonably comparable to the performance of mBERT, XLM-R, and DistilBERT, considering the size of their training data.

Model	**Precision**	**Recall**	**F1**	**Accuracy**
mBERT	85.67	87.67	86.66	97.96
XLM-R	84.54	85.67	85.10	97.81
DistilBERT	79.79	77.00	78.37	96.20
IgboBERT	76.44	79.50	77.94	95.61

Table 4.3: Performance of mBERT, XLM-R, DistilBERT, and IgboBERT: We display the fine-tuned results of the models after 20 epochs at 1e-4 learning rate.

Model	**Precision**	**Recall**	**F1**	**Accuracy**
mBERT	88.22	89.83	89.02	98.05
XLM-R	87.21	88.67	87.93	97.74
DistilBERT	81.26	79.50	80.37	96.67
IgboBERT	73.23	77.50	75.30	95.55

Table 4.4: Performance of mBERT, XLM-R, DistilBERT, and IgboBERT: We display the fine-tuned results of the models after 20 epochs at 2e-5 learning rate.

Figure 4.2 is an illustration of the relationship between training loss and validation loss, as well as precision, recall, and F1 scores for the models IgboBERT and the fine-tuned mBERT (IgboMBERT), XLM-R (IgboXML-R), and DistilledBert (IgboDistilBERT) over 20 epochs at a learning rate of 1e-4. Effective learning from the training data by all the models was demonstrated by the consistent decrease in training loss, but the moderate divergence of the validation loss after initial epochs depicts overfitting. The result also showed a steady performance progress over some time for all the models and a clear difference was seen in their highest scores.

The graphs in Figure 4.3 are a performance comparison during the training

and evaluation of the IgbomBERT, IgboDistilBERT, IgboBERT, and IgboXLM-R models at a learning rate of 2e-5. When compared with the output graph with a higher learning rate of 1e-4 in Figure 4.2, we depict that there was also effective learning from the training data by all the models as demonstrated by the consistent decrease in training loss, but at 2e−5, the validation loss of the models was lower and more stable, which shows its better generalization and stability. For both learning rates, the performance of IgbomBERT is seen to remain the most consistent. However, IgbBERT performed better with an F1-score of 77.94%, Recall 79.50%, Precision 76.44% at learning rate of 1e-4 which showed better performance over the learning rate of 2e-5 at an F1 score of 75.30%, Recall 77.50%, Precision 73.23%.

The metrics (precision, recall, F1, and accuracy) are calculated using **seqeval**, a Python library created specifically for sequence labelling tasks. The Hugging Face **evaluate** library wraps the seqeval and helps us use seqeval more easily inside the **Trainer class** for evaluation. During training, metrics were evaluated on the validation set per epoch. After training (at the end of 20 epochs), we retrieved the F1 score of the model based on its performance on the validation set in the final epoch. The internal formulas of seqeval follow the same strict matching principle as in CoNLL were an entity is correct only if the start index, end index, and the type match exactly. The strict entity-level(full span + type) formulas is as follows:
Precision = Correctly predicted named entities / Total predicted named entities
Recall = Correctly predicted named entities / Total actual named entities

F1 = 2 × (Precision × Recall) / (Precision + Recall)

Accuracy = Number of correctly predicted labels / Total labels
Accuracy is the percentage of exact matched entities.

4.6 Error analysis

Figure 4.4 provides the confusion matrix of the Igbo models which gives a holistic view of the performance of the models. The highest precision and recall throughout many entity classes is recorded by IgbomBERT. The confusion matrix showed that the non-entity token "O" dominates across all models. For B-PER, B-LOC, and B-ORG, all the models showed higher values along the diagonal (true positives), illustrating their ability to correctly predict the beginning of a person's name, location, and organization entity classes. However, lower values displayed by all the models in the following classes: I-PER, I-LOC, and I-ORG indicate misclassification. Classifying B-DATE and I-DATE is seen to be a struggle for all the models shown by the significant false negatives and lower recall, indicating a high ambiguity level in this class.

Figure 4.2: The TrainLoss vs ValidationLoss; Precision, Recall and F1-score of IgboBERT, IgboDistillBERT, IgbomBERT, IgboXLM-R at learning rate 1e-4.

4.7 Chapter Summary

RQ1 was addressed in this chapter by applying the human-annotated method which is one of the most accurate methods of creating a dataset. Additionally, the duplication of this dataset creation process addressed RQ3 of this thesis. In

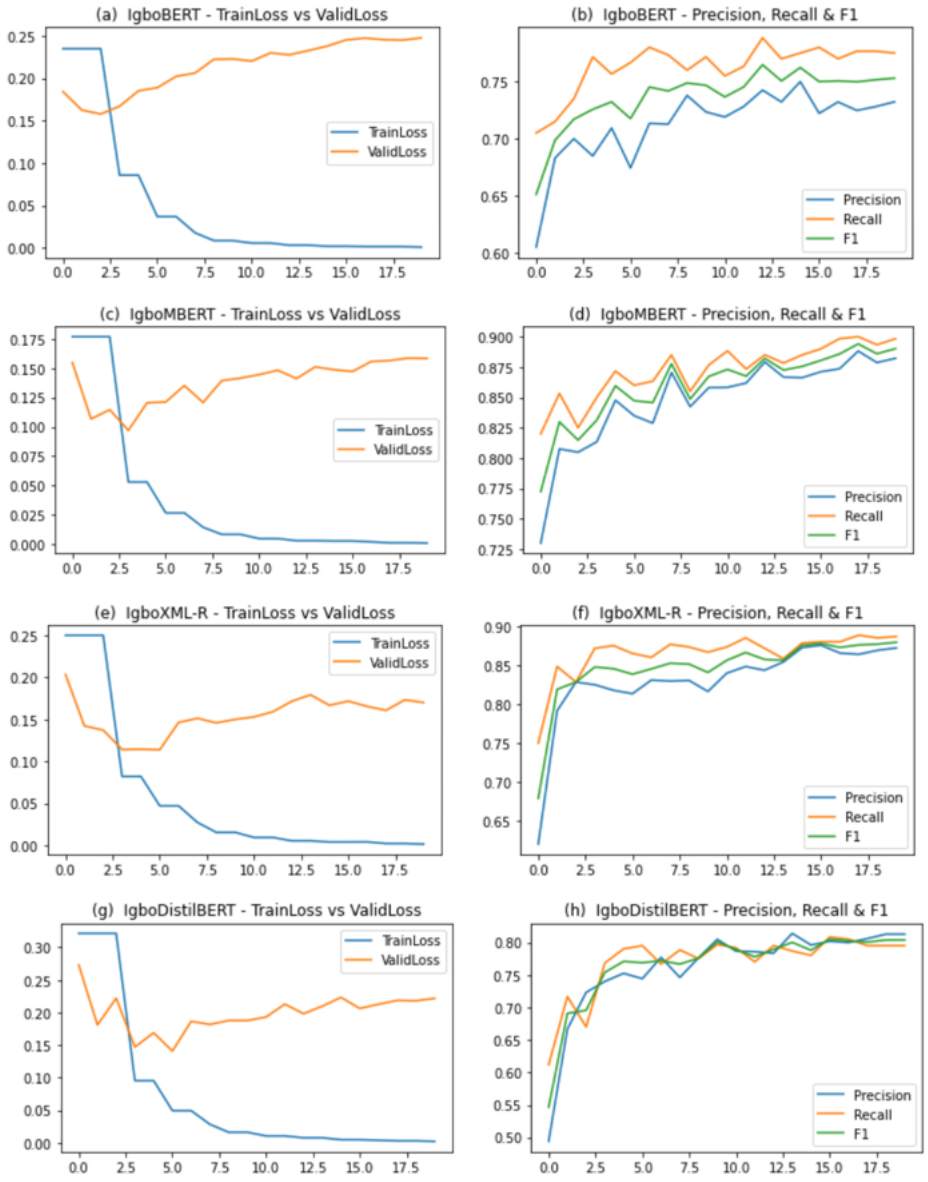


Figure 4.3: The TrainLoss vs ValidationLoss; Precision, Recall and F1-score of IgboBERT, IgboDistillBERT, IgbomBERT, IgboXLM-R at learning rate 2e-5.

this chapter, we developed an IgboBERT model[12], which to our knowledge is the first transformer-based language model pre-trained on the Igbo Language. We fine-tuned it on a downstream NER task with the MasakhaNER data set. Even though the IgboBERT was outperformed as shown by the various F1 scores results in the tables above, we can argue that IgboBERT achieved good performance based on that it was trained on a huge model of 84M parameters and it was pre-trained on relatively small raw data when compared to the millions of data used to pre-

[12]https://huggingface.co/chymaks/IgboBERT-NER-finetuned-Final-Version

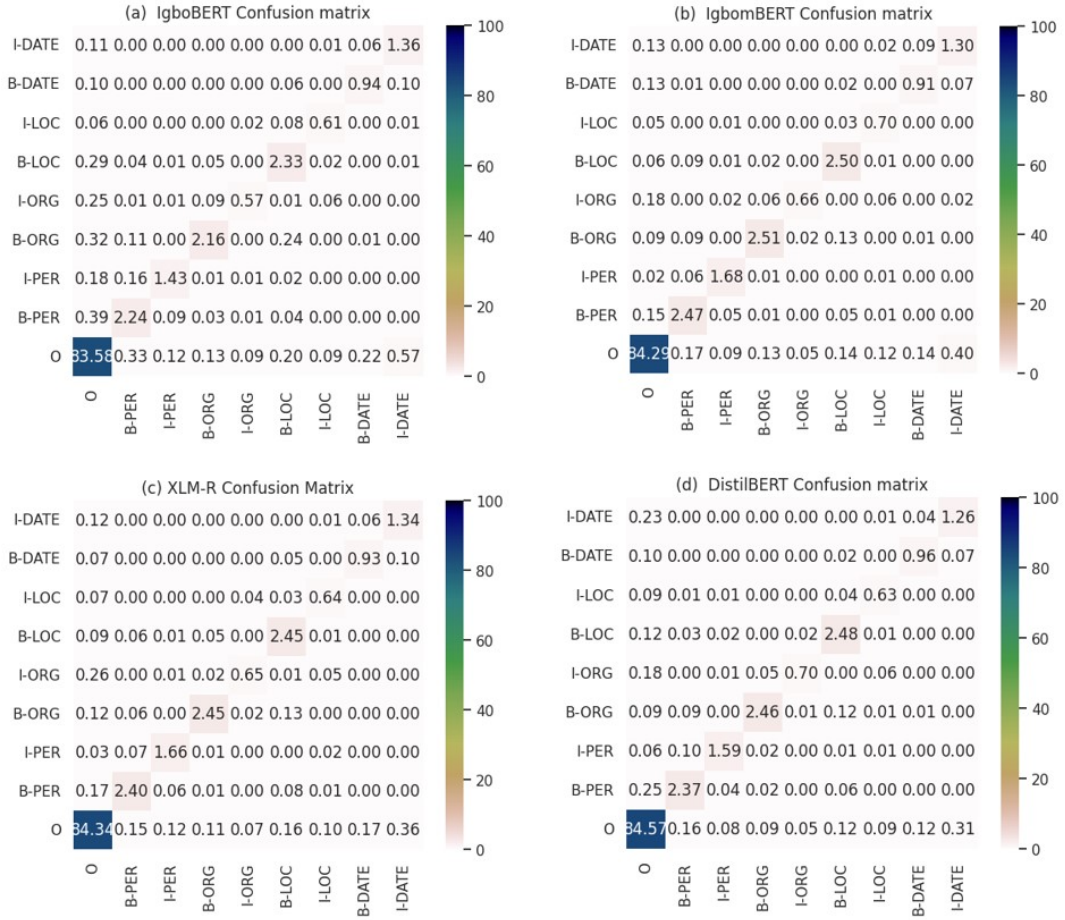


Figure 4.4: The confusion matrix of IgboBERT, IgbomBERT, IgboXML-R, IgboDistillBERT at learning rate 2e-5.

train mBERT, XLM-R, and DistilBERT. This resulted in no convergence in the training vs. validation loss (over-fitting). Given that IgboBERT achieves an F1 of 77.94 with such small data, the introduction of more data for fine-tuning may well improve the performance further.

Chapter 5

Expanding Named Entity Recognition Datasets Via Projection

5.1 Introduction

This Chapter is derived from the published paper titled "IGBONER 2.0: Expanding Named Entity Recognition Datasets via Projection" (C. I. Chukwuneke et al., 2023). In this chapter, a mapping dictionary was created and used to automatically generate and format an NER training dataset from the Igbo monolingual corpus. Research shows that pre-trained language models achieve improved results for many natural language processing (NLP) tasks like Machine Translation (Nekoto et al., 2020), Topic Classification (Michael A Hedderich, David Adelani, et al., 2020), Part of Speech Tagging (Kann, Lacroix, and Søgaard, 2020) and Named Entity Recognition (NER) (David Ifeoluwa Adelani, J. Abbott, et al., 2021; Ogueji, Zhu, and J. Lin, 2021). However, these models are known to require a lot of labeled data to perform well (Q. Xie et al., 2020). For NLP tasks, there is a crucial and constant need for annotated corpora, which is often a challenge for low-resource languages. The African language Igbo is low-resourced for NLP research in general as described in Chapter1, subsection 1.2.3, including for the NER task (David Ifeoluwa Adelani, J. Abbott, et al., 2021). This is a limitation for the IgboNLP research and in the training and evaluation of some state-of-the-art (SOTA) models built in other languages. This is portrayed in the evaluation result in the Chapter 4 of this thesis. IgboNER was created from only 3,190 sentences obtained from BBC Igbo news[1]. This forms our motivation in creating further annotated datasets for the low-resourced Igbo language. The crucial question then becomes how can we efficiently create more annotated data for Igbo NLP tasks?

[1]https://www.bbc.com/igbo

In this work, we adopt the cross-language projection method ((B. Li, Y. He, and Wenjin Xu, 2021); (Ehrmann, Turchi, and Steinberger, 2011); (J. Xie et al., 2018); (Bari, Joty, and Jwalapuram, 2020)) to automatically create more NER datasets as an alternative to the time-consuming and costly human-annotated method. The projection-based method generates a tagged corpus by projecting the tags from the source language to the target language. A parallel corpus is usually required for this method. This study aims to create more annotated NER data for the Igbo language using the projection-based method. Figure 5.1 is an example of a projection-based task. Kulshreshtha, Redondo Garcia, and Chang, 2020 showed that using parallel corpora is better suited for aligning contextual embeddings as the context of words, especially the ambiguous words are maintained within the sentences. English-Igbo parallel corpora were used in this work. Firstly, we chose the English Language because it is the lingua franca of Nigeria where Igbo is one of the official native languages. Secondly, English has been widely researched in NLP and this has produced a lot of NLP resources for English. We tagged the English sentences (source language) using an existing English annotation tool and then projected the tags onto the parallel Igbo sentences (target language). The Igbo language has historically faced a lot of disagreement with the adoption of an official written orthography (Oraka, 1983). This greatly affected the language and resulted in Igbo not having a long written tradition when compared to languages like Arabic, English, and French (Agbo, 2013). This has resulted in the scarcity of Igbo corpora online, and the few Igbo texts written with mixed orthographies. Also, the time, labour, and resource needed to manually annotate a large gold standard corpus are a severe constraint.

The contributions of this chapter include the creation of an additional NER dataset for Igbo Language via annotation projection using parallel corpora to augment the already existing IgboNER datasets. We also built a mapping dictionary containing all the unique entities identified by spaCy[2] with their tags. This to the best of our knowledge is the first IgboNER mapping dictionary, which is an important contribution to Igbo NLP and African NLP at large. Crucially, the dictionary was used to handle the issue of multi-words and spelling variation during the tag projection. Our experiments showed that there is an increase in model performance with increasing data size.

5.2 Related Work

Named Entity Recognition is a term defined as the task of identifying names of organisations, people, currency, time, percentage expression, and geographic locations in text and was introduced at the Sixth Message Understanding Conference (Grishman and B. Sundheim, 1995). With the data-hungry deep neural networks recently used in training the SOTA models for performing NLP tasks as

[2]https://spacy.io/models/en_core_web_sm

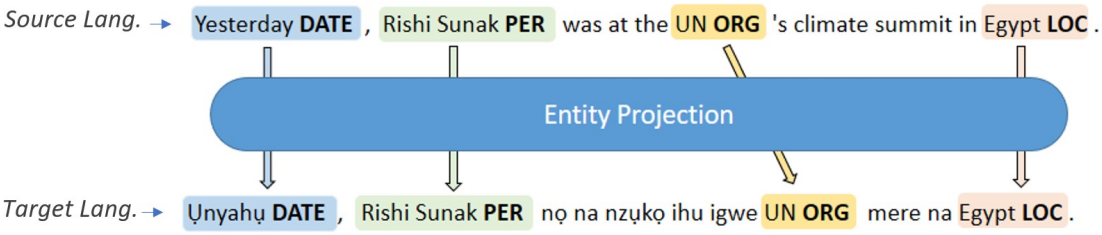


Figure 5.1: Projecting the tags from the source language to the target language.

NER, the need for ever larger annotated datasets in various languages is inevitable. Over the past few years, attention has been drawn to weak or distant supervision[3] and a considerable number of studies (Mintz et al., 2009; Xiao et al., 2020; Liang et al., 2020; Michael A. Hedderich, Lange, and Klakow, 2021; Shruti Rijhwani et al., 2020; Quirk and Poon, 2017; Meng et al., 2021; Hogan, 2022) applied this method to generate a large amount of auto-labelled data without human effort. This method has helped address the issue of data scarcity and showed improvements in performance as seen in the above studies however, the availability of external resources like an existing knowledge base, gazetteers or dictionaries of named entities is a problem for some low-resource languages like Igbo.

The work by Li et al. (B. Li, Y. He, and Wenjin Xu, 2021) built an entity alignment model on top of an XLM-R (Conneau et al., 2020) and projected the labelled entities on the source English language sentences of the parallel data to the target language sentences. They evaluated on four languages: Spanish, Chinese, German and Dutch. Their result showed Chinese to have the highest improvement in F1 score by 6.4%. Fei, M. Zhang, and Ji, 2020 implemented automatic corpus translation of the source gold-standard corpus to the target languages and then projected the labels on the source dataset to the translated target languages. Their evaluation result on the Universal Proposition Bank (UPB, v1.0)[4] dataset showed the translation-based method to be effective with an F1 score increase of 6.7%. Guo Guo and Roth, 2021 translated high-resource labelled sentences to the target language word-by-word with a dictionary. Then, they constructed target-language text from the source-language named entities with a pre-trained language model. The study achieved state-of-the-art performances on LORELEI (low-resource) languages and performed comparatively on CoNLL (high-resource) languages. Garcia et al. García-Ferrero, Agerri, and Rigau, 2022 automatically created data for the target language by translating the gold-labelled English data and then projected the gold labels from the source sentences to

[3]an automatic method of creating labelled data using an external source like an already existing knowledge base, gazetteers or dictionaries of named entities, etc

[4]https://github.com/System-T/UniversalPropositions

the translated sentences by leveraging automatic word alignments. Their result showed that data-based cross-lingual transfer approaches remain a useful option when high-capacity multilingual language models are not available. Enghoff et al. Enghoff, Harrison, and Agić, 2018 extensively studied annotation projection from multiple sources for low-resource NER. They showed that standalone multi-source annotation projection for NER works when the quality and availability of resources are high but suffers when parallel corpora are not available. THE PROJECTOR, an interactive graphical user interface tool that displays the sentence pair with all predicted and projected annotations was designed by Akbik and Vollgraf, 2017. This facilitates experiments and discussions in the research community. The study by Agerri et al., 2018 demonstrated the feasibility of automatically generating Named Entity Recognition (NER) taggers for a given language when no manual data is available. This they achieved with the use of parallel corpora, projecting the existing annotations from multiple source languages to the target language via strict match projections. Their evaluation showed that the automatically generated model outperforms the gold-standard trained model in an in-domain evaluation.

5.3 Data Collection and Annotation

5.3.1 Data collection

Our work used a parallel corpus (English-Igbo) that was created for a machine translation project (Ezeani, Rayson, et al., 2020a). The corpus was created by collecting English and Igbo sentences from local newspapers in Nigeria (e.g. Punch) and from BBC Igbo News website[5] respectively. The 5,630 English sentences collected were translated into Igbo and the 5,503 collected Igbo sentences to English via human translation to produce English-Igbo sentence pairs used to build a standard machine translation benchmark dataset for Igbo. We adopted and used this specific corpus because it contains more contemporary texts in both languages which will guarantee a significant representation of known named entities than the multilingual bible text data[6] often used in the parallel corpus research. In addition, it was assumed that the data quality would be good enough due to the human translation process it has gone through. Table 5.1 describes the data splits that make up the parallel corpus used in this work.

For, the gazetteer, we crawled the Wikipedia[7] and INEC [8] websites for the following entities: Person, Organization, and Location. Table 5.2 summarizes the gazetteers collected.

[5]https://www.bbc.com/igbo
[6]Mostly from https://www.jw.org and other sources
[7]https://en.wikipedia
[8]https://www.inecnigeria.org/

Type	Sentence pairs	Source
Igbo	5,503	BBC News
English	5,630	Local newspapers
Total	11,133	

Table 5.1: Data source. This describes the parallel data sources.

Entity Types	Number
PER	1,188
ORG	1,791
LOC	52,019
Total	54,998

Table 5.2: Total of collected gazetteer entities.

5.3.2 Data Preprocessing

Step 1: We began by manually inspecting the gathered data to identify the delimiter used for separating the sentences and to ensure that each sentence begins on a new line, which is crucial for parsing the parallel data. Upon observation, it was noted that the sentences were delimited by a full stop. However, approximately one-fourth of the sentences did not commence on a new line.
Step 2: A Python script was developed to relocate any sentence following a full stop to a new line. Subsequently, manual adjustments were made to rectify any sentences overlooked by the Python script.
Step 3: We generated sentence pairs using a Python script, pairing English sentences with their corresponding Igbo translations.
Step 4: The sentence pairs underwent manual cross-checking to verify that each English sentence corresponded accurately to its Igbo translation. This meticulous process is essential to ensure consistency and accuracy for further analysis and application.

5.3.3 Data annotation

The key to building the mapping dictionary in the next section is the ability to identify all unique named entities in the source language text. Therefore, the initial data annotation (i.e. NER tagging) process involved running a NER tool over the source language (English in this case) to identify and extract the named entities in the text and their tags.

ANNIE Named Entity Recognizer

The tagging process of the English sentences started with the use of ANNIE[9]-

[9]https://cloud.gate.ac.uk/shopfront/displayItem/annie-named-entity-recognizer

A Nearly-New Information Extraction system. ANNIE is a modular framework designed for information extraction within the General Architecture for Text Engineering (GATE)[10] ANNIE identifies various entity types by default, including Person, Location, Organization, Address, and Date. Additional annotations such as Money, Percent, Token, SpaceToken, and Sentence are also available if selected. ANNIE generates output files in both JSON and XML formats.

Our utilization of ANNIE encountered obstacles due to several factors. Firstly, we encountered limitations in the processing capacity, restricting us to handling less than 100 lines of sentences at a time. Additionally, the format of the annotation output generated by ANNIE did not align with our specific requirements, hindering our ability to extract the necessary information effectively. Furthermore, despite our efforts, we faced challenges in configuring and extending the ANNIE pipeline to accommodate our unique needs and preferences. Overall, these limitations impeded our ability to fully leverage the capabilities of ANNIE for our text-processing tasks. Figure5.2 displays a sample ANNIE output.

{"text":"Happily, Orji Uzor Kalu said in a Facebook post that he had joined the APC because of Buhari's promise that South-East contractors \nwould not go ahead with the project. before November. I joined the APC because Buhari has fulfilled his promise to me and I hope he will do \nbetter. As an award to join the APC, President Buhari handed over several road projects to SLOK Holding contractors. Also, the Niger \nDelta Development Commission provided funding to SLOK Holding to build a road from Mzuhali to Mzuakoli and from Uzuakoli to Ozuitem.\nIn the unfortunate case of Senator Orji Uzor Kalu, his court case with the Economic and Financial Crimes Commission could not be as expected.\nWith his conviction, the prospect of becoming president under the APC in 2023 has come to an abrupt end.","entities":{"Date":[{"indices":[176,184],"rule":"GazDate","ruleFinal":"DateOnlyFinal","kind":"date"},{"indices":[755,759],"rule":"YearContext1","ruleFinal":"DateOnlyFinal","kind":"date"}],"Location":[{"indices":[108,118],"kind":"locName","rule":"LocKey","ruleFinal":"LocFinal"},{"indices":[400,405],"name":"The Republic of Niger","ISO2":"NE","ISO3":"NER","locType":"country","rule":"Location1","ruleFinal":"LocFinal"}],"Organization":[{"indices":[34,42],"orgType":"company","rule":"GazOrganization","ruleFinal":"OrgFinal"},{"indices":[71,74],"rule ":"AcronymOrg","orgType ":"unknown","matches":[478,479,480,481]},{"indices":[199,202],"rule ":"AcronymOrg","orgType ":"unknown","matches":[478,479,480,481]},{"indices":[305,308],"rule ":"AcronymOrg","orgType ":"unknown","matches":[478,479,480,481]},{"indices":[407,435],"orgType":"unknown","rule":"OrgXBase","ruleFinal":"OrgFinal"},{"indices":[615,637],"orgType":"unknown","rule":"OrgXandYKey","ruleFinal":"OrgFinal"},{"indices":[638,655],"orgType":"unknown","rule":"OrgXBase","ruleFinal":"OrgFinal"},{"indices":[748,751],"rule ":"AcronymOrg","orgType ":"unknown","matches":[478,479,480,481]}],"Person":[{"indices":[9,23],"firstName":"Orji","gender":"male","surname":"Uzor","kind":"fullName","rule":"PersonFullDoubleBarrelled","ruleFinal":"PersonFinal"},{"indices":[86,92],"kind":"PN","rule":"Unknown","matches":[455,482,483],"NMRule":"Unknown"},{"indices":[211,217],"kind":"PN","rule":"Unknown","matches":[455,482,483],"NMRule":"Unknown"},{"indices":[310,326],"title":"President","gender":"unknown","surname":"Buhari","kind":"personName","rule":"PersonTitleGenderUnknown","ruleFinal":"PersonFinal","matches":[455,482,483]},{"indices":[567,589],"rule":"PersonTitle1"}]}}

Figure 5.2: JSON output of ANNIE annotation.

spaCy

The English spaCy[11] which, was used to extract the named entities in the English text for this work. The spaCy model pipeline used is the 'en_core_web_sm version 3.4.0' [12] and the model recognized the following tags or named entities describes; PERSON(PER), NORP (Nationalities or religious or political groups), FAC (Faculty) ORG (Organisation), GPE (Geopolitical entity), LOC (Location), PRODUCT, EVENT, WORK_OF_ART, LAW, LANGUAGE, DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL, CARDINAL. We passed our

[10]https://gate.ac.uk/

[11]SpaCy is a free open-source library for Natural Language Processing in Python. SpaCy features NER tagging, part-of-speech tagging, dependency parsing, word vectors, and more: https://spacy.io/

[12]Trained on written web text (blogs, news, comments), that includes vocabulary, syntax, and entities

11,133 English source sentences through SpaCy and it identified named entities in 7,894 sentences. This means that no entity was identified in 3,239 sentences. Table 5.3 describes our annotated data.

Entity Types	# Tagged entities	Cohen's Kappa	Disagreement(%)
PER	6,466	0.9971	0.3438
LOC	3,927	0.9968	0.2677
ORG	4,626	0.9821	2.4584
DATE	3,869	0.9966	3.2090

Table 5.3: Describes our dataset annotation including the entity types, number of tagged entities, Cohen's Kappa inter-annotation scores, and percentage disagreement.

AWESoME aligner

We utilized the AWESoME (Aligning Word Embedding Spaces Of Multilingual Encoders) tool, as outlined in the work by Dou and Graham Neubig (2021), to align our spaCy-annotated English sentences with their corresponding Igbo sentences. This alignment process operates token by token, aiming to transfer tags from English to Igbo seamlessly. However, we encountered a challenge during this process, particularly when aligning English words that correspond to multi-word expressions in Igbo. As a consequence, some sentences exhibited misalignment between the English and Igbo versions.
For clarity, Figure 5.3 provides a snippet of the output generated by the AWESoME aligner, presented in a two-column format. Each line pairs an English word with its aligned Igbo counterpart, separated by a comma. Misalignments are seen in an instance in Line 1 where "Judge Mohammed Idris" is aligned.
This led to the building of a mapping dictionary for our tag transfer.

5.4 Building the Mapping Dictionary

The mapping dictionary is a key contribution of this thesis and could be useful for other NLP tasks such as machine translation and other structured prediction systems. Its creation follows a semi-automatic process described in Figure 5.4. The key idea is based on the assumption that the majority (not all) of the named entities in the source language will either remain the same or translate to unique words or expressions in the target language. For example in Figure 5.1, `Rishi Sunak`, tagged **PER**, is always going to remain the same in both languages as it is a person's name while `Yesterday`, tagged **DATE**, is most likely going to be translated as `Ụnyahụ`. For this work, the mapping dictionary is an actual python dictionary object which has a `key:value` structure. The entities from the English

[Judge_Mohammed_Idris_Person],Orji	[Orji_Uzor_Kalu_Person], Ukwu
[Judge_Mohammed_Idris_Person],Uzor	[Orji_Uzor_Kalu_Person], nke
on,aka	[Orji_Uzor_Kalu_Person], Ikoyi,
the, chịburu	[Orji_Uzor_Kalu_Person], Lagos
[Ikoyi_High_Court_Organization], Financial	by,Akwụkwọ
[Ikoyi_High_Court_Organization], Crimes	the, Akwụkwọ
[Ikoyi_High_Court_Organization], steeti	spokesperson, ozi
[Ikoyi_High_Court_Organization],Abịa	of, nke
[Lagos_Location], Kalu	company,Ltd
[Lagos_Location],akara	[Wilson_Uwujaren_Person], Sun
Fighters,site	[Wilson_Uwujaren_Person], Naịjirịa
Fighters,ahụ,	[Wilson_Uwujaren_Person], onwunwe
[Economic_Economic_Crimes_Commission_Organization],n'okwu	marking, so
[Economic_Economic_Crimes_Commission_Organization],nke	property, Kālu

Figure 5.3: An output from AWESoME align.

sentences are the *keys* while the corresponding Igbo entities and tags are the *values*. For example, looking at the sentences from Figure 5.1, the entry to the dictionary may be as shown below:

```
{
"Yesterday": {ig:"Ụnyahụ", tag:"DATE"}, "Rishi Sunak": {ig:"Rishi Sunak",
tag:"PER"}, "UN": {ig:"UN", tag:"ORG"}, "Egypt" {ig:"Egypt", tag:"LOC"},
}
```

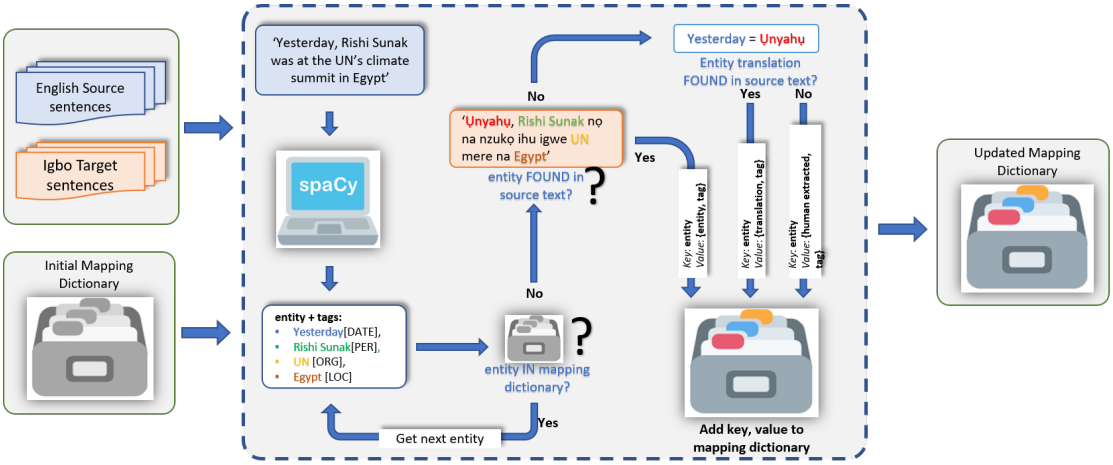


Figure 5.4: An illustration of the semi-automatic process used in creating the mapping dictionary.

As shown in Figure 5.4, the input to the process of building the mapping dictionary is the parallel corpus containing the sentence pairs in both languages as well as an initial mapping dictionary. The initial dictionary could be empty at

the start of the process but may also have been initialized by a previous run. The English sentences are passed through the spaCy pipeline as described in Section 5.3.3 to extract the NER-tagged entities. Each of the extracted entities is passed through the pipeline checking the following conditions:

- if the entity is already in the mapping dictionary, discard and get another entity.
- else if the entity is found in the target (Igbo) text, key = entity, value = entity, tag
- else if the entity's translation is found in the target (Igbo) text, key = entity, value = translated, tag
- else manually annotate the target (Igbo) entity, key = entity, value = annotated, tag

Our experiment with the parallel text used shows that over 95% of the extracted entities from the source language (English) were found in the corresponding target (Igbo) sentence making human translation and annotation a lot easier.

5.4.1 Annotation validation

Table 5.1 showed that our source data is from local news and newspapers which represents the daily communication of speakers of Igbo language. Because of the nature of our data (local news), spaCy's output was noisy. For example, an entity 'Sowore' was tagged as an ORG instead of a PERSON. The tagged spaCy output was therefore manually corrected by two people who are among the authors of this paper following the MUC-6 annotation guide[13] which we adopt in this work. The total number of misclassified entities that we manually corrected is eight hundred and seventy-four (874). The PER entity had five hundred and one (501) misclassification occurrences, which is the highest, followed by the LOC entity. This is because the majority of the PER and LOC are written in the Igbo language, which spaCy did not understand. Agreement was made by them where there were discrepancies in the tags. The inter-annotation agreement was calculated using Cohen's Kappa (Cohen, 1960), which uses one distribution for each rater and takes into account the possibility of the agreement occurring by chance. The inter-annotation percentage disagreement after the manual correction is in Table 5.3. Our dataset achieves an inter-annotator agreement of 0.985, i.e. excellent agreement. In section 3.2, we have the list of output tags by spaCy but for this work, we are using only the PER, LOC, ORG, and DATE tags for our dataset creation. We choose these four tags and replace all the 'GPE' tags with 'LOC', and 'TIME' tags we also replace them with 'DATE' following the previously created IgboNER dataset (C. Chukwuneke et al., 2022). A Python script that selects only

[13]https://cs.nyu.edu/ grishman/muc6.html

the entity to replace was written and used for all the manual corrections. We use the annotated dataset as a train split only, this is to increase the amount of training data for this work. As mentioned in section 5.3, our method requires a parallel corpus and we used the data described in Table 5.1.

5.4.2 Experimental settings

In this section, we describe the IgboBERT 2.0 model we pre-trained in the Igbo language and some state-of-the-art transformer models fine-tuned to the downstream Igbo NER task. We show the effect of dataset sizes on the performance by fine-tuning using various dataset sizes.

5.4.2.1 IgboBERT 2.0 Model

IgboBERT 2.0 is another version of IgboBERT (C. Chukwuneke et al., 2022) language model, which we pre-trained using the same corpus as IgboBERT plus the corpus described in Table 5.1 as an addition to increase the training sentences. A total number of 402,582 training sentences is used. We train IgboBERT 2.0 at a learning rate of 1e-4 for 5 epochs and a batch size of 16. Only 5 epoch training was carried out because of limited computing resources such as GPU at the time of the experiment.

5.4.2.2 Fine-tuned Models

We fine-tune the IgboBERT 2.0 model and some of the state-of-the-art (SOTA) transformer models on the IgboNER downstream task using the Igbo dataset created by C. Chukwuneke et al., 2022, MasakhaNER 2.0 dataset (Adelani et al., 2022) and the dataset created in this work. The fine-tuned SOTA models are:

- Multilingual BERT-base cased (mBERT-base cased) by Devlin et al., 2019.
- XLM-RoBERTa-base (XLM-R) by Conneau et al., 2020.
- DistilBERT by Sanh, 2019

To predict the entity types, we added a linear classification layer to the pre-trained transformer models. 30 epoch training with a batch size of 8 at a learning rate of 1e-4 was run.

5.4.3 Results

Table 5.4 shows the fine-tuned results of mBERT, XML-R, DistilBERT, and IgboBERT 2.0. We combine the 'train' split of the Igbo dataset created by C. Chukwuneke et al., 2022, MasakhaNER 2.0 dataset (Adelani et al., 2022), and the dataset we created in this work. We divide the combination of the 'train' split into 4 different data sizes: 25% (5726), 33% (7635), 50% (11,452), and 100%

(22,904). We then fine-tune the above-listed models with the 4 different data sizes. We validate and test the models with the same validation and test set. The comparative evaluation shows:

- increase in performance with increase in the data size.
- XML-R outperformed others with F1-score of 88.83.
- IgboBERT 2.0 with F1 of 79.25 with data size 22,904 is relatively comparable to the SOTA models in terms of the amount of data used in training these SOTA models.

		Percentage Data Sizes			
Models	**Scores**	**(25%)**	**(33%)**	**(50%)**	**(100%)**
IgBERT 2.0	F1	72.33	76.00	77.10	79.91
	Precision	71.49	74.29	75.93	79.25
	Recall	73.20	77.79	78.31	80.58
mBERT	F1	80.93	82.64	83.58	87.10
	Precision	79.71	80.68	82.54	85.71
	Recall	82.21	84.71	84.65	88.55
XLM-R	F1	84.05	83.29	86.56	88.83
	Precision	82.68	80.55	84.82	87.35
	Recall	85.47	86.22	88.37	90.35
DistilBERT	F1	76.68	80.32	80.19	83.77
	Precision	76.50	80.81	79.91	83.81
	Recall	76.86	79.83	80.47	83.72

Table 5.4: Performance of mBERT, XML-R, DistilBERT and IgboBERT 2.0. We display the fine-tuned results of the models after 30 epochs at 1e-4 learning rate with varying dataset sizes.

5.4.4 Challenges of the Projection Method

We encountered some errors during the process of transferring the tags from English to Igbo sentences. This is as a result of the use of various orthographies during the translation of the parallel corpus. Some of these errors include:

- Missing entities and words in the Igbo translated text. For example, the sentence: `"Chadwick Boseman who acted as T'Challa, king of Wakanda, who the movie 'Black Panther' was made for, led other actors to Los Angeles where the occasion was held."` has this as the Igbo translation `"duuru ndị ọzọ nọ n'ihe nkiri a gaa ebe emere mmeme a na los angeles"` missing out the first part of the sentence "Chadwick Boseman

who acted as T'Challa, king of Wakanda, who the movie 'Black Panther' was made for" in the translation.

- Variations in the spelling of various Igbo words. For example, the word `"Nigeria"` has the following translations `("Nigeria","Naijiria", "Naija")},`.

- Variations in translation of some words because of lack of standardized words for the Igbo words. The majority of these words are described in words in Igbo. For example, `"8:30am"` will be translated as `"O jiri nkeji iri atọ wee gafe elekere asatọ nke ụtụtụ"` .

We addressed these challenges by adding multiple values of these word variations to one dictionary key in the mapping dictionary. Different entities were identified with different colors.

5.5 Visualising IgboNER

This section demonstrates the visualisation of IgboNER using spaCy. This provides an interactive way to analyse and understand the performance of our results in Igbo text. Highlighting entities and their tags in different colors makes it easy to see the tagged words aiding comprehension for non-technical users. It can serve as an educational tool for linguists or learners of Igbo to understand how language structures are recognized computationally. Additionally, this allows for immediate feedback while evaluating Igbo NER models, making it easier to spot wrongly tagged names or missed entities.

We built a custom NER pipeline based on using spaCy's EntityRuler specifically for Igbo language. The EntityRuler implemented a pattern-based rule system to recognise the specific named entities (LOC, ORG, PER, DATE) from the dataset we created. Streamlit, an open source Python library, was used to deploy the interactive application on the Web. The following steps describe the code for creating our visualisation tool:

- Install **spaCy** and **Streamlit** libraries in the programming environment (Google Colab in our case)

- We create the Streamlit App and load spaCy.

- The entity lists (locations, organizations, persons, dates) saved in four separate text files are loaded.

- The lists are processed and labelled patterns (LOC, ORG, PER, DATE) for each entity are created.

- The spaCy English NLP pipeline model **(en _core _web _sm)** is initialised while disabling its built-in NER and parser components.
- A custom **EntityRuler** is added to the pipeline and entity patterns created in step 3 is added into it.
- Add a text input box using **st.text_area**.
- An Igbo text file is read in and processed with the modified NLP pipeline to identify the custom entities.
- Each named entity recognised in the text file is highlighted and visually displayed using **displacy.render()** function in the Colab environment.
- To render the output in Streamlit, convert **displacy** HTML to be safely embedded in Streamlit and Run the app.
- On the browser interface, input text and click enter to view the tagged entities with colors as shown in Figure 5.5

5.6 Chapter Summary

This chapter highlights the lack of adequate datasets for building NLP models for low-resource languages, emphasizing the named entity recognition task for the Igbo language. A novel and efficient approach was proposed that takes advantage of parallel data in English - a well-resourced language with highly developed NER tools - and Igbo addressing RQ1 and RQ5 This approach is based on the idea that the majority of name entities in the English text (mostly proper nouns) remain untranslated in the corresponding Igbo text. The concept of a mapping dictionary (see Figure 5.4) was introduced by creating a rule-based pipeline that enables the automatic transfer of the NER tags produced with the English NER tagger to the corresponding Igbo entities. With manual checks and corrections, this process not only speeds up the data creation effort but also produces the mapping dictionary which can serve as a key resource in other NLP tasks such as machine translation, and part-of-speech tagging. Our experiments show that model performance improves with an increase in the data size when fine-tuned in the NER downstream task. Our IgboBERT 2.0 though was outperformed by the other models as shown in the tables but the performance is comparative when compared to the huge amount of data used in pre-training these SOTA models. We hope to improve our model further by creating further IgboNER datasets with the help of the NER mapping dictionary[14] created in this work.

[14]https://github.com/Chiamakac/IgboNER-Models/tree/main/IgboNER_2.0

Figure 5.5: IgboNER visualisation.

Section C
NER for African Languages

Chapter 6

Named Entity Recognition for African languages

6.1 Introduction

This chapter is derived from the published paper titled "MasakhaNER: Named Entity for African Languages"(David Ifeoluwa Adelani, J. Abbott, et al., 2021) where I served as the team lead for dataset annotation for the Igbo language. We created the largest human-annotated NER dataset for 10 African languages including Igbo. Africa has over 2,000 spoken languages (David M. Eberhard, Gary F. Simons, and (eds.), 2023); however, these languages are scarcely represented in existing natural language processing (NLP) datasets, research, and tools (Martinus and J. Z. Abbott, 2019). Nekoto et al. (2020) investigate the reasons for these disparities by examining how NLP for low-resource languages is constrained by several societal factors. One of these factors is the geographical and language diversity of NLP researchers. For example, of the 2695 affiliations of authors whose works were published at the five major NLP conferences in 2019, only five were from African institutions (Caines, 2019). Conversely, many NLP tasks such as machine translation, text classification, part-of-speech tagging, and named entity recognition would benefit from the knowledge of native speakers who are involved in the development of datasets and models.

We focus on named entity recognition (NER)—one of the most impactful tasks in NLP (Sang and Meulder, 2003; Lample et al., 2016). NER is an important information extraction task and an essential component of numerous products including spell-checkers, localization of voice and dialogue systems, and conversational agents. It also enables the identification of African names, places, and organizations for information retrieval. African languages are under-represented in this crucial task due to the identification of a lack of datasets, reproducible results, and researchers who understand the challenges that such languages present for NER.

We take an initial step towards improving the representation of African languages for the NER task, making the following contributions:

1. We bring together language speakers, dataset curators, NLP practitioners, and evaluation experts to address the challenges facing NER for African languages. Based on the availability of online news corpora and language annotators, we develop NER datasets, models, and evaluations covering ten widely spoken African languages.

2. We curate NER datasets from local sources to ensure the relevance of future research for native speakers of the respective languages.

3. We train and evaluate multiple NER models for all ten languages. Our experiments provide insights into the transfer across languages and highlight open challenges.

4. We release the datasets, code, and models to facilitate future research on the specific challenges raised by NER for African languages.

6.2 Related Work

African NER datasets

NER is a well-studied sequence labeling task (V. Yadav and Bethard, 2018) and has been subject of many shared tasks in different languages (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003; K. Shaalan, 2014; Benikova, Biemann, and Reznicek, 2014). However, most of the available datasets are from high-resource languages. Although there have been efforts to create NER datasets for lower-resourced languages, such as the WikiAnn corpus (Pan et al., 2017) covering 282 languages, such datasets consist of "silver-standard" labels created by transferring annotations from English to other languages through cross-lingual links in knowledge bases. Because the WikiAnn corpus data comes from Wikipedia, it includes some African languages; though most have fewer than 10k tokens.
Other NER datasets for African languages are SADiLaR (Eiselen, 2016) for ten South African languages based on government data, and small corpora of less than 2K sentences for Yorùbá (Jesujoba Alabi et al., 2020) and Hausa (Michael A Hedderich, David Adelani, et al., 2020). Additionally, LORELEI language packs (Strassel and Tracey, 2016) were created for some African languages including Yorùbá, Hausa, Amharic, Somali, Twi, Swahili, Wolof, Kinyarwanda, and Zulu, but are not publicly available.

NER models

Popular sequence labeling models for NER are CRF (Lafferty, McCallum, and Pereira, 2001), CNN-BiLSTM (Chiu and E. Nichols, 2016), BiLSTM-CRF (Huang, Wei Xu, and Yu, 2015), and CNN-BiLSTM-CRF (Ma and Hovy, 2016). The traditional CRF makes use of hand-crafted features like part-of-speech tags, context words, and word capitalization. Neural network NER models are initialized with word embeddings like Word2Vec (Mikolov et al., 2013), GloVe (Pennington, Socher, and Manning, 2014) and FastText (Bojanowski et al., 2017). More recently, pre-trained language models such as BERT (Devlin et al., 2019), RoBERTa (Y. Liu et al., 2019), and LUKE (Yamada et al., 2020) produce state-of-the-art results for the NER task. Multilingual variants of these models like mBERT and XLM-RoBERTa (Conneau et al., 2020) make it possible to train NER models for several languages using transfer learning. Language-specific parameters and adaptation to unlabelled data of the target language have yielded further gains (Pfeiffer, Vuli, et al., 2020; Pfeiffer, Vulić, et al., 2020).

6.3 Focus Languages

Table 6.1 provides an overview of the languages considered in this work, their language family, number of speakers, and the regions in Africa where they are spoken. We chose to focus on these languages due to the availability of online news corpora, annotators, and most importantly because they are widely spoken native African languages. Both region and language family might indicate a notion of proximity for NER, either because of linguistic features shared within that family, or because data sources cover a common set of locally relevant entities. We highlight language specifics for each language to illustrate the diversity of this selection of languages in Section 6.3.1, and then showcase the differences in named entities across these languages in Section 6.3.2.

6.3.1 Language Characteristics

Amharic
(amh) uses the Fidel script consisting of 33 basic scripts ሀ (hä) ለ (lä) መ (mä) ሠ (šä) ...) with at least 7 vowel sequences (such as ሀ (hä) ሁ (hu) ሂ (hi) ሃ (ha) ሄ (he) ህ (hi) ሆ (ho)). This results in more than 231 characters or Fidels. Numbers and punctuation marks are also represented uniquely with specific Fidels (፩ (1), ፪ (2), ... and ። (.), !(!), ፤ (;),).

Hausa
(hau) has 23-25 consonants, depending on the dialect, and five short and five long vowels. Hausa has labialized phonemic consonants, as in /gw/ e.g. ('agwagwa'). As found in some African languages, implosive consonants also exist in Hausa,

Languages	**Family**	**Speakers**	**Region**
Amharic	Afro-Asiatic-Ethio-Semitic	33M	East
Hausa	Afro-Asiatic-Chadic	63M	West
Igbo	Niger-Congo-Volta-Niger	27M	West
Kinyarwanda	Niger-Congo-Bantu	12M	East
Luganda	Niger-Congo-Bantu	7M	East
Luo	Nilo Saharan	4M	East
Nigerian-Pidgin	English Creole	75M	West
Swahili	Niger-Congo-Bantu	98M	Central & East
Wolof	Niger-Congo-Senegambia	5M	West & NW
Yorùbá	Niger-Congo-Volta-Niger	42M	West

Table 6.1: Language, family, number of speaker (David M. Eberhard, Gary F. Simons, and (eds.), 2023), and regions in Africa.

(e.g. ‘b, ‘d, etc as in ‘barna’). Similarly, the Hausa approximant ‘r’ is realized in two distinct manners: roll and trill, as in ‘rai’ and ‘ra’ayi’, respectively.

Igbo
(`ibo`) is an agglutinative language, with frequent suffixes and prefixes (Emenanjo, 1978). A single stem can yield many word forms by the addition of affixes that extend its original meaning (Ikechukwu E Onyenwe and Hepple, 2016). Igbo is also tonal, with two distinctive tones (high and low) and a down-stepped high tone in some cases. The alphabet consists of 28 consonants and 8 vowels (A, E, I, Ị, O, Ọ, U, Ụ). In addition to the Latin letters (except *c*), Igbo contains the following digraphs: (ch, gb, gh, gw, kp, kw, nw, ny, sh).

Kinyarwanda
(`kin`) makes use of 24 Latin characters with 5 vowels similar to English and 19 consonants excluding q and x. Moreover, Kinyarwanda has 74 additional complex consonants (such as mb, mpw, and njyw), (Government, 2014). It is a tonal language with three tones: low (no diacritic), high (signaled by “/”), and falling (signaled by “ˆ”). The default word order is Subject-Verb-Object.

Luganda
(`lug`) is a tonal language with subject-verb-object word order. The Luganda alphabet is composed of 24 letters that include 17 consonants (p, v, f, m, d, t, l, r, n, z, s, j, c, g), 5 vowel sounds represented in the five alphabetical symbols (a, e, i, o, u) and 2 semi-vowels (w, y). It also has a special consonant ng'.

Luo
(`luo`) is a tonal language with 4 tones (high, low, falling, rising) although the tonality is not marked in orthography. It has 26 Latin consonants without Latin letters (c, q, v, x, and z) and additional consonants (ch, dh, mb, nd, ng', ng, ny, nj, th, sh). There are nine vowels (a, e, i, o, u, ɐ, , ɔ o, u, Ω,) which are distinguished primarily by advanced tongue root (ATR) harmony (De Pauw, Wagacha, and Abade, 2007).

Nigerian-Pidgin
(`pcm`) is a largely oral, national lingua franca with a distinct phonology from English, its lexifier language. Portuguese, French and especially indigenous languages form the substrate of lexical, phonological, syntactic, and semantic influence on Nigerian-Pidgin (NP). English lexical items absorbed by NP are often phonologically closer to indigenous Nigerian languages, notably in the realization of vowels. As a rapidly evolving language, NP orthography is undergoing codification and indigenization (Offiong Mensah (2012), Ojarikre (2013)).

Swahili
(`swa`) is the most widely spoken language on the African continent. It has 30 letters including 24 Latin letters without characters (q and x) and six additional consonants (ch, dh, gh, ng', sh, th) unique to Swahili pronunciation.

Wolof
(`wol`) has an alphabet similar to that of French. It consists of 29 characters, including all letters of the French alphabet except H, V, and Z. It also includes the characters Ɗ ("ng", lowercase: ŋ) and Ñ ("gn" as in Spanish). Accents are present, but limited in number (À, É, Ë, Ó). However, unlike many other Niger-Congo languages, Wolof is not a tonal language.

Yorùbá
(`yor`) has 25 Latin letters without the Latin characters (c, q, v, x, and z) and with additional letters (ẹ, gb, ṣ, ọ). Yorùbá is a tonal language with three tones: low ("\"), middle ("−", optional) and high ("/"). The tonal marks and underdots are referred to as diacritics and they are needed for the correct pronunciation of a word. Yorùbá is a highly isolating language and the sentence structure follows Subject-Verb-Object.

6.3.2 Named Entities

Most of the work on NER is centered around English, and it is unclear how well existing models can generalize to other languages in terms of sentence structure or surface forms. In J. Hu et al., 2020's evaluation on cross-lingual generalization for NER, only two African languages were considered and it was seen that transformer-based models particularly struggled to generalize to named entities in Swahili.

Language	Sentence
English	The Emir of Kano turbaned Zhang who has spent 18 years in Nigeria
Amharic	የካኖ ኢምር በናይጀርያ ፩፰ ዓመት ያሳለፈውን ዣንግን ዋና መሪ አደረጉት
Hausa	Sarkin Kano yayi wa Zhang wanda yayi shekara 18 a Nigeria sarauta
Igbo	Onye Emir nke Kano kpubere Zhang okpu onye nke nọgoro afọ iri na asatọ na Naijiria
Kinyarwanda	Emir w'i Kano yimitse Zhang wari umaze imyaka 18 muri Nijeriya
Luganda	Emir w'e Kano yatikkidde Zhang amaze emyaka 18 mu Nigeria
Luo	Emir mar Kano ne orwakone turban Zhang ma osedak Nigeria kwuom higni 18
Nigerian-Pidgin	Emir of Kano turban Zhang wey don spend 18 years for Nigeria
Swahili	Emir wa Kano alimvisha kilemba Zhang ambaye alikaa miaka 18 nchini Nigeria
Wolof	Emiiru Kanó dafa kaala kii di Zhang mii def Nigeria fukki at ak juróom ñett
Yorùbá	Ẹ́míà ìlú Kánò wé láwàní lé orí Zhang ẹni tí ó ti lo ọ́dún méjìdínlógún ní orílè-èdè Nàìjíríà

Figure 6.1: Example of named entities in different languages. PER, LOC, and DATE are in colors purple, orange, and green respectively. The original sentence is from BBC Pidgin.[2]

To highlight the differences across our focus languages, Figure 6.1 shows an English[1] example sentence, with color-coded PER, LOC, and DATE entities, and the corresponding translations. The following characteristics of the languages in our dataset could pose challenges for NER systems developed for English:

- Amharic shares no lexical overlap with the English source sentence.
- While "Zhang" is identical across all Latin-script languages, "Kano" features accents in Wolof and Yorùbá due to its localization.
- The Fidel script has no capitalization, which could hinder transfer from other languages.
- Igbo, Wolof, and Yorùbá all use diacritics, which are not present in the English alphabet.
- The surface form of named entities (NE) is the same in English and Nigerian-Pidgin, but there exist lexical differences (e.g. in terms of how time is realized).
- Between the 10 African languages, "Nigeria" is spelled in 6 different ways.
- Numerical "18": Igbo, Wolof and Yorùbá write out their numbers, resulting in different numbers of tokens for the entity span.

6.4 Data and Annotation Methodology

Our data was obtained from local news sources, to ensure relevance of the dataset for native speakers from those regions. The dataset was annotated using the ELISA

[1]Although the original sentence is from BBC Pidgin https://www.bbc.com/pidgin/tori-51702073

tool (Y. Lin et al., 2018) by native speakers who come from the same regions as the news sources and volunteered through the *Masakhane* community[3]. Annotators were not paid but are all part of the authors of the paper that this chapter is based on. The annotators were trained on how to perform NER annotation using the MUC-6 annotation guide[4]. We annotated four entity types: Personal name (`PER`), Location (`LOC`), Organization (`ORG`), and date & time (`DATE`). The annotated entities were inspired by the English CoNLL-2003 Corpus (Tjong Kim Sang, 2002). We replaced the `MISC` tag with the `DATE` tag following Jesujoba Alabi et al. (2020) as the `MISC` tag may be ill-defined and cause disagreement among non-expert annotators. We report the number of annotators and general statistics of the datasets in Table 6.2. For each language, we divided the annotated data into training, development, and test splits consisting of 70% training, 10%, and 20% of the data respectively.

Language	**Data Source**	**Train/ dev/ test**	**# Anno.**	**PER**	**ORG**	**LOC**	**DATE**	**% of Entities in Tokens**	**# Tokens**
Amharic	DW & BBC	1750/ 250/ 500	4	730	403	1,420	580	15.13	37,032
Hausa	VOA Hausa	1903/ 272/ 545	3	1,490	766	2,779	922	12.17	80,152
Igbo	BBC Igbo	2233/ 319/ 638	6	1,603	1,292	1,677	690	13.15	61,668
Kinyarwanda	IGIHE news	2110/ 301/ 604	2	1,366	1,038	2096	792	12.85	68,819
Luganda	BUKEDDE news	2003/ 200/ 401	3	1,868	838	943	574	14.81	46,615
Luo	Ramogi FM news	644/ 92/ 185	2	557	286	666	343	14.95	26,303
Nigerian-Pidgin	BBC Pidgin	2100/ 300/ 600	5	2,602	1,042	1,317	1,242	13.25	76,063
Swahili	VOA Swahili	2104/ 300/ 602	6	1,702	960	2,842	940	12.48	79,272
Wolof	Lu Defu Waxu & Saabal	1,871/ 267/ 536	2	731	245	836	206	6.02	52,872
Yorùbá	GV & VON news	2124/ 303/ 608	5	1,039	835	1,627	853	11.57	83,285

Table 6.2: Statistics of our datasets including their source, number of sentences in each split, number of annotators, number of entities of each label type, percentage of tokens that are named entities, and total number of tokens.

A key objective of our annotation procedure was to create high-quality datasets by ensuring a high annotator agreement. To achieve high agreement scores, we ran collaborative workshops for each language, which allowed annotators to discuss any disagreements. ELISA provides an entity-level F1 score and also an interface for annotators to correct their mistakes, making it easy to achieve inter-annotator agreement scores between 0.96 and 1.0 for all languages.

We report inter-annotator agreement scores in Table 7.2 using Fleiss' Kappa (Fleiss, 1971) at both the token and entity level. The latter considers each span an annotator proposed as an entity. As a result of our workshops, all our datasets have exceptionally high inter-annotator agreement. For Kinyarwanda, Luo, Swahili, and Wolof, we report perfect inter-annotator agreement scores ($\kappa = 1$). For each of these languages, two annotators annotated each token and were instructed to discuss and resolve conflicts among themselves. The Appendix provides a detailed

[3] `https://www.masakhane.io`

[4] `https://cs.nyu.edu/~grishman/muc6.html`

entity-level confusion matrix in Table A.1.

Dataset	Token Fleiss' Kappa	Entity Fleiss' Kappa	Disagreement from Type
amh	0.987	0.959	0.044
hau	0.988	0.962	0.097
ibo	0.995	0.983	0.071
kin	1.0	1.0	0.0
lug	0.997	0.99	0.023
luo	1.0	1.0	0.0
pcm	0.989	0.966	0.048
swa	1.0	1.0	0.0
wol	1.0	1.0	0.0
yor	0.99	0.964	0.079

Table 6.3: Inter-annotator agreement for our datasets calculated using Fleiss' kappa (κ) at the token and entity level.

6.5 Experimental Setup

6.5.1 NER baseline models

To evaluate baseline performance on our dataset, we experiment with three popular NER models: CNN-BiLSTM-CRF, multilingual BERT (mBERT), and XLM-RoBERTa (XLM-R). The latter two models are implemented using the HuggingFace transformers toolkit (Wolf, Debut, Victor Sanh, Chaumond, Delangue, Moi, Cistac, Rault, R'emi Louf, et al., 2019). For each language, we train the model on the in-language training data and evaluate its test data.

CNN-BiLSTM-CRF
This architecture was proposed for NER by (Ma and Hovy, 2016). For each input sequence, we first compute the vector representation for each word by concatenating character-level encodings from a CNN and vector embeddings for each word. Following Shruti Rijhwani et al. (2020), we use randomly initialized word embeddings since we do not have high-quality pre-trained embeddings for all the languages in our dataset. Our model is implemented using the DyNet toolkit (Graham Neubig et al., 2017).

mBERT
We fine-tune multilingual BERT (Devlin et al., 2019) on our NER corpus by adding a linear classification layer to the pre-trained transformer model, and train it end-to-end. mBERT was trained in 104 languages including only two African

languages: Swahili and Yorùbá. We use the mBERT-base cased model with 12-layer Transformer blocks consisting of 768-hidden size and 110M parameters.

XLM-R
XLM-R (Conneau et al., 2020) was trained on 100 languages including Amharic, Hausa, and Swahili. The major differences between XLM-R and mBERT are (1) XLM-R was trained on Common Crawl while mBERT was trained on Wikipedia; (2) XLM-R is based on Roberta, which is trained with a masked language model (MLM) objective while mBERT was additionally trained with a next sentence prediction objective. We make use of the XLM-R base and large models for the baseline models. The XLM-R-base model consists of 12 layers, with a hidden size of 768 and 270M parameters. On the other hand, the XLM-R-large has 24 layers, with a hidden size of 1024 and 550M parameters.

MeanE-BiLSTM
This is a simple BiLSTM model with an additional linear classifier. For each input sequence, we first extract a sentence embedding from the mBERT or XLM-R language model (LM) before passing it into the BiLSTM model. Following Reimers and Gurevych (2019), we make use of the mean of the 12-layer output embeddings of the LM (i.e. *MeanE*). This has been shown to provide better sentence representations than the embedding of the `[CLS]` token used for fine-tuning mBERT and XLM-R.

Language BERT
The mBERT and the XLM-R models only support two and three languages under study respectively. One effective approach to adapt the pre-trained transformer models to new domains is "domain-adaptive fine-tuning (Howard and Sebastian Ruder, 2018; Gururangan et al., 2020)—fine-tuning on unlabeled data in the new domain, which also works very well when adapting to a new language (Pfeiffer, Vuli, et al., 2020; Jesujoba Alabi et al., 2020). For each of the African languages, we performed *language-adaptive fine-tuning* on available unlabeled corpora mostly from JW300 (Agić and Vulić, 2019), indigenous news sources and XLM-R Common Crawl corpora (Conneau et al., 2020). The appendix provides the details of the unlabeled corpora in Table A.2. This approach is quite useful for languages whose scripts are not supported by the multi-lingual transformer models like Amharic where we replace the vocabulary of mBERT by an Amharic vocabulary before we perform language-adaptive fine-tuning, similar to (Jesujoba Alabi et al., 2020).

6.5.2 Improving the Baseline Models

In this section, we consider techniques to improve the baseline models such as utilizing gazetteers, transfer learning from other domains and languages, and aggregating NER datasets by regions. For these experiments, we focus on the `PER`,

ORG, and LOC categories, because the gazetteers from Wikipedia do not contain DATE entities and some source domains and languages that we transfer from do not have the DATE annotation. We apply these modifications to the XLM-R model because it generally outperforms mBERT in our experiments (see section 6.6).

6.5.2.1 Gazetteers for NER

Gazetteers are lists of named entities collected from manually crafted resources such as GeoNames or Wikipedia. Before the widespread adoption of neural networks, NER methods used gazetteers-based features to improve performance (Ratinov and Roth, 2009). These features are created for each n-gram in the dataset and are typically binary-valued, indicating whether the n-gram is present in the gazetteer.

Recently, Shruti Rijhwani et al., 2020 showed that augmenting the neural CNN-BiLSTM-CRF model with gazetteer features can improve NER performance for low-resource languages. We conduct similar experiments on the languages in our dataset, using entity lists from Wikipedia as gazetteers. For Luo and Nigerian-Pidgin, which do not have their own Wikipedia, we use entity lists from English Wikipedia.

6.5.2.2 Transfer Learning

Here, we focus on cross-domain transfer from Wikipedia to the news domain, and cross-lingual transfer from English and Swahili NER datasets to the other languages in our dataset.

Domain Adaptation from WikiAnn
We make use of the WikiAnn corpus (Pan et al., 2017), which is available for five of the languages in our dataset: Amharic, Igbo, Kinyarwanda, Swahili, and Yorùbá. For each language, the corpus contains 100 sentences in each of the training, development, and test splits except for Swahili, which contains 1K sentences in each split. For each language, we train on the corresponding WikiAnn training set and either zero-shot transfer to our respective test set or additionally fine-tune our training data.

Cross-lingual transfer
For training the cross-lingual transfer models, we use the CoNLL-2003[5] NER dataset in English with over 14K training sentences and our annotated corpus. The reason for CoNLL-2003 is that it is in the same news domain as our annotated corpus. We also make use of the languages that are supported by the XLM-R

[5]We also tried OntoNotes 5.0 by combining FAC & ORG as "ORG" and GPE & LOC as "LOC" and others as "O" except "PER", but it gave a lower performance in zero-shot transfer (19.38 F1) while CoNLL-2003 gave 37.15 F1.

model and are widely spoken in East and West Africa like Swahili and Hausa. The English corpus has been shown to transfer very well to low-resource languages (Michael A Hedderich, David Adelani, et al., 2020; Lauscher et al., 2020). We first train on either the English CoNLL-2003 data or our training data in Swahili, Hausa, or Nigerian-Pidgin before testing on the target African languages.

6.5.3 Aggregating Languages by Regions

As previously illustrated in Figure 6.1, several entities have the same form in different languages while some entities may be more common in the region where the language is spoken. To study the performance of NER models across geographical areas, we combine languages based on the region of Africa that they are spoken in (see Table 6.1): (1) East region with Kinyarwanda, Luganda, Luo, and Swahili; (2) West Region with Hausa, Igbo, Nigerian-Pidgin, Wolof, and Yorùbá languages, (3) East and West regions—all languages except Amharic because of its distinct writing system.

6.6 Results

6.6.1 Baseline Models

Table 6.4 gives the F1-score obtained by CNN-BiLSTM-CRF, mBERT, and XLM-R models on the test sets of the ten African languages when training on our in-language data. We additionally indicate whether the language is supported by the pre-trained language models (✓). The percentage of entities that are out-of-vocabulary (OOV; entities in the test set that are not present in the training set) is also reported alongside the results of the baseline models. In general, the datasets with greater numbers of OOV entities have lower performance with the CNN-BiLSTM-CRF model, while those with lower OOV rates (Hausa, Igbo, Swahili) have higher performance. We find that the CNN-BiLSTM-CRF model performs worse than fine-tuning mBERT and XLM-R models end-to-end (FTune). We expect performance to be better (e.g., for Amharic and Nigerian-Pidgin with over 18 F1 point difference) when using pre-trained word embeddings for the initialization of the BiLSTM model rather than random initialization (we leave this for future work as discussed in section 6.7).

Interestingly, the pre-trained language models (PLMs) have reasonable performance even in languages they were not trained on such as Igbo, Kinyarwanda, Luganda, Luo, and Wolof. However, languages supported by the PLM tend to have better performance overall. We observe that fine-tuned XLM-R-base models have significantly better performance in five languages; two of the languages (Amharic and Swahili) are supported by the pre-trained XLM-R. Similarly, fine-tuning mBERT has better performance for Yorùbá since the language is part of

the PLM's training corpus. Although mBERT is trained on Swahili, XLM-R-base shows better performance. This observation is consistent with Hu2020xtreme and could be because XLM-R is trained on more Swahili text (Common Crawl with 275M tokens) whereas mBERT is trained on a smaller corpus from Wikipedia (6M tokens[6]).

Lang.	In mBERT?	In XLM-R?	% OOV in Test Entities	CNN-BiLSTM CRF	mBERT-base MeanE / FTune	XLM-R-base MeanE / FTune	XLM-R Large FTune	lang. BERT FTune	lang. XLM-R FTune
amh	✗	✓	72.94	52.08	0.0 / 0.0	63.57 / 70.62	76.18	60.89	**77.97**
hau	✗	✓	33.40	83.52	81.49 / 86.65	86.06 / 89.50	90.54	91.31	**91.47**
ibo	✗	✗	46.56	80.02	76.17 / 85.19	73.47 / 84.78	84.12	86.75	**87.74**
kin	✗	✗	57.85	62.97	65.85 / 72.20	63.66 / 73.32	73.75	77.57	**77.76**
lug	✗	✗	61.12	74.67	70.38 / 80.36	68.15 / 79.69	81.57	83.44	**84.70**
luo	✗	✗	65.18	65.98	56.56 / 74.22	52.57 / 74.86	73.58	**75.59**	75.27
pcm	✗	✗	61.26	67.67	81.87 / 87.23	81.93 / 87.26	89.02	89.95	**90.00**
swa	✓	✓	40.97	78.24	83.08 / 86.80	84.33 / 87.37	89.36	89.36	**89.46**
wol	✗	✗	69.73	59.70	57.21 / 64.52	54.97 / 63.86	67.90	**69.43**	68.31
yor	✓	✗	65.99	67.44	74.28 / 78.97	67.45 / 78.26	78.89	82.58	**83.66**
avg	–	–	57.50	69.23	64.69 / 71.61	69.62 / 78.96	80.49	80.69	**82.63**
avg (excl. amh)	–	–	55.78	71.13	71.87 / 79.88	70.29 / 79.88	80.97	82.89	**83.15**

Table 6.4: NER model comparison, showing F1-score on the test sets after 50 epochs averaged over 5 runs. This result is for all 4 tags in the dataset: PER, ORG, LOC, DATE.

Another observation is that mBERT tends to have better performance for the non-Bantu Niger-Congo languages i.e., Igbo, Wolof, and Yorùbá, while XLM-R-base works better for Afro-Asiatic languages (i.e., Amharic and Hausa), Nilo-Saharan (i.e., Luo) and Bantu languages like Kinyarwanda and Swahili. We also note that the writing script is one of the primary factors influencing the transfer of knowledge in PLMs about the languages they were not trained on. For example, mBERT achieves an F1-score of 0.0 on Amharic because it has not encountered the script during pre-training. In general, we find the fine-tuned XLM-R-large (with 550M parameters) to be better than XLM-R-base (with 270M parameters) and mBERT (with 110 parameters) in almost all languages. However, mBERT models perform slightly better for Igbo, Luo, and Yorùbá despite having fewer parameters.

We further analyze the transfer abilities of mBERT and XLM-R by extracting sentence embeddings from the LMs to train a BiLSTM model (*MeanE-BiLSTM*) instead of fine-tuning them end-to-end. Table 6.4 shows that languages that are not supported by mBERT or XLM-R generally perform worse than the CNN-BiLSTM-CRF model (despite being randomly initialized) except for *kin*. Also, sentence embeddings extracted from mBERT often lead to better performance than XLM-R for languages they both do not support (like *ibo*, *kin*, *lug*, *luo*, and *wol*).

Lastly, we train NER models using *language BERT* models that have been adapted

[6] https://github.com/mayhewsw/multilingual-data-stats

to each of the African languages via language-specific fine-tuning on unlabeled text. In all cases, fine-tuning language BERT and language XLM-R models achieve a $1-7\%$ improvement in F1-score over fine-tuning mBERT-base and XLM-R-base respectively. This approach is still effective for small-sized pre-training corpora provided they are of good quality. For example, the Wolof monolingual corpus, which contains less than 50K sentences (see Table A.2 in the Appendix) still improves performance by over 4% F1. Further, we obtained over 60% improvement in performance for Amharic BERT because mBERT does not recognize the Amharic script.

6.6.2 Evaluation of Gazetteer Features

Table 6.5 shows the performance of the CNN-BiLSTM-CRF model with the addition of gazetteer features as described in subsubsection 6.5.2.1. On average, the model that uses gazetteer features performs better than the baseline. In general, languages with larger gazetteers, such as Swahili (16K entities in the gazetteer) and Nigerian-Pidgin (for which we use an English gazetteer with 2M entities), have more improvement in performance than those with fewer gazetteer entries, such as Amharic and Luganda (2K and 500 gazetteer entities respectively). This indicates that having high-coverage gazetteers is important for the model to take advantage of the gazetteer features.

Method	amh	hau	ibo	kin	lug	luo	pcm	swa	wol	yor	avg
CNN-BiLSTM-CRF	50.31	84.64	81.25	60.32	75.66	68.93	62.60	77.83	61.84	66.48	68.99
+ Gazetteers	49.51	**85.02**	80.40	**64.54**	73.85	65.44	**66.54**	**80.16**	**62.44**	65.49	**69.34**

Table 6.5: Improving NER models using Gazetteers. The result is only for 3 Tags: PER, ORG & LOC. Models trained for 50 epochs. The result is an average of over 5 runs.

6.6.3 Transfer Learning Experiments

Table 6.6 shows the result for the different transfer learning approaches, which we discuss individually in the following sections. We make use of the XLM-R-base model for all the experiments in this sub-section because the performance difference if we use XLM-R-large is small (<2%) as shown in Table 6.4 and because it is faster to train.

6.6.3.1 Cross-Lingual Transfer

Zero-shot

In the zero-shot setting, we evaluated NER models trained on the English eng-*CoNLL03* dataset, and on the Nigerian-Pidgin (pcm), Swahili (swa), and Hausa

Method	amh	hau	ibo	kin	lug	luo	pcm	swa	wol	yor	avg
XLM-R-base	69.71	91.03	86.16	73.76	80.51	75.81	86.87	**88.65**	69.56	78.05	77.30
WikiAnn zero-shot	27.68	–	21.90	9.56	–	–	–	36.91	–	10.42	–
eng-CoNLL zero-shot	–	67.52	47.71	38.17	39.45	34.19	67.27	76.40	24.33	39.04	37.15
pcm zero-shot	–	63.71	42.69	40.99	43.50	33.12	–	72.84	25.37	35.16	36.81
swa zero-shot	–	85.35*	55.37	58.44	57.65*	42.88*	72.87*	–	41.70	57.87*	52.32
hau zero-shot	–	–	58.41*	59.10*	59.78	42.81	70.74	83.19*	42.81*	55.97	53.14*
WikiAnn + finetune	**70.92**	–	85.24	72.84	–	–	–	87.90	–	76.78	–
eng-CoNLL + finetune	–	89.73	85.10	71.55	77.34	73.92	84.05	87.59	68.11	75.77	75.30
pcm + finetune	–	90.78	86.42	71.69	79.72	75.56	–	87.62	67.21	78.29	76.48
swa + finetune	–	91.50	87.11	74.84	80.21	74.49	86.74	–	68.47	**80.68**	77.63
hau + finetune	–	–	86.84	74.22	80.56	75.55	88.03	87.92	**70.20**	79.44	77.80
combined East Langs.	–	–	–	**75.65**	81.10	77.56	–	88.15	–	–	–
combined West Langs.	–	90.88	87.06	–	–	–	87.21	–	69.70	**80.68**	–
combined 9 Langs.	–	**91.64**	**87.94**	75.46	**81.29**	**78.12**	**88.12**	88.10	69.84	80.59	78.87

Table 6.6: Transfer Learning Result (i.e. F1-score). 3 Tags: PER, ORG & LOC. WikiAnn, eng-CoNLL, and the annotated datasets are trained for 50 epochs. Fine-tuning is only for 10 epochs. Results are averaged over 5 runs and the total average (avg) is computed over ibo, kin, lug, luo, wol, and yor languages. The overall highest F1-score is in **bold**, and the best F1-score in zero-shot settings is indicated with an asterisk (*).

(hau) annotated corpus. We excluded the MISC entity in the eng-*CoNLL03* corpus because it is absent in our target datasets. Table 6.6 shows the result for the (zero-shot) transfer performance. We observe that the closer the source and target languages are geographically, the better the performance. The pcm model (trained on only 2K sentences) obtains similar transfer performance as the eng-*CoNLL03* model (trained on 14K sentences). swa performs better than pcm and eng-*CoNLL03* with an improvement of over 14 F1 on average. We found that, on average, transferring from Hausa provided the best F1, with an improvement of over 16% and 1% compared to using the eng-*CoNLL* and swa data respectively. Per-entity analysis in Table 6.7 shows that the largest improvements are obtained for ORG. The pcm data was more effective in transferring to LOC and ORG, while swa and hau performed better when transferring to PER. In general, zero-shot transfer is most effective when transferring from Hausa and Swahili.

Source Language	PER	ORG	LOC
eng-CoNLL	36.17	27.00	50.50
pcm	21.50	65.33	68.17
swa	55.00	69.67	46.00
hau	52.67	57.50	48.50

Table 6.7: Average per-named entity F1-score for the zero-shot NER using the XLM-R model. The average is computed over ibo, kin, lug, luo, wol, yor languages.

Fine-tuning

We use the target language corpus to fine-tune the NER models previously trained on `eng`-*CoNLL*, `pcm`, and `swa`. On average, there is only a small improvement when compared to the XLM-R base model. In particular, we see significant improvement for Hausa, Igbo, Kinyarwanda, Nigerian-Pidgin, Wolof, and using either `swa` or `hau` as the source NER model.

6.6.4 Regional Influence on NER

We evaluate whether combining different language training datasets by region affects the performance of individual languages. Table 6.6 shows that all languages spoken in West Africa (`ibo`, `wol`, `pcm`, `yor`) except `hau` have slightly better performance (0.1–2.6 F1) when we train on their combined training data. However, for the East-African languages, the F1 score only improved (0.8–2.3 F1) for three languages (`kin`, `lug`, `luo`). Training the NER model on all nine languages leads to better performance on all languages except Swahili. On average over six languages (`ibo`, `kin`, `lug`, `luo`, `wol`, `yor`), the performance improves by 1.6 F1.

6.6.5 Error analysis

Finally, to better understand the types of entities that were successfully identified and those that were missed, we performed a fine-grained analysis of our baseline methods mBERT and XLM-R using the method of (Fu, P. Liu, and Graham Neubig, 2020), with results shown in Table 6.8. Specifically, we found that across all languages, entities that were not contained in the training data (zero-frequency entities), and entities consisting of more than three words (long entities) were particularly difficult in all languages; compared to the F1 score overall entities, the scores dropped by around 5 points when evaluated on zero-frequency entities, and by around 20 points when evaluated on long entities. Future work on low-resource NER or cross-lingual representation learning may further improve these hard cases.

Language	CNN-BiLSTM					mBERT-base					XLM-R-base				
	all	0-freq	0-freq Δ	long	long Δ	all	0-freq	0-freq Δ	long	long Δ	all	0-freq	0-freq Δ	long	long Δ
amh	52.89	40.98	-11.91	45.16	-7.73	–	–	–	–	–	70.96	68.91	-2.05	64.86	-6.10
hau	83.70	78.52	- 5.18	66.21	-17.49	87.34	79.41	-7.93	67.67	-19.67	89.44	85.48	-3.96	76.06	-13.38
ibo	78.48	70.57	- 7.91	53.93	-24.55	85.11	78.41	-6.70	60.46	-24.65	84.51	77.42	-7.09	59.52	-24.99
kin	64.61	55.89	- 8.72	40.00	-24.61	70.98	65.57	-5.41	55.39	-15.59	73.93	66.54	-7.39	54.96	-18.97
lug	74.31	67.99	- 6.32	58.33	-15.98	80.56	76.27	-4.29	65.67	-14.89	80.71	73.54	-7.17	63.77	-16.94
luo	66.42	58.93	- 7.49	54.17	-12.25	72.65	72.85	0.20	66.67	-5.98	75.14	72.34	-2.80	69.39	-5.75
pcm	66.43	59.73	- 6.70	47.80	-18.63	87.78	82.40	-5.38	77.12	-10.66	87.39	83.65	-3.74	74.67	-12.72
swa	79.26	64.74	-14.52	44.78	-34.48	86.37	78.77	-7.60	45.55	-40.82	87.55	80.91	-6.64	53.93	-33.62
wol	60.43	49.03	-11.40	26.92	-33.51	66.10	59.54	-6.56	19.05	-47.05	64.38	57.21	-7.17	38.89	-25.49
yor	67.07	56.33	-10.74	64.52	-2.55	78.64	73.41	-5.23	74.34	-4.30	77.58	72.01	-5.57	76.14	-1.44
avg (excl. amh)	69.36	60.27	- 9.09	50.18	-19.18	79.50	74.07	-5.43	59.10	-20.40	79.15	73.80	-5.36	63.22	-15.94

Table 6.8: F1 score for two varieties of hard-to-identify entities: zero-frequency entities that do not appear in the training corpus, and longer entities of four or more words.

6.7 Chapter Summary

This chapter is based on the collaborative work we did to address the NER task for African languages and I was the lead Igbo language annotator. We created a high-quality NER dataset for ten African languages. We evaluated multiple state-of-the-art NER models and established strong baselines. We released one of our best models that can recognize named entities in ten African languages on HuggingFace Model Hub[7]. RQ1 and RQ2 were addressed in this chapter. We also investigated cross-domain transfer with experiments on five languages with the WikiAnn dataset, along with cross-lingual transfer for low-resource NER using the English CoNLL-2003 dataset and other languages supported by XLM-R.

[7] `https://huggingface.co/Davlan/xlm-roberta-large-masakhaner`

Chapter 7

Africa-Centric Transfer Learning for Named Entity Recognition

I was part of the dataset annotators for Igbo language for this published paper from which the chapter is derived titled "MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition" which is an expansion of the publication "MasakhaNER: Named Entity for African Languages". We created the largest human-annotated NER dataset for 20 African languages, and we studied the behavior of state-of-the-art cross-lingual transfer methods in an Africa-centric setting, demonstrating that the choice of source language significantly affects performance. We showed that choosing the best transfer language improves zero-shot F1 scores by an average of 14 points across 20 languages compared to using English. Our results highlight the need for benchmark datasets and models that cover topologically diverse African languages.

7.1 Introduction

Many African languages are spoken by millions or tens of millions of speakers. However, these languages are poorly represented in NLP research, and the development of NLP systems for African languages is often limited by the lack of datasets for training and evaluation (David Ifeoluwa Adelani, J. Abbott, et al., 2021).

Additionally, while there has been much recent work in using zero-shot cross-lingual transfer (Ponti et al., 2020; Pfeiffer, Vuli, et al., 2020; Ebrahimi et al., 2022) to improve performance on tasks for low-resource languages with multilingual pre-trained language models (PLMs) (Devlin et al., 2019; Conneau et al., 2020), the settings under which contemporary transfer learning methods work best are still unclear (Pruksachatkun et al., 2020; Xia et al., 2020; Lauscher et al., 2020). For example, several methods use English as the source language because of the availability of training data across many tasks (J. Hu et al., 2020; Sebastian Ruder,

Constant, et al., 2021), but there is evidence that English is often not the best transfer language (Y.-H. Lin et al., 2019; Vries, Wieling, and Nissim, 2022; Oladipo et al., 2022), and the process of choosing the best source language to transfer from remains an open question.

There has been recent progress in creating benchmark datasets for training and evaluating models in African languages for several tasks such as machine translation (Nekoto et al., 2020; Reid et al., 2021; David Adelani et al., 2021; David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022; Abdulmumin et al., 2022), and sentiment analysis (Yimam et al., 2020; Shamsuddeen Hassan Muhammad et al., 2022). We focus on the standard NLP task of named entity recognition (NER) because of its utility in downstream applications such as question-answering and information extraction. For NER, annotated datasets exist only in a few African languages (David Ifeoluwa Adelani, J. Abbott, et al., 2021; Yohannes and Amagasa, 2022), the largest of which is the MasakhaNER dataset (David Ifeoluwa Adelani, J. Abbott, et al., 2021), which we call MasakhaNER 1.0 in the remainder of this chapter. While MasakhaNER 1.0 covers 10 African languages spoken mostly in West and East Africa, it does not include any languages spoken in Southern Africa, which have distinct syntactic and morphological characteristics and are spoken by 40 million people.

We tackle two current challenges in developing NER models for African languages: (1) the lack of typologically- and geographically diverse evaluation datasets for African languages; and (2) choosing the best transfer language for NER in an Africa-centric setting, which has not been previously explored in the literature.

To address the first challenge, we create the MasakhaNER 2.0 corpus, the largest human-annotated NER dataset for African languages. MasakhaNER 2.0 contains annotated text data from 20 languages widely spoken in Sub-Saharan Africa and is complementary to the languages present in previously existing datasets (e.g., David Ifeoluwa Adelani, J. Abbott, et al., 2021). We discuss our annotation methodology and perform benchmarking experiments on our dataset with state-of-the-art NER models based on multilingual PLMs.

In addition, to better understand the effect of source language on transfer learning, we extensively analyze different features that contribute to cross-lingual transfer, including linguistic characteristics of the languages (i.e., typological, geographical, and phylogenetic features) as well as data-dependent features such as entity overlap across source and target languages (Y.-H. Lin et al., 2019). We demonstrate that choosing the best transfer language(s) in both single-source and co-training setups leads to large improvements in NER performance in zero-shot settings; our experiments show an average of a 14-point increase in F1 score as compared to using English as a source language across 20 target African languages. We release

the data, code, and models on Github[1]

7.2 Related Work

African NER Datasets

There are some human-annotated NER datasets for African languages such as the SaDiLAR NER corpus (Eiselen, 2016) covering 10 South African languages, LORELEI (Strassel and Tracey, 2016), which covers nine African languages but is not open-sourced, and some individual language efforts for Amharic (Rijsbergen, 1979), (Jesujoba Alabi et al., 2020), Hausa (Michael A Hedderich, David Adelani, et al., 2020), and Tigrinya (Yohannes and Amagasa, 2022). Closest to our work is the MasakhaNER 1.0 corpus (David Ifeoluwa Adelani, J. Abbott, et al., 2021), which covers 10 widely spoken languages in the news domain, but excludes languages from the southern region of Africa like isiZulu, isiXhosa, and chiShona with distinct syntactic features (e.g., noun prefixes and capitalization in between words) which limits transfer learning from other languages. We include five languages from Southern Africa in our new corpus.

Cross-lingual Transfer

Leveraging cross-lingual transfer has the potential to drastically improve model performance without requiring large amounts of data in the target language (Conneau et al., 2020). Still, it is not always clear from which language we must transfer from (Y.-H. Lin et al., 2019; Vries, Wieling, and Nissim, 2022). To this end, recent work investigates methods for selecting good transfer languages and informative features. For instance, the token overlap between the source and target language is a useful predictor of transfer performance for some tasks (Y.-H. Lin et al., 2019; Wu and Dredze, 2019). Linguistic distance (Y.-H. Lin et al., 2019; Vries, Wieling, and Nissim, 2022), word order (K et al., 2020; Pires, Schlinger, and Garrette, 2019) and script differences (Vries, Wieling, and Nissim, 2022), and syntactic similarity (Karamolegkou and Stymne, 2021) have also been shown to impact performance. Another research direction attempts to build models of transfer performance that predict the best transfer language for a target language by using some linguistic and data-dependent features (Y.-H. Lin et al., 2019; Ahuja et al., 2022).

[1] `https://github.com/masakhane-io/masakhane-ner/tree/main/MasakhaNER2.0`

7.3 Languages and Their Characteristics

7.3.1 Focus Languages

Table 7.1 provides an overview of the languages in our MasakhaNER 2.0 corpus. We focus on 20 Sub-Saharan African languages[2] with varying numbers of speakers (between 1M–100M) that are spoken by over 500M people in around 27 countries in the Western, Eastern, Central and Southern regions of Africa. The selected languages cover four language families. 17 languages belong to the Niger-Congo language family, and one language belongs to each of the Afro-Asiatic (Hausa), Nilo-Saharan (Luo), and English Creole (Naija) families. Although many languages belong to the Niger-Congo language family, they have different linguistic characteristics. For instance, Bantu languages (eight in our selection) make extensive use of affixes, unlike many languages of non-Bantu subgroups such as Gur, Kwa, and Volta-Niger.

7.3.2 Language Characteristics

Script and Word Order

African languages mainly employ four major writing scripts: Latin, Arabic, N'ko and Ge'ez. Our focus languages mostly make use of the Latin script. While N'ko is still actively used by the Mande languages like Bambara, the most widely used writing script for the language is Latin. However, some languages use additional letters that go beyond the standard Latin script, e.g., "ɛ", "ɔ", "ŋ", "ẹ", and more than one character letters like "bv", "gb", "mpf", "ntsh". 17 of the languages are tonal except for Naija, Kiswahili, and Wolof. Nine of the languages make use of diacritics (e.g., é, ë, ñ). All languages use the SVO word order, while Bambara additionally uses the SOV word order.

Morphology and Noun classes

Many African languages are morphologically rich. According to the World Atlas of Language Structures "WALS; J. Nichols and Bickel, 2013", 16 of our languages employ strong prefixing or suffixing inflections. Niger-Congo languages are known for their system of noun classification. 12 of the languages *actively* make use of between 6–20 noun classes, including all Bantu languages, Ghomálá', Mossi, Akan and Wolof (MARTEN, 2005; Payne, Pacchiarotti, and Bosire, 2017; Bodomo and Marfo, 2002; Babou and Loporcaro, 2016). While noun classes are often marked using affixes on the head word in Bantu languages, some non-Bantu languages, e.g., Wolof make use of a dependent such as a determiner that is not attached to the head word. For the other Niger-Congo languages such as Fon, Ewe, Igbo, and

[2]Our selection was also constrained by the availability of volunteers that speak the languages in different NLP/AI communities in Africa.

Language	Family	African Region	No. of Speakers	Source	Train / dev / test	% Entities in Tokens	No. of Tokens
Bambara (bam)	NC / Mande	West	14M	David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	4462/ 638/ 1274	6.5	155,552
Ghomálá' (bbj)	NC / Grassfields	Central	1M	David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	3384/ 483/ 966	11.3	69,474
Éwé (ewe)	NC / Kwa	West	7M	David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	3505/ 501/ 1001	15.3	90420
Fon (fon)	NC / Volta-Niger	West	2M	David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	4343/ 621/ 1240	8.3	173,099
Hausa (hau)	Afro-Asiatic / Chadic	West	63M	Kano Focus and Freedom Radio	5716/ 816/ 1633	14.0	221,086
Igbo (ibo)	NC / Volta-Niger	West	27M	IgboRadio and Ka Ọdị Taa	7634/ 1090/ 2181	7.5	344,095
Kinyarwanda (kin)	NC / Bantu	East	10M	IGIHE, Rwanda	7825/ 1118/ 2235	12.6	245,933
Luganda (lug)	NC / Bantu	East	7M	David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	4942/ 706/ 1412	15.6	120,119
Luo (luo)	Nilo-Saharan	East	4M	David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	5161/ 737/ 1474	11.7	229,927
Mossi (mos)	NC / Gur	West	8M	David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	4532/ 648/ 1294	9.2	168,141
Naija (pcm)	English-Creole	West	75M	David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	5646/ 806/ 1613	9.4	206,404
Chichewa (nya)	NC / Bantu	South-East	14M	Nation Online Malawi	6250/ 893/ 1785	9.3	263,622
chiShona (sna)	NC / Bantu	South	12M	VOA Shona	6207/ 887/ 1773	16.2	195,834
Kiswahili (swa)	NC / Bantu	East & Central	98M	VOA Swahili	6593/ 942/ 1883	12.7	251,678
Setswana (tsn)	NC / Bantu	South	14M	David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	3489/ 499/ 996	8.8	141,069
Akan/Twi (twi)	NC / Kwa	West	9M	David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	4240/ 605/ 1211	6.3	155,985
Wolof (wol)	NC / Senegambia	West	5M	David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	4593/ 656/ 1312	7.4	181,048
isiXhosa (xho)	NC / Bantu	South	9M	Isolezwe Newspaper	5718/ 817/ 1633	15.1	127,222
Yorùbá (yor)	NC / Volta-Niger	West	42M	Voice of Nigeria and Asejere	6877/ 983/ 1964	11.4	244,144
isiZulu (zul)	NC / Bantu	South	27M	Isolezwe Newspaper	5848/ 836/ 1670	11.0	128,658

Table 7.1: **Languages and Data Splits for MasakhaNER 2.0 Corpus**. Language, family (NC: Niger-Congo), number of speakers, news source, and data split in number of sentences

Yorùbá, the use of noun classes is merely *vestigial* (Konoshenko and Shavarina, 2019). Three of our languages from the Southern Bantu family (chiShona, isiXhosa, and isiZulu) capitalize proper names after the noun class prefix as in the language names themselves. This characteristic may limit transfer from languages without this feature as NER models overfit on capitalization (Mayhew, Tsygankova, and Roth, 2019). section B.2 provides more details regarding the languages' linguistic characteristics.

7.4 MasakhaNER 2.0 Corpus

7.4.1 Data source and collection

We annotate news articles from local sources. The choice of the news domain is based on the availability of data for many African languages and the variety of named entity types (e.g., person names and locations) as illustrated by popular datasets such as CoNLL-03 (Tjong Kim Sang and De Meulder, 2003).[3] Table 7.1 shows the sources and sizes of the data we use for annotation. Overall, we collected between 4.8K–11K sentences per language from either a monolingual or a translation corpus.

Monolingual corpus

We collect a large monolingual corpus for nine languages, mostly from local news articles except for chiShona and Kiswahili texts, which were crawled from Voice of America (VOA) websites.[4]As the Yorùbá text was missing diacritics, we asked native speakers to add diacritics manually before annotation. During data collection, we ensured that the articles were from a variety of topics e.g. politics, sports, culture, technology, society, and education. In total, we collected between 8K–11K sentences per language.

Translation corpus

For the remaining languages for which we were unable to obtain sufficient amounts of monolingual data, we use a translation corpus, MAFAND-MT (David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022), which consists of French and English news articles translated into 11 languages. We note that translation may lead to undesired properties, e.g., unnaturalness. However, we did not observe serious issues during the annotation. The number of sentences is constrained by the size of the MAFAND-MT corpus, which is between 4,800–8,000.

[3]We also considered using Wikipedia as our data source, but did not due to quality issues (Jesujoba Alabi et al., 2020).

[4]`www.voashona.com/` and `www.voaswahili.com/`

7.4.2 NER Annotation Methodology

We annotated the collected monolingual texts with the ELISA annotation tool **lin** with four entity types: Personal name (`PER`), Location (`LOC`), Organization (`ORG`), and date and time (`DATE`), similar to MasakhaNER 1.0 (David Ifeoluwa Adelani, J. Abbott, et al., 2021). We made use of the MUC-6 annotation guide.[5] The annotation was carried out by three native speakers per language recruited from AI/NLP communities in Africa. To ensure high-quality annotation, we recruited a language coordinator to supervise annotation in each language. We organized two online workshops to train language coordinators on the NER annotation. As part of the training, each coordinator annotated 100 English sentences, which were verified. Each coordinator then trained three annotators in their team using both English and African language texts with the support of the workshop organizers. All annotators and language coordinators received appropriate remuneration.[6] At the end of annotation, language coordinators worked with their teams to resolve disagreements using the adjudication function of ELISA, which ensures a high inter-annotator agreement score.

Lang.	Fleiss' Kappa	QC flags fixed?	Lang.	Fleiss' Kappa	QC flags fixed?
bam	0.980	✗	pcm	0.966	✗
bbj	1.000	✓	nya	0.988	✓
ewe	0.991	✓	sna	0.957	✓
fon	0.941	✗	swa	0.974	✓
hau	0.950	✗	tsn	0.962	✗
ibo	0.965	✗	twi	0.932	✗
kin	0.943	✗	wol	0.979	✓
lug	0.950	✓	xho	0.945	✓
luo	0.907	✗	yor	0.950	✓
mos	0.927	✗	zul	0.953	✓

Table 7.2: Inter-annotator agreement for our datasets calculated using Fleiss' kappa κ at the entity level before adjudication. QC flags (✓) are the languages that fixed the annotations for all **Q**uality **C**ontrol flagged tokens.

[5] `https://cs.nyu.edu/~grishman/muc6.html`

[6] $10 per hour, annotating about 200 sentences per hour.

7.4.3 Quality Control

As discussed in subsection 7.4.2, language coordinators helped resolve several disagreements in annotation before quality control. Table 7.2 reports the Fleiss Kappa score after the intervention of language coordinators (i.e. post-intervention score). The pre-intervention Fleiss Kappa score was much lower. For example, for `pcm`, the pre-intervention Fleiss Kappa score was 0.648 and improved to 0.966 after the language coordinator discussed the disagreements with the annotators.

For quality control, annotations were automatically adjudicated when there was agreement but were flagged for further review when annotators disagreed on mention spans or types. The process for reviewing and fixing quality control issues was voluntary so not all languages were further reviewed (see Table 7.2).

We automatically identified positions in the annotation that were more likely to be annotation errors and flagged them for further review and correction. The automatic process flags tokens that are commonly annotated as a named entity but were not marked as a named entity in a specific position. For example, the token *Province* may appear commonly as part of a named entity and infrequently not as a named entity, so when it is seen as not marked it is flagged. Similarly, we flagged tokens that had near-zero entropy with regard to a certain entity type, for example, a token almost always annotated as ORG but very rarely annotated as PER. We also flagged potential sentence boundary errors by identifying sentences with few tokens or sentences that end in a token that appears to be an abbreviation or acronym. As shown in Table 7.2, before further adjudication and correction there was already relatively high inter-annotator agreement measured by Fleiss' Kappa at the mentioned level.

After quality control, we divided the annotation into training, development, and test splits consisting of 70%, 10%, and 20% of the data respectively. section B.1 provide details on the number of tokens per entity (PER, LOC, ORG, and DATE) and the fraction of entities in the tokens.

PLM	**# Lang.**	**Languages in MasakhaNER 2.0**
mBERT-cased (110M)	104	**swa**, **yor**
XLM-R-base/large (270M / 550M)	100	**hau**, **swa**, **xho**
mDeBERTaV3 (276M)	100	**hau**, **swa**, **xho**
RemBERT (575M)	110	**hau**, **ibo**, **nya**, **sna**, **swa**, **xho**, **yor**, **zul**
AfriBERTa (126M)	11	**hau**, **ibo**, **kin**, **pcm**, **swa**, **yor**
AfroXLMR-base/large (270M/550M)	20	**hau**, **ibo**, **kin**, **nya**, **pcm**, **sna**, **swa**, **xho**, **yor**, **zul**

Table 7.3: Language coverage and size for PLMs.

7.5 Baseline Experiments

7.5.1 Baseline Models

As baselines, we fine-tune several multilingual PLMs including mBERT (Devlin et al., 2019), XLM-R (base & large; Conneau et al., 2020), mDeBERTaV3 (P. He, Gao, and Chen, 2021), AfriBERTa (Ogueji, Zhu, and J. Lin, 2021), RemBERT (Chung et al., 2021), and AfroXLM-R base & large; (Jesujoba O. Alabi et al., 2022). We fine-tune the PLMs on each language's training data and evaluate performance on the test set using the HuggingFace Transformers (Wolf, Debut, Victor Sanh, Chaumond, Delangue, Moi, Cistac, Rault, Remi Louf, et al., 2020).

Massively multilingual PLMs

Table 7.3 shows the language coverage and size of different massively multilingual PLMs trained on 100–110 languages. mBERT was pre-trained using masked language modeling (MLM) and next-sentence prediction on 104 languages, including `swa` and `yor`. RemBERT was trained with a similar objective but makes use of a larger output embedding size during pre-training and covers more African languages. XLM-R was trained only with MLM on 100 languages and on a larger pre-training corpus. mDeBERTaV3 makes use of ELECTRA-style (K. Clark et al., 2020) pre-training, i.e., a replaced token detection (RTD) objective instead of MLM.

Africa-centric multilingual PLMs

We also obtained NER models by fine-tuning two PLMs that are pre-trained on African languages. AfriBERTa (Ogueji, Zhu, and J. Lin, 2021) was pre-trained on less than 1 GB of text covering 11 African languages, including six of our focus languages, and has shown impressive performance on NER and sentiment classification for languages in its pre-training data (David Ifeoluwa Adelani, J. Abbott, et al., 2021; Shamsuddeen Hassan Muhammad et al., 2022). AfroXLM-R (Jesujoba O. Alabi et al., 2022) is a language-adapted (Pfeiffer, Vuli, et al., 2020) version of XLM-R that was fine-tuned on 17 African languages and three high-resource languages widely spoken in Africa ("eng", "fra", and "ara"). section B.10 provides the model hyper-parameters for fine-tuning the PLMs.

7.5.2 Baseline Results

Table 7.4 shows the results of training NER models on each language using the eight multilingual and Africa-centric PLMs. All PLMs provided good performance in general. However, we observed worse results for mBERT and AfriBERTa especially for languages they were not pre-trained on. For instance, both models performed between 6–12 F1 worse for `bbj`, `wol` or `zul` compared to XLM-R-base.

Model	bam	bbj	ewe	fon	hau	ibo	kin	lug	luo	mos	nya	pcm	sna	swa	tsn	twi	wol	xho	yor	zul	AVG
PLM pre-trained on 100+ world languages																					
mBERT	78.9	60.6	86.9	79.9	85.2	87.3	83.2	85.5	80.3	71.4	88.6	87.1	92.4	92.1	86.4	75.7	79.9	85.0	87.7	81.7	$82.8_{\pm 0.2}$
XLM-R-base	78.7	72.3	88.5	81.9	83.8	87.8	82.5	86.7	79.3	72.7	89.9	88.5	93.6	92.2	86.1	78.7	82.3	87.0	85.8	84.6	$84.1_{\pm 0.1}$
XLM-R-large	79.4	**75.2**	89.1	81.6	86.3	87.2	84.3	88.1	80.8	74.9	90.5	89.2	94.2	92.6	85.9	79.8	82.0	88.1	86.6	86.7	$85.1_{\pm 0.5}$
RemBERT	80.1	74.2	89.2	82.2	84.7	86.4	85.2	87.1	80.4	72.7	91.4	89.5	94.8	92.0	87.0	78.5	83.6	88.3	87.2	85.5	$85.0_{\pm 0.2}$
mDeBERTaV3	80.2	73.5	89.8	81.8	85.4	88.8	86.4	88.7	80.3	**76.4**	92.0	**90.1**	95.5	92.5	86.5	79.4	83.6	88.1	86.7	88.3	$85.7_{\pm 0.2}$
PLM pre-trained on African languages																					
AfriBERTa	78.6	71.0	86.9	79.9	85.2	87.3	83.2	85.5	78.4	71.4	88.6	87.1	92.4	92.1	83.2	75.7	79.9	85.0	87.7	81.7	$83.0_{\pm 0.2}$
AfroXLMR-base	79.6	73.3	89.2	82.3	86.6	88.5	86.1	88.1	80.8	74.4	91.9	89.3	95.7	92.3	87.7	78.9	84.9	88.6	88.3	88.4	$85.7_{\pm 0.1}$
AfroXLMR-large	**82.2**	74.8	**90.3**	**82.7**	**87.4**	**89.6**	**87.5**	**89.6**	**82.2**	**76.4**	**92.4**	89.7	**96.2**	**92.7**	**89.4**	**81.1**	**86.8**	**89.9**	**89.3**	**90.6**	$\mathbf{87.0_{\pm 0.2}}$

Table 7.4: **NER Baselines on MasakhaNER 2.0**. We compare several multilingual PLMs including the ones trained on African languages. Average is over 5 runs.

Train Lang.	**Data**	**bam**	**bbj**	**ewe**	**fon**	**hau**	**ibo**	**kin**	**lug**	**luo**	**mos**	**nya**	**pcm**	**sna**	**swa**	**tsn**	**twi**	**wol**	**xho**	**yor**	**zul**	**AVG**
Language in MasakhaNER 1.0?	X	X	X	X						X	X		X		X	X				X	-	
Evaluation on MasakhaNER 2.0 test set																						
(a) MasakhaNER 1.0	MasakhaNER 1.0	52.2	48.4	78.3	52.9	76.9	86.0	77.6	83.2	68.6	55.0	82.1	86.7	49.6	89.4	80.0	56.6	73.6	56.9	69.4	69.9	$69.7_{\pm 0.6}$
(b) MasakhaNER 1.0	MasakhaNER 2.0	50.9	49.8	76.2	57.1	88.7	90.1	87.6	90.0	82.7	49.6	80.4	90.2	42.5	93.1	79.4	57.3	87.0	47.4	89.7	64.3	$72.7_{\pm 0.6}$
(c) MasakhaNER 2.0	MasakhaNER 2.0	**82.3**	**75.5**	**89.5**	**83.2**	**87.7**	**92.3**	**87.2**	**89.1**	**81.8**	**75.3**	**92.2**	**89.9**	**95.9**	**93.1**	**89.5**	**78.8**	**86.4**	**89.7**	**89.1**	**90.7**	$\mathbf{87.0}_{\pm 1.2}$
Evaluation on MasakhaNER 1.0 test set																						
(a) MasakhaNER 1.0	MasakhaNER 1.0	–	–	–	–	92.1	89.2	79.1	86.0	80.0	–	–	91.2	–	89.5	–	–	70.8	–	85.0	–	$84.8_{\pm 0.3}$
(b) MasakhaNER 1.0	MasakhaNER 2.0	–	–	–	–	80.8	84.6	77.7	79.0	67.0	–	–	88.0	–	86.3	–	–	71.6	–	85.0	–	$80.0_{\pm 0.3}$
(c) MasakhaNER 2.0	MasakhaNER 2.0	–	–	–	–	80.4	84.3	77.0	79.8	67.6	–	–	87.9	–	86.5	–	–	72.1	–	84.8	–	$80.1_{\pm 0.8}$

Table 7.5: **Multilingual evaluation on African NER datasets**. We compare the performance of AfroXLM-R-large trained on languages of MasakhaNER 2.0 and MasakhaNER 1.0 and evaluated both on the same and on the other dataset. The first column indicate the languages used for training (the 10 languages from MasakhaNER or the 20 languages from MasakhaNER 2.0). The second column indicates the training data. The average is over 5 runs.

We hypothesize that the performance drop is largely due to the small number of African languages covered by mBERT as well as AfriBERTa's comparatively small model capacity. XLM-R-base gave much better performance (> 1.0 F1) on average compared to mBERT and AfriBERTa. We found the larger variants of mBERT and XLM-R, i.e., RemBERT and XLM-R-large to give much better performance (> 2.0 F1) than the smaller models. Their larger capacity facilitates positive transfer, yielding better performance for unseen languages. Surprisingly, mDeBERTaV3 provided slightly better results than XLM-R-large and RemBERT despite its smaller size, demonstrating the benefits of the RTD pre-training (K. Clark et al., 2020).

The best PLM is AfroXLM-R-large, which outperforms mDeBERTaV3, RemBERT and AfriBERTa by +1.3 F1, +2.0 F1 and +4.0 F1 respectively. Even the performance of its smaller variant, AfroXLM-R-base is comparable to mDeBERTaV3. Overall, our baseline results highlight that large PLMs, PLM with improved pre-training objectives, and PLMs pre-trained on the target African languages can achieve reasonable baseline performance. Combining these criteria provides improved performance, such as AfroXLM-R-large, a large PLM trained on several African languages.

7.5.3 Entity-level Analysis of MasakhaNER 2.0

7.5.3.1 Error Analysis with ExplainaBoard

Furthermore, using ExplainaBoard (P. Liu et al., 2021), we analyzed the best three baseline NER models: AfroXLM-R-large, mDeBERTaV3, and XLM-R-large. We discovered that 2-token entities were easier to predict accurately than lengthier entities (4 or more words). Moreover, the result shows that all the models have difficulty predicting zero-frequency entities effectively (entities with no occurrences in the training set). Interestingly, AfroXLMR-large is significantly better than other models for zero-frequency entities, suggesting that training PLMs on African languages promotes generalization to unseen entities. Finally, we observed that the three models perform better when predicting PER and LOC entities compared to ORG and DATE entities by up to (+5%). section B.4 provides more details on the error analysis.

7.5.3.2 Dataset Geography of Entities

Next, we analyze the geographical representativeness of the entities in our dataset, specifically, we measure the count of entities based on the countries they originate from. Following the approach of (Faisal, Y. Wang, and Anastasopoulos, 2022), we first performed entity linking of named entities present in our dataset to Wikidata IDs using mGenre (De Cao et al., 2022), followed by mapping Wikidata IDs to countries. Figure 7.1 shows the result of the number of entities per continent

and the top 10 countries with the largest representation of entities. Over 50% of the entities are from Africa, followed by Europe. This shows that the entities of MasakhaNER 2.0 properly represent the African continent. Seven out of the top 10 countries are from Africa but also include the USA, the United Kingdom, and France.

7.5.4 Transfer Between African NER Datasets

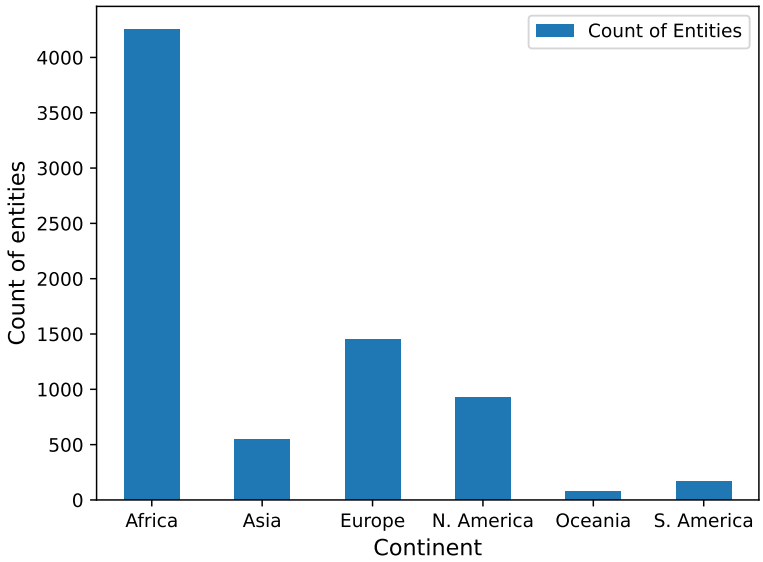
African languages have a diverse set of linguistic characteristics. To demonstrate this heterogeneity, we perform a transfer learning experiment where we compare the performance of multilingual NER models jointly trained on the languages of MasakhaNER 1.0 or MasakhaNER 2.0 and perform zero-shot evaluation on both test sets. We consider three experimental settings:

1. Train on all languages in MasakhaNER 1.0 using MasakhaNER 1.0 training data.
2. Train on the languages in MasakhaNER 1.0 (excl. "amh") using the MasakhaNER 2.0 training data.
3. Train on all languages in MasakhaNER 2.0 using MasakhaNER 2.0 training data.

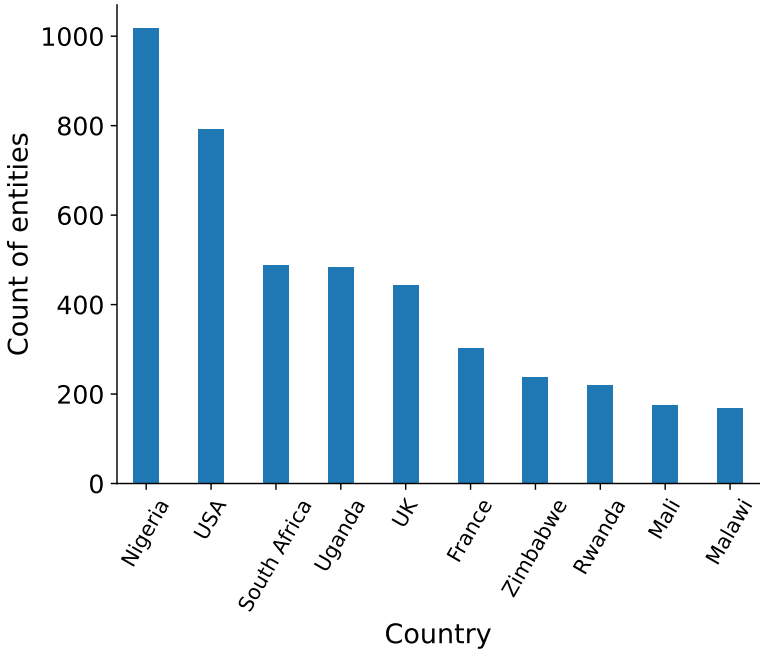
Table 7.5 shows the result of the three settings. When evaluating on the MasakhaNER 2.0 test set in setting (a), the performance is mostly high (> 65 F1) for languages in MasakhaNER 1.0. Most of the languages that are not in MasakhaNER 1.0 have worse zero-shot performance, typically between $48 - 60$ F1 except for `ewe`, `nya`, `tsn`, and `zul` with over 69 F1. Making use of a larger dataset, i.e., setting (b) from MasakhaNER 2.0 only provides a small improvement (+3 F1). The evaluation on setting (c) shows a large gap of about 15 F1 and 17 F1 compared to settings (b) and (a) on the MasakhaNER 2.0 test set respectively, especially for Southern Bantu languages like `sna` and `xho`. On the MasakhaNER 1.0 test set, training on the in-distribution MasakhaNER 1.0 languages and training set achieves the best performance. However, the performance gap compared to training on the MasakhaNER 2.0 data is much smaller. Overall, these results demonstrate the need to create large benchmark datasets (like MasakhaNER 2.0) covering diverse languages with different linguistic characteristics, particularly for Africa.

7.6 Cross-Lingual Transfer

The success of cross-lingual transfer either in zero or few-shot settings depends on several factors, including an appropriate selection of the best source language. Several attempts at cross-lingual transfer make use of English as the source



(a) Number of entities per continent



(b) Top-10 countries

Figure 7.1: Number of entities per continent and the top-10 countries with the largest number of entities

language due to its availability of training data. However, English is unrepresentative of African languages and transfer performance is often lower for distant languages (David Ifeoluwa Adelani, J. Abbott, et al., 2021).

7.6.1 Choosing Transfer Languages for NER

Here, we follow the approach of Y.-H. Lin et al. (2019), `LangRank`, that uses source-target transfer evaluation scores and data-dependent features such as dataset size and entity overlap, and six different linguistic distance measures based on `lang2vec` (Littell et al., 2017) such as geographic distance (d_{geo}), genetic distance (d_{gen}), inventory distance (d_{inv}), syntactic distance (d_{syn}), phonological distance (d_{pho}), and featural distance (d_{fea}).

We provide definitions of the features in section B.5. `LangRank` is trained using these features to determine the best transfer language in a leave-one-out setting where, for each target language, we train on all other languages except the target language. We compute transfer F1 scores from a set of N transfer (source) languages and evaluate on N target languages, yielding $N \times N$ transfer scores.

Choice of Transfer Languages

We selected 22 human-annotated NER datasets of diverse languages by searching the web and HuggingFace Dataset Hub (Lhoest et al., 2021). We required each dataset to contain at least the PER, ORG, and LOC types, and we limited our analysis to these types. We also added our MasakhaNER 2.0 dataset with 20 languages. In total, the datasets cover 42 languages (21 African). Each language is associated with a single dataset. section B.3 provides details about the languages, datasets, and data splits. To compute zero-shot transfer scores, we fine-tune mDeBERTaV3 on the NER dataset of a source language and perform zero-shot transfer to the target languages. We chose mDeBERTaV3 because it supports 100 languages and has the best performance among the PLMs trained on a similar number of languages.

7.6.2 Single-source Transfer Results

Figure 7.2 shows the zero-shot evaluation of training on 42 NER datasets and evaluation on the test sets of the 20 MasakhaNER 2.0 languages. On average, we find the transfer from non-African languages to be slightly worse (51.7 F1) than the transfer from African languages (57.3 F1). The worst transfer result is using `bbj` as source language (41.0 F1) while the best is using `sna` (64 F1), followed by `yor` (63 F1).

Target Lang.	Top-2 Transf. Lang	Top-2 LangRank Model	Top-3 features selected by LangRank model Lang 1; Lang 2	Target Lang. F1	Top-1 LangRank Lang. F1	Top-2 LangRank Lang. F1	Top-2 Transf. Lang. F1	Best Transf. F1	Second Best Transf. F1	eng Tranf. F1
bam	twi, fon	wol, fon	(d_{geo}, d_{inv}, sr); (d_{geo}, sr, d_{pho})	80.4	47.1	52.8	55.1	54.3	53.0	38.4
bbj	fon, ewe	twi, ewe	$(s_{tf}, d_{syn}, d_{geo})$; (s_{tf}, d_{geo}, sr)	72.9	53.9	58.8	60.1	59.8	58.4	45.8
ewe	swa, twi	pcm, swa	(d_{geo}, s_{tf}, sr); (eo, d_{geo}, s_{tf})	91.7	78.1	81.1	**83.9**	81.6	81.5	76.4
fon	mos, bbj	yor, ewe	(d_{syn}, d_{geo}, sr); $(s_{tf}, d_{geo}, d_{gen})$	84.9	58.4	64.9	69.9	65.4	62.0	50.6
hau	pcm, yor	yor, swa	(d_{geo}, sr, eo); (eo, sr, s_{tf})	86.9	74.3	74.8	**77.4**	75.9	74.3	72.4
ibo	sna, yor	pcm, kin	(eo, d_{geo}, s_{tf}); (d_{geo}, sr, eo)	91.0	64.2	63.9	**77.1**	70.4	66.0	61.4
kin	hau, swa	sna, yor	(eo, d_{geo}, s_{tf}); (eo, s_{tf}, sr)	89.5	69.2	71.8	**74.0**	71.1	70.6	67.4
lug	kin, nya	luo, zul	(d_{geo}, sr, eo); (d_{syn}, d_{geo}, sr)	91.5	75.9	78.1	**82.1**	81.1	80.0	76.5
luo	swa, hau	lug, sna	(d_{geo}, sr, eo); (d_{geo}, eo, sr)	81.2	54.9	61.6	61.1	60.4	59.5	53.4
mos	fon, ewe	yor, fon	(d_{geo}, d_{inv}, sr); (d_{geo}, s_{tf}, sr)	78.9	50.8	62.5	**65.6**	64.2	60.4	45.4
nya	swa, nld	zul, sna	(eo, d_{geo}, sr); (d_{geo}, eo, d_{syn})	93.5	65.5	81.5	**81.8**	81.8	81.7	80.1
pcm	hau, yor	eng, yor	(eo, d_{gen}, d_{syn}); (eo, d_{geo}, sr)	89.9	75.5	79.9	**81.8**	80.5	79.1	75.5
sna	zul, xho	swa, zul	(eo, sr, s_{tf}); (d_{geo}, sr, eo)	96.0	32.4	80.0	**80.0**	77.5	74.5	37.1
swa	deu, ara	ita, nld	(sr, d_{inv}, eo); (eo, s_{tf}, sr)	94.6	84.5	86.0	**89.6**	88.7	88.1	87.9
tsn	deu, swa	swa, nya	(eo, d_{inv}, s_{tf}); $(d_{inv}, d_{geo}, d_{gen})$	88.7	73.1	73.4	**74.0**	73.3	73.1	65.8
twi	swa, nya	swa, ewe	(eo, s_{tf}, d_{geo}); (d_{geo}, s_{tf}, sr)	82.0	61.9	57.2	64.3	61.0	61.9	49.5
wol	fon, mos	fon, yor	(d_{geo}, sr, s_{tf}); (sr, d_{geo}, d_{syn})	85.2	62.0	59.4	**63.0**	62.0	58.9	44.8
xho	zul, sna	zul, pcm	(eo, d_{geo}, d_{gen}); (eo, s_{tf}, d_{inv})	90.8	83.7	83.0	**84.3**	83.7	74.0	24.5
yor	hau, pcm	fon, pcm	$(d_{geo}, d_{inv}, d_{syn})$; (eo, d_{geo}, d_{inv})	88.3	37.3	43.2	50.3	50.3	48.8	40.4
zul	xho, sna	xho, sna	(eo, d_{gen}, d_{geo}); (d_{syn}, sr, d_{geo})	88.6	82.1	85.5	**85.5**	82.1	69.4	44.7
AVG	–			87.3	64.2	69.8	**73.1**	71.3	68.8	56.9

Table 7.6: **Best Transfer Languages for NER.** The best zero-shot result is **bolded**, numbers that are not significantly different are underlined. The ranking model features are based on the definitions in Y.-H. Lin et al., 2019 like: geographic distance (d_{geo}), genetic distance (d_{gen}), inventory distance (d_{inv}), syntactic distance (d_{syn}), phonological distance (d_{pho}), transfer language dataset size (s_{tf}), transfer over target size ratio (sr), and entity overlap (eo). The languages highlighted in gray have very good transfer performance ($> 70\%$) using the best transfer language.

We identify German (deu) and Finnish (fin) as the top-2 transfer languages among the non-African languages. In most cases, languages that are geographically and syntactically close tend to benefit most from each other. For example, sna, xho, and zul have very good transfer among themselves due to both syntactic and geographical closeness. Similarly, for Nigerian languages (hau, ibo, pcm, yor) and East African languages (kin, lug, luo, swa), geographical proximity plays an important role. While most African languages prefer transfer from another African language, there are few exceptions, like swa preferring transfer from deu or ara. The latter can be explained by the presence of Arabic loanwords in Swahili (Versteegh, 2001). Similarly, nya and tsn also prefer deu. section B.7 provides results for transfer to non-African languages.

7.6.3 LangRank and Co-training Results

We also investigate the benefit of training on the second-best language in addition to the languages selected by LangRank. We jointly train on the combined data of the top-2 transfer languages or the top-2 languages predicted by LangRank and evaluate their zero-shot performance on the target language. Table 7.6 shows the result for the top-2 transfer languages using the best from 42×42 transfer F1-scores and LangRank model predictions. LangRank predicted the right language as one of the top 2 best transfer languages in 13 target languages. The target languages with incorrect predictions are fon, ibo, kin, lug, luo, nya, and swa. The transfer languages predicted as alternatives are often in the top 5 transfer languages or are less than (-5 F1) worse than the best transfer language. For example, the best transfer language for lug is kin (81 F1) but LangRank predicted luo (76 F1). section B.8 gives results for non-African languages.

Features that are important for transfer

The most important features for the selection of best language by LangRank are geographic distance The d_{geo} is influential because named entities (e.g. name of a politician or a city) are often similar to languages spoken in the same country (e.g. Nigeria with 4 languages in MasakhaNER 2.0) or region (e.g. East African languages). Similarly, we find entity overlap to have a positive Spearman correlation ($R = 0.6$) to transfer the F1-score. section B.6 provides more details on the correlation results. d_{geo} occurred as part of the top-3 features for 15 best transfer languages and 16 second-best languages. Similarly, for *eo*, it appeared 11–13 times for the top 2 transfer languages. Interestingly, dataset size was not among the most important features, highlighting the need for typologically diverse training data.

Best Transfer Language Outperforms English

We compare the zero-shot transfer performance of the top-2 transfer languages

to using eng as the transfer language. They significantly outperform the eng average of 56.9 by +14 and +12 F1 for the first and second-best source language, respectively.

Co-training of Top-2 Transfer Languages Improves Performance

We find that co-training the top-2 transfer languages further improves zero-shot performance over the best transfer by around +3 F1. It is most significant for fon, ibo, kin and twi with 3–7 F1 improvement. Co-training the top-2 transfer languages predicted by LangRank is better than using the second-best transfer language, but often performs worse than the best transfer language.

7.6.4 Sample Efficiency Results

Figure 7.3 shows the performance when the model is trained on a few target language samples compared to when the best transfer language is used prior to fine-tuning on the same number of target language samples. We show the results for four languages (which reflect common patterns across all languages) and an average (ave) over the 20 languages. As seen in the figure, models achieve less than 50 F1 when we train on 100 sentences and over 75 F1 when training on 500 sentences. In practice, annotating 100 sentences takes about 30 minutes while annotating 500 sentences takes around 2 hours and 30 minutes; therefore, slightly more annotation effort can yield a substantial quality improvement. We also find that using the best transfer language in zero-shot settings gives a performance very close to annotating 500 samples in most cases, showing the importance of transfer language selection. By additionally fine-tuning the model on 100 or 500 target language samples, we can further improve the NER performance. section B.9 provides the sample efficiency results for individual languages.

7.7 Limitations

Some Language families not covered

While we try to cover 20 topologically diverse languages and language families, a few locations in Africa and smaller language family groups were not covered. For example, languages from the Khoisan and Austronesian (like Malagasy) family were not covered. Also, languages spoken in central Africa like South Sudan, Chad, and DRC were not covered.

News Domain Data

As the data we annotated belonged to the news domain, models trained from this data may not generalize well to other domains. In particular, the models may

not perform well on more casual text that may use different vocabulary, discuss different entities, and contain more orthographic variation. This limitation also applies to the English NER Corpus.

Generalizability of Transfer Learning Findings

As we only experimented with one task (NER), our findings regarding effective approaches to transfer learning for African languages and PLMs may not generalize to other tasks (e.g. machine translation, part of speech tagging); other features of language similarity may be more important for other tasks.

Explaining Transfer Learning Findings

We found that the `LangRank` model could not predict the top transfer languages with 100% accuracy. This suggests that there are other, unknown factors that could affect transfer performance, which we did not explore. For example, there is still work to be done to understand the sociolinguistic connections and language contact conditions that may correlate with effective transfer.

7.8 Chapter Summary

In this chapter, we present the creation of MasakhaNER 2.0, the largest NER dataset for 20 diverse African languages, and provide strong baseline results on the corpus by fine-tuning multilingual PLMs on in-language NER and multilingual datasets. Additionally, we analyze cross-lingual transfer in an Africa-centric setting, showing the importance of choosing the best transfer language in both zero-shot and few-shot scenarios. Using English as the default transfer language can have detrimental effects, and choosing a more appropriate language substantially improves fine-tuned NER models. By analyzing data-dependent, geographical, and typological features for transfer in NER, we conclude that geographical distance and entity overlap contribute most effectively to transfer performance. RQ1 and RQ2 were addressed in this chapter.

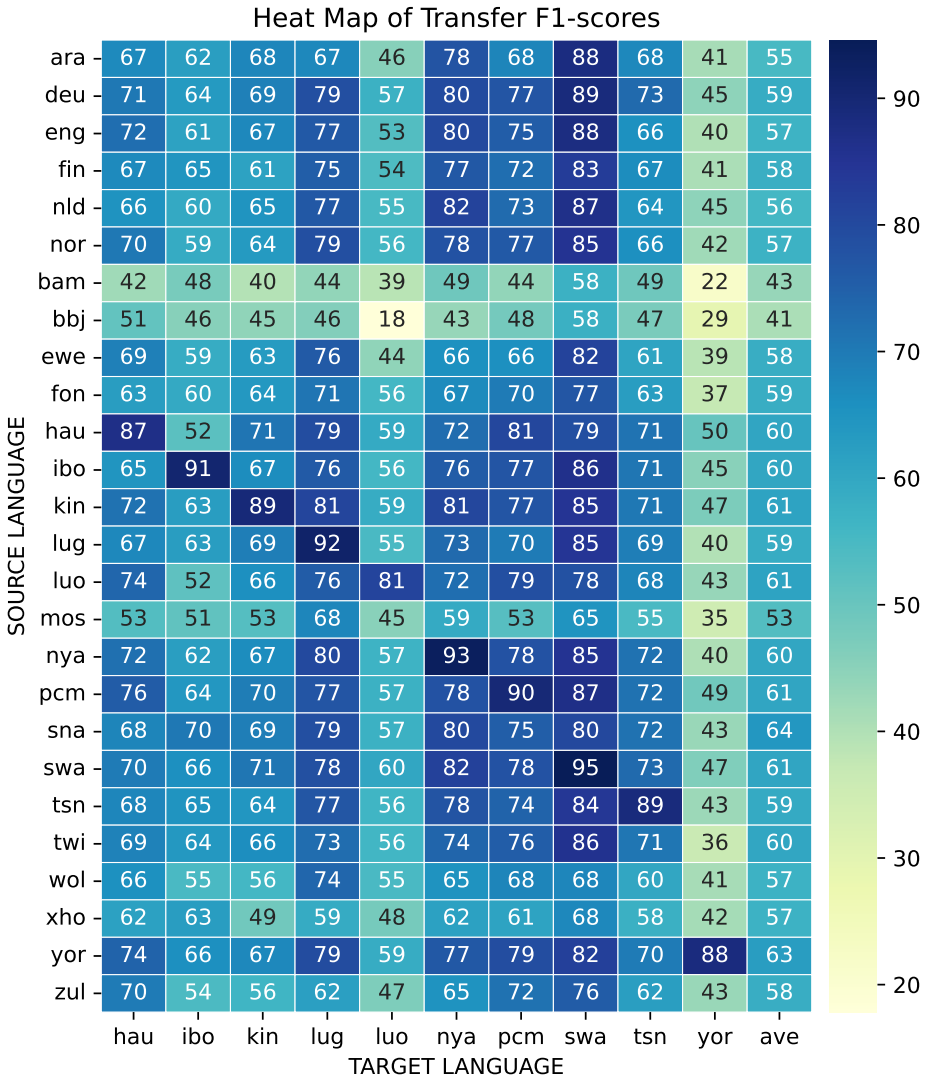


Figure 7.2: **Zero-shot Transfer** from several source languages to African languages for 10 languages in MasakhaNER 2.0 and the average (ave) over all 20 languages. Appendix B.7 shows results for each of the 20 languages.

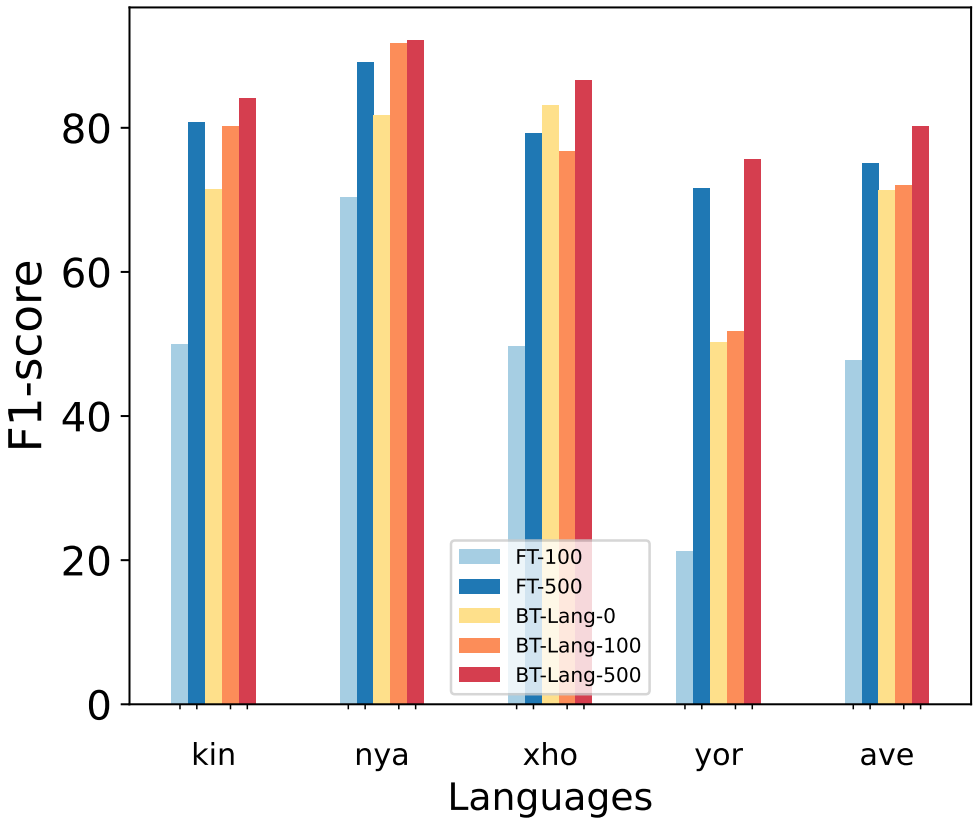


Figure 7.3: **Sample Efficiency Results** for 100 and 500 samples in the target language, model fine-tuned from a PLM (e.g. FT-100 – trained on 100 samples from the target language) or fine-tuned from the best transfer language NER model (e.g BT-Lang-0 – trained on 0 samples from the target language or zero-shot)

Section D
Conclusion

Chapter 8

Conclusion

8.1 Thesis Summary

In this thesis, we tackle the challenge of the under-representation of African languages in NLP research and system development by focusing on Named Entity Recognition (NER) for the Igbo language. Our efforts are directed towards enhancing the resources and techniques for Igbo, contributing significantly to the advancement of NLP for African languages.

Chapter 1 introduces the topic of Named Entity Recognition (NER). It explains the motivation for focusing on NER, particularly for Igbo, an African language, highlighting its importance in natural language processing tasks and its necessity for under-represented languages. The chapter then outlines the research questions that guided the study. Furthermore, it details the contributions of this thesis to the development of NER for the Igbo language, the broader field of African NLP, and the entire NLP community.

Chapter 2 provides an in-depth overview of the focus language, Igbo. It begins with a historical account of its writing system before colonization, which utilized unique symbols. The discussion then transitions to the development of Igbo orthography, highlighting the controversies and debates surrounding its adoption. The chapter also explains the impact of diacritics on the language and its orthography.

Furthermore, the chapter explores the current state of NLP research in Igbo under the IgboNLP initiative, detailing the resources available for the language. It also delves into Named Entity Recognition (NER), describing the techniques commonly applied to NER tasks and the tools researchers use. Additionally, it reviews the NER benchmark datasets that are widely utilized in the field, the evaluation metrics employed, and the datasets featuring African languages before this thesis. The chapter discusses the annotation schemes and NER label sets, providing a comprehensive overview of the methodologies. It also presents several

studies on NER in low-resource settings, examining the approaches previously used for both African and non-African languages. This offers a broader context for understanding the challenges and solutions in this domain.

Chapter 3 outlines the high-level framework utilized in this study for developing NLP resources specifically for the Igbo language, including NER datasets, models, and mapping dictionaries. This framework is structured in phases, each detailing a systematic approach that can be extended to other languages with limited resources by substituting Igbo with the target language. These phases encompass data collection, annotation, model training, and evaluation, ensuring a comprehensive and adaptable methodology. This chapter answered RQ3.

Chapter 4 This chapter is based on the published paper titled "IgboBERT Models: Building and Training Transformer Models for the Igbo Language" (C. Chukwuneke et al., 2022). In this work, we developed IgboBERT, which is the first transformer-based language model pre-trained from scratch specifically for the Igbo language. We evaluated IgboBERT in comparison to mBERT, XML-R, and DistilledBERT by fine-tuning these models on a downstream NER task using the MasakhaNER dataset (David Ifeoluwa Adelani, J. Abbott, et al., 2021). The MasakhaNER dataset is a collaborative project aimed at creating NER datasets for 10 African languages, and I contributed to this effort as an Igbo language annotator.

The evaluation results indicated that although IgboBERT was outperformed by the other models, it still achieved a respectable F1 score of 77.94. This is particularly noteworthy given that IgboBERT was pre-trained on a relatively small amount of raw data, unlike mBERT, XML-R, and DistilBERT, which were trained on millions of data points. This suggests that with additional data for fine-tuning, IgboBERT's performance could be further enhanced. This chapter answered the RQ1

Chapter 5 is based on the published paper "IgboNER 2.0: Expanding Named Entity Recognition Datasets via Projection" (C. Chukwuneke et al., 2022). This work proposes a novel and efficient approach to address the scarcity of adequate datasets for building NLP models for low-resource languages, focusing specifically on the named entity recognition (NER) task for the Igbo language. The project expanded the existing Igbo dataset by leveraging parallel data in English — a well-resourced language with advanced NER tools — and Igbo. This approach takes advantage of the fact that Igbo, written in the Latin script, often incorporates English words, allowing for more effective data transfer.

We introduced the concept of a mapping dictionary by creating a semi-automatic pipeline to transfer NER tags generated by the spaCy English NER tagger to corresponding Igbo entities. This mapping dictionary will also serve as a valuable

resource for other NLP tasks, such as machine translation and part-of-speech tagging. Our experiments demonstrate that model performance improves with increased data size when fine-tuned on the NER downstream task. We aim to continue expanding our mapping dictionary and dataset to enhance our model's performance further. This chapter answered RQ1 and RQ4.

Chapter 6 is derived from the publication "MasakhaNER: Named Entity Recognition for African Languages," a collaborative effort aimed at addressing the significant under-representation of African languages in NLP research. I contributed as the lead Igbo language annotator. This study presents the first large-scale, publicly available, high-quality dataset for named entity recognition (NER) across ten African languages. We provided detailed linguistic characteristics of these languages to help researchers and practitioners better understand the unique challenges they present for NER tasks. The RQ1 and RQ2 were addressed in this chapter.

In our study, we analyzed the datasets and conducted an extensive empirical evaluation of state-of-the-art methods. This included cross-domain transfer experiments on five languages using the WikiAnn dataset and cross-lingual transfer experiments for low-resource NER using the English CoNLL-2003 dataset, along with other languages supported by XLM-R. Our comprehensive analysis highlights the complexities and opportunities in applying these methods to African languages. Additionally, we have made the data, code, and models publicly available to encourage and facilitate future research in African NLP. By providing these resources, we aim to inspire further advancements in the field and contribute to a more inclusive representation of African languages in NLP research.

Chapter 7 is based on the publication "MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition" (Adelani et al., 2022). This work is a collaborative effort, in which I contributed as an Igbo language annotator and also assisted with model evaluations for the Igbo language. In this study, we expanded our NER dataset creation to encompass 20 diverse African languages and provided robust baseline results by fine-tuning multilingual pre-trained language models (PLMs) on both in-language NER and multilingual datasets.

We also conducted an in-depth analysis of cross-lingual transfer in an Africa-centric context, demonstrating the critical importance of selecting the most suitable transfer language in both zero-shot and few-shot scenarios. Our findings indicate that choosing a more appropriate transfer language significantly enhances the performance of fine-tuned NER models. Through our analysis of data-dependent, geographical, and typological features for transfer in NER, we concluded that geographical proximity and entity overlap are the most effective factors contributing to improved transfer performance. By examining these features, we provided valuable insights into optimizing cross-lingual transfer for NER tasks in African languages. The RQ1 and RQ2 were addressed in this chapter.

8.2 Contributions beyond NER

This section outlines the contributions that extend beyond NER in the course of this PhD journey to advance research in African languages.

1. **Afriqa: Cross-lingual open-retrieval question answering for African languages (Ogundepo et al., 2023)**
   In this study, we introduce AFRIQA, the pioneering cross-lingual question-answering dataset designed specifically for African languages. This initiative aims to narrow the information gap between native speakers of numerous African languages and the extensive digital content available online. AFRIQA encompasses over 12,000 questions in 10 African languages, targeting open-retrieval question answering. We assess the dataset through cross-lingual retrieval and reading comprehension tasks.

   Our goal is to enhance access to pertinent information for African language speakers. This work marks a significant advance towards democratizing information access and empowering underrepresented African communities by providing tools that enable engagement with digital content in their native languages. Our experiments reveal the limitations of current automatic translation and multilingual retrieval methods, indicating that AFRIQA presents substantial challenges for leading QA models. We hope this dataset will catalyze the development of more equitable question-answering technologies. I contributed to this work by assisting in the paper writing and serving as one of the Igbo language annotators.

2. **IgboSum1500-Introducing the Igbo Text Summarization Dataset (MBONU et al., 2022)**.
   This ongoing work aims to address a significant gap in IgboNLP research: the lack of a text summarization tool for Igbo. In this paper, we detail our efforts in creating the IgboSum1500 dataset, the first standardized, high-quality, publicly available Igbo text summarization dataset. This dataset will serve as a crucial foundation for developing Igbo text summarization resources and expanding both Igbo and African NLP. This contribution is particularly significant for Igbo and AfricanNLP, and more broadly for low-resource NLP, especially in the areas of natural language understanding and text generation. The dataset will be released, and future work will focus on experimenting with and fine-tuning state-of-the-art neural models for Igbo summarization, with an emphasis on the abstractive approach. I assisted with the paper writing.

3. **Findings of the 1st Shared Task on Multi-lingual Multi-task Information Retrieval at MRL 2023 (Tinner et al., 2023)**. This paper details our findings from participating in the Multi-lingual Representation Learning (MRL 2023) shared task. The main objective of this shared task is to evaluate and understand the multilingual inference capabilities of language

models, particularly their ability to comprehend and generate language based on logical, factual, or causal relationships over long text contexts, especially in low-resource settings.

The shared task comprises two essential subtasks for information retrieval: Named Entity Recognition (NER) and Reading Comprehension (RC), conducted in seven low-resource languages: Azerbaijani, Igbo, Indonesian, Swiss German, Turkish, Uzbek, and Yorùbá. These languages previously lacked annotated resources for information retrieval tasks. Our evaluation of leading LLMs reveals that despite their competitive performance, they still have notable weaknesses, such as producing output in non-target languages or providing counterfactual information that cannot be inferred from the context.

In developing the multilingual evaluation benchmark for information retrieval, we relied on Wikipedia. We found that using Wikipedia has inherent limitations, such as variations in content and styles across languages, making it challenging to ensure a uniform difficulty level for comprehension questions. As more advanced models are developed, this benchmark will remain essential for supporting fairness and applicability in information retrieval systems. I contributed to both subtasks by creating questions, labeling the answers, and annotating named entities for the Igbo language.

4. **AfriMTE and AfriCOMET: Enhancing COMET to Embrace Under-resourced African Languages (J. Wang et al., 2023)**.
In this study, we address the challenges of adapting the COMET metric for machine translation evaluation in various under-resourced African languages. We have developed a simplified MQM annotation guideline and created the AFRIMTE dataset, which includes 13 typologically diverse African languages. We also established benchmark COMET systems, known as AFRICOMET, to tackle critical issues in this field. Our experimental results indicate that it is feasible to use transfer learning techniques from existing, well-resourced Direct Assessment data and to utilize multilingual pre-trained language models enhanced with African languages to build MT evaluation systems for these languages. All datasets, codes, and models are released to support ongoing research and development in machine translation evaluation. I contributed to the dataset annotation for the Igbo language.

5. **MasakhaNEWS: News Topic Classification for African languages (David Ifeoluwa Adelani, Masiak, et al., 2023)**.
In this work, we developed MasakhaNEWS—the largest dataset for news topic classification across 16 widely spoken African languages. We conducted extensive evaluations using both fully-supervised and few-shot learning settings. Additionally, we examined various methods of adapting prompt-based tuning and non-prompt approaches of language models (LMs) for African languages. Our experimental results indicate that prompting

large language models (LLMs) like ChatGPT performs poorly on the straightforward task of text classification for several under-resourced African languages, particularly those with non-Latin scripts. Furthermore, we demonstrated the potential of prompt-based few-shot learning approaches, such as PET (based on smaller LMs), for African languages. Our findings show that existing supervised methods are effective for all African languages and that language models can achieve competitive performance with only a few supervised samples, underscoring the applicability of current NLP techniques for African languages. I contributed to the dataset annotation for the Igbo language.

8.3 Achieved Contributions

In this research, the contributions that have been achieved are:

- The development of the first Igbo transformer-based language model from scratch[1]. IgboNER,a basline model for Igbo language was developed.

- Creation of new IgboNER dataset[2]. NER dataset using a parallel corpus applying projection method was created.

- Creation of a mapping dictionary for Igbo entities[3]. This contains a list of English entities, their Igbo translations, and their tags and will enhance NLP systems' accuracy, efficiency, and interpretability across various applications and domains.

- Developed a framework for the creation of NER resources for different languages[4]. This framework can be applied to any language by replacing the Igbo and English corpus used in this work with any language of your choice.

- Creation of IgboNER visualisation tool[5]. We use an open-source python framework Streamlit, to create and deploy web visualisation for Igbo NER.

- Advanced NER research for low-resource languages by collaborating in the creation of the largest NER dataset for some African languages[6].

[1]https://openreview.net/pdf?id=tHUS9-vmUfC
[2]https://github.com/Chiamakac/IgboNER-Models/tree/main/IgboNER_2.0
[3]https://openreview.net/pdf?id=tHUS9-vmUfC
[4]chapter 3
[5]section5.5
[6]https://aclanthology.org/2022.emnlp-main.298/

8.4 Limitations of this study

The limitations encountered in the course of this work include:

- **Non-Availability of dataset** - At the beginning of this PhD research, there was a significant absence of high-quality NER datasets for the Igbo language. We therefore began with a data collection process to create a dataset for IgboNER. The digital scarcity of Igbo text across various genres further complicated this effort, restricting us to approximately 85.95% of data from news sources, 14% from Bible texts, and only 0.05% from novels.
- **Dialectal and Orthographic Variations in Igbo** - Various languages, including Igbo, have many dialects. Additionally, the lack of a standardized orthography complicates this task, as multiple orthographies are used in different texts. This makes it difficult to develop a model that effectively covers all dialectal and orthographic variations, particularly when data for some dialects are limited or entirely unavailable.
- **Computational Resources** - Training transformer-based models require significant computational power, which may not be readily accessible for researchers working with low-resource languages. This can limit the capability to experiment with and optimize models effectively.

8.5 Future Work

This thesis highlights the digitally disadvantaged state of Igbo and African languages at large in NLP research and outlines efforts taken to address this issue. The resources, tools, and results from this thesis provide a foundation for further research in Igbo NLP. The following are some areas of future research:

- **Expansion of Annotated Datasets**: Chapter 2 and Chapter 8.2 outlines the few existing annotated datasets available in Igbo for various NLP tasks such as NER, part-of-speech, question answering, machine translation, diacritic restoration, news topic classification. There is a need to expand on the aforementioned tasks and to address more tasks such as summarization, sentiment classification, hate-speech detection, natural language inference, causal commonsense reasoning, slot-filling, intent detection, create datasets that address the various dialects of Igbo and multimodal data collection and annotation (e.g., audio, video) to enhance speech recognition and synthesis for Igbo language.
- **Speech and Audio Processing**: Enhance automatic speech recognition (ASR) and text-to-speech (TTS) systems for Igbo, focusing on tonal accuracy. To develop conversational agents and voice assistants that can understand and respond in Igbo.

- **Educational Tools**: Develop NLP-powered educational tools to support language learning, and literacy in Igbo e.g. reading pen [7] , search engine, and grammatical error correction.
- **Language Models**: Develop and fine-tune more advanced pre-trained language models specifically for Igbo, such as IgboBERT to improve performance on downstream tasks. Models that can understand and generate contextually accurate Igbo text, including idiomatic expressions and cultural references, and models that address the various dialects of Igbo, ensuring broad applicability and inclusivity.

[7]Reading pens are a piece of assistive technology that can help people with dyslexia or other reading difficulties to have printed words read back to them.

Appendix A

Named entity recognition

A.1 Annotator Agreement

To shed more light on the few cases where annotators disagreed, we provide entity-level confusion matrices across all ten languages in Table A.1. The most common disagreement is between organizations and locations.

	DATE	**LOC**	**ORG**	**PER**
DATE	32,978	-	-	-
LOC	10	70,610	-	-
ORG	0	52	35,336	-
PER	2	48	12	64,216

Table A.1: Entity-level confusion matrix between annotators, calculated over all ten languages.

A.2 Model Hyper-parameters for Reproducibility

For fine-tuning mBERT and XLM-R, we used the base and large models with maximum sequence length of 164 for mBERT and 200 for XLM-R, batch size of 32, learning rate of 5e-5, and number of epochs 50. For the MeanE-BiLSTM model, the hyper-parameters are similar to fine-tuning the LM except for the learning rate that we set to be 5e-4, the BiLSTM hyper-parameters are: input dimension is 768 (since the embedding size from mBERT and XLM-R is 768) in each direction of LSTM, one hidden layer, hidden layer size of 64, and drop-out probability of 0.3 before the last linear layer. All the experiments were performed on a single GPU (Nvidia V100).

A.3 Monolingual Corpora for Language Adaptive Fine-tuning

Table A.2 shows the monolingual corpus we used for the language adaptive fine-tuning. We provide the details of the source of the data, and their sizes. For most of the languages, we make use of JW300[1] and CC-100[2]. In some cases CC-Aligned (El-Kishky et al., 2020) was used, in such a case, we removed duplicated sentences from CC-100. For fine-tuning the language model, we make use of the HuggingFace (Wolf, Debut, Victor Sanh, Chaumond, Delangue, Moi, Cistac, Rault, R'emi Louf, et al., 2019) code with learning rate 5e-5. However, for the Amharic BERT, we make use of a smaller learning rate of 5e-6 since the multilingual BERT vocabulary was replaced by Amharic vocabulary, so that we can slowly adapt the mBERT LM to understand Amharic texts. All language BERT models were pre-trained for 3 epochs ("ibo", "kin","lug","luo", "pcm","swa","yor") or 10 epochs ("amh", "hau","wol") depending on their convergence. The models can be found on HuggingFace Model Hub[3].

[1] https://opus.nlpl.eu/

[2] http://data.statmt.org/cc-100/

[3] https://huggingface.co/Davlan

Language	Source	Size (MB)	No. sentences
amh	CC-100 (Conneau et al., 2020)	889.7MB	3,124,760
hau	CC-100	318.4MB	3,182,277
ibo	JW300 (Agić and Vulić, 2019), CC-100, CC-Aligned (El-Kishky et al., 2020), and IgboNLP (Ezeani, Rayson, et al., 2020a)	118.3MB	1,068,263
kin	JW300, KIRNEWS (Niyongabo et al., 2020), and BBC Gahuza	123.4MB	726,801
lug	JW300, CC-100, and BUKEDDE News	54.0MB	506,523
luo	JW300	12.8MB	160,904
pcm	JW300, and BBC Pidgin	56.9MB	207,532
swa	CC-100	1,800MB	12,664,787
wol	OPUS (Tiedemann, 2012) (excl. CC-Aligned), Wolof Bible **wolof_bible**, and news corpora(Lu Defu Waxu, Saabal, and Wolof Online)	3.8MB	42,621
yor	JW300, Yoruba Embedding Corpus (Jesujoba Alabi et al., 2020), MENYO-20k (David Ifeoluwa Adelani, Ruiter, et al., 2021), CC-100, CC-Aligned, and news corpora (BBC Yoruba, Asejere, and Alaroye).	117.6MB	910,628

Table A.2: Monolingual Corpora, their sources, size, and number of sentences

Appendix B

Africa Centric Transfer Learning for Named Entity Recognition

B.1 Data Source and Splits

Table B.1 shows the MasakhaNER 2.0 language, data source, train/dev/test split, and the number of tokens per entity type.

B.2 Language Characteristics

Table B.2 provides the details about the language characteristics.

B.2.1 Morphology and Noun classes

Many African languages are morphologically rich. According to the World Atlas of Language Structures (WALS; J. Nichols and Bickel, 2013), 16 of our languages employ strong prefixing or suffixing inflections. Niger-Congo languages are known for their system of noun classification. 12 of the languages *actively* make use of between 6–20 noun classes, including all Bantu languages and Ghomálá', Mossi, Akan and Wolof (Nurse and Philippson, 2006; Payne, Pacchiarotti, and Bosire, 2017; Bodomo and Marfo, 2002; Babou and Loporcaro, 2016). While noun classes are often marked using affixes on the head word in Bantu languages, some non-Bantu languages, e.g., Wolof make use of a dependent such as a determiner that is not attached to the headword. For the other Niger-Congo languages such as Fon, Ewe, Igbo, and Yorùbá, the use of noun classes is merely *vestigial* (Konoshenko and Shavarina, 2019). For example, Yorùbá only distinguishes between human and non-human nouns. Bambara is the only Niger-Congo language without noun classes, and some have argued that the Mande family should be regarded as an independent language family. Three of our languages from the Southern Bantu family (chiShona, isiXhosa and isiZulu) capitalize proper names after the noun class prefix as in the language names themselves. This characteristic limits the

			# Tokens				% Entities in Tokens	
Language	**Data Source**	**Train / dev / test**	**PER**	**LOC**	**ORG**	**DATE**		**#Tokens**
Bambara (bam)	MAFAND-MT David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	4462/ 638/ 1274	4281	2557	429	2898	6.5	155,552
Ghomálá (bbj)	MAFAND-MT David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	3384/ 483/ 966	2464	1371	1586	2457	11.3	69,474
Éwé (ewe)	MAFAND-MT David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	3505/ 501/ 1001	3931	5168	2064	2665	15.3	90420
Fon (fon)	MAFAND-MTDavid Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	4343/ 621/ 1240	3572	2595	3082	5120	8.3	173,099
Hausa (hau)	Kano Focus and Freedom Radio	5716/ 816/ 1633	9853	6759	7089	7251	14.0	221,086
Igbo (ibo)	IgboRadio and Ka Ọdị Taa	7634/ 1090/ 2181	8532	7077	5418	4727	7.5	344,095
Kinyarwanda (kin)	IGIHE, Rwanda	7825/ 1118/ 2235	6889	8960	7012	8187	12.6	245,933
Luganda (lug)	MAFAND-MT David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	4942/ 706/ 1412	6058	3706	5441	3484	15.6	120,119
Luo (luo)	MAFAND-MT David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	5161/ 737/ 1474	6806	5605	7099	7359	11.7	229,927
Mossi (mos)	MAFAND-MT David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	4532/ 648/ 1294	2804	3044	3209	6334	9.2	168,141
Naija (pcm)	MAFAND-MT David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	5646/ 806/ 1613	4711	5077	5940	3654	9.4	206,404
Chichewa (nya)	Nation Online Malawi	6250/ 893/ 1785	9657	4600	5924	4308	9.3	263,622
Shona (sna)	VOA Shona	6207/ 887/ 1773	10667	5289	12418	3423	16.2	195,834
Swahili (swa)	VOA Swahili	6593/ 942/ 1883	9510	10515	6515	5331	12.7	251,678
Setswana (tsn)	MAFAND-MT David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	3489/ 499/ 996	3991	2285	2905	3190	8.8	141,069
Akan/Twi (twi)	MAFAND-MT David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	4240/ 605/ 1211	3588	2474	2375	1433	6.3	155,985
Wolof (wol)	MAFAND-MT David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, et al., 2022	4593/ 656/ 1312	3588	2474	2375	1433	7.4	181,048
isiXhosa (xho)	Isolezwe Newspaper	5718/ 817/ 1633	8098	3087	5633	2433	15.1	127,222
Yorùbá (yor)	Voice of Nigeria and Asejere	6877/ 983/ 1964	8537	5819	6998	6372	11.4	244,144
isiZulu (zul)	Isolezwe Newspaper	5848/ 836/ 1670	5050	1900	5229	2012	11.0	128,658

Table B.1: **Languages and Data Splits for MasakhaNER 2.0 Corpus**. Distribution of the number of entities

Language	No. of Letters	Latin Letters Omitted	Letters added	Tonality	diacritics	Word Order	Morphological typology	Inflectional Morphology (WALS)	Noun Classes
Bambara (bam)	27	q,v,x	ɔ, ɛ, ŋ	yes, 2 tones	yes	SVO & SOV	isolating	strong suffixing	absent
Ghomálá' (bbj)	40	q, w, x, y	bv, dz, ɛ, a, ɔ, gh, ny, nt, ŋ, pk, ɔ, pf, mpf, sh, ts, ɔ, zh, '	yes, 5 tones	yes	SVO	agglutinative	strong prefixing	active, 6
Éwé (ewe)	35	c, j, q	ɔ, dz, ɛ, f, gb, ɔ, kp, ny, ŋ, ɔ, ts,	yes, 3 tones	yes	SVO	isolating	equal prefixing and suffixing	vestigial
Fon (fon)	33	q	ɔ, gb, hw, kp, ny, ɔ, xw	yes, 3 tones	yes	SVO	isolating	little affixation	vestigial
Hausa (hau)	44	p,q,v,x	ɓ, ɗ, ƙ, kw, w, gw, ky, y, gy, sh, ts	yes, 2 tones	no	SVO	agglutinative	little affixation	absent
Igbo (ibo)	34	c, q, x	ch, gb, gh, gw, kp, kw, nw, ny, ọ, ọ, sh, ụ	yes, 2 tones	yes	SVO	agglutinative	little affixation	vestigial
Kinyarwanda (kin)	30	q, x	cy, jy, nk, nt, ny, sh	yes, 2 tones	no	SVO	agglutinative	strong prefixing	active, 16
Luganda (lug)	25	h, q, x	ŋ, ny	yes, 3 tones	no	SVO	agglutinative	strong prefixing	active, 20
Luo (luo)	31	c, q, x, v, z	ch, dh, mb, nd, ng', ng, ny, ŋ, th, sh	yes, 4 tones	no	SVO	agglutinative	equal prefixing and suffixing	absent
Mossi (mos)	26	c, j, q, x	', ɔ, ɛ	yes, 2 tones	yes	SVO	isolating	strongly suffixing	active, 11
Chichewa (nya)	31	q, x, y	ch, kh, ng, ŋ, ph, tch, th, ŵ	yes, 2 tones	no	SVO	agglutinative	strong prefixing	active, 17
Naija (pcm)	26	–	–	no	no	SVO	mostly analytic	strongly suffixing	absent
Shona (sna)	29	c, l, q, x	bh, ch, dh, nh, sh, vh, zh	yes, 2 tones	no	SVO	agglutinative	strong prefixing	active, 20
Swahili (swa)	33	x, q	ch, dh, gh, kh, ng', ny, sh, th, ts	no	yes	SVO	agglutinative	strong suffixing	active, 18
Setswana (tsn)	36	c, q, v, x, z	ê, kg, kh, ng, ny, ô, ph, š, th, tl, tlh, ts, tsh, ts, tsh	yes, 2 tones	no	SVO	agglutinative	strong prefixing	active, 18
Akan/Twi (twi)	22	c,j,q,v,x,z	ɔ	yes, 5 tones	no	SVO	isolating	strong prefixing	active, 6
Wolof (wol)	29	h,v,z	ŋ, à, é, ë, ó, ñ	no	yes	SVO	agglutinative	strong suffixing	active, 10
isiXhosa (xho)	68	–	bh, ch, dl, dy, dz, gc, gq, gr, gx, hh, hl, kh, kr, lh, mh, ng, ngc, ngh, ngq, ngx, nkq, nkx, nh, nkc, nx, ny, nyh, ph, qh, rh, sh, th, ths, thsh, ts, tsh, ty, tyh, wh, xh, yh, zh	yes, 2 tones	no	SVO	agglutinative	strong prefixing	active, 17
Yorùbá (yor)	25	c, q, v, x, z	ẹ, gb, ṣ, ọ	yes, 3 tones	yes	SVO	isolating	little affixation	vestigial, 2
isiZulu (zul)	55	–	nx, ts, nq, ph, hh, ny, gq, hl, bh, nj, ch, ngc, ngq, th, ngx, kl, ntsh, sh, kh, tsh, ng, nk, gx, xh, gc, mb, dl, nc, qh	yes, 3 tones	no	SVO	agglutinative	strong prefixing	active, 17

Table B.2: Linguistic Characteristics of the Languages

transfer learning of NER from languages without this feature, since NER models overfit on capitalization (Mayhew, Tsygankova, and Roth, 2019).

B.2.2 IsiXhosa and isiZulu morphological structure

IsiXhosa and isiZulu are agglutinative languages with a complex morphology. Each root or stem can attach a variety of affixes to form new inflections and derivations, with a variety of affixes added to root and stem morphemes to vary their meaning and convey syntactic agreement. The noun class system and the concord agreement system are the foundations of isiXhosa and isiZulu noun grammar. This section offers an overview of these two principles and their applicability to the realization of NEs. First, we briefly describe the noun class system, after which we discuss prefixing and capitalization work for both languages.

According to the Meinhoff system Melzian (1933), nouns in African languages are classified into one of 18 numbered classes based on their prefix. As shown in the following example, singular nouns in class 1 take the prefix um-, while associated plural nouns in class 2 take the prefix aba-.

B.2.2.1 Prefix

Even though all named entities are nouns since they designate a distinct entity, noun class designations are critical in identifying NEs. According to Oosthuysen (2016), the purpose of the noun class prefix is to distinguish the class to which it belongs. It shows whether the noun is singular or plural. The derivation of all significant prefixes and cordial agreements is based on this.

In isiXhosa, named entities referring to personal nouns with the prefix um- belongs to noun class 1 with noun class 2 being its plural form. Named entities such as jobs, objects and concepts belong to noun class 3, e.g. umpheki (cook) and umthwalo (burden). Lastly in isiXhosa, borrowed words from English and Afrikaans such as ibhanka (bank) and ihamire (hammer), belong to class 9. In isiZulu, noun class 1 is a singular class which uses the prefix umu-/um-. The allomorph umu- occurs when the noun stem consists of one syllable, e.g. umuntu (person) and the allomorph um- occurs when the noun stem has more than one syllable, e.g. umfana (boy). The noun class 2 is a plural class, with its singular in class 1. Noun class 2 uses the prefix aba-/ab-, e.g. abantu (people), abafana (boys). Noun classes 1 and 2 are a personal class only containing personal nouns.
Noun class 1a is a subclass of noun class 1. This class contains personal nouns referring to family relationships, professions, proper names and personalized nouns. This class uses the prefix u- with no allomorphs, e.g. ugogo (grandmother), unesi (nurse) or uSipho (personal name). The noun class 2a is a regular plural of class 1a which uses the prefix o-, e.g. ogogo (grandmothers), onesi (nurses) or oSipho (Sipho and company).

Language	Data Source	# Train	# dev	# test
Amharic (amh)	MasakhaNER 1.0 (David Ifeoluwa Adelani, J. Abbott, et al., 2021)	1,750	250	500
Arabic (ara)	ANERcorp (Benajiba, Rosso, and Benedíruiz, 2007; Obeid et al., 2020)	3,472	500	924
Danish (dan)	DANE (Hvingelby et al., 2020)	4,383	564	565
German (deu)	CoNLL03 (Tjong Kim Sang and De Meulder, 2003)	12,152	2,867	3,005
English (eng)	CoNLL03 (Tjong Kim Sang and De Meulder, 2003)	14,041	3,250	3,453
Spanish (spa)	CoNLL02 (Tjong Kim Sang, 2002)	8,322	1,914	1,516
Farsi (fas)	PersoNER (Poostchi et al., 2016)	4,121	1,000	2,560
Finnish (fin)	FINER (Ruokolainen et al., 2020)	13,497	986	3,512
French (fra)	Europeana (Neudecker, 2016)	9,546	2,045	2,047
Hungarian (hun)	Hungarian MTI (Szarvas et al., 2006)	4,532	648	1,294
Indonesia (ind)	(Khairunnisa, Imankulova, and Komachi, 2020)	6,707	1,437	1,438
Italian (ita)	I-CAB EVALITA 2007 & 2009 (Magnini et al., 2008)	11,227	4,136	2,068
Korean (kor)	KLUE (S. Park et al., 2021)	20,008	1,000	5,000
Latvian (lav)	(Gruzitis et al., 2018)	7,997	1,713	1,715
Nepali (nep)	(Singh, Padia, and Joshi, 2019)	2,301	328	659
Dutch (nld)	CoNLL02 (Tjong Kim Sang, 2002)	15,806	2,895	5,195
Norwegian (nor)	(Johansen, 2019)	15,696	2,410	1,939
Portuguese (por)	Second HAREM (Freitas et al., 2010) & Paramopama Júnior et al., 2015	11,258	2,412	2,414
Romanian (ron)	RONEC (Dumitrescu and Avram, 2020)	5,886	1,000	2,453
Swedish (swe)	“swedish_ner_corpus” on HuggingFace Datasets (Lhoest et al., 2021)	9,000	1,330	2,000
Ukrainian (ukr)	“benjamin/ner-uk” on HuggingFace Datasets (Lhoest et al., 2021)	10,833	1,307	668
Chinese (zho)	“msra_ner” on HuggingFace Datasets (Lhoest et al., 2021)	45,057	3,442	1,721

Table B.3: **Languages and Data Splits for Other NER Datasets**.

Language	XLM-R-large					mDeBERTaV3-base					AfroXLMR-large				
	all	0-freq	Δ 0-freq	long	Δ long	all	0-freq	Δ 0-freq	long	Δ long	all	0-freq	Δ 0-freq	long	Δ long
bam	79.4	62.3	-17.1	74.7	-4.7	81.3	66.3	-15.0	78.6	-2.7	82.1	67.2	-14.9	81.1	-1.0
bbj	74.8	66.1	-8.7	87.4	12.6	75.0	65.8	-9.2	63.9	-11.1	76.5	65.8	-10.7	80.0	3.5
ewe	89.5	75.6	-13.9	70.6	-18.9	90.0	76.9	-13.1	70	-20.0	91.0	79.7	-11.3	74.2	-16.8
fon	81.5	71.2	-10.3	69.6	-11.9	83.3	74.5	-8.8	68.1	-15.2	82.8	73.6	-9.2	68.7	-14.1
hau	87.4	83.8	-3.6	77.6	-9.8	84.8	80.0	-4.8	72.2	-12.6	87.8	84.6	-3.2	78.1	-9.7
ibo	87.0	77.4	-9.6	75.6	-11.4	89.7	82.6	-7.1	71.8	-17.9	89.1	80.9	-8.2	64.0	-25.1
kin	84.1	74.9	-9.2	75.3	-8.8	86.2	79.0	-7.2	75.3	-10.9	87.8	81.7	-6.1	77.1	-10.7
lug	87.3	75.3	-12.0	74.1	-13.2	88.7	77.4	-11.3	78.6	-10.1	89.4	79.7	-9.7	74.7	-14.7
mos	77.1	69.5	-7.6	55.8	-21.3	78.0	71.2	-6.8	58.9	-19.1	77.5	70.2	-7.3	60.1	-17.4
nya	89.7	82.0	-7.7	81.6	-8.1	91.9	86.5	-5.4	86.7	-5.2	92.2	87.3	-4.9	87.1	-5.1
pcm	89.8	84.5	-5.3	76.8	-13.0	90.2	84.9	-5.3	79.7	-10.5	90.4	86.1	-4.3	79.1	-11.3
sna	94.9	89.9	-5.0	93.3	-1.6	95.3	91.4	-3.9	92.4	-2.9	96.3	93.9	-2.4	93.9	-2.4
swa	92.8	84.1	-8.7	73.0	-19.8	92.4	82.8	-9.6	65.1	-27.3	92.3	83.0	-9.3	65.9	-26.4
tsn	86.4	74.9	-11.5	34.5	-51.9	87.0	75.8	-11.2	45.7	-41.3	89.8	80.9	-8.9	42.9	-46.9
twi	77.9	65.5	-12.4	52.2	-25.7	80.4	70.9	-9.5	62.3	-18.1	81.4	72.3	-9.1	63.2	-18.2
wol	83.3	65.9	-17.4	59.1	-24.2	83.3	67.2	-16.1	58.6	-24.7	86.2	72.0	-14.2	62.2	-24.0
xho	88.0	83.2	-4.8	76.7	-11.3	88.0	83.8	-4.2	76.2	-11.8	90.1	86.5	-3.6	78.5	-11.6
yor	86.4	78.2	-8.2	67.0	-19.4	86.8	79.2	-7.6	74.4	-12.4	90.2	85.0	-5.2	74.0	-16.2
zul	86.4	83.2	-3.2	69.5	-16.9	89.4	86.1	-3.3	68.8	-20.6	90.1	87.5	-2.6	67.1	-23.0
avg	85.5	76.2	-9.3	70.8	-14.7	86.4	78.0	-8.4	70.9	-15.5	87.5	79.9	-7.6	72.2	-15.3

Table B.4: F1 score for two varieties of hard-to-identify entities: zero-frequency entities that do not appear in the training corpus, and longer entities of four or more words.

Language	XLM-R-large				mDeBERTaV3-base				AfroXLMR-large			
	DATE	LOC	ORG	PER	DATE	LOC	ORG	PER	DATE	LOC	ORG	PER
bam	90.3	83.2	80.7	87.1	90.1	86.4	79.2	88.4	92.6	87.7	82.4	86.1
bbj	87.6	82.9	79.4	83.6	79.9	86.4	72.5	87.2	85.7	87.0	75.2	84.7
ewe	91.8	96.8	85.5	95.9	91.8	96.4	88.6	97.1	92.0	97.8	85.6	98.6
fon	85.4	89.2	86.9	94.6	86.8	93.3	89.3	94.3	85.9	91.9	86.4	94.6
hau	86.8	90.0	92.5	98.0	86.4	89.2	89.1	98.0	87.4	91	92.2	98.2
ibo	84.5	91.6	83.5	97.7	85.4	95.6	82.5	99.1	87.2	96.5	73.4	98.8
kin	88.4	92.7	84.0	94.8	87.4	95.0	87.8	97.7	88.1	95.6	89.1	99.1
lug	78.2	93.1	94.2	95.8	80.2	95.1	94.3	96.0	81.7	93.1	95.1	97.3
mos	80.3	92.7	74.4	93.1	81.6	92.1	78.9	88.3	83.2	93.7	75.4	88.9
pcm	96.6	91.1	89.7	96.9	96.1	93.1	90.9	97.3	95.6	92.4	90.9	97.1
nya	89.1	94.1	94.2	94.4	89.6	96.7	96.0	94.9	89.1	96.2	94.8	95.6
sna	95.6	95.6	96.1	98.1	96.0	95.1	96.5	98.7	96.6	95.4	97.4	99.3
swa	92.2	97.0	95.2	98.8	91.5	96.9	94.6	98.8	91.5	97.4	93.7	98.2
tsn	88.1	88.3	89.1	97.1	87.8	90.0	89.0	97.6	90.5	94.8	92.2	98.6
twi	66.7	89.3	79.4	96.1	76.5	90.4	82.9	97.5	75.7	91.4	85.1	97.7
wol	80.6	84.9	87.0	95.9	80.8	88.2	88.4	95.0	82.6	91.9	88.0	97.0
xho	90.7	91.6	93.1	96.9	89.7	92.0	93.4	98.1	91.1	93.5	95.0	98.3
yor	89.6	94.0	90.3	93.6	89.6	92.1	91.4	94.6	91.3	95.8	92.5	96.4
zul	85.0	90.1	87.8	97.1	92.2	95.5	88.1	97.1	90.8	96.2	91.8	97.2
avg	86.7	91.0	87.5	95.0	87.3	92.6	88.1	95.6	88.4	93.7	88.2	95.9

Table B.5: F1 score for the different entity types.

B.2.2.2 Capitalization

Capitalization is a very common feature for a number of natural language processing tools, such as named entity recognition systems that identify people's

names, and locations (Louis, De Waal, and Venter, 2006). Following are the four different types of the usage of capitalization in isiXhosa and isiZulu (Priatama et al., 2022):

1. Initial capitalization of words in which only the initial letter is capitalized;
2. Mixed capitalization of words in which the initial letter of the prefix is capitalized as well as the initial letter of the main word;
3. Internal capitalization in words which are found in the middle of a sentence where the prefix remains lower case and the first letter of the main word is capitalized.
4. All CAPS in words that are fully capitalized. These are usually abbreviations or acronyms;

B.3 Other NER Corpus

Table B.3 provides the NER corpus found online that we make use for determining the best transfer languages

B.4 Error Analysis of NER

Table B.4 and Table B.5 provides error analysis of MasakhaNER 2.0 based on performance on zero-frequency entities, long entities and distribution by named entity tags.

B.5 LangRank Feature Descriptions

The following definitions are listed here, originally from Y.-H. Lin et al. (2019).

Geographic distance (d_{geo}) based on the orthodromic distance between language locations obtained from Glottolog (Hammarström, Forkel, and Haspelmath, 2018).

Genetic distance (d_{gen}) based on the genealogical distance of Glottolog language tree.

Inventory distance (d_{inv}) based on the cosine distance between phonological feature vectors obtained from PHOIBLE database (Moran, McCloy, and Wright, 2014).

Syntactic distance (d_{syn}) based on cosine distance between feature vectors obtained from syntactic structures derived from WALS database (J. Nichols and Bickel, 2013). database (**wals**)."

Phonological distance (d_{pho}) based on the cosine distance between phonological feature vectors obtained from WALS and Ethnologue databases (Lewis and Linguistics, 2009).

Featural distance (d_{fea}) based on the cosine distance between feature vectors combining all 5 features mentioned above.

Transfer language dataset size (s_{tf}) The size of the transfer language's dataset.

Target language dataset size (s_{tg}) The size of the target language's dataset.

Transfer over target size ratio (sr) The size of the transfer language's dataset is divided by the size of the target language's dataset.

Entity Overlap (eo) The number of unique words that overlap between the source and target languages' training datasets.

B.6 Overlap Results

In Figure B.1, we examine the word overlap between different languages, and how this correlates with the transfer performance. In general, these two quantities are strongly correlated (Spearman's $R = 0.6, p < 0.05$), echoing a similar result described by Michael Beukman (2022). Note that the entity overlap feature used by the ranking model in the main text was calculated in a slightly different way; namely, considering *all* tokens instead of just the 4 named entities and not normalizing the overlap. This case still shows a positive correlation, although it is slightly smaller with Spearman's $R = 0.49$.

B.7 Zero-shot Transfer

Figure B.2 shows $N \times N$ transfer results to languages in MasakhaNER 2.0. We see that English is not the best transfer language in general. It is better to choose a more geographically close African language.
Figure B.3 shows $N \times N$ transfer results to languages not in MasakhaNER 2.0. We see that English appears to be the best transfer on average, which is not the case for African languages. The reason for this is that many of the non-African languages we evaluated on are from the Indo-European, similar to English.

B.8 Best Transfer Language for Other Languages

Table B.6 provides the result of the best transfer language for other languages not in MasakhaNER 2.0.

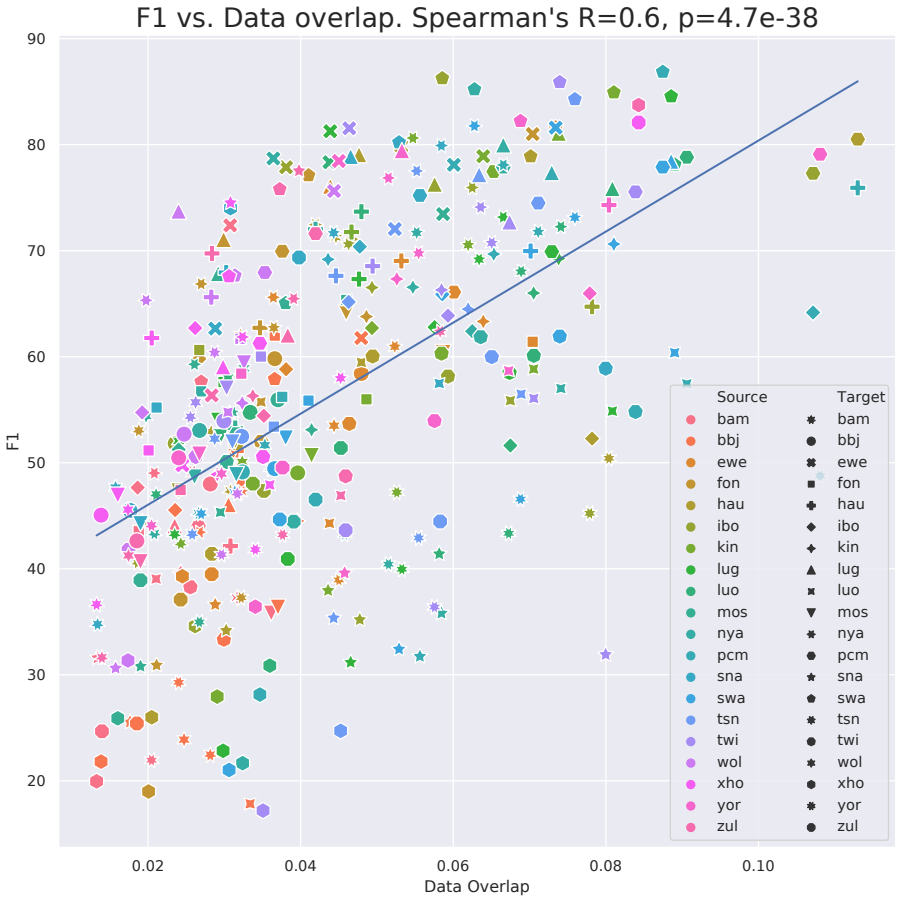


Figure B.1: The correlation between the data overlap and F1 transfer performance. For source language X and target language Y, denote the set of unique named entities (PER, ORG, LOC, DATE) by T_X and T_Y respectively. The overlap here was calculated as $\frac{|T_X \cap T_Y|}{|T_X|+|T_Y|}$, as in Y.-H. Lin et al. (2019).

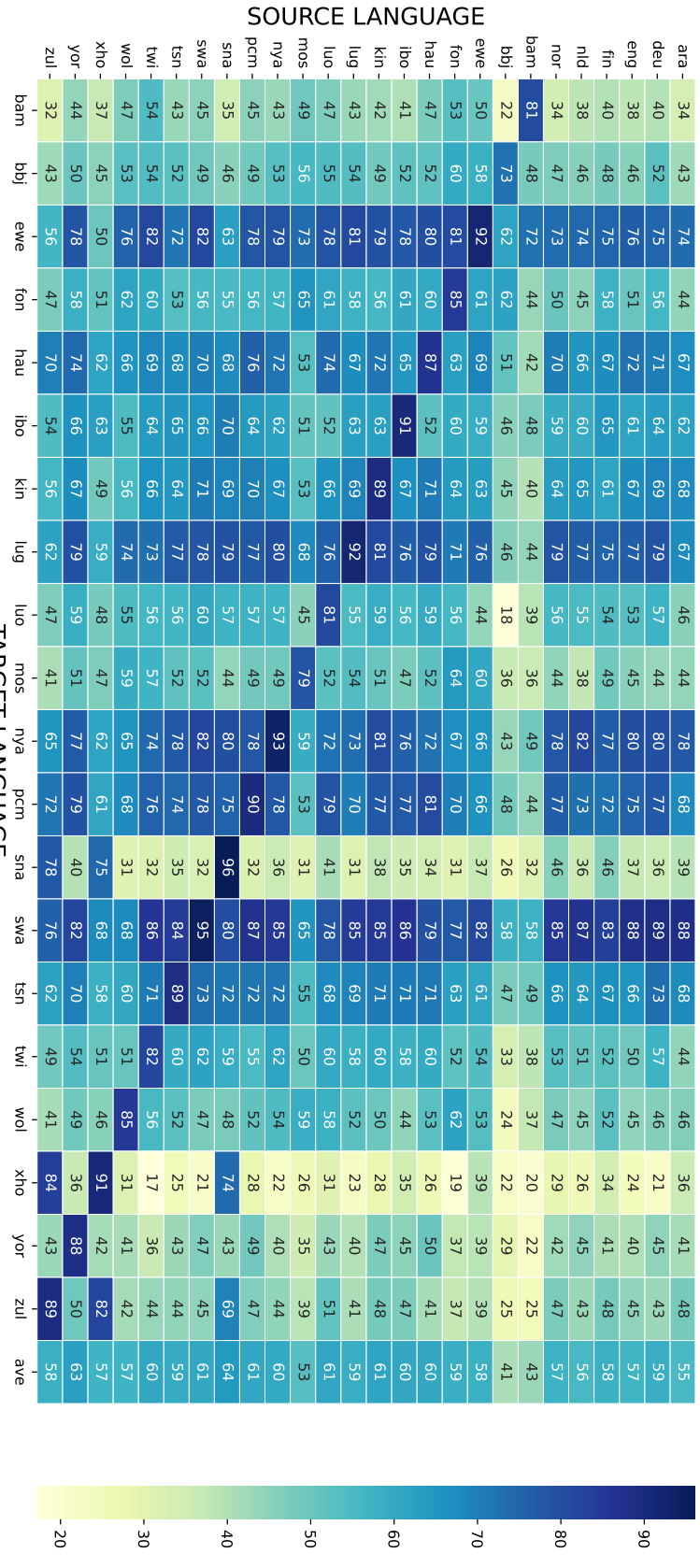


Figure B.2: **Zero-shot Transfer** from several source languages to African languages in MasakhaNER 2.0.

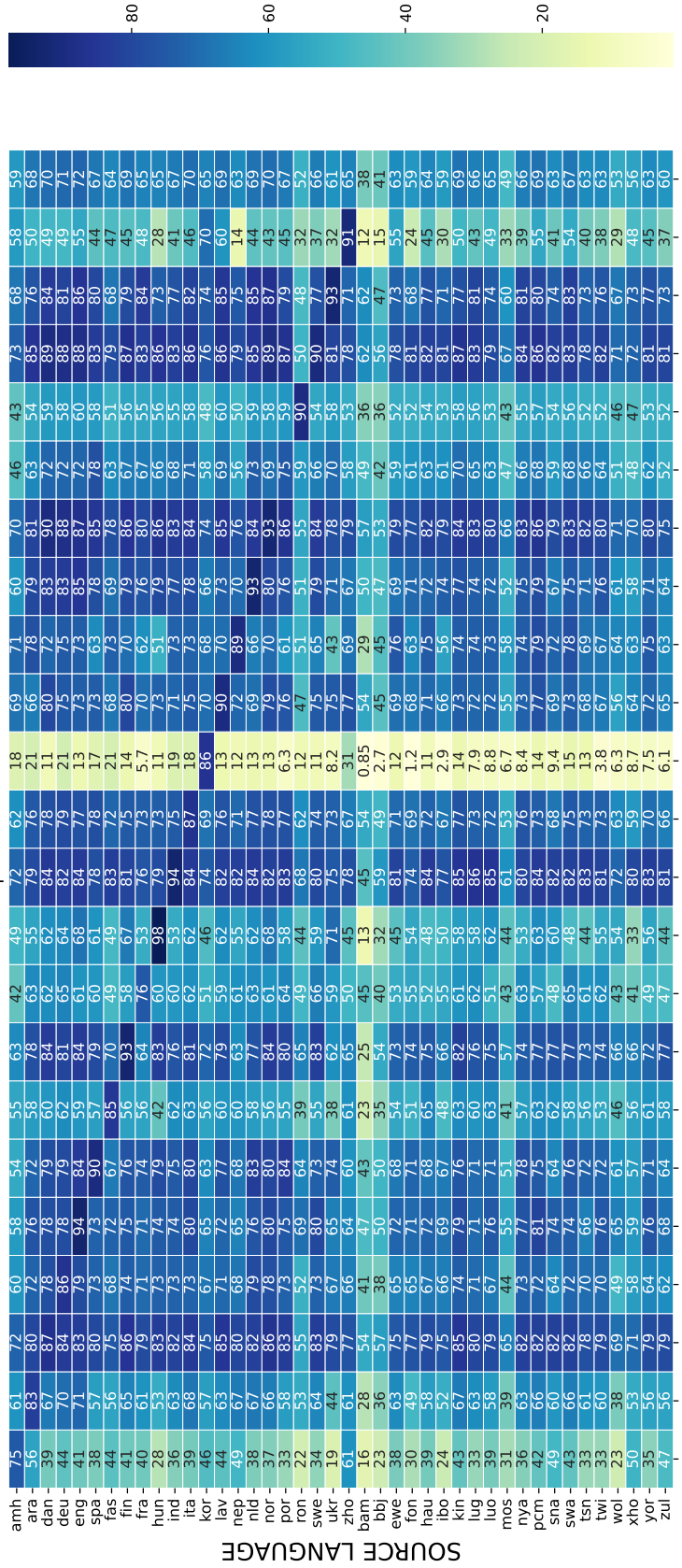


Figure B.3: **Zero-shot Transfer** from several source languages to other languages not in MasakhaNER 2.0

B.9 Sample Efficiency Results

Figure B.4 shows the result of training NER models using 100 and 500 samples for each language.

B.10 Model Hyper-parameters for Reproducibility

For training NER models, we *fine-tune* PLM, we make use of a maximum sequence length of 200, batch size of 16, gradient accumulation of 2, learning rate of 5e-5, and number of epochs 50. The experiments of the large PLMs were performed on using Nvidia V100 GPU. For AfriBERTa and mBERT, we make use of Nvidia GeForce RTX-2080Ti. For evaluation, we make use of the micro-averaged F1 score.

Target Lang.	Top-2 Transf. Lang	Top-2 LangRank Model	Top-3 features selected by the LangRank Model Lang 1; Lang 2	Target Lang. F1	Best Transf. F1	Second Best Transf. F1	eng Tranf. F1	LangRank First Lang F1	LangRank Second Lang F1
African languages									
amh	zho, ara	pcm, luo	$(s_{tf}, s_{tg}, sr); (s_{tf}, d_{geo}, sr)$	75.0	**61.0**	55.9	40.6	42.5	38.6
bam	twi, fon	wol, fon	$(d_{geo}, d_{inv}, sr); (d_{geo}, sr, d_{pho})$	80.4	**54.3**	53.0	38.4	47.1	53.0
bbj	fon, ewe	twi, ewe	$(s_{tf}, d_{syn}, d_{geo}); (s_{tf}, d_{geo}, sr)$	72.9	**59.8**	58.4	45.8	53.9	58.4
ewe	swa, twi	pcm, swa	$(d_{geo}, s_{tf}, sr); (eo, d_{geo}, s_{tf})$	91.7	**81.6**	81.5	76.4	78.1	**81.6**
fon	mos, bbj	yor, ewe	$(d_{geo}, d_{syn}, sr); (s_{tf}, d_{geo}, d_{gen})$	84.9	**65.4**	62.0	50.6	58.4	61.4
hau	pcm, yor	yor, swa	$(d_{geo}, sr, eo); (eo, sr, s_{tf})$	86.9	75.9	**74.3**	72.4	74.3	70.0
ibo	sna, yor	pcm, kin	$(eo, d_{geo}, s_{tf}); (d_{geo}, sr, eo)$	91.0	**70.4**	66.0	61.4	64.2	62.7
kin	hau, swa	sna, yor	$(eo, d_{geo}, s_{tf}); (eo, s_{tf}, sr)$	89.5	**71.1**	70.6	67.4	69.2	67.3
lug	kin, nya	luo, zul	$(d_{geo}, sr, eo); (d_{syn}, d_{geo}, sr)$	91.5	**81.1**	80.0	76.5	75.9	62.0
luo	swa, hau	lug, sna	$(d_{geo}, sr, eo); (d_{geo}, eo, sr)$	81.2	**60.4**	59.5	53.4	54.9	57.5
mos	fon, ewe	yor, fon	$(d_{geo}, d_{inv}, sr); (d_{geo}, s_{tf}, sr)$	78.9	**64.2**	60.4	45.4	50.8	**64.2**
nya	swa, nld	zul, sna	$(eo, d_{geo}, sr); (d_{geo}, eo, d_{syn})$	93.5	**81.8**	81.7	80.1	65.5	79.9
pcm	hau, yor	eng, yor	$(eo, d_{gen}, d_{syn}); (eo, d_{geo}, sr)$	89.9	**80.5**	79.1	75.5	75.5	79.1
sna	zul, xho	swa, zul	$(eo, sr, s_{tf}); (d_{geo}, sr, eo)$	96.0	**77.5**	74.5	37.1	32.4	77.5
swa	deu, ara	ita, nld	$(sr, d_{inv}, eo); (eo, s_{tf}, sr)$	94.6	**88.7**	88.1	87.9	84.5	86.6
tsn	deu, swa	swa, nya	$(eo, d_{inv}, s_{tf}); (d_{inv}, d_{geo}, d_{gen})$	88.7	**73.3**	73.1	65.8	73.1	71.7
twi	swa, nya	swa, ewe	$(eo, s_{tf}, d_{geo}); (d_{geo}, s_{tf}, sr)$	82.0	61.0	**61.9**	49.5	**61.9**	53.7
wol	fon, mos	fon, yor	$(d_{geo}, sr, s_{tf}); (sr, d_{geo}, d_{syn})$	85.2	**62.0**	58.9	44.8	**62.0**	49.0
xho	zul, sna	zul, pcm	$(eo, d_{geo}, d_{gen}); (eo, s_{tf}, d_{inv})$	90.8	**83.7**	74.0	24.5	83.7	28.1
yor	hau, pcm	fon, pcm	$(d_{geo}, d_{inv}, d_{syn}); (eo, d_{geo}, d_{inv})$	88.3	**50.3**	48.8	40.1	37.3	48.8
zul	xho, sna	xho, sna	$(eo, d_{gen}, d_{geo}); (d_{syn}, sr, d_{geo})$	88.6	**82.1**	69.4	44.7	**82.1**	69.4
Non-African languages									
ara	eng, deu	fas, pcm	$(eo, d_{inv}, d_{syn}); (d_{syn}, sr, d_{inv})$	82.8	**71.5**	69.9	**71.5**	55.7	57.9
dan	nor, fin	swe, nor	$(eo, d_{gen}, d_{geo}); (eo, d_{geo}, d_{syn})$	87.1	**86.3**	85.6	83.1	82.8	86.3
deu	nld, eng	dan, nld	$(d_{geo}, eo, s_{tf}, d_{syn}); (eo, d_{syn}, d_{geo})$	86.5	**79.3**	78.8	78.8	79.3	79.3
eng	pcm, swe	nld, pcm	$(eo, d_{geo}, d_{syn}); (eo, d_{gen} d_{pho})$	93.5	81.3	79.7	**93.5**	76.0	81.3
fas	hau, pcm	ara, eng	$(d_{syn}, d_{inv}, eo); (d_{syn}, d_{geo}, s_{tf})$	84.8	**64.8**	63.4	59.3	57.9	59.2
fin	dan, eng	deu, eng	$(eo, s_{tf}, d_{geo}); (d_{syn}, d_{geo}, eo)$	93.4	**83.7**	83.6	83.6	80.8	83.6
fra	swe, swa	nld, deu	$(eo, d_{syn}, d_{geo}); (d_{geo}, eo, sr)$	75.5	**66.3**	65.4	60.6	63.3	64.9
hun	ukr, eng	deu, ron	$(d_{geo}, d_{syn}, eo); (d_{geo}, eo, d_{syn})$	98.0	**70.7**	68.4	68.4	63.6	43.8
ind	lug, luo	zho, nld	$(s_{tg}, s_{tf}, sr); (d_{syn}, s_{tf}, eo)$	93.7	**85.9**	85.2	83.9	78.6	84.1
ita	deu, spa	nld, eng	$(d_{syn}, eo, d_{geo}); (eo, d_{syn}, d_{geo})$	86.7	**79.1**	78.2	77.0	77.1	77.1
kor	zho, ind	ara, nep	$(sr, s_{tf}, d_{syn}); (d_{inv}, d_{syn}, s_{tf})$	85.7	**31.1**	21.5	12.7	21.3	11.9
lav	fin, dan	eng, nld	$(s_{tf}, d_{syn}, sr); (s_{tf}, d_{syn}, d_{geo})$	89.7	**80.4**	80.1	73.5	73.5	69.5
nep	pcm, swa	kor, zho	$(d_{syn}, s_{tf}, d_{pho}); (s_{tf}, sr, d_{geo})$	89.5	**79.0**	77.7	73.4	68.2	68.5
nld	eng, deu	eng, nor	$(eo, d_{geo}, d_{syn}); (eo, d_{geo}, s_{tf})$	93.4	**85.4**	83.7	85.4	**85.4**	79.9
nor	dan, deu	dan, eng	$(eo, d_{geo}, s_{tf}); (eo, d_{geo}, sr)$	92.5	**89.8**	87.8	87.3	89.8	87.2
por	es, nld	spa, eng	$(eo, d_{syn}, d_{gen}); (eo, d_{syn}, d_{geo})$	75.0	**77.8**	73.5	72.0	77.8	72.0
ron	lav, eng	eng, ita	$(eo, d_{syn}, d_{geo}); (eo, d_{geo}, d_{syn})$	89.6	**59.6**	59.5	59.5	59.5	57.8
spa	eng, por	por, lav	$(eo, d_{geo}, d_{syn}); (d_{syn}, eo, d_{geo})$	89.6	**83.9**	83.6	**83.9**	83.6	77.3
swe	dan, nor	nor, nld	$(eo, d_{syn}, d_{geo}); (d_{syn}, d_{geo}, eo)$	90.3	**89.4**	89.1	88.1	89.3	85.2
ukr	nor, eng	deu, eng	$(d_{geo}, d_{syn}, sr); (d_{syn}, d_{geo}, s_{tf})$	92.6	**87.2**	85.6	85.6	81.5	85.6
zho	lav, amh	pcm, deu	$(d_{syn}, s_{tf}, s_{geo}); (d_{syn}, s_{tf}, d_{pho})$	91.4	**60.2**	58.3	54.7	54.7	48.9
AVG	–			87.7	73.3	71.2	64.6	67.3	66.2

Table B.6: **Best Transfer Language for NER.** The ranking model features are based on the definitions in Y.-H. Lin et al., 2019 like: geographic distance (d_{geo}), genetic distance (d_{gen}), inventory distance (d_{inv}), syntactic distance (d_{syn}), phonological distance (d_{pho}), transfer language dataset size (s_{tf}), target language dataset size(s_{tg}), transfer over target size ratio (sr), and entity overlap (eo).

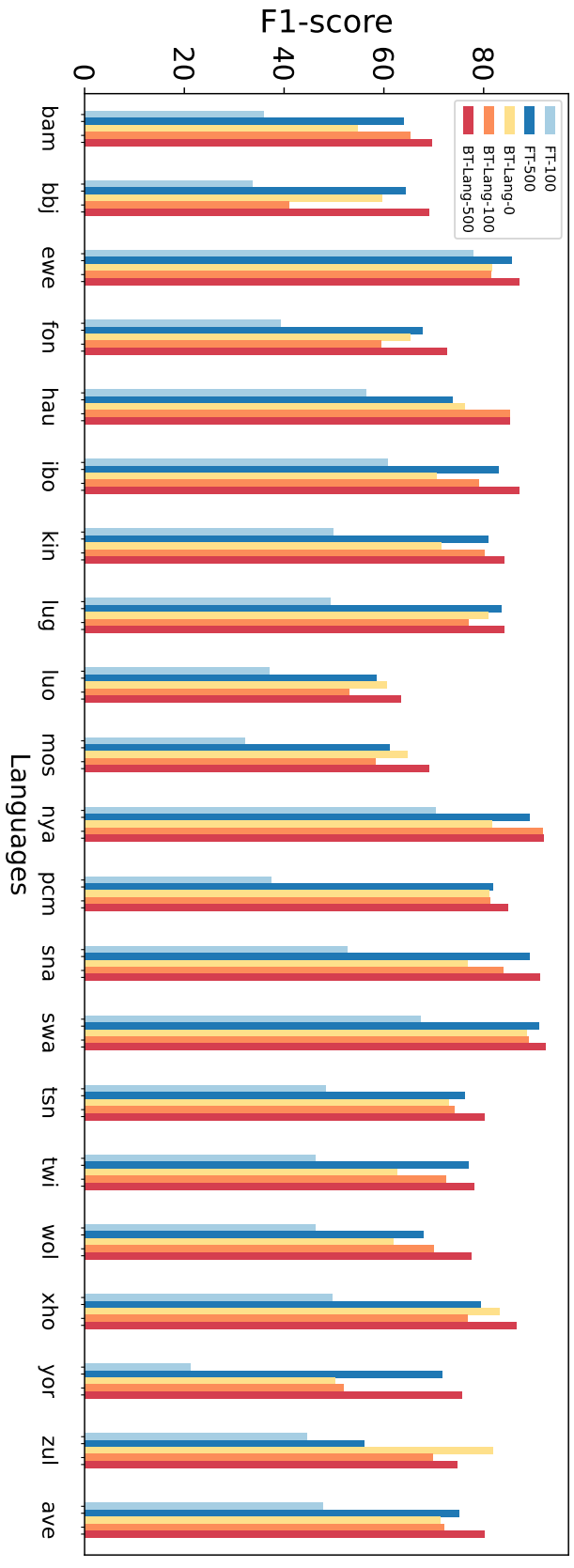


Figure B.4: **Sample Efficiency Results** for 100 and 500 samples in the target language, model fine-tuned on a PLM (e.g. FT-100 – trained on 100 samples from the target language) or fine-tuned on the best transfer language NER model (e.g. BT-Lang-0 – trained on 0 samples from the target language or zero-shot)

Appendix C

Ethics application

C.1 Ethics application

Here is the ethical approval application as approved by Lancaster University FST Ethics Committee[1].

[1] 1https://www.lancaster.ac.uk/sci-tech/research/ethics

Lancaster University

Faculty of Science and Technology Research Ethics Committee (FSTREC)
Lancaster University

Application for Ethical Approval for Research

This form should be used for all projects by staff and research students, whether funded or not, which have not been reviewed by any external research ethics committee. If your project is or has been reviewed by another committee (e.g. from another University), please contact the FST research ethics officer for further guidance.

In addition to the completed form, you need to submit **research materials** such as:

i. Participant information sheets
ii. Consent forms
iii. Debriefing sheets iv. Advertising materials (posters, e-mails)
v. Letters/emails of invitation to participate
vi. Questionnaires, surveys, demographic sheets that are non-standard
vii. Interview schedules, interview question guides, focus group scripts

Please note that **you DO NOT need to submit pre-existing questionnaires or standardized tests** that support your work, but which cannot be amended following ethical review. These should simply be referred to in your application form.

Please submit this form and any relevant materials **by email as** a **SINGLE attachment** to *fstethics@lancaster.ac.uk*

Section One

Applicant and Project Information

Name of Researcher: *Chukwuneke Chiamaka Ijeoma*

Project Title: *Named Entity Recognition for African Languages: A focus on Igbo*

Level: ***PhD***

Supervisor (if applicable): *Prof. Paul Rayson, Dr. Mahmoud El-Haj and Dr. Ignatius Ezeani*

Researcher's Email address: *c.chukwuneke@lancaster.ac.uk*

Telephone:
Address: *School of Computing and Communications, InfoLab21, Lancaster University*

Names and appointments/position of all further members of the research team:

Is this research externally funded? If yes,

ACP ID number: *Not Applicable*
Funding source: *Tertiary Education Trust Fund (TETFund), Nigeria*
Grant code : *TETF/ES/UNIV/ANNAMBRA STATE/TSAS/2019*

Does your research project involve any of the following?

☐ Human participants (including all types of interviews, questionnaires, focus groups, records relating to humans, use of internet or other secondary data, observation etc.)

☐ Animals - the term animals shall be taken to include any non-human vertebrates or cephalopods.

☐ Risk to members of the research team e.g. lone working, travel to areas where researchers may be at risk, risk of emotional distress

☐ Human cells or tissues other than those established in laboratory cultures

☐ Risk to the environment

☐ Conflict of interest

☐ Research or a funding source that could be considered controversial

☑ Social media and/or data from internet sources that could be considered private

☐ any other ethical considerations

Yes – complete the rest of this form

No – your project does not require ethical review or submission of this form

Section Two

Type of study

☐ Includes *direct* involvement by human subjects. ***Complete all sections apart from Section 3***

☐

☑ Involves *existing documents/data only*, or the evaluation of an existing project with no direct contact with human participants. ***Complete all sections apart from Section 4.***

If your research involves data from chat rooms and similar online spaces where privacy and anonymity are contentious, please complete all sections

Project Details

1. Anticipated project dates (month and year) Start date: Oct. 2020 End date: Sept. 2023

2. Please briefly describe the background to the research (no more than 150 words, in lay-person's language): Named Entity Recognition (NER) can be defined as the task of identifying names of organizations, people, currency, time, percentage expression and geographic locations in text. For instance, when we search in Google, this tool is used to identify the entities and used to locate them in the database. We propose to build this tool for Igbo an Africa Language as none is existence as of the inception of this research.

3. Please state the aims and objectives of the project (no more than 150 words, in lay-person's language):

The aim is to build NER tool for Igbo language

Objectives:

- Critically review NER works for low-resource languages
- Create high quality IgboNER human labelled datasets
- Train IgboNER models on the datasets based on the best performing methods from literature
- Perform experiments to evaluate performance across models eg effect of different data size, domains, quality

4. Methodology and Analysis:

We will use the state-of-the-art models identified in literature like:

- Robustly Optimized Bidirectional Encoder Representations from Transformers (XLM-RoBERTa)
- Multilingual Bidirectional Encoder Representations from Transformers (mBERT).

We will train the model and perform experiment to know how it will perform on Igbo Language using various data size, data from different domain etc

Section Three

Secondary Data Analysis

Complete this section if your project involves *existing documents/data only*, or the evaluation of an existing project with no direct contact with human participants

1. Please describe briefly the data or records to be studied, or the evaluation to be undertaken.

    - Jehovah Witness 300 Corpus (https://opus.nlpl.eu/JW300.php) - the license is CC-BY-NC-SA.
    - Igbo local news sites (https://igboradio.com/)- Received permission from Chidi Igwe (Assistant Professsor of French) on the 22nd July, 2021.
    - Igbo local news sites (https://kaoditaa.com/)- Received permission from the MD/CEO (Chuka Nnabuife) on the 17th, August, 2021.
    - BBC-Igbo News (https://www.bbc.com/igbo)- Permitted under copyright exceptions in the UK-such as non-commercial research or text and data mining
    - Gazzetteers (https://www.geonames.org/, INEC name lists, Igbo names) - This work is licensed under a Creative Commons Attribution 4.0 License

2. How will any data or records be obtained?

    - From publicly available sources and websites, permissions have already been checked and obtained as listed above

3. Confidentiality and Anonymity: If your study involves re-analysis and potential publication of existing data but which was gathered as part of a previous project involving direct contact with human beings, how will you ensure that your re-analysis of this data maintains confidentiality and anonymity as guaranteed in the original study?

    - There is no requirement for confidentiality and anonymity from anything that we are adding to these existing datasets.

4. What plan is in place for the storage of data (electronic, digital, paper, etc)? Please ensure that your plans comply with the General Data Protection Regulation (GDPR) and the (UK) Data Protection Act 2018.

    - Data will be stored electronically on university laptops and OneDrive using encrypted drives and password protection from university accounts.

5. What are the plans for dissemination of findings from the research?

- Documented in my Thesis and Paper publications in academic conferences and journals in the Natural Language Processing community, e.g. ACL, LREC, TACL.

6a. Is the secondary data you will be using in the public domain? Yes
6b. If NO, please indicate the original purpose for which the data was collected, and comment on whether consent was gathered for additional later use of the data.

7.What other ethical considerations (if any), not previously noted on this application, do you think there are in the proposed study? How will these issues be addressed?

- Nil

8a. Will you be gathering data from discussion forums, on-line 'chat-rooms' and similar online spaces where privacy and anonymity are contentious?

- No

If yes, your project requires full ethics review. Please complete all sections.

Section Four

Participant Information

Complete this section if your project includes *direct* involvement by human subjects.

1. Please describe briefly the **intended human participants** (including number, age, gender, and any other relevant characteristics):

2. How will participants be **recruited** and from where?

3. Briefly describe your **data collection methods**, drawing particular attention to any potential ethical issues.

   This data will be scraped from their website which is permitted for research purposes.

4. **Consent**

4a. Will you take all necessary steps to **obtain the voluntary and informed consent** of the prospective participant(s) or, in the case of individual(s) not capable of giving informed consent, the permission of a legally authorised representative in accordance with applicable law? **YES/ NO** If yes, please go to question 4b. If no, please go to question 4c.

4b. Please explain the procedure you will use for **obtaining consent?** If applicable, please explain the procedures you intend to use to gain permission on behalf of participants who are unable to give informed consent.

4c. If it will be necessary for participants to take part in the study **without their knowledge and consent at the time**, please explain why (for example covert observations may be necessary in some settings; some experiments require use of deception or partial deception – not telling participants everything about the experiment).

5. Could participation cause **discomfort** (physical and psychological eg distressing, sensitive or embarrassing topics), **inconvenience or danger beyond the risks encountered in normal life**? Please indicate plans to address these potential risks. State the timescales within which participants may withdraw from the study, noting your reasons.

6. How will you protect participants' **confidentiality and/or anonymity** in data collection (e.g. interviews), data storage, data analysis, presentation of findings and publications?

7. Do you anticipate any ethical constraints relating to **power imbalances or dependent relationships**, either with participants or with or within the research team? If yes, please explain how you intend to address these?

8. What potential **risks may exist for the researcher** and/or research team? Please indicate plans to address such risks (for example, noting the support available to you/the researcher; counselling considerations arising from the sensitive or distressing nature of the research/topic; details of the lone worker plan you or any researchers will follow, in particular when working abroad.

9. Whilst there may not be any significant direct **benefits to participants** as a result of this research, please state here any that may result from participation in the study.

10. Please explain the **rationale for any incentives/payments** (including out-of-pocket expenses) made to participants:

11. What are your plans for the **storage of data** (electronic, digital, paper, etc.)? Please ensure that your plans comply with the General Data Protection Regulation (GDPR) and the (UK) Data Protection Act 2018.

12. Please answer the following question *only* if you have not completed a Data Management Plan for an external funder.
    12.a How will you make your data available under open access requirements?

    12b. Are there any restrictions on sharing your data for open access purposes?

13. Will **audio or video recording** take place? ☐ no ☐ audio ☐ video
    13a. Please confirm that portable devices (laptop, USB drive etc) will be **encrypted** where they are used for identifiable data. If it is not possible to encrypt your portable devices, please comment on the steps you will take to protect the data.

13b. What arrangements have been made for **audio/video data storage**? At what point in the research will tapes/digital recordings/files be destroyed?

13c. If your study includes video recordings, what are the implications for participants' anonymity? Can anonymity be guaranteed and if so, how? If participants are identifiable on the recordings, how will you explain to them what you will do with the recordings? How will you seek consent from them?

14. What are the plans for dissemination of findings from the research? If you are a student, mention here your thesis. Please also include any impact activities and potential ethical issues these may raise.

15. What particular ethical considerations, not previously noted on this application, do you think there are in the proposed study? Are there any matters about which you wish to seek guidance from the FSTREC?

Section Five

Additional information required by the university insurers

If the research involves either the nuclear industry or an aircraft or the aircraft industry (other than for transport), please provide details below:

- N/A

Section Six

Declaration and Signatures

I understand that as Principal Investigator/researcher/PhD candidate I have overall responsibility for the ethical management of the project and confirm the following:

- I have read the Code of Practice, Research Ethics at Lancaster: a code of practice and I am willing to abide by it in relation to the current proposal.
- I will manage the project in an ethically appropriate manner according to: (a) the subject matter involved and (b) the Code of Practice and Procedures of the University.
- On behalf of the University I accept responsibility for the project in relation to promoting good research practice and the prevention of misconduct (including plagiarism and fabrication or misrepresentation of results).
- On behalf of the University I accept responsibility for the project in relation to the observance of the rules for the exploitation of intellectual property.

- If applicable, I will give all staff and students involved in the project guidance on the good practice and ethical standards expected in the project in accordance with the University Code of Practice. (Online Research Integrity training is available for staff and students here.)
- If applicable, I will take steps to ensure that no students or staff involved in the project will be exposed to inappropriate situations.
- I confirm that I have completed all risk assessments and other Health and Safety requirements as advised by my departmental Safety Officer.

Confirmed

Please note: If you are not able to confirm the statement above please contact the FST Research Ethics Committee and provide an explanation.

Student applicants:

Please tick to confirm that you have discussed this application with your supervisor, and that they agree to the application being submitted for ethical review

Students must submit this application from your Lancaster University email address, and copy your supervisor in to the email in which you submit this application

All Staff and Research Students must complete this declaration:

I confirm that I have sent a copy of this application to my Head of Department (or their delegated representative) . **Tick here to confirm**

Name of Head of Department *(or their delegated representative)*

Nigel Davies / Mark Rouncefield

Applicant electronic signature: *amakachi* Date 28/09/2021

References

Abdulmumin, Idris et al. (June 2022). "Hausa Visual Genome: A Dataset for Multi-Modal English to Hausa Machine Translation." In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 6471–6479. URL: https://aclanthology.org/2022.lrec-1.694.

Adelani, David et al. (Aug. 2021). "The Effect of Domain and Diacritics in Yoruba–English Neural Machine Translation." In: *Proceedings of Machine Translation Summit XVIII: Research Track*. Ed. by Kevin Duh and Francisco Guzmán. Virtual: Association for Machine Translation in the Americas, pp. 61–75. URL: https://aclanthology.org/2021.mtsummit-research.6.

Adelani, David Ifeoluwa, Jade Abbott, et al. (2021). "MasakhaNER: Named entity recognition for African languages." In: *Transactions of the Association for Computational Linguistics* 9, pp. 1116–1131.

Adelani, David Ifeoluwa, Jesujoba Oluwadara Alabi, et al. (July 2022). "A Few Thousand Translations Go a Long Way! Leveraging Pre-trained Models for African News Translation." In: *NAACL-HLT*. URL: https://openreview.net/forum?id=EtZ9h4Lqs5-.

Adelani, David Ifeoluwa, Michael A Hedderich, et al. (2020). "Distant Supervision and Noisy Label Learning for Low Resource Named Entity Recognition: A Study on Hausa and Yor\ub\'a." In: *arXiv preprint arXiv:2003.08370*.

Adelani, David Ifeoluwa, Marek Masiak, et al. (2023). "MasakhaNEWS: News Topic Classification for African languages." In: *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 144–159.

Adelani, David Ifeoluwa, Dana Ruiter, et al. (2021). "MENYO-20k: A Multi-domain English-Yorùbá Corpus for Machine Translation and Domain Adaptation." In: *2nd AfricaNLP Workshop Proceedings, AfricaNLP@EACL 2021, Virtual Event, April 19, 2021*. Ed. by Kathleen Siminyu et al. URL: https://arxiv.org/abs/2103.08647.

Adelani, DI et al. (2022). "MasakhaNER 2.0: Africa-centric Transfer Learning for Named Entity Recognition." In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*. Association for Computational Linguistics, pp. 4488–4508.

Agbo, Maduabuchi Sennen (2013). "Orthography Ttheories and the Sstandard Iigbo Oorthography." In: *Language in India* 13.4.

Agerri, Rodrigo et al. (2018). "Building named entity recognition taggers via parallel corpora." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Agić, Željko 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*, pp. 3204–3210.

Ahuja, Kabir et al. (May 2022). "Multi Task Learning For Zero Shot Performance Prediction of Multilingual Models." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 5454–5467. DOI: 10.18653/v1/2022.acl-long.374. URL: https://aclanthology.org/2022.acl-long.374.

Akbik, Alan, Tanja Bergmann, et al. (2019). "FLAIR: An easy-to-use framework for state-of-the-art NLP." In: *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59.

Akbik, Alan and Roland Vollgraf (Sept. 2017). "The Projector: An Interactive Annotation Projection Visualization Tool." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Lucia Specia, Matt Post, and Michael Paul. Copenhagen, Denmark: Association for Computational Linguistics, pp. 43–48. DOI: 10.18653/v1/D17-2008. URL: https://aclanthology.org/D17-2008.

Alabi, Jesujoba et al. (May 2020). "Massive vs. Curated Embeddings for Low-Resourced Languages: the Case of Yorùbá and Twi." English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 2754–2762. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.335.

Alabi, Jesujoba O. et al. (Oct. 2022). "Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning." In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by Nicoletta Calzolari et al. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 4336–4349. URL: https://aclanthology.org/2022.coling-1.382.

Alfred, R. et al. (2013). "A Rule-Based Named-Entity Recognition for Malay Articles." In: *International Conference on Advanced Data Mining and Applications*. URL: https://api.semanticscholar.org/CorpusID:26697777.

Alsentzer, Emily et al. (2019). "Publicly available clinical BERT embeddings." In: *arXiv preprint arXiv:1904.03323*.

Azarine, Indira Suri, Moch Arif Bijaksana, and Ibnu Asror (2019). "Named Entity Recognition on Indonesian Tweets using Hidden Markov Model." In: *2019 7th International Conference on Information and Communication Technology*

(ICoICT), pp. 1–5. URL: https://api.semanticscholar.org/CorpusID:202687331.

Babou, Cheikh Anta and Michele Loporcaro (2016). In: *Journal of African Languages and Linguistics* 37.1, pp. 1–57. DOI: doi:10.1515/jall-2016-0001. URL: https://doi.org/10.1515/jall-2016-0001.

Bari, M Saiful, Shafiq Joty, and Prathyusha Jwalapuram (2020). "Zero-resource cross-lingual named entity recognition." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 05, pp. 7415–7423.

Benajiba, Yassine, Paolo Rosso, and José Miguel Benedíruiz (2007). "Anersys: An arabic named entity recognition system based on maximum entropy." In: *Computational Linguistics and Intelligent Text Processing: 8th International Conference, CICLing 2007, Mexico City, Mexico, February 18-24, 2007. Proceedings 8*. Springer, pp. 143–153.

Benikova, Darina, Chris Biemann, and Marc Reznicek (May 2014). "NoSta-D Named Entity Annotation for German: Guidelines and Dataset." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 2524–2531. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/276_Paper.pdf.

Beukman, Michael (2022). "Analysing the effects of transfer learning on low-resourced named entity recognition performance." In: *3rd Workshop on African Natural Language Processing*.

Bharathi, A et al. (2024). "A Hybrid Named Entity Recognition System for Aviation Text." In: *EAI Endorsed Transactions on Scalable Information Systems* 11.1.

Bird, S and E Loper (2004). *NLTK: the natural language toolkit. ACL 2004*.

Bodomo, Adams and Charles Ofosu Marfo (2002). "The Morphophonology of Noun Classes in Dagaare and Akan." In: URL: https://api.semanticscholar.org/CorpusID:59748217.

Bojanowski, Piotr et al. (2017). "Enriching Word Vectors with Subword Information." In: *Transactions of the Association for Computational Linguistics* 5, pp. 135–146. ISSN: 2307-387X.

Cai, Jiong et al. (July 2023). "Improving Low-resource Named Entity Recognition with Graph Propagated Data Augmentation." In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 110–118. DOI: 10.18653/v1/2023.acl-short.11. URL: https://aclanthology.org/2023.acl-short.11.

Caines, Andrew (Oct. 2019). *The Geographic Diversity of NLP Conferences*. URL: http://www.marekrei.com/blog/geographic-diversity-of-nlp-conferences/.

Chiu, Jason P.C. and Eric Nichols (2016). "Named Entity Recognition with Bidirectional LSTM-CNNs." In: *Transactions of the Association for Computational*

Linguistics 4, pp. 357–370. DOI: 10.1162/tacl_a_00104. URL: https://www.aclweb.org/anthology/Q16-1026.

Chukwuneke, Chiamaka et al. (2022). "IgboBERT Models: Building and Training Transformer Models for the Igbo Language." In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 5114–5122.

Chukwuneke, Chiamaka Ijeoma et al. (2023). "IGBONER 2.0: EXPANDING NAMED ENTITY RECOGNITION DATASETS VIA PROJECTION." In: *4th Workshop on African Natural Language Processing*. URL: https://openreview.net/forum?id=tHUS9-vmUfC.

Chung, Hyung Won et al. (2021). "Rethinking Embedding Coupling in Pre-trained Language Models." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=xpFFI_NtgpW.

Clark, Kevin et al. (2020). "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators." In: *International Conference on Learning Representations*.

Cohen, Jacob (1960). "A coefficient of agreement for nominal scales." In: *Educational and psychological measurement* 20.1, pp. 37–46.

Collier, Nigel et al. (Aug. 2004). "Introduction to the Bio-entity Recognition Task at JNLPBA." In: *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*. Ed. by Nigel Collier, Patrick Ruch, and Adeline Nazarenko. Geneva, Switzerland: COLING, pp. 73–78. URL: https://aclanthology.org/W04-1213.

Conneau, Alexis et al. (July 2020). "Unsupervised Cross-lingual Representation Learning at Scale." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://www.aclweb.org/anthology/2020.acl-main.747.

Cunningham, Hamish, Yorick Wilks, and Robert J. Gaizauskas (1996). "GATE-a General Architecture for Text Engineering." In: *16th International Conference on Computational Linguistics, Proceedings of the Conference, COLING 1996, Center for Sprogteknologi, Copenhagen, Denmark, August 5-9, 1996*, pp. 1057–1060. URL: https://aclanthology.org/C96-2187/.

Das, Sarkar Snigdha Sarathi et al. (2022). "CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6338–6353.

De Cao, Nicola et al. (Mar. 2022). "Multilingual Autoregressive Entity Linking." In: *Transactions of the Association for Computational Linguistics* 10, pp. 274–290. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00460. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00460/2004070/tacl_a_00460.pdf. URL: https://doi.org/10.1162/tacl%5C_a%5C_00460.

De Pauw, Guy, Peter W Wagacha, and Dorothy Atieno Abade (2007). "Unsupervised Induction of Dholuo Word Classes using Maximum Entropy Learning." In:

Proceedings of the First International Computer Science and ICT Conference, p. 8. DOI: http://hdl.handle.net/11295/44250.

Derczynski, Leon et al. (Sept. 2017). "Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition." In: *Proceedings of the 3rd Workshop on Noisy User-generated Text*. Ed. by Leon Derczynski et al. Copenhagen, Denmark: Association for Computational Linguistics, pp. 140–147. DOI: 10.18653/v1/W17-4418. URL: https://aclanthology.org/W17-4418.

Devlin, Jacob et al. (June 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)*. Minneapolis, Minnesota: Association for Computational Linguistics. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Doddington, George et al. (May 2004). "The Automatic Content Extraction (ACE) Program – Tasks, Data, and Evaluation." In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. Ed. by Maria Teresa Lino et al. Lisbon, Portugal: European Language Resources Association (ELRA). URL: http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf.

Dou, Zi-Yi and Graham Neubig (2021). "Word Alignment by Fine-tuning Embeddings on Parallel Corpora." In: *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Du, Jingcheng et al. (2021). "Extracting postmarketing adverse events from safety reports in the vaccine adverse event reporting system (VAERS) using deep learning." In: *Journal of the American Medical Informatics Association* 28.7, pp. 1393–1400.

Dumitrescu, Stefan Daniel and Andrei-Marius Avram (May 2020). "Introducing RONEC - the Romanian Named Entity Corpus." In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 4436–4443. URL: https://aclanthology.org/2020.lrec-1.546.

Eberhard, David M, Gary F Simons, and Charles D Fennig (2024). *Ethnologue: Languages of the World. Twenty-seventh edition.* SIL International.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.) (2023). *Ethnologue: Languages of the World. Twenty-sixth edition.* SIL International. URL: http://www.ethnologue.com.

Ebrahimi, Abteen et al. (May 2022). "AmericasNLI: Evaluating Zero-shot Natural Language Understanding of Pretrained Multilingual Models in Truly Low-resource Languages." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 6279–6299. URL: https://aclanthology.org/2022.acl-long.435.

Ehrmann, Maud, Marco Turchi, and Ralf Steinberger (Sept. 2011). "Building a Multilingual Named Entity-Annotated Corpus Using Annotation Projection."

In: *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Ed. by Ruslan Mitkov and Galia Angelova. Hissar, Bulgaria: Association for Computational Linguistics, pp. 118–124. URL: https://aclanthology.org/R11-1017.

Eiselen, Roald (May 2016). "Government Domain Named Entity Recognition for South African Languages." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 3344–3348. URL: https://www.aclweb.org/anthology/L16-1533.

Emenanjo, Nolue (1978). *Elements of Modern Igbo Grammar - a descriptive approach*. Ibadan, Nigeria: Oxford University Press.

Enghoff, Jan Vium, Søren Harrison, and Željko Agić (2018). "Low-resource named entity recognition via multi-source projection: Not quite there yet?" In: *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pp. 195–201.

Ezeani, Ignatius (2019). "Corpus-based approaches to Igbo diacritic restoration." PhD thesis. University of Sheffield.

Ezeani, Ignatius, Mark Hepple, and Ikechukwu Onyenwe (2016). "Automatic restoration of diacritics for Igbo language." In: *International Conference on Text, Speech, and Dialogue*. Springer, pp. 198–205.

— (2017). "Lexical Disambiguation of Igbo through Diacritic Restoration." In: *SENSE 2017*, p. 53.

Ezeani, Ignatius, Paul Rayson, et al. (2020a). *Igbo-English Machine Translation: An Evaluation Benchmark*. arXiv: 2004.00648 [cs.CL]. URL: https://arxiv.org/abs/2004.00648.

— (2020b). "Igbo-english machine translation: An evaluation benchmark." In: *arXiv preprint arXiv:2004.00648*.

Faisal, Fahim, Yinkai Wang, and Antonios Anastasopoulos (May 2022). "Dataset Geography: Mapping Language Data to Language Users." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, pp. 3381–3411. DOI: 10.18653/v1/2022.acl-long.239. URL: https://aclanthology.org/2022.acl-long.239.

Fei, Hao, Meishan Zhang, and Donghong Ji (2020). "Cross-Lingual Semantic Role Labeling with High-Quality Translated Training Corpus." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7014–7026.

Feng, Zhang (2023). "Chinese medical named entity recognition of long text based on deep learning." In: *Research Square*. DOI: 10.21203/rs.3.rs-2796269/v1. URL: https://doi.org/10.21203/rs.3.rs-2796269/v1.

Fleiss, Joseph L (1971). "Measuring nominal scale agreement among many raters." In: *Psychological bulletin* 76.5, p. 378.

Freitas, Cláudia et al. (2010). "Second harem: advancing the state of the art of named entity recognition in portuguese. In quot." In: *Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis*.

Fu, Jinlan, Pengfei Liu, and Graham Neubig (Nov. 2020). "Interpretable Multi-dataset Evaluation for Named Entity Recognition." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 6058–6069. DOI: 10.18653/v1/2020.emnlp-main.489. URL: https://www.aclweb.org/anthology/2020.emnlp-main.489.

García-Ferrero, Iker, Rodrigo Agerri, and German Rigau (Dec. 2022). "Model and Data Transfer for Cross-Lingual Sequence Labelling in Zero-Resource Settings." In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Ed. by Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 6403–6416. DOI: 10.18653/v1/2022.findings-emnlp.478. URL: https://aclanthology.org/2022.findings-emnlp.478.

Government, Rwanda (2014). *Official Gazette number 41 bis of 13/10/2014*. URL: https://gazettes.africa/archive/rw/2014/rw-government-gazette-dated-2014-10-13-no-41%20bis.pdf.

Goyal, Archana, Vishal Gupta, and M.Anjan Kumar (2018). "Recent Named Entity Recognition and Classification techniques: A systematic review." In: *Comput. Sci. Rev.* 29, pp. 21–43.

Grishman, Ralph and B Sundheim (1995). "Message Understanding Conference 6." In: *Proceedings of the 16th conference on Computational linguistics*. Vol. 1, p. 466.

Grishman, Ralph and Beth M. Sundheim (1996). "Message Understanding Conference-6: A Brief History." In: *International Conference on Computational Linguistics*. URL: https://api.semanticscholar.org/CorpusID:11986411.

Gruzitis, Normunds et al. (May 2018). "Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Ed. by Nicoletta Calzolari et al. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1714.

Guo, Ruohao and Dan Roth (2021). "Constrained labeled data generation for low-resource named entity recognition." In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4519–4533.

Gururangan, Suchin et al. (2020). "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360.

Haider, Ali et al. (2023). "A Hybrid Approach for Food Name Recognition in Restaurant Reviews." In: *2023 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE, pp. 1–6.

Hammarström, Harald, Robert Forkel, and Martin Haspelmath (2018). "Glottolog 3.0." In: *Max Planck Institute for the Science of Human History*.

Haq, Rafiul et al. (2023). "Urdu named entity recognition system using deep learning approaches." In: *The Computer Journal* 66.8, pp. 1856–1869.

He, Pengcheng, Jianfeng Gao, and Weizhu Chen (2021). "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing." In: *arXiv preprint arXiv:2111.09543*.

Hedderich, Michael A, David Adelani, et al. (2020). "Transfer learning and distant supervision for multilingual transformer models: A study on African languages." In: *arXiv preprint arXiv:2010.03179*.

Hedderich, Michael A, Lukas Lange, et al. (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*, pp. 2545–2568.

Hedderich, Michael A., Lukas Lange, and Dietrich Klakow (2021). "ANEA: Distant Supervision for Low-Resource Named Entity Recognition." In: *CoRR* abs/2102.13129. arXiv: 2102.13129. URL: https://arxiv.org/abs/2102.13129.

Hogan, William (2022). "An overview of distant supervision for relation extraction with a focus on denoising and pre-training methods." In: *arXiv preprint arXiv:2207.08286*.

Honnibal, Matthew and Ines Montani (2017). "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing." In: *To appear* 7.1, pp. 411–420.

Howard, Jeremy and Sebastian Ruder (2018). "Universal Language Model Fine-tuning for Text Classification." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339.

Hu, Junjie et al. (2020). "XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization." In: *Proceedings of ICML 2020*. arXiv: arXiv:2003.11080v1.

Hu, Shulin et al. (2022). "Chinese Named Entity Recognition based on BERT-CRF Model." In: *2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS)*. IEEE, pp. 105–108.

Hu, Yan et al. (2024). *Improving Large Language Models for Clinical Named Entity Recognition via Prompt Engineering*. arXiv: 2303.16416 [cs.CL].

Huang, Zhiheng, Wei Xu, and Kai Yu (2015). *Bidirectional LSTM-CRF Models for Sequence Tagging*. eprint: arXiv:1508.01991.

Hvingelby, Rasmus et al. (May 2020). "DaNE: A Named Entity Resource for Danish." English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 4597–4604. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.565.

Iovine, Andrea et al. (2022). "CycleNER: An Unsupervised Training Approach for Named Entity Recognition." In: *Proceedings of the ACM Web Conference 2022*. WWW '22. Virtual Event, Lyon, France: Association for Computing

Machinery, pp. 2916–2924. ISBN: 9781450390965. DOI: 10.1145/3485447.3512012. URL: https://doi.org/10.1145/3485447.3512012.

Jain, Arti, Divakar Yadav, and Dev Tayal (2014). "NER for Hindi language using association rules." In: *2014 International Conference on Data Mining and Intelligent Computing (ICDMIC)*, pp. 1–5. URL: https://api.semanticscholar.org/CorpusID:15118454.

James, A. Lloyd (1928). "The Practical Orthography of African Languages." In: *Africa* 1.1, pp. 125–129. DOI: 10.2307/1155869.

Jibril, Ebrahim Chekol and A. Cüneyd Tantuğ (2022). "ANEC: An Amharic Named Entity Corpus and Transformer Based Recognizer." In: *IEEE Access* 11, pp. 15799–15815. URL: https://api.semanticscholar.org/CorpusID:250264963.

— (2023). "ANEC: An Amharic Named Entity Corpus and Transformer Based Recognizer." In: *IEEE Access* 11, pp. 15799–15815. DOI: 10.1109/ACCESS.2023.3243468.

Johansen, Bjarte (2019). "Named-entity recognition for Norwegian." In: *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pp. 222–231.

Júnior, C Mendonça et al. (2015). "Paramopama: a brazilian-portuguese corpus for named entity recognition." In: *Encontro Nac. de Int. Artificial e Computacional*.

Jurafsky, Dan (2020). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*.

K, Karthikeyan et al. (2020). "Cross-Lingual Ability of Multilingual BERT: An Empirical Study." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=HJeT3yrtDr.

Kann, Katharina, Ophélie Lacroix, and Anders Søgaard (2020). "Weakly Supervised POS Taggers Perform Poorly on Truly Low-Resource Languages." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05, pp. 8066–8073.

Karamolegkou, Antonia and Sara Stymne (May 2021). "Investigation of Transfer Languages for Parsing Latin: Italic Branch vs. Hellenic Branch." In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Ed. by Simon Dobnik and Lilja Øvrelid. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, pp. 315–320. URL: https://aclanthology.org/2021.nodalida-main.32.

Keraghel, Imed, Stanislas Morbieu, and Mohamed Nadif (2024). "A survey on recent advances in named entity recognition." In: *arXiv preprint arXiv:2401.10825*.

Khairunnisa, Siti Oryza, Aizhan Imankulova, and Mamoru Komachi (Dec. 2020). "Towards a Standardized Dataset on Indonesian Named Entity Recognition." In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*. Ed. by Boaz Shmueli and Yin Jou Huang. Suzhou, China: Association for Computational Linguistics, pp. 64–71. URL: https://aclanthology.org/2020.aacl-srw.10.

Kim, Jin-Dong et al. (2003). "GENIA corpus - a semantically annotated corpus for bio-textmining." In: *Bioinformatics* 19 Suppl 1, pp. i180–2. URL: https://api.semanticscholar.org/CorpusID:11522524.

El-Kishky, Ahmed et al. (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)*, pp. 5960–5969.

Konoshenko, Maria Yu and Dasha Shavarina (2019). "A microtypological survey of noun classes in Kwa." In: *Journal of African Languages and Linguistics* 40, pp. 114–75. URL: https://api.semanticscholar.org/CorpusID:198490426.

Kulshreshtha, Saurabh, Jose Luis Redondo Garcia, and Ching-Yun Chang (Nov. 2020). "Cross-lingual Alignment Methods for Multilingual BERT: A Comparative Study." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 933–942. DOI: 10.18653/v1/2020.findings-emnlp.83. URL: https://aclanthology.org/2020.findings-emnlp.83.

Kwartler, Ted (2017). "The OpenNLP Project." In: *Text Mining in Practice with R; John Wiley & Sons, Inc.: Hoboken, NJ, USA*, pp. 237–269.

Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira (2001). "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data." In: *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 282–289. ISBN: 1-55860-778-1. URL: http://dl.acm.org/citation.cfm?id=645530.655813.

Lample, Guillaume et al. (June 2016). "Neural Architectures for Named Entity Recognition." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 260–270. DOI: 10.18653/v1/N16-1030. URL: https://www.aclweb.org/anthology/N16-1030.

Lauscher, Anne et al. (Nov. 2020). "From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4483–4499. DOI: 10.18653/v1/2020.emnlp-main.363. URL: https://www.aclweb.org/anthology/2020.emnlp-main.363.

Lewis, M.P. and Summer Institute of Linguistics (2009). *Ethnologue: Languages of the World*. Ethnologue: Languages of the World. SIL International. ISBN: 9781556712166. URL: https://books.google.co.uk/books?id=FVVFPgAACAAJ.

Lhoest, Quentin et al. (Nov. 2021). "Datasets: A Community Library for Natural Language Processing." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Heike Adel and Shuming Shi. Online and Punta Cana, Dominican Republic:

Association for Computational Linguistics, pp. 175–184. DOI: 10.18653/v1/2021.emnlp-demo.21. URL: https://aclanthology.org/2021.emnlp-demo.21.

Li, Bing, Yujie He, and Wenjin Xu (2021). "Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment." In: *arXiv preprint arXiv:2101.11112*.

Li, Jing et al. (2020). "A survey on deep learning for named entity recognition." In: *IEEE Transactions on Knowledge and Data Engineering*.

Liang, Chen et al. (2020). "Bond: Bert-assisted open-domain named entity recognition with distant supervision." In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 1054–1064.

Lin, Yu-Hsiang et al. (July 2019). "Choosing Transfer Languages for Cross-Lingual Learning." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 3125–3135. DOI: 10.18653/v1/P19-1301. URL: https://aclanthology.org/P19-1301.

Lin, Ying et al. (July 2018). "Platforms for Non-speakers Annotating Names in Any Language." In: *Proceedings of ACL 2018, System Demonstrations*. Melbourne, Australia: Association for Computational Linguistics, pp. 1–6. DOI: 10.18653/v1/P18-4001. URL: https://www.aclweb.org/anthology/P18-4001.

Littell, Patrick et al. (Apr. 2017). "URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, pp. 8–14. URL: https://aclanthology.org/E17-2002.

Liu, Pengfei et al. (Aug. 2021). "ExplainaBoard: An Explainable Leaderboard for NLP." In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. Ed. by Heng Ji, Jong C. Park, and Rui Xia. Online: Association for Computational Linguistics, pp. 280–289. DOI: 10.18653/v1/2021.acl-demo.34. URL: https://aclanthology.org/2021.acl-demo.34.

Liu, Yinhan et al. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. eprint: arXiv:1907.11692.

Louis, Anita, Alta De Waal, and Cobus Venter (2006). "Named entity recognition in a South African context." In: *Proceedings of the 2006 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries*. SAICSIT '06. Somerset West, South Africa: South African Institute for Computer Scientists and Information Technologists, pp. 170–179. ISBN: 1595935673. DOI: 10.1145/1216262.1216281. URL: https://doi.org/10.1145/1216262.1216281.

Ma, Xuezhe and Eduard Hovy (Aug. 2016). "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Katrin Erk and Noah A. Smith. Berlin, Germany: Association for Computational Linguistics, pp. 1064–1074. DOI: 10.18653/v1/P16-1101. URL: https://aclanthology.org/P16-1101.

Macgregor, J. K. (1909). "Some Notes on Nsibidi." In: *The Journal of the Royal Anthropological Institute of Great Britain and Ireland* 39, pp. 209–219. ISSN: 03073114. URL: http://www.jstor.org/stable/2843292 (visited on 01/12/2024).

Magnini, Bernardo et al. (2008). "Evaluation of Natural Language Tools for Italian: EVALITA 2007." In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

MARTEN, LUTZ (2005). "DEREK NURSE AND GÉRARD PHILIPPSON (eds): The Bantu Languages. (Routledge Language Family Series.) London: Routledge, 2003. xvii, 708 pages. £170." In: *Bulletin of the School of Oriental and African Studies* 68.3. DOI: 10.1017/S0041977X05490278.

Martinus, Laura and Jade Z Abbott (2019). "A focus on neural machine translation for African languages." In: *arXiv preprint arXiv:1906.05685*. URL: https://arxiv.org/abs/1906.05685.

Mayhew, Stephen, Tatiana Tsygankova, and Dan Roth (Nov. 2019). "ner and pos when nothing is capitalized." 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)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 6256–6261. DOI: 10.18653/v1/D19-1650. URL: https://aclanthology.org/D19-1650.

MBONU, CHINEDU EMMANUEL et al. (2022). "IgboSum1500-Introducing the Igbo Text Summarization Dataset." In: *3rd Workshop on African Natural Language Processing*.

Melzian, Hans J. (1933). "Introduction to the Phonology of the Bantu Languages. By Carl Meinhof. Translated, revised, and enlarged in collaboration with the author and DrAlice Werner, by N. J. v. Warmelo. pp. 248, 1 map. Berlin: Dietrich Reimer (Ernst Vohsen). London: Williams 38;amp; Norgate, Ltd., 1932." In: *Bulletin of the School of Oriental and African Studies* 7.1, pp. 246–247. DOI: 10.1017/S0041977X00105828.

Meng, Yu et al. (Nov. 2021). "Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens et al. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 10367–10378. DOI: 10.18653/v1/2021.emnlp-main.810. URL: https://aclanthology.org/2021.emnlp-main.810.

Mengliev, Davlatyor B. et al. (2023). "Developing Rule-Based and Gazetteer Lists for Named Entity Recognition in Uzbek Language: Geographical Names." eng. In: *2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE)*. IEEE, pp. 1500–1504. ISBN: 9798350330885.

Mikolov, Tomas et al. (2013). "Distributed Representations of Words and Phrases and their Compositionality." In: *Advances in Neural Information Processing Systems*. Ed. by C. J. C. Burges et al. Vol. 26. Curran Associates, Inc., pp. 3111–3119. URL: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Mintz, Mike et al. (2009). "Distant supervision for relation extraction without labeled data." In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 1003–1011.

Moran, Steven, Daniel McCloy, and Richard Wright (2014). "PHOIBLE online." In:

Muhammad, Shamsuddeen Hassan et al. (June 2022). "NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis." In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 590–602. URL: https://aclanthology.org/2022.lrec-1.63.

Nadeau, David and Satoshi Sekine (2007). "A survey of named entity recognition and classification." In: *Lingvisticae Investigationes* 30.1, pp. 3–26.

Nekoto, Wilhelmina et al. (2020). "Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages." In: *arXiv preprint arXiv:2010.02353*.

Neubig, Graham et al. (2017). "DyNet: The Dynamic Neural Network Toolkit." In: *ArXiv* abs/1701.03980.

Neudecker, Clemens (May 2016). "An Open Corpus for Named Entity Recognition in Historic Newspapers." In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Ed. by Nicoletta Calzolari et al. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 4348–4352. URL: https://aclanthology.org/L16-1689.

Nichols, Johanna and Balthasar Bickel (2013). "Possessive Classification (v2020.3)." In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Zenodo. DOI: 10.5281/zenodo.7385533. URL: https://doi.org/10.5281/zenodo.7385533.

Niyongabo, Rubungo Andre et al. (Dec. 2020). "KINNEWS and KIRNEWS: Benchmarking Cross-Lingual Text Classification for Kinyarwanda and Kirundi." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 5507–5521. DOI: 10.18653/v1/2020.coling-main.480. URL: https://www.aclweb.org/anthology/2020.coling-main.480.

Nothman, Joel et al. (2013). "Learning multilingual named entity recognition from Wikipedia." In: *Artificial Intelligence* 194, pp. 151–175.

Nurse, Derek and Gérard Philippson (2006). *The bantu languages*. Vol. 4. Routledge.

Nwagu, Nkemjika Bernardine (2023). "IGBO ENTREPRENEURSHIP AND ECONOMIC DEVELOPMENT: A CATALYST FOR GROWTH AND PROGRESS." In: *Nnadiebube Journal of Social Sciences* 3.1.

Obeid, Ossama et al. (May 2020). "CAMeL Tools: An Open Source Python Toolkit for Arabic Natural Language Processing." English. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, pp. 7022–7032. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.868.

Offiong Mensah, Eyo (2012). "Grammaticalization in Nigerian Pidgin." In: *Íkala, revista de lenguaje y cultura* 17.2, pp. 167–179.

Ogueji, Kelechi, Yuxin Zhu, and Jimmy Lin (Nov. 2021). "Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages." In: *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Ed. by Duygu Ataman et al. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 116–126. DOI: 10.18653/v1/2021.mrl-1.11. URL: https://aclanthology.org/2021.mrl-1.11.

Ogundepo, Odunayo et al. (2023). "Cross-lingual Open-Retrieval Question Answering for African Languages." In: *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14957–14972.

Ojarikre, Anthony (2013). "Perspectives and problems of codifying Nigerian pidgin English orthography." In: *Perspectives* 3.12.

Oladipo, Akintunde et al. (2022). "An Exploration of Vocabulary Size and Transfer Effects in Multilingual Language Models for African Languages." In: *3rd Workshop on African Natural Language Processing*. URL: https://openreview.net/forum?id=HOZmF9MV8Wc.

Onyenwe, Ignatius Ezeani Mark Hepple Ikechukwu and Chioma Enemuo (2018). "Transferred Embeddings for Igbo Similarity, Analogy and Diacritic Restoration Tasks." In: *COLING 2018*, p. 30.

Onyenwe, Ikechukwu et al. (2015). "Use of transformation-based learning in annotation pipeline of igbo, an african language." In: *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*. Association for Computational Linguistics, pp. 24–33.

Onyenwe, Ikechukwu E and Mark Hepple (2016). "Predicting Morphologically-Complex Unknown Words in Igbo." In: *international conference on text, speech, and dialogue*. Springer, pp. 206–214.

Onyenwe, Ikechukwu E, Mark Hepple, Uchechukwu Chinedu, et al. (2018). "A Basic Language Resource Kit Implementation for the Igbo NLP Project."

In: *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 17.2, pp. 1–23.

Onyenwe, Ikechukwu E, Mark Hepple, Uchechukwu Chinedu, et al. (2019). "Toward an effective igbo part-of-speech tagger." In: *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 18.4, pp. 1–26.

Onyenwe, Ikechukwu E, Chinedu Uchechukwu, and Mark Hepple (2014). "Part-of-speech Tagset and Corpus Development for Igbo, an African." In: *LAW VIII*, p. 93.

Onyenwe, Ikechukwu Ekene (2017). "Developing methods and resources for automated processing of the African language Igbo." PhD thesis. University of Sheffield.

Oosthuysen, Jacobus Christiaan (2016). *The grammar of isiXhosa*. African Sun Media.

Oraka, Louis Nnamdi (1983). *The foundations of Igbo studies: A short history of the study of Igbo language and culture*. University Publishing Company.

Pakhale, Kalyani (2023). "Comprehensive overview of named Entity Recognition: Models, Domain-Specific applications and challenges." In: *arXiv preprint arXiv:2309.14084*.

Pan, Xiaoman et al. (July 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)*. Ed. by Regina Barzilay and Min-Yen Kan. Vancouver, Canada: Association for Computational Linguistics, pp. 1946–1958. DOI: 10.18653/v1/P17-1178. URL: https://aclanthology.org/P17-1178.

Pande, Sandeep Dwarkanath, R Kishore Kanna, Imran Qureshi, et al. (2022). "Natural language processing based on name entity with n-gram classifier machine learning process through ge-based hidden markov model." In: *Machine Learning Applications in Engineering Education and Management* 2.1, pp. 30–39.

Pandey, Priya and Bharti Nathani (2024). "State-of-art approach for Indian Language based on NER: Comprehensive Review." In: *Research Square*. DOI: 10.21203/rs.3.rs-3827718/v1. URL: https://doi.org/10.21203/rs.3.rs-3827718/v1.

Park, Cheoneum, Seohyeong Jeong, and Juae Kim (2023). "ADMit: Improving NER in automotive domain with domain adversarial training and multi-task learning." In: *Expert Systems with Applications* 225, p. 120007.

Park, Sungjoon et al. (2021). "KLUE: Korean Language Understanding Evaluation." In: *ArXiv* abs/2105.09680. URL: https://api.semanticscholar.org/CorpusID:234790338.

Patil, Nita, Ajay Patil, and B.V. Pawar (2020). "Named Entity Recognition using Conditional Random Fields." In: *Procedia Computer Science* 167. International Conference on Computational Intelligence and Data Science, pp. 1181–1188. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2020.03.

431. URL: https://www.sciencedirect.com/science/article/pii/S1877050920308978.

Payne, Doris L., Sara Pacchiarotti, and Mokaya Bosire, eds. (2017). *Diversity in African languages. Selected papers from the 46th Annual Conference on African Linguistics*. Contemporary African Linguistics 1. Berlin: Language Science Press. DOI: 10.17169/langsci.b121.280.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://www.aclweb.org/anthology/D14-1162.

Pfeiffer, Jonas, Ivan Vuli, et al. (2020). "MAD-X: An Adapter-based Framework for Multi-task Cross-lingual Transfer." In: *Proceedings of EMNLP 2020*.

Pfeiffer, Jonas, Ivan Vulić, et al. (2020). "UNKs Everywhere: Adapting Multilingual Language Models to New Scripts." In: *arXiv preprint arXiv:2012.15562*.

Pires, Telmo, Eva Schlinger, and Dan Garrette (July 2019). "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David Traum, and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, pp. 4996–5001. DOI: 10.18653/v1/P19-1493. URL: https://aclanthology.org/P19-1493.

Ponti, Edoardo Maria et al. (Nov. 2020). "XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, pp. 2362–2376. DOI: 10.18653/v1/2020.emnlp-main.185. URL: https://aclanthology.org/2020.emnlp-main.185.

Poostchi, Hanieh et al. (Dec. 2016). "PersoNER: Persian Named-Entity Recognition." In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Ed. by Yuji Matsumoto and Rashmi Prasad. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 3381–3389. URL: https://aclanthology.org/C16-1319.

Priatama, Aditya Rizky et al. (2022). "Regression models for estimating aboveground biomass and stand volume using landsat-based indices in post-mining area." In: *Jurnal Manajemen Hutan Tropika* 28.1, pp. 1–14.

Pruksachatkun, Yada et al. (July 2020). "Intermediate-Task Transfer Learning with Pretrained Language Models: When and Why Does It Work?" In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 5231–5247. DOI: 10.18653/v1/2020.acl-main.467. URL: https://aclanthology.org/2020.acl-main.467.

Puccetti, Giovanni et al. (2023). "Technology identification from patent texts: A novel named entity recognition method." In: *Technological Forecasting and Social Change* 186, p. 122160.

Quirk, Chris and Hoifung Poon (Apr. 2017). "Distant Supervision for Relation Extraction beyond the Sentence Boundary." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Ed. by Mirella Lapata, Phil Blunsom, and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, pp. 1171–1182. URL: https://aclanthology.org/E17-1110.

Ratinov, Lev and Dan Roth (June 2009). "Design Challenges and Misconceptions in Named Entity Recognition." In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*. Boulder, Colorado: Association for Computational Linguistics, pp. 147–155. URL: https://www.aclweb.org/anthology/W09-1119.

Reid, Machel et al. (Nov. 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*. Ed. by Marie-Francine Moens et al. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1306–1320. DOI: 10.18653/v1/2021.emnlp-main.99. URL: https://aclanthology.org/2021.emnlp-main.99.

Reimers, Nils and Iryna Gurevych (Nov. 2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. URL: https://arxiv.org/abs/1908.10084.

Rijhwani, Shruti et al. (July 2020). "Soft Gazetteers for Low-Resource Named Entity Recognition." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8118–8123. DOI: 10.18653/v1/2020.acl-main.722. URL: https://www.aclweb.org/anthology/2020.acl-main.722.

Rijsbergen, C. J. Van (1979). *Information Retrieval*. 2nd. USA: Butterworth-Heinemann. ISBN: 0408709294.

Ruder, Sebastian (2020). *Why You Should Do NLP Beyond English*. http://ruder.io/nlp-beyond-english.

Ruder, Sebastian, Noah Constant, et al. (Nov. 2021). "XTREME-R: Towards More Challenging and Nuanced Multilingual Evaluation." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens et al. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 10215–10245. DOI: 10.18653/v1/2021.emnlp-main.802. URL: https://aclanthology.org/2021.emnlp-main.802.

Ruder, Sebastian, Anders Søgaard, and Ivan Vulić (2019). "Unsupervised Cross-Lingual Representation Learning." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 31–38.

Ruokolainen, Teemu et al. (2020). "A Finnish news corpus for named entity recognition." In: *Language Resources and Evaluation* 54, pp. 247–272.

Sang, Erik F. Tjong Kim and Fien De Meulder (2003). *Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition*. arXiv: cs/0306050 [cs.CL].

Sanh, V (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." In: *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NIPS2019)*.

Sanh, Victor et al. (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." In: *ArXiv* abs/1910.01108.

Segura-Bedmar, Isabel, Paloma Martínez, and María Herrero-Zazo (June 2013). "SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013)." In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Ed. by Suresh Manandhar and Deniz Yuret. Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 341–350. URL: https://aclanthology.org/S13-2056.

Shaalan, K. (2014). "A Survey of Arabic Named Entity Recognition and Classification." In: *Computational Linguistics* 40, pp. 469–510.

Shaalan, Khaled F. and Hafsa Raza (2009). "NERA: Named Entity Recognition for Arabic." In: *J. Assoc. Inf. Sci. Technol.* 60, pp. 1652–1663. URL: https://api.semanticscholar.org/CorpusID:16387993.

Shrestha, Sabita (2024). "Named Entity Recognition for Nepali Text Using Pre-Trained BERT-Based Model." masterthesis. Hochschule Rhein-Waal.

Singh, Oyesh Mann, Ankur Padia, and Anupam Joshi (2019). "Named Entity Recognition for Nepali Language." In: *2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)*, pp. 184–190. DOI: 10.1109/CIC48465.2019.00031.

Strassel, Stephanie and Jennifer Tracey (May 2016). "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)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 3273–3280. URL: https://www.aclweb.org/anthology/L16-1521.

Szarvas, György et al. (2006). "A highly accurate Named Entity corpus for Hungarian." In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*.

Tiedemann, Jörg (2012). "Parallel Data, Tools and Interfaces in OPUS." In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 2214–2218.

Tinner, Francesco et al. (2023). "Findings of the 1st Shared Task on Multi-lingual Multi-task Information Retrieval at MRL 2023." In: *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pp. 310–323.

Tjong Kim Sang, Erik F. (2002). "Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition." In: *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. URL: https://www.aclweb.org/anthology/W02-2024.

Tjong Kim Sang, Erik F. and Sabine Buchholz (2000). "Introduction to the CoNLL-2000 Shared Task Chunking." In: *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*. URL: https://aclanthology.org/W00-0726.

Tjong Kim Sang, Erik F. and Fien De Meulder (2003). "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition." In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147. URL: https://aclanthology.org/W03-0419.

Tkachenko, Alexander, Timo Petmanson, and Sven Laur (Aug. 2013). "Named Entity Recognition in Estonian." In: *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*. Ed. by Jakub Piskorski et al. Sofia, Bulgaria: Association for Computational Linguistics, pp. 78–83. URL: https://aclanthology.org/W13-2412.

Tsygankova, Tatiana et al. (2021). "Building Low-Resource NER Models Using Non-Speaker Annotations." In: *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pp. 62–69.

Uchechukwu, Chinedu (2005). "The representation of Igbo with the appropriate keyboard." In:

Vaswani, Ashish et al. (2017). "Attention is All you Need." In: *Neural Information Processing Systems*. URL: https://api.semanticscholar.org/CorpusID:13756489.

Versteegh, Kees (2001). "Linguistic Contacts Between Arabic and Other Languages." In: *Arabica* 48.4, pp. 470–508. DOI: 10.1163/157005801323163825. URL: https://brill.com/view/journals/arab/48/4/article-p470_3.xml.

Vries, Wietse de, Martijn Wieling, and Malvina Nissim (May 2022). "Make the Best of Cross-lingual Transfer: Evidence from POS Tagging with over 100 Languages." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 7676–7685. URL: https://aclanthology.org/2022.acl-long.529.

Wan, Q et al. (2020). "A self-attention based neural architecture for Chinese medical named entity recognition." In: *Mathematical Biosciences and Engineering: MBE* 17.4, pp. 3498–3511.

Wang, Jiayi et al. (2023). "AfriMTE and AfriCOMET: Empowering COMET to Embrace Under-resourced African Languages." In: *arXiv preprint arXiv:2311.09828*.

Weischedel, Ralph M. et al. (2017). "OntoNotes : A Large Training Corpus for Enhanced Processing." In: URL: https://api.semanticscholar.org/CorpusID:204845447.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, et al. (2019). "HuggingFace's Transformers: State-of-the-art Natural Language Processing." In: *ArXiv* abs/1910.03771.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, et al. (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Qun Liu and David Schlangen. Online: Association for Computational Linguistics, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6.

Wu, Shijie and Mark Dredze (Nov. 2019). "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT." 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)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, pp. 833–844. DOI: 10.18653/v1/D19-1077. URL: https://aclanthology.org/D19-1077.

Xia, Mengzhou et al. (July 2020). "Predicting Performance for Natural Language Processing Tasks." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, pp. 8625–8646. DOI: 10.18653/v1/2020.acl-main.764. URL: https://aclanthology.org/2020.acl-main.764.

Xiao, Chaojun et al. (Nov. 2020). "Denoising Relation Extraction from Document-level Distant Supervision." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, pp. 3683–3688. DOI: 10.18653/v1/2020.emnlp-main.300. URL: https://aclanthology.org/2020.emnlp-main.300.

Xie, Jiateng et al. (2018). "Neural Cross-Lingual Named Entity Recognition with Minimal Resources." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 369–379.

Xie, Qizhe et al. (2020). "Unsupervised data augmentation for consistency training." In: *Advances in neural information processing systems* 33, pp. 6256–6268.

Yadav, Vikas and Steven Bethard (Aug. 2018). "A Survey on Recent Advances in Named Entity Recognition from Deep Learning models." In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2145–2158. URL: https://www.aclweb.org/anthology/C18-1182.

Yamada, Ikuya et al. (Nov. 2020). "LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 6442–6454. DOI: 10.

18653/v1/2020.emnlp-main.523. URL: https://www.aclweb.org/anthology/2020.emnlp-main.523.

Yimam, Seid Muhie et al. (Dec. 2020). "Exploring Amharic Sentiment Analysis from Social Media Texts: Building Annotation Tools and Classification Models." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by Donia Scott, Nuria Bel, and Chengqing Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 1048–1060. DOI: 10.18653/v1/2020.coling-main.91. URL: https://aclanthology.org/2020.coling-main.91.

Yohannes, Hailemariam Mehari and Toshiyuki Amagasa (2022). "Named-entity recognition for a low-resource language using pre-trained language model." In: *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*. URL: https://api.semanticscholar.org/CorpusID:248545746.

Zevallos, Rodolfo et al. (2022). "Introducing qubert: A large monolingual corpus and bert model for southern quechua." In: *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pp. 1–13.

Zhang, Shaodian and Noémie Elhadad (2013). "Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts." In: *Journal of Biomedical Informatics* 46.6. Special Section: Social Media Environments, pp. 1088–1098. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2013.08.004. URL: https://www.sciencedirect.com/science/article/pii/S1532046413001196.

Zhou, Wenxuan et al. (2023). "Universalner: Targeted distillation from large language models for open named entity recognition." In: *arXiv preprint arXiv:2308.03279*.