FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Low-Resource Machine Translation for Emakhuwa: Transfer Learning, Data Augmentation, and Lexical Resource Integration

Felermino Dário Mário António Ali



Programa Doutoral em Engenharia Informática (PRODEI)

Supervisor: Professor Doutor Henrique Lopes Cardoso

Co-Supervisor: Professor Doutor Rui Manuel Sousa Silva

March 11, 2026

© Felermino Dário Mário António Ali, 2026

Low-Resource Machine Translation for Emakhuwa: Transfer Learning, Data Augmentation, and Lexical Resource Integration

Felermino Dário Mário António Ali

Programa Doutoral em Engenharia Informática (PRODEI)

Approved in oral examination by the committee on February 20, 2026:

Jury Chair: Professor Pedro Nuno Ferreira da Rosa da Cruz Diniz

External Member: Professor Maarit Tuulikki Koponen

External Member: Professor Maria Luísa Torres Ribeiro Marques da Silva Coheur

Internal Member: Professor Sérgio Sobral Nunes

Supervisor: Professor Henrique Lopes Cardoso

March 11, 2026

Resumo

Esta pesquisa explora a sub-representação de línguas de poucos recursos no campo da tradução automática, com foco específico no Emakhuwa, a língua local mais falada em Moçambique. Apesar de contar com mais de 7 milhões de falantes nativos, o Emakhuwa continua sub-representado tanto na esfera acadêmica quanto na tecnológica devido à falta de recursos digitais e ferramentas linguísticas. Para preencher essa lacuna, desenvolvemos os primeiros recursos significativos de tradução automática para o par de línguas Português–Emakhuwa. Nossas contribuições incluem a criação de um corpus paralelo por meio da tradução manual de textos jornalísticos, a digitalização de materiais existentes e a tradução de benchmarks estabelecidos de avaliação em tradução automática. Avaliamos três estratégias centrais para melhorar o desempenho da tradução automática nesse contexto de poucos recursos: (1) aprendizagem por transferência utilizando modelos multilíngues e centrados na África, (2) aumento de dados por meio de retrotradução e (3) integração de recursos linguísticos externos, como glossários de empréstimos e dicionários bilíngues. Os resultados mostram que modelos encoder-decoder, particularmente arquiteturas otimizadas para tradução como NLLB e M2M-100, apresentam desempenho equivalente ou superior a modelos maiores apenas com decoder, mantendo a eficiência computacional. A retrotradução oferece melhorias modestas, e a integração de empréstimos e recursos de dicionário, especialmente na direção português–Emakhuwa, melhora significativamente a qualidade da tradução, sobretudo com o uso de LLMs. Este trabalho estabelece as bases para pesquisas futuras em PLN para línguas sub-representadas e demonstra caminhos práticos para o desenvolvimento de sistemas de tradução automática em contextos com recursos limitados.

Mukwasu

Muthokoreryo ola onithokorerya orumeeliwa wa mattaava vakhaani mmuttettheni wa otaphuleliwa wootthokiheya, voohimya sa Makhuwa, nttaava na muttetthe onilavuliwa vanceene oMosampiikhi. Nnaamwi okhalana atthu anivikana imilyawu 7 anilavula nttaava na elapo aya, Makhuwa nlelo khonihimmwa ni atthu anceene mwaha wa ohikhalana ikaruma sa ekaruma ni ikaruma sa mattaava. Wira nilokiherye mwaha ola, hiyo noopaka marehereryo oopacerya a otaphulela wookhweya wa nttaava na ekunya. Soovaha sahu sinihela muhina opakiwa wa ekorupu emosa yoovirikana erumeeliwaka otaphuleliwa ni matata wa soolepa sa ihapari, otaphuleliwa wa iliivuru sikhanle, ni otaphuleliwa wa ipreferensiya sootthokiheya mmuteko wootaphulela. Noothokorerya inamuna tthaaru sootthuneya wira nincererye ovara muteko wa otaphulela wootthokiha mmuttettheni wa mireerelo vakhaani: (1) wiixutta orumeelaka mathoonyeryo a mattaava manceene ni aniphwanyaneya eriyari ya elapo ya Afirika, (2) wuncererya mihupi orumeelaka otthikela otthokiherya ni (3) oheliwa muhina mwa mireerelo sa mattaava a elapo ekina, ntoko ilatarato sa sookopiha ni disionaario a mattaava manli. Soophwanyaneya sinooniherya wiira imodelo sa encoder-decoder, vanceenexa makhalelo otthokiheriwe saana wa mutthokiheryo ntoko NLLB ni M2M-100, soonihierya ovara muteko wolikana wala ovikana imodelo suulupale ni decoder paahi, okhalana okhumela saana wa mitthenkeso. Otaphuleliwa wookiserya onnikumiherya mirukunuxo vakhaani, ni oheliwa muhina wa sookophiwa ni mireerelo sa disionaario, otepexa wene muhina mwa mukhalelo wa nttaava na ekunya Emakhuwa, onnincererya vanceene mukhalelo wa otaphuleliwa, otepexa wene orumeeliwa wa LLMs. Muteko ola onnikumiherya mathipelo a otokosa wa muhoolo muhina mwa PLN wa mattaava oowaatta anithoonyeriwa ni onooniherya iphiro sookhaliherya wiira sipakiwe mitthenkeso sa otaphulela wootthokiheya muhina mwa makhalelo arina mireerelo vakhaani.

***Note**: This abstract was automatically generated using the fine-tuned NLLB-200 model developed as part of this dissertation*

Abstract

This research explores the underrepresentation of low-resource languages in the field of machine translation, with a specific focus on Emakhuwa, the most widely spoken local language in Mozambique. Despite over 7 million native speakers, Emakhuwa remains underrepresented in both academic and technological spheres due to a lack of digital resources and language tools.

To bridge this gap, we develop the first significant machine translation resources for the Portuguese–Emakhuwa language pair. Our contributions include the creation of a parallel corpus through manual translation of news texts, digitization of existing materials, and the translation of established machine translation evaluation benchmarks.

We evaluate three core strategies for improving MT performance in this low-resource context: (1) transfer learning using multilingual and Africa-centric models, (2) data augmentation through back-translation, and (3) integration of external linguistic resources, such as loanword glossaries and bilingual dictionaries. The results show that encoder-decoder models, particularly translation-optimized architectures such as NLLB and M2M-100, perform on par or better than larger decoder-only models while maintaining computational efficiency. Back-translation offers modest improvements, and the integration of loanwords and dictionary resources, especially in the Portuguese-Emakhuwa direction, significantly enhances translation quality particularly with the use of LLMs. This work lays the groundwork for further research on NLP for underrepresented languages and demonstrates practical pathways for developing machine translation systems in resource-constrained settings.

Acknowledgments

First and foremost, I would like to thank God. This PhD journey has been long and challenging, and without him, it would not have been possible.

I extend my deepest gratitude to my advisors, Professor Henrique Lopes Cardoso and Professor Rui Sousa-Silva. Over the past four years, their consistent mentorship, insightful guidance, and encouragement have been invaluable in shaping this work and supporting my academic growth.

A special thank you to Shamsuddeen Hassan Muhammad, whose immense support was vital in helping me integrate into the Masakhane community. His encouragement opened the doors to a vibrant and collaborative space that greatly enriched this project.

I am also profoundly grateful to the Fundação para a Ciência e a Tecnologia (FCT) I.P./MCTES, who sponsored this PhD, with the grant with reference SFRH/BD/151435/2021. Supported by UID/00027 – Artificial Intelligence and Computer Science Laboratory (LIACC), as well as by Base (UIDB/00022/2020) and Programmatic (UIDP/00022/2020) projects of the Center for Linguistics of the University of Porto. Such initiatives are essential for advancing research, particularly in areas with untapped potential to create meaningful impact in low-income countries. I sincerely hope this support continues and expands.

I also thank the Lacuna Fund for their generous sponsorship, which enabled the collection of datasets and provided the computing resources to run our experiments.

Big thanks, to all the translators and linguists who contributed to this project. Their dedication, expertise, and commitment were essential; without them none of this would have been possible. This accomplishment is as much theirs as it is mine. As we say in Emakhuwa, "***Muttu mekhawe khanteka empa***"—"a person alone cannot build a house." This proverb, deeply rooted in Emakhuwa culture, reflects the belief that true success comes through collaboration.

I would also like to express special appreciation to António Miguel Ndapassoa for his invaluable support throughout this journey. He played a key role in facilitating data collection, connecting me with the right people, and creating platforms to share this work. His invitations to speak on Radio Mozambique Scientific Conferences and at the Center for African Studies Seminars at Eduardo Mondlane University greatly contributed to raising awareness about the value of working with Mozambican local languages in the context of natural language processing.

Thank you to all LIACC members (Bernardo, Afonso, Tomás, and Andre) for the regular discussions and feedback on my research projects.

To my family, thank you for your continuous support, patience, and understanding of the path I have chosen. I am especially thankful to my wife, Noémia Viegas, whose unwavering presence and support have been a source of strength throughout this journey. Her companionship in difficult moments has been invaluable.

Finally, I would like to express my deepest gratitude to three people who have played a pivotal role in shaping my life and guiding me at crucial moments. First, "Mano Gil", my high school tutor, who greatly supported me in science classes and helped me stay on the right path during challenging times in my youth. Second, Mr. Pondja (in memoriam), believed in my potential and

continually encouraged my passion for science; his faith in me was fundamental in helping me reach university. And finally, Eládio Cuellar, my lecturer during my undergraduate years, whose constant encouragement inspired me to pursue a research career. Without the guidance and support of these remarkable individuals, my journey would have taken a very different path. I will always be deeply grateful for their impact on my life.

Felermino Ali

“Dedicated to grandmother "Adelaide Terra" - ***Taciana*** *(1952 – 2025) ”*

Contents

Acknowledgments iv

1 Introduction 1
- 1.1 Motivation . . . 2
- 1.2 Research Questions . . . 3
- 1.3 Contributions . . . 4
  - 1.3.1 Resources . . . 4
  - 1.3.2 Publications . . . 4
- 1.4 Structure . . . 5

2 Machine Translation 7
- 2.1 Background . . . 7
- 2.2 Evaluation . . . 11
- 2.3 The State of Machine Translation for African Languages . . . 12
- 2.4 Summary . . . 21

3 Emakhuwa 22
- 3.1 A Brief Introduction of Emakhuwa . . . 22
- 3.2 How low-resource is Emakhuwa? . . . 25
  - 3.2.1 Existing linguistic knowledge . . . 25
  - 3.2.2 Data availability . . . 26
  - 3.2.3 Technological support . . . 26
- 3.3 Challenges . . . 26
  - 3.3.1 Data Scarcity . . . 26
  - 3.3.2 Complex Morphology and Lexical variations . . . 27
  - 3.3.3 Direction . . . 28
  - 3.3.4 Gender Neutrality . . . 28
- 3.4 Summary . . . 29

4 Data Collection 30
- 4.1 Introduction . . . 30
- 4.2 Collecting Data from Online Sources . . . 31
  - 4.2.1 Parallel Data . . . 31
  - 4.2.2 Monolingual data . . . 31
- 4.3 Collecting Data from Professional Translation . . . 32
  - 4.3.1 News Domain Data . . . 32
  - 4.3.2 FLORES+ Evaluation Set . . . 33
- 4.4 Collecting Data from Other Sources . . . 38

4.5 Datasets . . . 38
4.6 Summary . . . 40

5 Benchmarks **41**
5.1 Baseline System . . . 41
5.2 Error Analysis . . . 43
5.2.1 Emakhuwa Translations . . . 44
5.2.2 Portuguese Translations . . . 45
5.3 Summary . . . 47

6 Transfer Learning **48**
6.1 Related Work . . . 48
6.1.1 MT as a Downstream Task . . . 49
6.1.2 Pre-trained Models for MT and African Languages . . . 49
6.2 Research Objectives . . . 50
6.3 Methodology . . . 50
6.3.1 Training . . . 51
6.3.2 Evaluation . . . 52
6.4 Results and Discussion . . . 53
6.5 Discussion . . . 59
6.6 Summary . . . 61

7 Data Augmentation **62**
7.1 Related Work . . . 62
7.2 Research Question . . . 63
7.3 Methodology . . . 63
7.3.1 Monolingual Data . . . 64
7.3.2 Experimental Setup . . . 64
7.4 Results . . . 65
7.5 Summary . . . 66

8 Leveraging External Information **68**
8.1 Related Work . . . 69
8.1.1 Lexical handling in MT . . . 69
8.1.2 Loanwords Challenge in MT . . . 70
8.1.3 Loanword Handling in MT . . . 71
8.2 Research Questions . . . 72
8.3 Methodology . . . 73
8.4 Approach 1: Constrained Neural Machine Translation with loanwords . . . 74
8.4.1 Replacement . . . 75
8.4.2 Prompting . . . 75
8.4.3 Experimental Setup . . . 76
8.4.4 Results . . . 80
8.4.5 Case Study . . . 83
8.5 Approach 2: Incorporating Dictionary Information . . . 86
8.5.1 Dictionary . . . 87
8.5.2 Matching Dictionary . . . 87
8.5.3 Integration of Glossary . . . 89
8.5.4 Experimental Setup . . . 90

8.5.5 Results 91
8.6 Summary 94

9 Conclusion and Future Work **96**

References **100**

A Appendix **125**
A.1 Quality Assessment 125
A.2 Base Models 126
A.3 Hyperparameters 126

List of Figures

2.1 Transformer Architecture (Vaswani et al., 2017) . . . 8
2.2 Trend of MT research on African languages: number of publications (Y-axis) per year . . . 15
2.3 African languages under study for low-resource MT . . . 16
2.4 Limitations and challenges in Low-Resource Machine Translation (Tafa et al., 2025) 18
2.5 Limitations and challenges in Low-Resource African Language Machine Translation 19

3.1 Geographic distribution of Emakhuwa variants in Mozambique (Ngunga and Faquir, 2012) . . . 23
3.2 Emakhuwa's alphabet (Ngunga et al., 2022) . . . 23
3.3 Negative conjugations of the verb *osuwela* ("to know"), showing several valid prefix forms. Examples include: *prefix-suwel-a* ("it is not known"), *prefix-suwel-ana* ("we do not know each other"), *prefix-suwel-exa* ("do not know each other well"), and *prefix-suwel-iha* ("does not make itself known"). . . . 27

4.1 Screenshot of a spelling report. The first column lists the segment ID along with any potential translation issues (i.e., punctuation, source-target length ratio flag, number mismatch, loanwords not annotated, case mismatch, etc.). The second column displays the source text at the top and its translation below. In the translation, potential misspellings are highlighted in yellow, indicating that suggestions for corrections are available, and red, indicating that no suggestions exist. Additionally, the report lists, at the bottom, all words that translators have annotated as loanwords from Portuguese, using the format <*donor sequence in Portuguese*>:<*recipient sequence in Emakhuwa*>. . . . 34
4.2 Screenshot of the Computer Assisted Translation Tool . . . 35
4.3 Direct Assessment adequacy scores per annotator on *dev* set . . . 36
4.4 Direct Assessment adequacy scores per annotator on *devtest* set . . . 36
4.5 Direct Assessment adequacy scores per annotator on *control* set . . . 37
4.6 Assessment of standard orthography usage on the control set. . . . 37

5.1 Cross-entropy loss over training steps between *pt-vmw* and *vmw-pt* directions. . . . 42
5.2 Accuracy over training steps between *pt-vmw* and *vmw-pt* directions. . . . 42

6.1 Prefix and Prompt templates for MT . . . 54
6.2 Pairwise bootstrap resampling significance test results. The heat map cells highlight where the performance of the system in the row is significantly higher and different than the system in the column ($p < 0.05$). The value inside each cell corresponds to the p-value. Afrocentric systems are indicated with * . . . 59

6.3 Pairwise bootstrap resampling significance test results. The heat map cells highlight where the performance of the system in the row is significantly higher and different than the system in the column ($p < 0.05$). The value inside each cell corresponds to the p-value. Afrocentric systems are indicated with * 60

7.1 Synthetic source sentence generation via backtranslation as in Sennrich et al. (2015) 63
7.2 Augmenting parallel data with backtranslation 64

8.1 Methodology 73
8.2 Constrained Neural Machine Translation with loanwords 74
8.3 Prompt template for standard input (a) and with CNMT (b) 76
8.4 Prompt template. The placeholders are substituted with source and target languages and corresponding sentences. 78
8.5 Generating Loanword Constraint Pairs Using the Fine-Tuned *gpt-4o-mini-2024-07-18* Model 78
8.6 Labeling scheme for tagging source and target sentence tokens with "LOAN" or "O" category. 79
8.7 Model rankings based on human evaluations. Each bar plot shows the average direct assessment across the evaluated 200 sample from test set. 83
8.8 Constrained Neural Machine Translation with loanwords and dictionary 86
8.9 Visualization of the process for extracting dictionary entries. Step-by-step operation of Algorithm 1, followed by OCR and the subsequent JSON formatting. . . . 87
8.10 Prompt template with glossary integration. The translation task is augmented by incorporating structured glossary information, which includes dictionary terms and loanword adaptations. 90

A.1 Annotation Tool User Interface. 125

List of Tables

2.1 Search queries used for different databases. . . . 14
2.2 MT studies for low-resource African languages, categorized by language pair: parallel porpora between English, French, or Portuguese and African languages (XX), and between African languages (XX–XX). . . . 17
2.3 Strategies and methodologies . . . 20

3.1 Variants of Emakhuwa . . . 24
3.2 Verb structure in Bantu languages (Mutaka and Tamanji, 2000). NM = negation marker; VP = verbal prefix; TM = tense marker; AM = aspect marker; OM = object marker; IMP = imperative; EXT = extension. . . . 24
3.3 Examples of agglutination in Emakhuwa . . . 24
3.4 Example of Gender-neutrality in Emakhuwa . . . 29

4.1 Matecat's quality report post revision. . . . 35
4.2 Intraclass Correlation Coefficient (ICC) and Confidence Interval (CI) Results for Adequacy and Orthography usage annotation. . . . 37
4.3 Summary of the collected data. Ali-2021 refers to the training subset from Ali et al. (2021). News Data consists of sentence pairs obtained through manual translation. OCR-ed Data includes sentences extracted using OCR. Mono denotes monolingual data, and FLORES+ corresponds to the manually translated FLORES evaluation set. . . . 39
4.4 Dataset splits and distribution of examples across different topics . . . 40

5.1 Evaluation scores on the test set for the MT models . . . 43
5.2 Error Analysis Evaluation on the *pt-vmw* model . . . 45
5.3 Error Analysis Evaluation on the *vmw-pt* model . . . 46

6.1 Overview of pretrained language models and their coverage of African languages. Models optimized for machine translation are indicated with *. . . . 51
6.2 Pre-trained Models used in our experiments. Prefixes and prompts followed the standard format typically used with each model. M2M-100, NLLB, and mBART models do not require a prefix; instead, the source and target languages are specified directly through the tokenizer. . . . 53
6.3 Evaluation results on the **DEV** set. CHRF underlined if significantly better than baseline ($p < 0.05$). COMET refers to the AfriCOMET metric. . . . 55
6.4 Evaluation results on the **DEVTEST** set. CHRF underlined if significantly better than baseline ($p < 0.05$). COMET refers to the AfriCOMET metric. . . . 56
6.5 Evaluation results on the **MULTI-SOURCE** set. CHRF underlined if significantly better than baseline ($p < 0.05$). COMET refers to the AfriCOMET metric. . . . 57

6.6 Error Analysis Evaluation on the *pt-vmw* model. Note, only *byT5* is shown, as it outperformed *AfriByT5* in this translation direction. 58
6.7 Error Analysis Evaluation on the *vmw-pt* model. Note, only *AfriByT5* is shown, as it outperformed *byT5* in this translation direction. 58

7.1 Evaluation results on the *DEV* set. CHRF underlined if significantly better than baseline (see Table 6.3) ($p < 0.05$). 66
7.2 Evaluation results on the *DEVTEST* set. CHRF underlined if significantly better than baseline (see Table 6.4) ($p < 0.05$). 66
7.3 Evaluation results on the **MULTI-SOURCE** set. CHRF underlined if significantly better than baseline (see Table 6.5) ($p < 0.05$). 66

8.1 Loanword adaptations by fine-tuned translation models for Portuguese into Emakhuwa 72
8.2 Standard input and CNMT input with replacement strategy. 75
8.3 Training and validation Portuguese-Emakhuwa bitext data (Table 4.4). Note: Some strings in the news category did not contain loanword constraints, meaning translations were performed entirely using native words in the target language. 77
8.4 Evaluation results of loanword constraints generation model 79
8.5 Details of the evaluation sets 79
8.6 Evaluation results on the **DEV** set. CHRF underlined if significantly better than baseline (Table 6.3) ($p < 0.05$). 82
8.7 Evaluation results on the **DEVTEST** set. CHRF underlined if significantly better than baseline (Table 6.4) ($p < 0.05$). 82
8.8 Robustness evaluation on the **DEV** set. 82
8.9 Robustness evaluation on the **DEVTEST** set. 82
8.10 A sample from DEV set showcasing the improvement over the baseline translation in Table 8.1 84
8.11 A sample from DEVTEST set, showcasing the improvement over the baseline translation with CNMT with loaword injection 85
8.12 Standard input and CNMT input with appending strategy. 89
8.13 Statistics of the training, validation, and test sets. Note that the average number of injected loanwords from glossaries is computed using a threshold of $\theta = 85$. . . . 90
8.14 Results of Portuguese-to-Emakhuwa translation when incorporating additional context using a dictionary ($\theta = 85$) on *FLORES+ DEV* and *DEVTEST* sets. Variations refer to the results in Table 8.6 and Table 8.7. 91
8.15 Example 93
8.16 Portuguese(pt) →Emakhuwa(vmw) performance of results of standard prompts on LOAN+DICT setup. For CHRF values underlined indicate significant evidence that they outperform the baseline ($p < 0.05$), based on pairwise randomized significance tests (Koehn, 2004) 94
8.17 Portuguese(pt) →Emakhuwa(vmw) performance of results, threshold of $\theta = 80$. 94

A.1 Base models hugging-face checkpoints. 126
A.2 Text-to-Text models Hyperparameters for fine-tuning. 126
A.3 NLLB Hyperparameters for fine-tuning. 127
A.4 M2M Hyperparameters for fine-tuning. 127
A.5 mBART50 Hyperparameters for fine-tuning. 127
A.6 OPUS Hyperparameters for fine-tuning. 127
A.7 InkubaLM Hyperparameters for fine-tuning. 127

A.8 LLMs Hyperparameters for fine-tuning (with Unsloth). 127
A.9 TowerInstruct-v0.1 Hyperparameters for fine-tuning. 128

Acronyms

LLM Large Language Models. 10, 70

MT Machine Translation. viii, x, xii, 2–7, 10–22, 25–30, 33, 37, 38, 41, 43, 44, 48, 49, 52, 54, 55, 57, 62, 63, 66–74, 94–96, 98, 99

NLP Natural Language Processing. 1–3, 7, 16, 22, 25, 26, 39, 48, 49, 99

NMT Neural Machine Translation. 7, 47

OCR Optical Character Recognition. xii, 31, 38, 39

SMT Statistical Machine Translation. 7

Chapter 1

Introduction

The field of Natural Language Processing (NLP) has seen rapid progress over the last decade, largely driven by advances in machine learning and the availability of large-scale text data. However, these advancements have disproportionately benefited high-resource languages—such as English, French, Spanish, and Mandarin—leaving the vast majority of the world's linguistic diversity underrepresented. African languages, in particular, have long been marginalized in NLP research due to a combination of systemic factors.

Within this broader context of global linguistic underrepresentation, this thesis places a specific emphasis on African languages, which remain among the most underserved despite their rich diversity and cultural importance.

In recent years, there has been a welcome shift. Research interest in African languages is growing, boosted by local initiatives, international collaborations, and increased funding for underrepresented languages. Efforts such as dataset creation, community workshops, and open-source model development have focused particularly on African languages spoken in Anglophone and Francophone regions. Languages like Swahili, Yoruba, Wolof, Hausa, and Xhosa now benefit from the expanding ecosystem of linguistic tools and resources.

Despite notable progress, advancements in NLP remain uneven across African regions. In particular, languages spoken in Portuguese-speaking African countries (PALOPs), such as Mozambique, Angola, Guinea-Bissau, Cape Verde, and São Tomé and Príncipe, continue to be significantly underrepresented in the global landscape of NLP. These countries, home to over 79 million people and marked by rich linguistic diversity, have seen minimal research efforts focused on their local languages. This exclusion not only hinders research and technological progress but also perpetuates digital division, limiting access to information.

In this thesis, we focus on the Mozambique case study—a country of over 30 million people with a distinctive linguistic diversity, including at least 42 Bantu languages spoken across different regions. Among these, Emakhuwa is the most widely spoken, with over 7 million speakers primarily concentrated in the northern region. Despite this rich linguistic landscape and the cultural significance of those languages, Portuguese remains the sole official language of Mozambique. While the policy of designating Portuguese as the official language so as to to unify the nation and

facilitate administrative and educational communication, this policy has inadvertently contributed to the gradual marginalization of local languages. Therefore, local languages are largely confined to informal and domestic settings. As a consequence, there has been limited academic and technological investment in the advancement of local languages in Mozambique. Resources remain scarce, both in print and digital formats, and there is a notable absence of language technology tools specifically designed for these languages. Furthermore, these languages are minimally represented on the web, which is currently the dominant source for gaining access to data used in NLP.

This thesis aims to address this resource scarcity gap by developing resources and Machine Translation (MT) tools for the Portuguese–Emakhuwa language pair. We believe that by doing so, we will enable bidirectional knowledge transfer, facilitating content creation and accessibility in both languages.

In this dissertation, the term MT refers broadly to computational systems that automatically translate text from one language to another. While traditionally MT research focused on encoder–decoder neural architectures, recent advances in large language models have expanded the range of approaches used for translation. Accordingly, this work considers both conventional neural machine translation systems and decoder-only large language models that can perform translation through prompting or fine-tuning.

1.1 Motivation

Developing MT systems for low-resource languages like Emakhuwa presents both a demanding social need and a crucial research opportunity. One key area where MT could have a tangible impact is in information access. In Mozambique, public broadcasters like Radio of Mozambique are legally required to broadcast part of their programming in national languages. However, this is typically carried out by manually translating content originally produced in Portuguese. This manual process delays the dissemination of crucial information to the majority of the population, who do not speak Portuguese. A reliable MT system will help automate this process and make multilingual communication more efficient and inclusive.

Education is another domain where the need for Emakhuwa language technology is critical. Since 2017, Mozambique has expanded bilingual education initiatives, which begin instruction in local languages before transitioning to Portuguese. Studies show that this improves literacy rates and student engagement (Nakamura et al., 2017). However, these programs struggle with a severe lack of translated teaching materials. The shortage of qualified translators and the absence of computational tools make it difficult and costly to produce localized content, threatening the sustainability of these programs (Lopes and Pinto, 2017). MT could significantly accelerate the production of educational resources, supporting inclusive, linguistically relevant learning environments.

Beyond clear social benefits, a compelling economic opportunity is emerging from the increased focus on African languages. As internet penetration and digital adoption rise in Africa,

especially among young people, the demand for language technologies, such as voice assistants, chatbots, and spell/grammar checkers, is increasing. Among these technologies, MT stands out as a particularly powerful enabler. MT enables the transfer of knowledge across languages, breaking down linguistic barriers and making information more accessible to diverse communities. For low-resource languages like Emakhuwa, MT has the potential to catalyze the rapid creation of digital content in the local language, which is essential for building large-scale textual corpora. This, in turn, can support the development of more advanced NLP tools customized to the language. When paired with human-in-the-loop validation, machine-translated content can reach higher levels of adequacy, fluency, and cultural relevance. Ultimately, such efforts contribute not only to the preservation of Emakhuwa but also to empowering the language to grow and adapt, securing its place as a living, evolving part of the digital age.

To address this gap, this research is guided by the following objectives:

- Enriching Data: Collect a corpus of Emakhuwa texts by aggregating open-source datasets, digitizing print materials, and generating new parallel data through translation efforts.
- Leveraging multilingual models: Investigate the effectiveness of transfer learning by fine-tuning existing multilingual models, including Africa-centric ones, to assess improvement in Emakhuwa MT quality.
- Leveraging linguistic resources: Explore the integration of dictionaries, glossaries, and other lexical resources to augment translation accuracy and consistency.

With these goals in mind, this work focuses on developing MT for the Portuguese-Emakhuwa language pair and contributing to the general body of research on low-resource languages.

1.2 Research Questions

To address the above objectives, the research is structured around four core research questions:

How does expanding the size and diversity of training data influence the performance of Portuguese–Emakhuwa machine translation systems? Low-resource language pairs like Portuguese–Emakhuwa suffer from acute data scarcity, with few available parallel sentences, leading to poor performance of Neural Machine Translation systems. Expanding data size and diversity can mitigate this problem as evidenced by gains of similar Bantu language pairs, e.g., Swahili–English (Arivazhagan et al., 2019; Adelani et al., 2022b; Elmadany et al., 2024a) when scaling the size of sentence pairs.

How can fine-tuning pre-trained multilingual language models (e.g., Afrocentric or universal) with limited Portuguese–Emakhuwa data improve translation performance? Fine-tuning, with a small amount of high-quality parallel data, can allow the model to adjust its internal representations to better account for specific morphology, syntax, and vocabulary. This adaptation can mitigate common zero-shot errors and lead to measurable improvements in translation quality.

To what extent can data augmentation techniques, particularly leveraging monolingual Emakhuwa corpora, improve the performance of Portuguese–Emakhuwa machine translation systems? The availability of Emakhuwa monolingual corpora is considerably higher than that of parallel data, which presents a unique opportunity for improving machine translation models. By using data augmentation techniques, such as back-translation, we can generate additional training examples that enrich the dataset (Emezue and Dossou, 2021b; NLLB Team et al., 2022; Kimera et al., 2025).

How can external resources, such as bilingual dictionaries, glossaries, or loanword mappings, be integrated into Portuguese–Emakhuwa machine translation systems to guide translation choices and improve output quality? Emakhuwa features a significant number of Portuguese loanwords (e.g., escola → exikola), which raises challenges for machine translation. This issue is further compounded by the lack of a standardized orthography and the language's agglutinative nature, both of which hinder consistent tokenization and alignment. To address these challenges, external lexical resources, such as bilingual dictionaries, glossaries, and loanword mappings, can be integrated into translation models through mechanisms like soft constraints, guided decoding, or copy-based strategies. These approaches can help disambiguate translation choices, reinforce lexical consistency, and improve overall output quality.

1.3 Contributions

This section presents the main contributions of this research, including the resources developed and publications.

1.3.1 Resources

Throughout the course of this research, we developed various artifacts to advance the state-of-the-art in MT for Emakhuwa and contributed to broader low-resource African language research. Our contributions include both datasets and models, all of which have been made publicly available under an open and permissive license (CC BY 4.0) on the Hugging Face platform. These resources can be accessed through the LIACC Hugging Face repository under the Makhuwa-NLP collection[1].

1.3.2 Publications

The following publications have resulted from or are closely related to our research:

1. **Felermino D. M. A. Ali**, Rui Sousa-Silva, Henrique Lopes Cardoso (2025, submitted). *Lexicon-Guided Machine Translation for Emakhuwa*, Language Resources and Evaluation Journal, special issue on Machine Translation for Low-Resource Languages.

[1] https://huggingface.co/collections/LIACC/makhuwa-nlp-66a93ea22df7f4b31e96a5ab

2. Senyu Li, **Felermino D. M. A. Ali** , Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, Colin Cherry, and David Ifeoluwa Adelani. (2025), *Evaluating WMT 2025 Metrics Shared Task Submissions on the SSA-MTE African Challenge Set*, Conference on Machine Translation - WMT2025 workshop, The 2025 Conference on Empirical Methods in Natural Language Processing

3. Senyu Li, Jiayi Wang, **Felermino D. M. A. Ali**, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. (2025), *SSA-COMET: Do LLMs outperform learned metrics in evaluating mt for under-resourced African languages?*, The 2025 Conference on Empirical Methods in Natural Language Processing

4. **Felermino D. M. A. Ali**, Rui Sousa-Silva, Henrique Lopes Cardoso (2025). *Leveraging Loanword Constraints for Improving Machine Translation in a Low-Resource Multilingual Context*, The 2025 Conference on Empirical Methods in Natural Language Processing

5. **Felermino D. M. A. Ali**, Rui Sousa-Silva, Henrique Lopes Cardoso (2024). *Expanding FLORES+ Benchmark for more Low-Resource Settings: Portuguese-Emakhuwa Machine Translation Evaluation WMT 2024* - Open Language Data Initiative shared tasks, Miami, Florida, US

6. **Felermino D. M. A. Ali**, Rui Sousa-Silva, Henrique Lopes Cardoso (2024). *Building Resources for Emakhuwa: Machine Translation and News Classification Benchmarks*, The 2024 Conference on Empirical Methods in Natural Language Processing, Miami, US

7. **Felermino D. M. A. Ali**, Rui Sousa-Silva, Henrique Lopes Cardoso (2024). *Detecting Loanwords in Emakhuwa: An Extremely Low-Resource Bantu Language Exhibiting Significant Borrowing From Portuguese*, 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Torino, Italy

8. **Felermino D. M. A. Ali**, Andrew Caines, Jaimito LA Malavi (2021). *Towards a parallel corpus of Portuguese and the Bantu language Emakhuwa of Mozambique*. AfricanNLP Workshop 2021 arXiv preprint arXiv:2104.05753

1.4 Structure

The remaining chapters in this document are organized as follows:

- **Chapter 2** provides a review of the current state of MT, with a focus on challenges and advances in low-resource settings, particularly for African languages.

- **Chapter 3** introduces Emakhuwa, describing its linguistic features, how Emakhuwa stands within the low-resource context, and its importance to this research.

- **Chapter 4** details the training and test data collection and preparation efforts, including corpus creation, digitization, and alignment for the Portuguese–Emakhuwa language pair.
- **Chapter 5** presents a benchmark evaluation of neural MT systems for Portuguese–Emakhuwa.
- **Chapter 6** explores the application of transfer learning by fine-tuning various pretrained multilingual models and evaluating their performance in the Portuguese–Emakhuwa language pair.
- **Chapter 7** investigates data augmentation techniques, particularly backtranslation, as a method for improving MT performance.
- **Chapter 8** demonstrates how linguistic knowledge, including Portuguese loanwords and bilingual dictionaries, can be leveraged to enhance Portuguese–Emakhuwa MT.
- **Chapter 9** presents the conclusion and future works

Chapter 2

Machine Translation

This chapter explores the evolving field of MT, a discipline that interests NLP and translation studies. The chapter begins by providing the background of the MT field. Then, the chapter continues addressing the critical and complex task of evaluating MT system performance, outlining traditional human assessment methods and the diverse array of automatic evaluation metrics developed to evaluate translation systems. The last part of the chapter is dedicated to the challenges and advancements in applying MT to low-resource languages, particularly from the African continent. We examine the current research landscape, including which languages are being studied, the common high-resource language pairings, the inherent difficulties posed by low-resource settings, and the strategies being used to overcome these limitations.

2.1 Background

MT as a NLP's subtask is dedicated to automatically translating text from one language into another, with or without human assistance. First proposed by Warren Weaver in 1947 (Hutchins, 1997), early MT systems relied on hand-crafted rules and linguistic expertise. However, the inherent complexity and ambiguity of natural language quickly revealed the limitations of rule-based approaches. This led to a paradigm shift toward data-driven methods as **Statistical Machine Translation (SMT)** (Osborne, 2010). SMT models treat translation as a probabilistic process, trained on large corpora of aligned sentence pairs to calculate the most likely rendering of a given input. While this eliminated the need for extensive ontologies and manually designed grammars, SMT remained largely sentence-bound, struggling with broader contextual dependencies and unseen sentences or phrases. Despite these shortcomings, SMT dominated the field for many years, laying the groundwork for the emergence of **Neural Machine Translation (NMT)**, which sought to overcome these very challenges (Ataman et al., 2025).

NMT has advanced through several stages, and each major stage of development has addressed specific limitations of its predecessors, gradually expanding the capabilities of MT systems.

Architectures Early encoder–decoder models represented entire input sentences as fixed-length vectors, which often led to significant information loss, especially for long and complex sentences. To overcome this, researchers introduced recurrent neural network architectures such as Long Short-Term Memory (LSTM) (Sutskever et al., 2014) and Gated Recurrent Unit (GRU) models (Cho et al., 2014). These recurrent models were better equipped to capture sequential dependencies, but they still struggled with long-range relationships and were hampered by slow, sequential training (Sutskever et al., 2014; Cho et al., 2014). The introduction of attention mechanisms (Bahdanau et al., 2015) marked a major breakthrough, allowing models to dynamically focus on relevant parts of the sentence during translation. This innovation not only improved accuracy but also made training more efficient (Bahdanau et al., 2015). Building on these advances, the transformer architecture (See Figure 2.1) replaced recurrence entirely with parallel self-attention and positional encoding, resulting in improved scalability, faster training, and more effective handling of long-distance dependencies within text (Vaswani et al., 2017) .

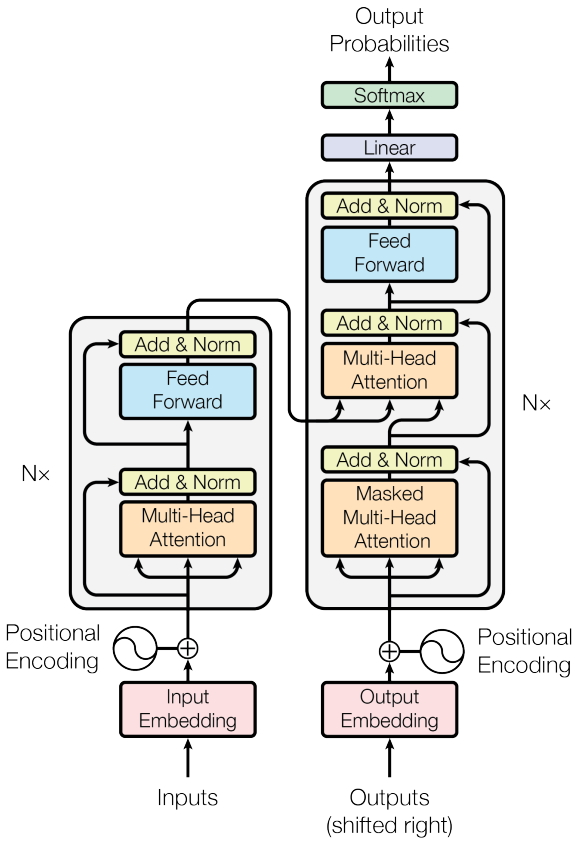


Figure 2.1: Transformer Architecture (Vaswani et al., 2017)

Vocabulary and Representation Despite these architectural advances, handling rare or unseen words remains a persistent challenge in NMT. Early systems relied heavily on fixed word-level vocabularies, which led to poor coverage and high rates of out-of-vocabulary errors, particularly

in languages with rich morphology. To address these issues, models began using subword units—such as those produced by Byte Pair Encoding (BPE) or SentencePiece—which allow systems to cover a broader range of word forms and reduce vocabulary size (Sennrich et al., 2016b; Kudo and Richardson, 2018). However, subword segmentation can sometimes break apart meaningful morphological units, potentially impacting translation quality. Character-level modeling provides an even finer-grained approach, capturing morphological details, but at the cost of increased computational demands and slower training times (Luong and Manning, 2016; Costa-Jussà and Fonollosa, 2016; Lee et al., 2017).

Building on these advances in vocabulary handling, researchers began to explore how representation strategies could be extended across multiple languages. This shift laid the groundwork for models that not only address word-level challenges within a single language but also leverage shared structures across languages to improve efficiency and generalization.

Multilingual NMT To further improve resource sharing, multilingual NMT models have emerged, enabling the translation of many languages within a single unified system. By sharing parameters across languages, these models can perform zero-shot translation between pairs not seen during training (Aharoni et al., 2019; Arivazhagan et al., 2019). However, as the number of supported languages increases, competition for model capacity can lead to declining performance, especially for low-resource directions. Strategies such as selective parameter sharing, language-specific vocabularies, and modular components, combined with careful data sampling, have been shown to mitigate these issues and enhance transfer learning for underrepresented languages (Conneau et al., 2020). Large-scale projects have further expanded coverage to hundreds of languages using data mining and mixture-of-experts architectures (NLLB Team et al., 2022), yet significant imbalances remain between high- and low-resource languages .

Pretraining and Fine-Tuning Paradigm Building on these multilingual foundations, the pretrain–finetune paradigm has introduced greater flexibility. By first training on general cross-lingual objectives, such as denoising auto-encoding or masked language modeling, models learn robust representations that can then be adapted to specific translation tasks using parallel data (Liu et al., 2020; Wang and Li, 2021). This two-step process improves generalization and enables better transfer across languages, although it still relies on supervised fine-tuning for optimal performance.

Large Language Models in Translation In parallel to these developments, large decoder-only language models have recently demonstrated strong translation capabilities through prompt-based methods, even without direct task-specific training (Zhang et al., 2023b). The effectiveness of these models depends heavily on prompt design, including the clarity of instructions, the use of examples, and their ordering. While such models excel at producing fluent output, they may sometimes prioritize fluency over accuracy, leading to omissions or hallucinations. Notably, built-in biases toward English can facilitate indirect translation via English prompts, but genuinely low-resource translation directions remain challenging. Although these models rival strong supervised

baselines in many cases, they often fall short of the best commercial systems, and there are ongoing questions about their ability to faithfully interpret meaning and follow nuanced instructions (Kocmi et al., 2024).

The emergence of Large Language Models (LLM)s has created new opportunities for integrating explicit grammatical information directly into translation tasks, offering a flexible platform where translation instructions and linguistic rules can be combined, an approach that is reminiscent of traditional machine translation systems (Tanzer et al., 2023; Zhang et al., 2024a). This fusion allows for expert-driven interventions that can guide the model's outputs, potentially improving linguistic accuracy by leveraging rule-based methods alongside the strengths of neural architectures. As a result, there is growing interest in revisiting hybrid rule-based techniques within LLM-driven machine translation (Tanzer et al., 2023; Zhang et al., 2024a), suggesting that a thoughtful combination of data-driven and expert-informed strategies may be necessary to address challenges in grammaticality and semantic fidelity.

Evaluation Challenges Finally, evaluating progress in NMT remains an open challenge. Common metrics such as BLEU emphasize lexical overlap and may fail to capture improvements in syntactic or semantic adequacy. In high-resource settings, explicit modeling of linguistic structure often yields limited gains under these metrics, highlighting the need for evaluation frameworks that better reflect grammatical correctness, semantic fidelity, and real-world translation quality.

Data Availability and Low-Resource settings The limitations imposed by data availability present another major hurdle for scaling NMT systems, particularly for resource constraint settings. For example, while some languages benefit from large parallel corpora, many languages lack sufficient training data. To supplement parallel data, synthetic examples generated by back-translating monolingual text have proven useful (Sennrich et al., 2016a). Although this approach works best when the initial translation system is already strong, which is a condition that favors well-resourced languages . Alternatively, unsupervised translation methods attempt to align monolingual representations through reconstruction losses and adversarial training. While promising, these techniques often struggle when faced with distant language pairs that share little linguistic structure (Artetxe et al., 2018; Lample et al., 2018). Another strategy, pivot-based translation (also known as Bridging), routes translations through a third language, typically English, which can help bridge gaps but sometimes compounds errors across the translation chain (Cheng et al., 2017).

The next sections provide an in-depth discussion of evaluation and the state-of-the-art of African language MT, highlighting ongoing efforts to design metrics that better capture semantic meaning. In addition, we examine how current methodologies address the persistent challenges of low-resource translation.

2.2 Evaluation

MT evaluation is crucial for assessing the quality of translation generated by MT systems. Effective evaluation helps determine if the system conveys meaning accurately and fluently. Due to the inherent multiplicity of valid translations for a given input, MT evaluation remains a challenging task. One common approach involves human evaluation, which can be conducted through several methods:

1. **Adequacy and Fluency**: This traditional approach assesses translation quality based on two key criteria: adequacy and fluency (Moorkens et al., 2018). Adequacy measures how well the translation preserves the meaning of the source text, while fluency evaluates the grammatical correctness and naturalness of the translated text (Koehn, 2020).

2. **Human Translation Edit Rate (HTER)**: This method incorporates human post-editing into the evaluation process by comparing the machine-generated output with a version revised by a professional translator. For example, observations by Snover et al. (2006) that post-editing has become a central quality-assurance stage in contemporary translation workflows.

3. **Multidimensional Quality Metrics (MQM)**: The MQM framework (Lommel et al., 2013, 2024) offers a detailed and flexible model for assessing translation quality. It categorizes and weights translation errors across seven high-level dimensions: Terminology, Accuracy, Fluency, Style, Locale Conventions, Audience Appropriateness, and Design.

Despite its valuable insights, human evaluation is often costly, time-consuming, and susceptible to inconsistencies arising from subjective interpretation. To address these limitations, different automatic evaluation metrics have been proposed as faster and more objective alternatives to human assessment. Over the years, significant efforts have been made to enhance their alignment with human judgments by integrating factors such as semantic similarity, contextual embeddings, and reference-free evaluation methods. Below, we describe how these metrics can be broadly categorized within the following taxonomy:

- **Lexical/Character Overlap Metrics:** Metrics such as BLEU (Papineni et al., 2002) and CHRF (Popović, 2015) rely on n-gram overlap between the MT output and reference translations. These metrics are efficient and widely used but struggle with synonyms, paraphrasing, and word reordering.

- **Embedding-based Similarity Metrics:** Metrics like BERTScore (Zhang et al., 2020) and YiSi-1 (Lo, 2019) compare contextualized word embeddings from pre-trained language models to capture semantic similarity beyond surface word forms.

- **MT Model-Based Metrics:** PrismRef (Thompson and Post, 2020a,b) represents this category, leveraging MT models to evaluate translation quality by assessing model confidence or reconstruction capabilities.

- **Neural Fine-Tuned Metrics:** These metrics leverage neural models fine-tuned on human-labeled quality judgments. They can be divided into two categories: **Reference-Based metrics**, for example, COMET-22 (Rei et al., 2022a) or BLEURT-20 (Sellam et al., 2020), which use both the MT output and one or more reference translations; and **Reference-Free** or **Quality Estimation** (QE) metrics, for instance CometKiwi (Rei et al., 2022b), XCOMET-QE (Guerreiro et al., 2024), or MetricX-24-Hybrid-QE (Juraska et al., 2024), which assess translation quality without requiring any reference. While Reference-based metrics tend to offer higher accuracy due to the availability of a gold standard, reference-free metrics are particularly valuable in low-resource scenarios or real-time applications where references are unavailable.

- **LLM as a Judge:** GEMBA (Kocmi and Federmann, 2023) and Remedy (Tan and Monz, 2025) exemplify a growing trend of using large language models to simulate human judgment, offering flexible and powerful evaluations without manual annotation.

Each of these metric categories presents distinct trade-offs. Lexical metrics are simple and reproducible, but often misrepresent semantic quality (Ataman et al., 2025). Neural and LLM-based methods capture deeper linguistic information, but are resource-intensive. Reference-free metrics offer scalability in real-world settings where high-quality references are unavailable (Freitag et al., 2024). As MT systems become more diverse and complex, combining multiple metrics and aligning them with human judgments continues to be an open and active area of research.

2.3 The State of Machine Translation for African Languages

Africa is a vast and culturally vibrant continent, home to a remarkable diversity of languages. According to Ethnologue (Eberhard et al., 2022), over 2000 languages are spoken across Africa, reflecting a rich linguistic and cultural heritage. However, many African languages have struggled to develop robust written systems or standardized documentation. Several factors, including historical, political, and socio-cultural, have influenced how much support African languages receive in formal settings. In the post-independence period, several countries retained colonial languages as a means of fostering national unity, particularly in contexts where multiple ethnic groups spoke different indigenous languages (Vigouroux and Mufwene, 2008). Additionally, socio-cultural perceptions about the roles of African languages, such as viewing them as less useful or less prestigious, have also shaped language policies and practices (Wolff, 2017). As a result, the majority of national languages in Africa remain primarily oral, passed down through generations within communities but lacking extensive written materials and formal documentation (Zell, 2018).

In recent years, the rapid development of language technologies has highlighted a significant imbalance in representation. Many languages, particularly those spoken in Africa, remain vastly underrepresented in state-of-the-art language technology systems. This disparity is especially evident in the field of MT, where progress is often limited by the scarcity of digital resources. Although there have been some research initiatives aimed at supporting African languages, these

efforts still insufficient when compared to the level of performance and scalability achieved for high-resource languages (Ranathunga et al., 2023).

African languages are often classified as low-resource due to the limited availability of linguistic resources necessary for developing NLP tools. Classifications often used interchangeably with under-resourced languages include low-density, resource-poor, or less-resourced languages. However, as Nigatu et al. (2024) argue, such broad categorizations often oversimplify the complex realities of languages under constrained settings. Therefore, they recommend a more explicit and structured approach to defining a language's level of resourceness by considering three key dimensions: existing linguistic knowledge, data availability, and technological support.

An undeniable reality is that most African languages rank poorly across all three dimensions of resourceness proposed by Nigatu et al. (2024). However, not all languages are equally underserved, as some are in a more critical state than others. There is a clear regional disparity in the development of language technologies across the African continent, which can be partly attributed to differences in how communities preserve and promote their local languages. In some regions, stronger cultural and institutional efforts have helped maintain linguistic vitality, while in others, local languages face greater decline and neglect.

MT for African languages remains at an early stage of development, as evidenced by the limited number of languages supported by mainstream online MT systems. Although platforms like Google Translate have limited language coverage. Despite this, recent years have seen growing efforts to bridge the digital gap, leading to the emergence of new strategies, models, and datasets aimed at advancing MT for African languages. Community-driven initiatives like Masakhane[1] have played a key role in this progress by fostering research collaborations, curating datasets, organizing workshops, and disseminating models specifically designed for linguistic African contexts.

In the following section, we explore research questions in the field of MT for African languages. Our goal is to assess the current state-of-the-art and identify emerging trends in this rapidly evolving domain. Specifically, we investigate: which language pairs are being studied; what techniques are being employed; what challenges and limitations researchers face; and what strategies have been developed to improve MT for low-resource languages. To this end, we conducted a systematic search across multiple academic databases. Table 2.1 outlines the specific search queries used. Specific inclusion and exclusion criteria were established to identify relevant peer-reviewed studies published between 2020 and 2025, excluding those not directly related to MT for African low-resource languages. From the selected studies, we extracted essential data, including the languages addressed, MT models applied, datasets used, evaluation metrics, and key findings. We used qualitative analysis to uncover recurring themes, dominant trends, and research gaps.

As a result of our systematic search, we identified a total of 58 relevant research articles published across various scientific conferences and peer-reviewed journals. These studies reflect a

[1] https://www.masakhane.io/ (accessed on 28 November 2025)

Table 2.1: Search queries used for different databases.

Database	**Query**
IEEE Xplore	(("African languages" OR "low-resource African languages") AND ("machine translation" OR "neural machine translation" OR "NMT" OR "MT" OR "translation systems"))
Scopus	TITLE-ABS-KEY (("African languages" OR "low-resource African languages") AND ("machine translation" OR "neural machine translation" OR "MT" OR "NMT" OR "translation systems")) AND PUBYEAR > 2019 AND PUBYEAR < 2026
ScienceDirect	(("African languages" OR "low-resource African languages") AND ("machine translation" OR "neural machine translation" OR "NMT" OR "MT" OR "translation systems"))
Web of Science	TS=(("African languages" OR "low-resource African languages") AND ("machine translation" OR "neural machine translation" OR "NMT" OR "MT" OR "translation systems"))
ACM Digital Library	(("African languages" OR "low-resource African languages") AND ("machine translation" OR "neural machine translation" OR "NMT" OR "MT" OR "translation systems"))

growing interest and effort in the field of machine translation for African languages (see Figure 2.2). This body of work served as the basis for answering the following questions.

Which African languages are being studied in MT research for low-resource settings?

Figure 2.3 presents a list of African languages that have been the focus of research in the field of MT. The data reveal that MT research has been conducted mainly from the Bantu and Niger-Congo language families. Zulu ("zul") leads with 28 publications, followed closely by Swahili ("swa"), Yoruba ("yor"), Xhosa ("xho"), and Igbo ("ibo"), each of which has been featured in over 20 studies. Other languages that have attracted significant research interest include Setswana ("tsn"), Siswati ("ssw"), Tsonga ("tso"), and Afrikaans ("afr").

From a linguistic perspective, Bantu languages, spoken throughout Southern, Eastern, and Central Africa, dominate the MT research landscape. This dominance can be attributed, in part, to the fact that many of the most frequently studied languages (i.e., Figure 2.3's top 10) are spoken in South Africa – a country that has actively promoted the preservation and development of its national languages by granting them official status and supporting linguistic research, and thus more MT resources are available for these languages. Swahili is another widely studied Bantu language due to its status as a lingua franca across several African nations, which has further stimulated the availability of MT resources. Similarly, languages of the Niger-Congo family, particularly those spoken in West Africa, such as Yoruba ("your") and Igbo ("ibo"), have also drawn substantial attention, highlighting a growing interest in developing MT systems across diverse African

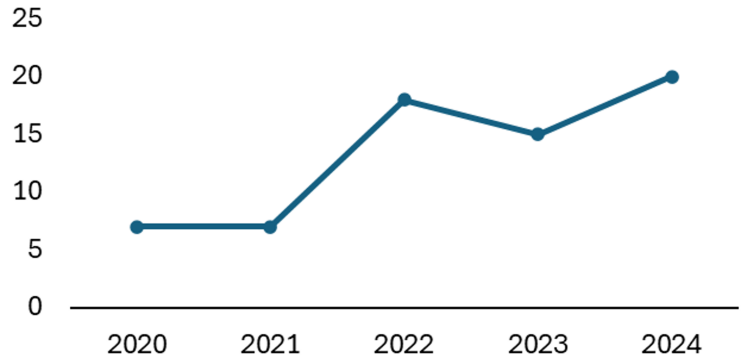


Figure 2.2: Trend of MT research on African languages: number of publications (Y-axis) per year

linguistic regions.

Another noteworthy observation is that the languages receiving the most attention in MT research are not necessarily those with the largest numbers of speakers. For instance, Zulu, which has a comparatively smaller speaker base (approximately 12 milion), is far better resourced in terms of available MT datasets and tools than Chichewa (approximately 21 milion), despite the latter being spoken by a significantly larger population across several Southern African countries.

Nevertheless, when aggregating all the studies, it becomes evident that the number of languages investigated remains limited, representing only a small fraction of the continent's vast linguistic diversity.

What are the predominant language pairs explored in MT research on African languages?

Table 2.2 categorizes the reviewed studies into four identified predominant language pairs in MT research on African languages: English-XX (English paired with African languages), French-XX (French paired with African languages), Portuguese-XX (Portuguese paired with African languages), and XX-XX (pairs involving African languages).

- **English-XX**: The findings indicate a significant emphasis on MT works between English and African languages. This category showcases the largest volume of research, underscoring English's role as a primary pivot or target language in the development of MT systems for the continent.

- **French-XX**: Following English, French-XX also constitutes a notable area of study. We identified 8 studies, suggesting a considerable, albeit smaller, research focus compared to English.

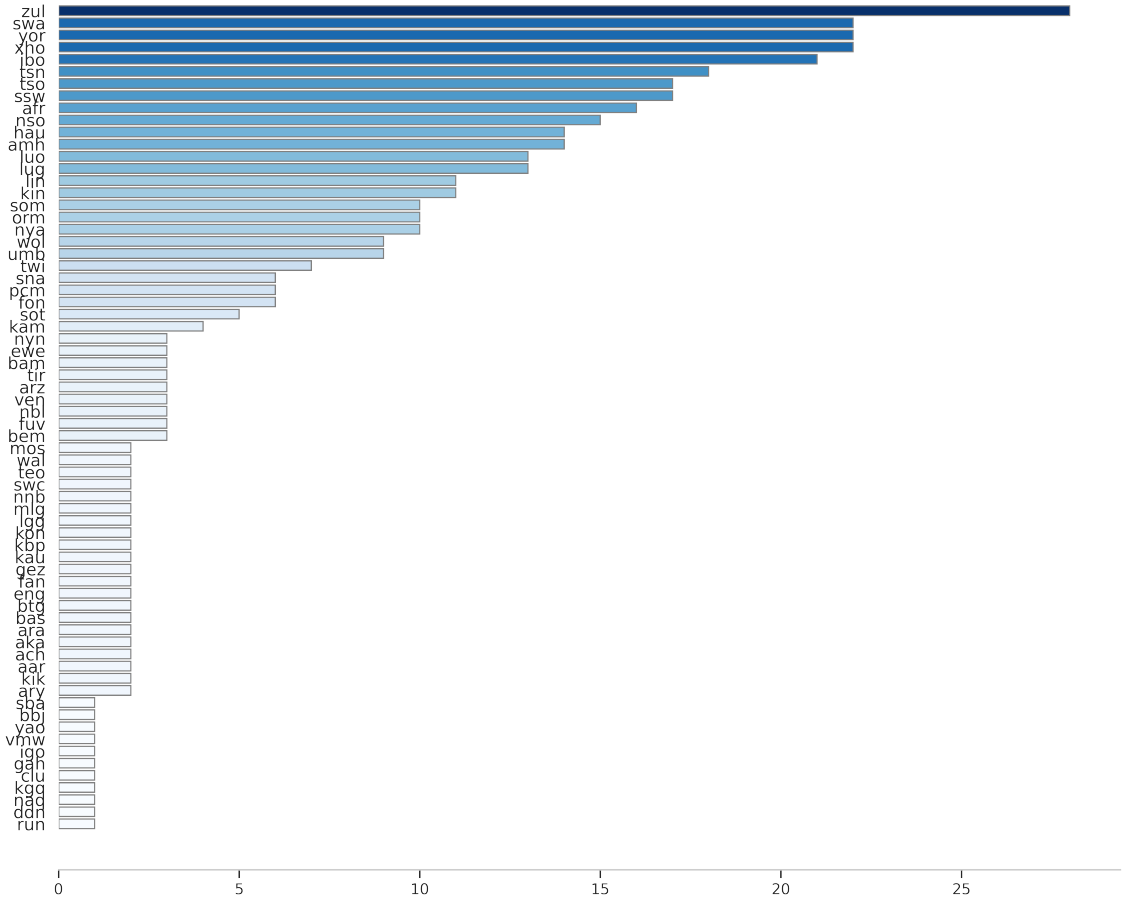


Figure 2.3: African languages under study for low-resource MT

- **Portuguese–XX**: The table reveals a clear scarcity of research focused on Portuguese as a pairing language. With only two studies identified, this group is severely underrepresented. Notably, both of these studies originate from the research conducted in this dissertation and its associated publications. This highlights a critical gap in MT development, especially considering the more than 76 million people living in Portuguese-speaking African countries, which are linguistically rich and diverse, with around 100 other local languages spoken.

- **XX-XX**: Research on translating directly between different African languages is also documented in six studies. Although this demonstrates an emerging and important interest in fostering intra-African linguistic capabilities through MT, the volume of work is considerably lower. These studies are crucial for promoting regional integration and preserving linguistic diversity by enabling communication directly between African language communities.

The results reflect factors such as the availability of emerging parallel data, historical ties, and the status of European languages in the NLP spectrum. However, the limited attention given to

Table 2.2: MT studies for low-resource African languages, categorized by language pair: parallel porpora between English, French, or Portuguese and African languages (XX), and between African languages (XX–XX).

Language-Pairs	**Studies**
English-XX	Kimera et al. (2025); Abdulmumin et al. (2022); Elmadani and Buys (2024); Robinson et al. (2023); Adelani et al. (2022b); Ogundepo et al. (2022); Wang et al. (2024b); Gyasi and Schlippe (2023); Alam and Anastasopoulos (2022); Jiao et al. (2022); Adelani et al. (2022c); Kamboj et al. (2022)
French-XX	Mbaye et al. (2024); Kamboj et al. (2022); Gyasi and Schlippe (2023); Emezue and Dossou (2021a); Wang et al. (2024c); Alam and Anastasopoulos (2022); Jiao et al. (2022); Adelani et al. (2022c)
Portuguese-XX	Ali et al. (2024c,b)
XX-XX	Elmadani and Buys (2024), McKellar and Puttkammer (2020), Emezue and Dossou (2021a), Chimoto and Bassett (2022), Ogundepo et al. (2022), Elmadany et al. (2024a)

Portuguese-African language pairs and the still-developing field of intra-African language translation point to significant areas requiring further investment and scholarly attention to ensure more equitable and comprehensive MT coverage for the diverse linguistic landscape of Africa.

What are the challenges and limitations in MT research for low-resource settings in Africa?

In line with the findings of Tafa et al. (2025), low-resource MT, including in African contexts, faces a range of deeply interrelated challenges that restrict system performance and research development, which are categorized as follows:

1. **Lack of sufficient training data and poor quality of pseudo-parallel corpora:** Relates to the availability and quality of data necessary for effective machine translation.
2. **Structural variability and language pairs difficulty:** Significant grammatical and syntactic differences between source and target languages (e.g., Amharic) complicate the learning of accurate translation mappings.
3. **Complex morphological structures and lexical diversity:** Many African languages are agglutinative, exhibiting rich morphological structures that increase data sparsity and complicate model learning.
4. **Model generalization issues and training complexity:** Low-resource conditions exacerbate problems such as error propagation and unstable training dynamics, reducing model robustness.
5. **Error propagation, training instability and catastrophic forgetting:** Issues that arise during the training process.

6. **Language complexity:** Limited lexical coverage, rich morphology, and divergent linguistic structures further constrain translation quality and model transferability.

7. **Tokenization problems:** Standard tokenization techniques often fail to handle complex morphological patterns in agglutinative or polysynthetic languages, leading to fragmented or inconsistent subword representations.

8. **Textual challenges:** The absence of standardized orthographies, frequent spelling variations, dialectal mixing, and widespread code-switching reduces data consistency and affects the perceived acceptability of MT outputs among end users.

9. **Feature integration challenges:** Effectively incorporating linguistic, morphological, and contextual features into MT models remains a significant technical and theoretical challenge.

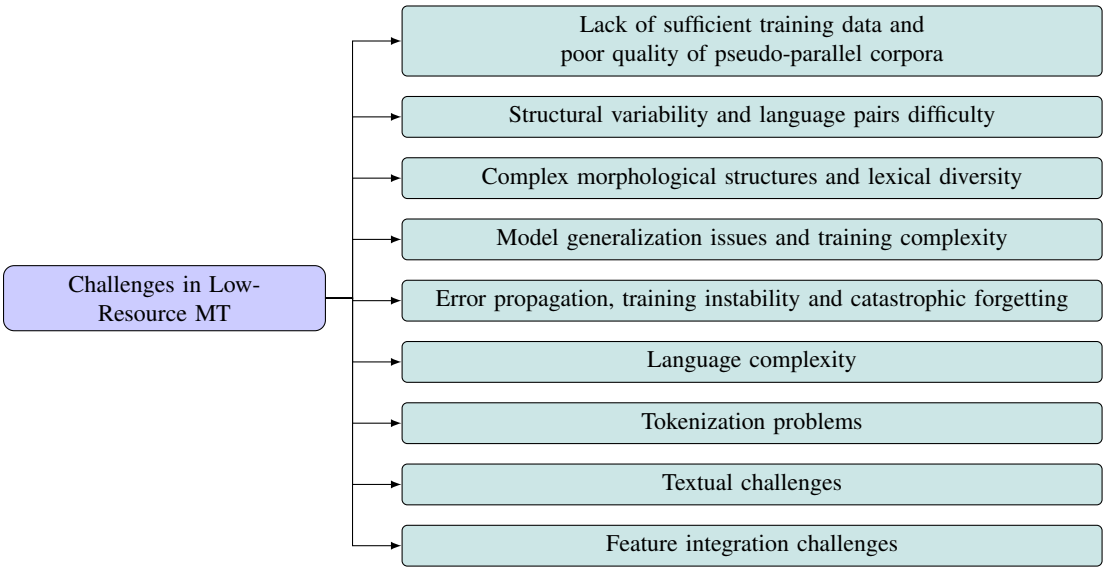


Figure 2.4: Limitations and challenges in Low-Resource Machine Translation (Tafa et al., 2025)

Building on the taxonomy of Tafa et al. (2025), we extend the analysis to reviewed studies focusing on African MT. In our view, these challenges can be collapsed into four top-level categories, as depicted in Figure 2.5: (1) **Data**, (2) **Linguistic complexity**, (3) **Modeling**, and (4) **Evaluation**. The latter represents an expansion of Tafa et al. (2025) taxonomy, since evaluation—while a general challenge in MT—becomes particularly necessary in African contexts due to the scarcity of standardized benchmarks, annotated datasets, and culturally appropriate evaluation metrics (Li et al., 2025a).

As shown in Table 2.5, the most recurrent challenge across studies is the scarcity of high-quality training data. This includes both parallel corpora and auxiliary resources such as dictionaries, monolingual corpora, or aligned comparable texts. Without adequate data, achieving effective generalization remains difficult. Moreover, most African languages are primarily oral,

with limited standardization and orthographic conventions (Adebara and Abdul-Mageed, 2022). This leads to significant spelling variations and dialectal differences, which further complicate data collection and annotation.

The morphologically rich nature of African languages also poses challenges for tokenization, especially when relying on methods designed for less agglutinative or isolating languages. Current tokenization approaches often fail to capture such linguistic phenomena, resulting in suboptimal MT performance.

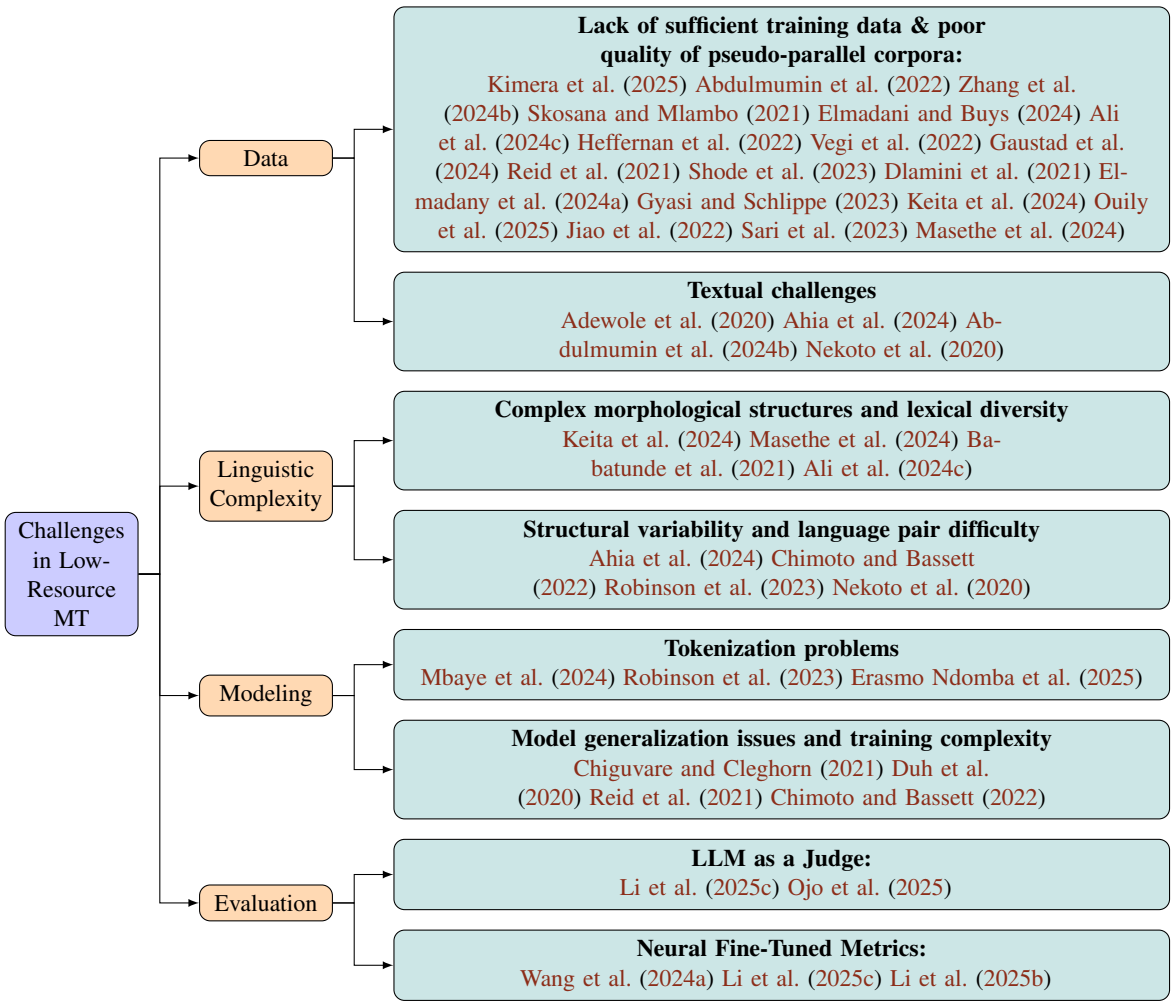


Figure 2.5: Limitations and challenges in Low-Resource African Language Machine Translation

What strategies and methodologies have been proposed to enhance translation quality, especially in the absence of large-scale parallel corpora?

Tafa et al. (2025) identified a set of strategies aimed at improving translation quality in low-resource settings, including African languages. We can broadly categorize these approaches into the following:

- **Data-centric methods**, such as data augmentation, preprocessing, backtranslation, oversampling, and sentence/document alignment, which aim to enrich or better structure available training data;
- **Model-centric techniques**, including the use of transformer encoder-decoder architectures, decoder parameter optimization, and hybrid systems, which enhance the translation model itself;
- **Learning paradigms**, like active learning, transfer learning, and BERT-based supervision, which leverage different training strategies to improve performance with limited data;
- **Multilingual and cross-lingual strategies**, such as multilingual models and embedding alignment, which enable knowledge transfer across languages;
- **Linguistic and semantic enhancements**, including the use of lexical constraints and modeling of semantic and linguistic information at multiple levels.

Table 2.3 summarizes the studies based on the strategies used to enhance MT for low-resource African languages. Most studies adopt data-centric approaches to address the core challenge posed by transformer-based architectures, which typically require extensive parallel corpora, resources that are scarce for many African languages. Multilingual models also feature prominently, as they leverage existing multilingual language resources and enable knowledge transfer through techniques like transfer learning, thereby mitigating data scarcity. However, certain strategies remain underexplored for African languages, such as lexical constraints or multi-level linguistic modeling. This gap likely stems from the limited availability of external linguistic resources for these languages.

Table 2.3: Strategies and methodologies

Category	**Studies**
Data-centric methods	Kimera et al. (2025), Ali et al. (2024c), Oppong (2023), Emezue and Dossou (2021a), Abdulmumin et al. (2022), Heffernan et al. (2022)
Model-centric techniques	Skosana and Mlambo (2021) Kimera et al. (2025), Ali et al. (2024c), Oppong (2023), Emezue and Dossou (2021a), Abdulmumin et al. (2022), Zhang et al. (2024b), Heffernan et al. (2022)
Learning paradigms	Dipta and Vallurupalli (2024)
Multilingual and cross-lingual strategies	Ogundepo et al. (2022), Elmadany et al. (2024a), Elmadani and Buys (2024), Ali et al. (2024c)

2.4 Summary

MT is a long-established field that has evolved significantly, with various approaches emerging to enhance translation quality and efficiency. Despite these advancements, the field continues to face the challenge of language underrepresentation, particularly for low-resource languages. Recent years, however, have seen a growing research interest in addressing this disparity, with increased efforts to develop MT systems for underrepresented linguistic communities. While progress has been made, significant gaps remain, particularly for African languages. Among these are the languages from Portuguese-speaking African countries.

This chapter has provided an overview of the current state of MT, focusing on the challenges and advancements in low-resource settings, particularly for African languages. The next chapter will futher explore into the case study of Emakhuwa, a language spoken in Mozambique, illustrating the specific limitations and opportunities in developing MT systems for this language.

Chapter 3

Emakhuwa

The focus of this thesis is Emakhuwa, the most widely spoken language in Mozambique. Despite its widespread use, Emakhuwa remains underrepresented in digital and computational resources, which limits its inclusion in modern language technologies. This chapter provides an overview of Emakhuwa's linguistic characteristics, standardization efforts, and classification as a low-resource language. In addition, it discusses the challenges faced in developing NLP tools and MT systems for Emakhuwa, setting the stage for subsequent chapters focused on addressing these gaps.

3.1 A Brief Introduction of Emakhuwa

Emakhuwa (also known as Makua, Macua, or Makhuwa) is a Bantu language predominantly spoken in the northern regions of Mozambique, namely Niassa, Cabo Delgado, and Nampula, as well as in parts of the central region, particularly the Zambezia province. Emakhuwa holds the status of being the most widely spoken language in Mozambique, with estimates suggesting that approximately 25% of the country's population (over 7 million people) use it in their daily lives, often alongside Portuguese, the country's official language (Ronaldo Rodrigues de Paula, 2016).

Emakhuwa has eight main variants (Ngunga and Faquir, 2012): Elomwe, Esankaci, Esaaka, Echirima, Emarevoni, Emeetto, Enahara, and Central. These variants show slight differences in accent, vocabulary, and usage. The Central variant is established as the standard form and is predominantly spoken in Nampula Province, extending into some districts of eastern Niassa that border Nampula Province. Other variants are distributed across neighboring regions: Elomwe and Emarevoni are primarily found in Zambezia; Esaaka, Esankaci, and Enahara are located in Nampula; Echirima is found in Niassa; and Emeetto is found in Cabo Delgado. Together, these variants showcase the rich linguistic and cultural diversity of northern and central Mozambique (see Figure 3.1). Examples of lexical variation among them are presented in Table 3.1.

Emakhuwa is primarily a spoken language, but over the years, Mozambican linguists have made significant efforts to formalize it through written documentation Ngunga and Faquir (2012).

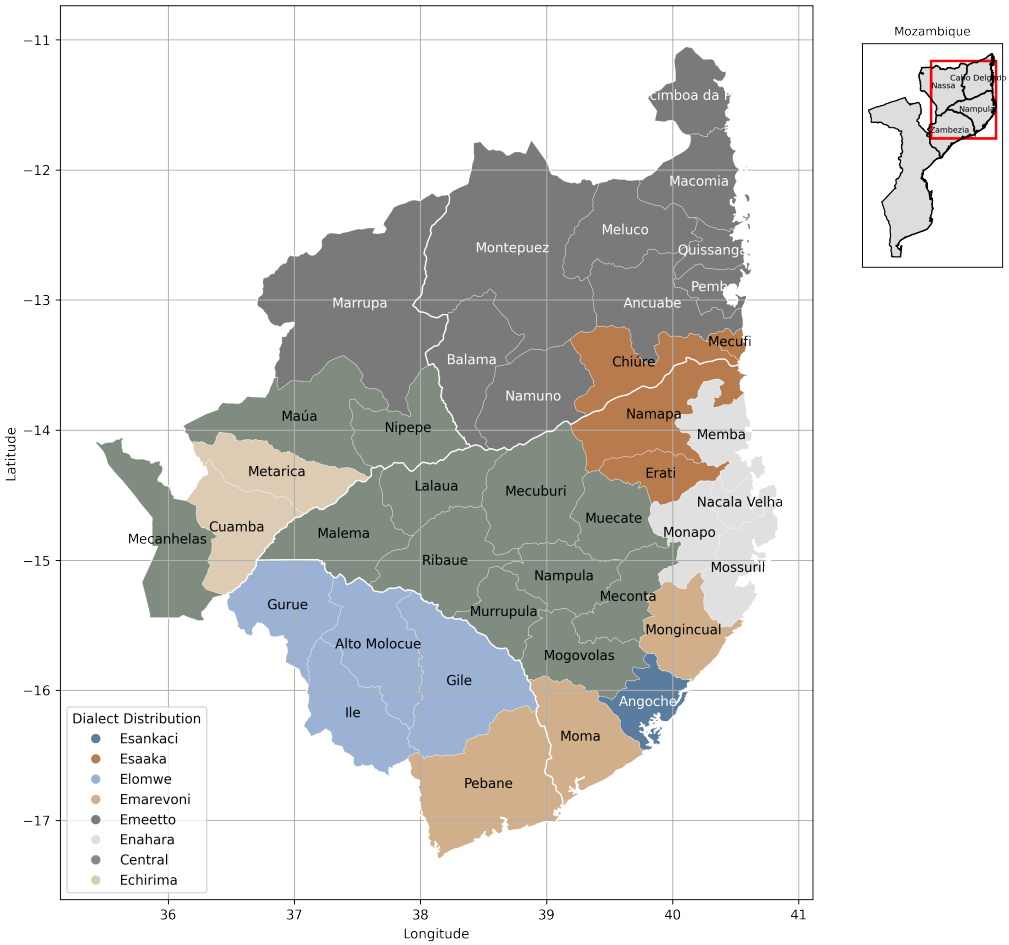


Figure 3.1: Geographic distribution of Emakhuwa variants in Mozambique (Ngunga and Faquir, 2012)

The first standardization of the language was proposed several decades ago and has since undergone four revisions, the most recent of which was published in 2022 by Ngunga et al. (2022). As part of the standardization process, the Central variant of Emakhuwa was given standard status. The alphabet was thoroughly defined, and some orthographic rules were established. The language uses the Latin script, but several commonly used letters in other Latin-based alphabets, such as “b,” “d,” “g,” “j,” and “q,” are absent from Emakhuwa (Ngunga et al., 2022). Additionally, new characters were introduced to better represent sounds specific to the language, as shown in Table 3.2.

Figure 3.2: Emakhuwa’s alphabet (Ngunga et al., 2022)

a	aa	c	e	ee	f	h	i	ii	k	kh	l	ly	m	n	ny	ng
o	oo	p	ph	r	s	t	th	tt	tth	u	uu	v	w	x	y	z

Emakhuwa follows the Subject-Verb-Object (SVO) word order (Van der Wal, 2009). It also features a rich morphological system characterized by agglutination, where words are formed by

Table 3.1: Variants of Emakhuwa

Category	**Elomwe**	**Esankaci**	**Esaaka**	**Echirima**	**Emarevoni**	**Emeeto**	**Enahara**	**Central**
ISO-639	ngl	–	xsq	vmk	xmc	mgh	–	vmw
“sleep”	okona	udupa	orupa	okona	orupa	urupa	orupa	orupa

attaching multiple affixes to a root. This allows for nuanced expression, as prefixes and suffixes can be added to convey tense, aspect, negation, plurality, and other grammatical features (see Table 3.2). As a result, a single word in Emakhuwa can encapsulate information that would require an entire phrase in less morphologically complex languages. Examples of this agglutinative structure are presented in Table 3.3.

Table 3.2: Verb structure in Bantu languages (Mutaka and Tamanji, 2000). NM = negation marker; VP = verbal prefix; TM = tense marker; AM = aspect marker; OM = object marker; IMP = imperative; EXT = extension.

Prefixes					**Stem**		**Suffixes**			
Pre-preliminary	**Initial**	**Post-initial**	**Formative**	**Post-formative**	**Infix**	**Radical**	**Suffix**	**Pre-final**	**Final**	**Post-final**
NM	VP	NM	TM	AM	OM	ROOT	EXT	AM	TM	IMP

Emakhuwa is also a tonal language, meaning that differences in pitch can distinguish word meanings and grammatical functions (Ngunga et al., 2022). In tonal languages, the same sequence of consonants and vowels may convey different meanings depending on the tone pattern used when the word is pronounced. However, tone is generally not marked in standard Emakhuwa orthography, which can make these distinctions difficult to observe directly in written text.

For example, the verbs *omala* (“to finish”) and *omaala* (“to silence”) differ in pronunciation and meaning. In this case, the spelling *aa* reflects a vowel length distinction that helps differentiate the words in writing, while tone differences are typically conveyed only through speech. This illustrates how phonological features such as tone and vowel length can play an important role in meaning, even though they are not always consistently represented in the orthography.

Table 3.3: Examples of agglutination in Emakhuwa

Type	English	Emakhuwa	segmentation in *prefix - base - suffix*
Verb affixes	writing	Kinlepa	**Ki** - **n** - lepa
	I bought	Kinthumaaka	**Ki** - **n** - thumaa - **ka**
Noun affixes	They cannot get	khamuniphwanya	**Kha** - **mu** - **ni** - phwanya
	hoe	ehipa	**e** - hipa
	knee	nikutha	**ni** - kutha
	bones	makhuva	**ma** - khuva
Locative affixes	In our bones	mmakhuvani	**m** - **ma** - khuva - **ni**
	school	oxikola	**o** - xikola
	Spain	wEspanha	**w** - Espanha

Furthermore, Emakhuwa is a gender-neutral language, meaning that it does not mark grammatical gender in the way many Indo-European languages do. Nouns typically have a generic form that applies to both males and females, without the need for gender-specific modifications. Like many other Bantu languages, Emakhuwa is considered a low-resource language in the field of NLP, primarily due to the scarcity of linguistic artifacts and digital resources necessary for developing language technologies.

In the following section, we will explore this classification in more detail and discuss its implications within the context of MT.

3.2 How low-resource is Emakhuwa?

To address this question, we adopt the definition proposed by Nigatu et al. (2024), which characterizes a language as low-resource based on three key dimensions:

1. Linguistic Knowledge – Assessed using the UNESCO language vitality scale (Brenzinger et al., 2003).

2. Data Availability – Measured using the resource classification scale from Joshi et al. (2020), which categorizes languages based on the quantity and quality of available digital text and speech data.

3. Technological Support – Evaluated according to the framework developed by Simons et al. (2022), which considers the availability of computational tools, language technologies, and digital infrastructure for the language.

3.2.1 Existing linguistic knowledge

Although Emakhuwa is primarily an oral language, it has undergone significant formalization efforts over the years. Mozambican linguists have played a central role in documenting the language and developing written resources, including orthographic standards and grammars. While some aspects of the linguistic material remain subject to scholarly debate, these initiatives represent a substantial step toward the language's institutional recognition and development. One tangible outcome of these efforts is the Mozambican government's introduction of a pilot bilingual education program in 2017. As part of this initiative, selected schools in remote areas began delivering instruction in local languages, including Emakhuwa. This policy shift triggered the creation and translation of educational materials, such as textbooks, dictionaries, and pedagogical resources.

Based on the six-level scale proposed by Brenzinger et al. (2003) and adopted by Nigatu et al. (2024), Emakhuwa can be classified at Level 2—Fragmentary. This classification indicates that while there are some grammatical sketches, word lists, and sample texts available for limited linguistic research, the existing documentation remains incomplete. Audio and video recordings may exist, but their quality and degree of annotation vary considerably.

3.2.2 Data availability

Taking the categorization proposed by Joshi et al. (2020), Emakhuwa stands in the 1-Scraping-Bys category since the language has some amount of unlabeled data available in physical or digital form, including in Wikipedia[1]. Also, we foresee that the amount of digital data will grow due to the increasing number of people with Emakhuwa's written skills, thanks to the introduction of bilingual education in Mozambique. Lastly, organized communities are appearing and giving awareness about the necessity to develop content in Emakhuwa, for example, the Ethale Publishing group[2] that is dedicated to producing content in Emakhuwa and publishes in Wikipedia[3].

3.2.3 Technological support

In terms of technological support, Emakhuwa remains classified in the 0-Still category of the taxonomy proposed by Joshi et al. (2020), indicating that the language currently lacks sufficient technological infrastructure to support robust NLP applications. This classification reflects the absence or severe underdevelopment of foundational NLP tools critical for computational processing of the language. These tools include, but are not limited to, tokenizers for segmenting text into meaningful units, part-of-speech taggers for grammatical categorization, spell checkers for error detection and correction, and language models for tasks such as text generation, translation, speech-to-text, sentiment analysis, etc.

3.3 Challenges

Like many other low-resource languages, Emakhuwa faces several challenges that hinder the development of effective MT systems. Below, we outline some of the most critical obstacles currently impeding progress in this area.

3.3.1 Data Scarcity

A major challenge in developing MT systems for Emakhuwa is the limited availability of digital text data. This scarcity affects both the quantity and quality of training resources necessary for building reliable models.

The majority of text written in Emakhuwa is concentrated in the religious domain, and consists particularly of translations of the Bible. This trend is common across many African languages, where religious texts often constitute the bulk of available data. However, as highlighted by Adelani et al. (2022a), relying solely on such data significantly limits the generalization capabilities of MT models. Religious texts tend to follow a specific style and vocabulary, making it difficult

[1] https://incubator.wikimedia.org/wiki/Category:Wp/vmw

[2] https://www.youtube.com/channel/UC45gddDGlA-QjUuZOGLhGxg/featured

[3] https://moleskinefoundation.org/moleskine-foundation-and-ethale-publishing-to-launch-first-ever-articles-in-the-emakhuwa-language-on-wikipedia/

for models trained exclusively on this data to perform well in other domains, such as news, education, or everyday conversation. To develop robust and versatile MT systems for Emakhuwa, it is essential to diversify the training data by incorporating content from a wide range of domains and communicative settings.

3.3.2 Complex Morphology and Lexical variations

Emakhuwa presents unique linguistic challenges due to its tonal system and agglutinative morphology. Tonal markers are essential for meaning disambiguation but remain inconsistently used across texts. Some sources indicate tone using lengthened vowels (e.g., *aa*, *ee*, *ii*), while others use diacritics (e.g., *à*, *á*, *è*, *é*). In some cases, both methods are combined within the same text. Inconsistencies also occur in the representation of consonants. For instance, the sounds "*kh*" and "*nn*" may appear as simply "*k*" and "*n*" depending on the source. The agglutinative nature of Emakhuwa further contributes to spelling variability. Morpheme-level tonal variation and extensive inflection result in multiple valid spellings for the same word. This is especially evident in verb conjugation. Figure 3.3 illustrates negative conjugations of the verb *osuwela* ("to know"). Prefixes such as *ka-* and *ni-* can be written in numerous forms—*kàni*, *kanii*, *kaanì*, *khani*, etc.—depending on how tone and morphology are represented. When combined with suffixes, these variants result in a high number of orthographically distinct but semantically equivalent forms.

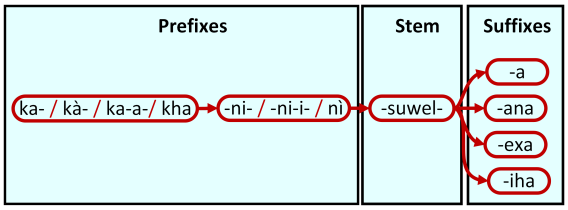


Figure 3.3: Negative conjugations of the verb *osuwela* (“to know”), showing several valid prefix forms. Examples include: *prefix-suwel-a* (“it is not known”), *prefix-suwel-ana* (“we do not know each other”), *prefix-suwel-exa* (“do not know each other well”), and *prefix-suwel-iha* (“does not make itself known”).

These orthographic inconsistencies pose significant challenges for MT, as they lead to vocabulary inflation, increased data sparsity, and out-of-vocabulary issues. In other words, spelling variability, influenced by tonal and morphological complexities, often leads to vocabulary inflation, where a single word can appear in several valid written forms. This variation increases data sparsity, as the same meaning is spread across multiple spellings, reducing the frequency of each form in the corpus. As a result, training robust MT models becomes more difficult, since the system must learn to handle these diverse forms without fragmenting the vocabulary. Consequently, MT systems may struggle with out-of-vocabulary (OOV) issues, where unseen word forms in test data are not represented in the training set, leading to poor translation accuracy.

Loanwords Lexical borrowing represents another source of lexical variation in Emakhuwa, particularly when translating modern or technical terms for which no direct equivalents exist in the

language. Concepts such as *internet* or *digital* are typically adapted either phonetically or through the construction of semantically equivalent expressions. These borrowings frequently result in neologisms or context-specific innovations, with significant variation in spelling and usage depending on the orthographic conventions adopted by different speakers or authors.

Unlike human translators, who can apply context-sensitive strategies—such as phonetic transliteration, calques (literal translations), or culturally appropriate substitutions—MT systems often struggle to handle such nuances effectively. This challenge is further exacerbated by the lack of terminology databases or orthography consensus in Emakhuwa, which leads to inconsistencies in the representation of borrowed terms across texts.

An additional layer of complexity arises from the need to integrate borrowed words into Emakhuwa's agglutinative morphological system. Borrowed nouns, for example, must conform to native pluralization patterns, while verbs require appropriate conjugation to align with syntactic constraints. Although human translators intuitively apply these grammatical transformations, MT models must learn to generalize such adaptations from limited and often inconsistent data. These factors contribute to vocabulary inflation, increased sparsity, and added difficulty in producing fluent and accurate translations.

3.3.3 Direction

Most available Emakhuwa texts are translations from Portuguese. This trend mirrors a broader pattern observed in low-resource language contexts, where dominant languages, such as English, French, or Portuguese, serve primarily as source languages, reflecting their socio-political status (Nigatu et al., 2024). However, this unidirectional approach limits the utility and inclusivity of MT systems, particularly in scenarios where the goal is to extract knowledge or content from local languages into high-resource languages. Moreover, when MT systems are trained in the reverse direction, from Emakhuwa to Portuguese, the process often ends up shaping Emakhuwa according to the syntactic and semantic norms of Portuguese, which can be detrimental to the linguistic integrity of Emakhuwa. Developing bidirectional translation, supporting both Emakhuwa-to-Portuguese and Portuguese-to-Emakhuwa, is essential for building inclusive language technologies. Despite this need, the Emakhuwa-to-Portuguese translation remains less developed due to the unavailability of resources.

This imbalance also affects evaluation practices. Most existing benchmarks are designed for a single-direction translation dataset, making it difficult to rely upon their results. To foster robust MT systems for Emakhuwa, there is a need to create balanced, bidirectional datasets and models that treat both languages as equally important in the translation pipeline.

3.3.4 Gender Neutrality

As mentioned in Section 3.1, Emakhuwa is a gender-neutral language. This means that nouns and pronouns do not generally carry explicit grammatical gender markers. For instance, the word

mwan'aka can refer to both "my son" and "my daughter," as demonstrated in Table 3.4. This inherent gender neutrality in Emakhuwa presents a unique challenge for MT. When translating from a gendered source language (e.g., Portuguese) into Emakhuwa, MT systems must learn to produce gender-neutral output, avoiding the insertion of gender distinctions that do not exist in the target language. Conversely, when translating from Emakhuwa into a gendered target language, the system often lacks explicit gender information in the source text and must rely on contextual cues or make assumptions to produce grammatically correct and semantically appropriate gendered translations.

The absence of explicit gender marking in Emakhuwa can lead to ambiguity or incorrect gender assignments in translations, especially in scenarios where the context is insufficient to infer the intended gender. This challenge highlights the importance of linguistically informed approaches to MT that respect the inherent grammatical structures and nuances of low-resource languages like Emakhuwa.

Table 3.4: Example of Gender-neutrality in Emakhuwa

Emakhuwa	Portuguese	English
Mwan'aka t'oniwereiwa	É o meu **filho** que está doente	It's my **son** who is sick
	É a minha **filha** que está doente	It's my **daughter** who is sick

3.4 Summary

This chapter introduced Emakhuwa, a major Bantu language in Mozambique, highlighting its diverse variants, standardized alphabet, and core linguistic features like agglutinative morphology, tonality, and gender neutrality. Despite its widespread use, Emakhuwa is a low-resource language for NLP, facing significant challenges in developing MT systems due to data scarcity (especially beyond religious texts), complex morphology leading to inconsistent spellings, an imbalance in translation direction favoring Portuguese, and difficulties managing its gender-neutral nature when translating to or from gendered languages. Overcoming these hurdles is essential for advancing Emakhuwa's digital presence and MT capabilities.

Chapter 4

Data Collection

This chapter outlines the data collection process. In the previous chapter, we highlighted the crucial role of data and discussed how the scarcity of linguistic resources limits the development of MT systems for low-resource languages such as Emakhuwa. To address this challenge, we focus on collecting high-quality parallel data produced by human translators to support the training and evaluation of MT models for the Portuguese–Emakhuwa language pair. We describe the dataset creation pipeline, including the quality assurance mechanisms implemented to ensure the development of reliable benchmark datasets, which are later used to assess translation quality in subsequent chapters.

4.1 Introduction

As discussed in Section 3.3.1, one of the key challenges in building machine translation systems between Portuguese and Emakhuwa is the limited availability of parallel training data. Although some resources can be found online, many remain inaccessible, often locked in non-machine-readable formats such as scanned documents or printed texts. These sources pose additional difficulties, including outdated spelling, domain imbalances, and preprocessing complexities.

Given the limited availability of Emakhuwa language resources, making the most of every available resource is essential. Therefore, our data collection strategy consisted of: (1) increasing the volume of available data; (2) ensuring diversity in terms of content type, domain, and orthographic variation; (3) maintaining high-quality standards; and (4) capturing supplementary resources beyond bitext(e.g., dictionaries, glossaries) that can further enhance MT performance.

Our first step involved scraping the web. As, discussed in Section 3.3.1, the web resources primarily consisted of religious texts, such as the Bible text, which often reflect older or non-standard orthographic styles. Although valuable, such materials provide only a narrow view of the language and tend to introduce a bias towards religious domains. To expand coverage, we placed a strong emphasis on acquiring high-quality, manually translated content. We collaborated with professional translators who are either formally trained in Bantu linguistics or have substantial experience working with Emakhuwa. Their expertise was crucial in producing translations that

adhere to a more standardized and formalized orthography. This step not only enriched the linguistic quality of our dataset but also helped fill gaps left by domain-specific or orthographically inconsistent sources. In parallel, we also sourced data from scanned books and printed materials. While these texts often exhibit a mix of orthographic styles, reflecting both historical and contemporary usage, they contribute valuable linguistic diversity to the dataset. Extracting this data involved Optical Character Recognition (OCR) and post-processing to correct errors introduced during digitization.

By combining these sources—web-scraped content, professionally translated texts, and scanned documents—we were able to compile a rich and representative parallel corpus. This dataset serves as the foundation for training and evaluating our machine translation models, supporting both linguistic diversity and quality in our system development. The following sections provide a detailed overview of each data collection effort.

4.2 Collecting Data from Online Sources

We gathered a considerable amount of data from online sources, including both parallel and monolingual corpora, as described below.

4.2.1 Parallel Data

Parallel data were sourced from two online platforms:

- **Jehovah's Witnesses (JW) Website**[1]: This website offers a vast collection of religious texts translated into multiple languages, including Portuguese and Emakhuwa. It has been widely used in the development of parallel datasets across Africa (Agić and Vulić, 2019). We extracted content from various magazines published between 2016 and 2021, including "The Watchtower—Study Edition", "The Watchtower", and "Awake!", as well as translations of 27 New Testament books.

- **African Storybook (ASB) platform**[2]: The ASB website provides openly licensed children's stories in many African languages. We identified 17 short stories available in Emakhuwa, each paired with Portuguese translations. Although the volume of data from this source is relatively small, the stories not only diversify the domain of our corpus but also offer culturally relevant and accessible language material.

4.2.2 Monolingual data

A considerable amount of monolingual data is also available exclusively in Emakhuwa. Below, we describe two sources from which we have collected the monolingual data:

[1] https://www.jw.org
[2] https://www.africanstorybook.org

- ***Wòniherya* News**: An online news magazine[3] with monthly publications, all exclusively written in Emakhuwa. Our dataset includes content from 32 issues published between 2021 and 2024.
- **Wikimedia data**: The Wikimedia Incubator[4], currently with a total of 1005 articles written in Emakhuwa.

In both sources, we observed that some of the text contained mixed words from other languages. Therefore, we performed preprocessing to clean the data by following these steps:

1. **Segmenting into smaller sentences**: To achieve this, we used the NLTK[5] sentence tokenizer. The Portuguese NLTK sentence tokenizer was chosen due to the structural similarities between Portuguese and Emakhuwa at the sentence segmentation level.
2. **Filtering Emakhuwa sentences**: To ensure that only Emakhuwa sentences were included in the final dataset, we applied GlotLID, a language identification system (Kargaran et al., 2023), to each sentence. Sentences with fewer than three tokens or identified as languages other than Emakhuwa were excluded.

4.3 Collecting Data from Professional Translation

Manual translation, while both costly and time-consuming, remains one of the most reliable methods for obtaining high-quality parallel data, particularly for low-resource languages such as Emakhuwa. In this context, professional translation becomes a highly valuable resource, offering well-structured, accurate, and contextually appropriate datasets that reflect contemporary usage of Emakhuwa. Recognizing these benefits, our methodology prioritized the collection of professionally translated parallel sentences from (1) the news domain and (2) the FLORES+ evaluation set. The following subsections provide a detailed description of each source and the methodology used in the collection and preparation of the data.

4.3.1 News Domain Data

We expanded the data coverage to the news text domain, and this effort led to the creation of the first news translation corpus between Portuguese and Emakhuwa (Ali et al., 2024c). To achieve this, we hired professional translators to translate a subset of news articles from the MOT (Palen-Michel et al., 2022) dataset, comprising of news articles sourced from the VOA (Voice of America) platform published between 2001 and 2021. Our data collection process involved the following steps:

[3] https://emakhuwa.org.mz/
[4] https://incubator.wikimedia.org/wiki/Category:Wp/vmw
[5] https://www.nltk.org/

Preparation: We sampled 1,897 Portuguese news articles divided into seven categories: politics (288 articles), economy (287 articles), culture (329 articles), sports (345 articles), health (204 articles), society (198 articles), and world news (246 articles). We chose articles that encouraged lexical diversity within each category and maintained contextual relevance for Mozambique. During translation, we have supplied translators with bilingual dictionaries compiled from multiple existing dictionaries as well as provided them with the Ndapassoa (2015) glossary, which provides terminologies for News translations in Mozambican languages. We also provided translation guidelines where we emphasized the following: (1) The translated text should convey the same meaning as the source; (2) Adhere to the latest orthography standards; (3) Personal names were not to be translated; (4) Loanwords adapted into Emakhuwa were to be annotated in the comments section of each segment (see Figure 4.2).

Recruitment: We recruited translators based on their proficiency in both Portuguese and the Central variant (*vmw*) of Emakhuwa and their proven experience in Emakhuwa translation. In total, we worked with 10 translators: two were final-year bachelor's students majoring in Bantu Linguistics at the Faculty of Arts and Social Sciences (*Faculdade de Letras e Ciências Sociais da Universidade Mondlane*- FLCS), seven held degrees in Bantu Linguistics and worked as professional translators, and one was a Radio Mozambique host with vast experience in translating and broadcasting content in Emakhuwa. The translation/revision work was paid at competitive rates and budgeted based on the word count of the source text.

Translation Workflow: To manage the entire translation workflow we used MateCat[6] (see Figure 4.2). The process involved iterative cycles with the following phases:

- **Translation**: Translators were assigned Portuguese text segments to translate into Emakhuwa.
- **Revision**: Translated documents were automatically checked using our custom-developed spelling tool (see Figure 4.1). The revision reports were then sent back to the translators so that they could revise their work and resubmit after completing the revisions.
- **Quality Assurance Meetings**: We conducted workshops that all translators were required to attend. During these workshops, we discussed potential new glossary entries and collaboratively resolved some major issues that led to translation disagreements.

4.3.2 FLORES+ Evaluation Set

Recognizing the imperative for diverse evaluation data to effectively assess the generality of MT systems, our data collection was expanded to include translations of the FLORES evaluation set. This popular multi-way benchmark, first introduced by Guzmán et al. (2019), is progressively increasing its linguistic coverage, particularly for low-resource languages (Goyal et al., 2022; NLLB

[6]https://www.matecat.com

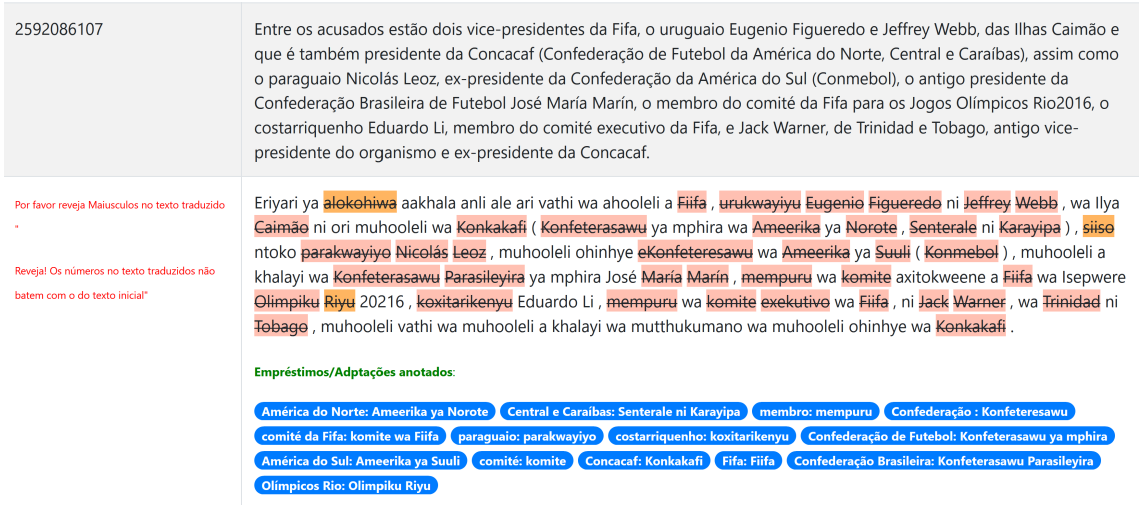


Figure 4.1: Screenshot of a spelling report. The first column lists the segment ID along with any potential translation issues (i.e., punctuation, source-target length ratio flag, number mismatch, loanwords not annotated, case mismatch, etc.). The second column displays the source text at the top and its translation below. In the translation, potential misspellings are highlighted in yellow, indicating that suggestions for corrections are available, and red, indicating that no suggestions exist. Additionally, the report lists, at the bottom, all words that translators have annotated as loanwords from Portuguese, using the format *<donor sequence in Portuguese>:<recipient sequence in Emakhuwa>*.

Team et al., 2022; Doumbouya et al., 2023; Gala et al., 2023; Abdulmumin et al., 2024a). As part of this initiative, we contributed to the FLORES+ expansion by adding support for Emakhuwa, resulting in the creation of a new benchmark between Portuguese and Emakhuwa (Ali et al., 2024b). Specifically, we translated the *dev* and *devtest* splits, using the Portuguese version as the source and Emakhuwa as the target language. Because this data is intended for evaluation purposes, we adopted a slightly more rigorous translation process compared to the News domain translations described in Section 4.3.1. In particular, we introduced an additional quality control step (i.e., Judgments), as outlined below.

Translation: The translation tasks were split between two professional translators: one handled the *dev* set and the other the *devtest* set. After completing their initial translations, the texts were run through an automated spell-checking system (see Figure 4.1) to flag potential spelling errors. Based on the results, we provided feedback to the translators, who then revised and improved their translations as required.

Revision: Following the initial translation phase, we performed a peer-review process by swapping the two sets between the translators. Each translator post-edited their colleague's work within the Matecat platform, with the aid of the previously generated spelling reports. This collaborative revision ensured a second layer of quality control. Table 4.1 provides the Quality Report generated by Matecat, which includes various metrics used to evaluate the translation based on the revisions

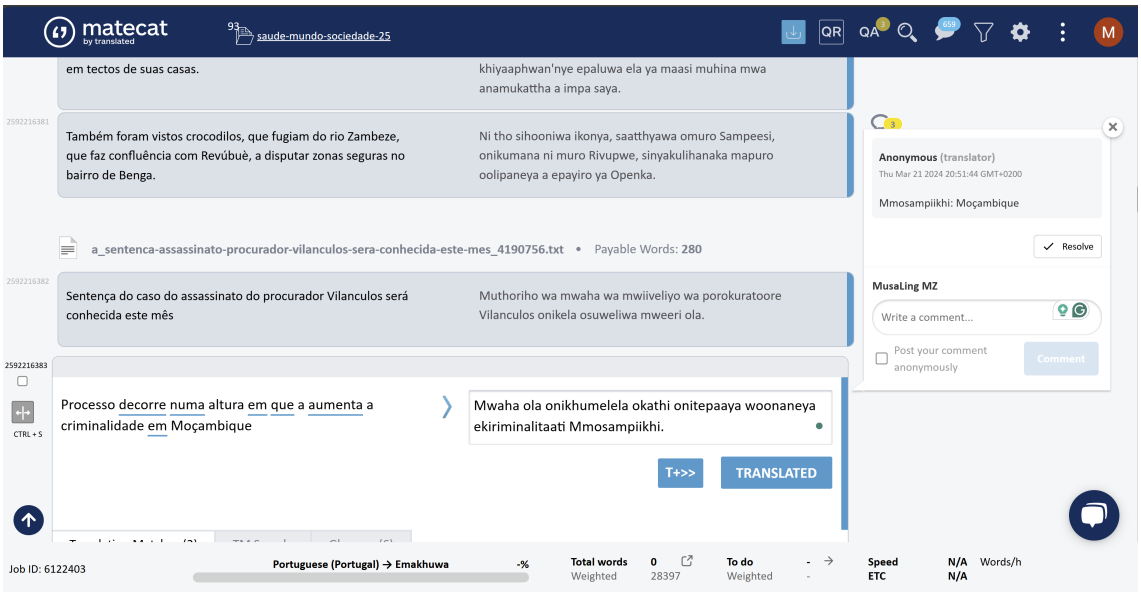


Figure 4.2: Screenshot of the Computer Assisted Translation Tool

made. The report shows that the reviewer working on the *devtest* set performed a greater amount of post-editing.

Table 4.1: Matecat's quality report post revision.

	dev	**devtest**
Post-Editing Effort	99%	95%
Time to edit	02m38s	05m42s
Quality score	23.31	54.22
Avg. Edit Distance	0.23 ± 1.77	7.09 ± 11.94

Judgments Following the revision phase, we conducted a final quality assessment using a Direct Assessment (DA) pipeline inspired by Guzmán et al. (2019). Three independent raters evaluated each translation for adequacy on a 0–100 scale. Translations scoring below an average of 70 were returned to the translators for reworking. The scoring thresholds follow those proposed by Wang et al. (2024a):

- 0: No meaning preserved;
- 1–34: Some meaning is preserved, but major content is lost;
- 35–67: Most meaning preserved;
- 68–99: Meaning consistent with the source;
- 100: Perfect translation.

Evaluations were conducted using our custom-developed annotation tool (see Figure A.1 in the Appendix A). To improve rater reliability and attentiveness, 200 **control** instances were added. These consisted of incorrect Portuguese–Emakhuwa translation pairs generated automatically using the Madlad-400-3bt model[7] (Kudugunta et al., 2024), known to perform poorly on this language pair. In addition to adequacy, raters assessed adherence to **standard Emakhuwa orthography**. Each translation was scored from 1 to 5:

- 1: Not written in standard orthography;
- 2: Minimal adherence;
- 3: Moderate adherence;
- 4: Mostly adherent;
- 5: Fully compliant with standard orthography.

What is the quality of the translations, and how reliable are the DA evaluations?

Figures 4.3 and 4.4 show the raw scores per annotator for Direct Assessments. Given the mean scores, in both the *test* and *devtest* sets, Annotator 1 and Annotator 2 awarded higher quality scores, while Annotator 3 was more critical but still within the spectrum of acceptable translations. This suggests a generally positive perception in terms of Adequacy. Figure 4.5 displays the Direct Assessment scores on the control set. Annotator 1 and Annotator 3 have median scores below the threshold of 34 points, which suggests that, as expected, they have generally assessed the control translations as of low quality. Annotator 2, however, has a median score above the threshold, suggesting he was less attentive than others. Nevertheless, the reliability results for adequacy in Table 4.2 in Inter-Class Correlation are 0.67 for *dev* and 0.66 for *devtest*, suggesting moderate reliability of DA evaluations.

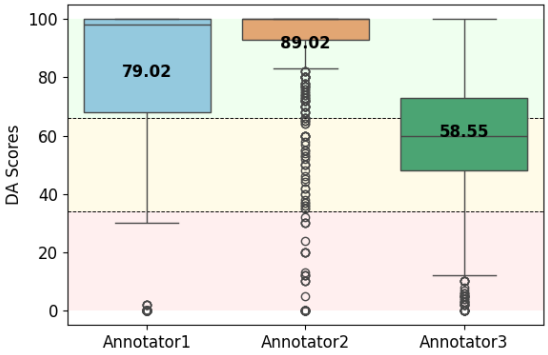


Figure 4.3: Direct Assessment adequacy scores per annotator on *dev* set

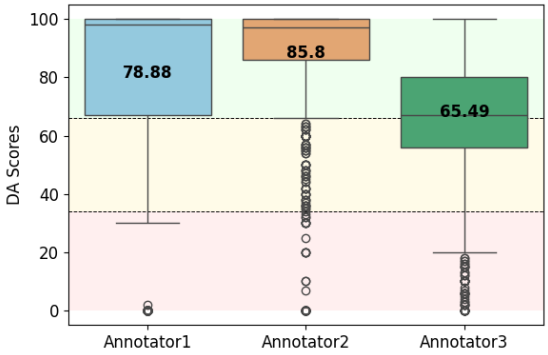


Figure 4.4: Direct Assessment adequacy scores per annotator on *devtest* set

[7] https://huggingface.co/google/madlad400-3b-mt

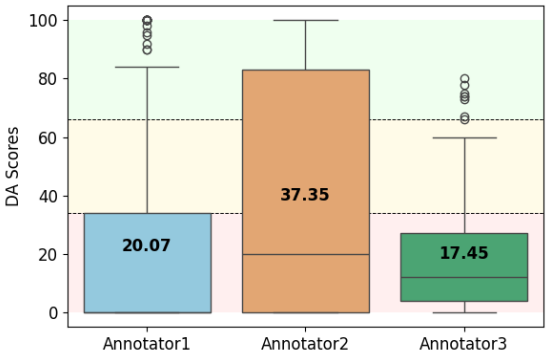


Figure 4.5: Direct Assessment adequacy scores per annotator on *control* set

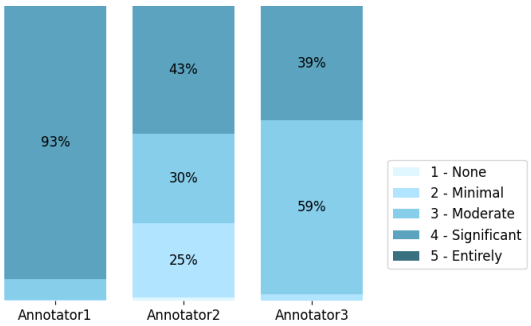


Figure 4.6: Assessment of standard orthography usage on the control set.

Do the translations comply with standard orthography?

The evaluation of orthographic compliance yielded lower inter-annotator agreement. As shown in Figure 4.6, raters varied significantly in their assessments of adherence to Emakhuwa's standard orthography. This variation is reflected in the corresponding ICC scores: 0.35 for the *dev* set and 0.27 for the *devtest* set.

These low agreement scores highlight a broader and ongoing challenge in standardizing Emakhuwa orthography. Despite the annotators' extensive experience and formal training in writing the language, their assessments diverged significantly. This discrepancy may be attributed to several factors, including inconsistent perceptions of orthographic norms, a lack of widely accepted or fully formalized spelling conventions, and individual linguistic intuition shaped by regional or dialectal influences.

The results raise important questions about user satisfaction with the orthographic output of MT systems. If trained annotators cannot consistently agree on what constitutes correct orthography, it is likely that end users will also have varying expectations. We argue that MT for Emakhuwa and other low-resource languages with orthographic variation must incorporate personalization features, as allowing users to select or influence the orthographic style of the output could improve both perceived quality and user satisfaction. Such adaptability is especially crucial in contexts where orthographic norms are still evolving or contested.

Table 4.2: Intraclass Correlation Coefficient (ICC) and Confidence Interval (CI) Results for Adequacy and Orthography usage annotation.

	Adequacy		**Orthography**	
	dev	devtest	dev	devtest
ICC	0.67	0.66	0.35	0.27
CI	[0.63, 0.71]	[0.62, 0.7]	[0.27, 0.42]	[0.18, 0.35]

4.4 Collecting Data from Other Sources

A significant volume of text in Emakhuwa exists solely in hardcopy format, making it hardly accessible for machine processing. We have collected this data type by digitizing printed documents with Google Vision OCR tool and manually preprocessing them. We follow a similar methodology as in Rijhwani et al. (2020), consisting of:

- **Data Extraction**: First, we scanned the pages as images.
- **Post-Correction**: After OCR, we manually corrected any errors with the help of two volunteers, who reviewed the text in the images and corrected potential OCR errors. Post-correction was facilitated using the free tier of Label Studio[8].

We extracted data from: (1) "Método Macua book", which is rich in cultural narratives, with tales intrinsic to Emakhuwa culture; (2) Translation of the Mozambican Land Law; (3) Translation of the Universal Declaration of Human Rights; (4) Catechism books from Paul VI Catechetical Centre; and (5) Bilingual dictionaries.

4.5 Datasets

Our data collection efforts culminated in the creation of the first comprehensive, high-quality, representative, and diverse parallel and monolingual corpus for Emakhuwa. This dataset serves as a foundational resource for machine translation and other NLP tasks, addressing the critical gap in available linguistic resources for Emakhuwa. Below, we provide a detailed breakdown of the collected data, topic coverage, and its potential applications.

Table 4.3 presents an overview of the sentences collected from each source discussed in the previous sections. The parallel corpus totals **68,596 sentence pairs**, while the monolingual Emakhuwa corpus comprises **14,463 sentences**, thereby ensuring broad coverage across multiple domains and linguistic styles.

Splits for Machine Translation For training and evaluation of MT systems, we partitioned the parallel corpus into Training Set, Validation Set, and Test Set. Table 4.4 details the distribution of sentences across splits and their respective domains. This partitioning ensures balanced representation, minimizing domain bias during model assessment.

FLORES+ evaluation This evaluation consists of two splits: *dev* and *devtest*, containing 997 and 1,012 sentences each, respectively. The dataset spans multiple domains, with content sourced from WikiNews, WikiJunior, and WikiVoyage, and encompasses a wide range of sub-topics, including Travel (1,529 sentences), Politics (341 sentences), Science (325 sentences), Crime (313 sentences), Sports (162 sentences), Entertainment (68 sentences), and others such as Disasters, Health, Geography, and Nature.

[8]https://labelstud.io/

Table 4.3: Summary of the collected data. Ali-2021 refers to the training subset from Ali et al. (2021). News Data consists of sentence pairs obtained through manual translation. OCR-ed Data includes sentences extracted using OCR. Mono denotes monolingual data, and FLORES+ corresponds to the manually translated FLORES evaluation set.

		Ali-2021	**OCR-ed**	**Mono**	**News**	**FLORES+**
	Topic	# Sent. pairs	# Sent. pairs	# Sent.	# Sent. pairs	# Sent. pairs
	Religious	45,983	-	-	-	-
	News Politics	-	-	-	2,084	-
	News Economy	-	-	-	2,274	-
	News Culture	-	-	-	3,822	-
	News Sports	-	-	-	3,148	-
	News Health	-	-	-	2,423	-
	News Society	-	-	-	2,099	-
	News World news	-	-	-	2,681	-
	Tales	204	1,911	-	-	-
	Legal	917	-	-	-	-
	History	93	-	-	-	-
	Wikipedia	-	-	-	-	2,009
	Miscellaneous	80	-	14,463	-	-
	Total	**47,273**	**1,911**	**14,463**	**17,403**	**2,009**
Toks.	*vmw*	951,520	12,283	27,2559	541,598	39,684

Supplementary datasets Beyond the primary parallel and monolingual corpora, our data collection efforts have also led to the creation of several supplementary datasets. These resources can be invaluable for a range of other NLP tasks, providing rich material for further research and development in Emakhuwa. These additional datasets include:

- **Loanwords Dataset**: The dataset contains lexical borrowing mappings between Portuguese and Emakhuwa.
- **OCR Post-processing Dataset**: Derived from our work with scanned and printed materials, this dataset contains examples of text before and after OCR correction, enabling the development and evaluation of OCR post-correction tools.
- **Bilingual Dictionary**: Compiled from various sources, this dictionary provides a valuable resource for lexical lookup between Portuguese and Emakhuwa.
- **News Classification Dataset**: Based on the news articles collected, this dataset is suitable for training and evaluating models for text classification tasks, particularly within the news domain.
- **Translation Quality Assessment Dataset**: Stemming from our rigorous quality control processes for the FLORES+ evaluation set, this dataset provides human judgments on translation quality.

Table 4.4: Dataset splits and distribution of examples across different topics

Topic	**TRAIN**	**DEV**	**TEST**	Ali-2021	News Data	OCR-ed
Religion	45,386	290	307	✓	-	-
Tales	2,057	20	36	✓	-	✓
Legal	877	24	16	✓	-	-
Miscellaneous	43	65	71	✓	-	-
News politics	1,919	67	67	-	✓	-
News economy	2,089	65	90	-	✓	-
News culture	3,585	108	103	-	✓	-
News sports	3,005	63	55	-	✓	-
News health	2,231	83	84	-	✓	-
New society	1,891	90	87	-	✓	-
News world	2,490	89	77	-	✓	-
Total	65,574	964	993			

4.6 Summary

This chapter details the comprehensive data collection efforts undertaken to create resources for Portuguese-Emakhuwa machine translation, a task that is complicated by the scarcity of existing parallel data for Emakhuwa. The strategy involved augmenting data volumes, ensuring diversity in content and orthography, securing high-quality evaluation data, and incorporating additional useful external data. Data was sourced by scraping online platforms like the Jehovah's Witnesses website and the African Storybook platform for parallel texts, and *Wòniherya* News and Wikimedia for monolingual Emakhuwa data. A significant emphasis was placed on acquiring high-quality, manually translated content, particularly in the news domain, by collaborating with professional translators. This involved translating news articles and the FLORES+ evaluation set, with a rigorous workflow including preparation, recruitment, translation using CAT tools, revision aided by custom spelling tools, workshops, and a judgment phase for the FLORES+ data to assess adequacy and orthographic consistency. Additionally, data was extracted from printed materials using OCR and manual post-correction.

These collection efforts resulted in a substantial and diverse dataset: a parallel corpus of 68,596 sentence pairs and a monolingual Emakhuwa corpus of 14,463 sentences, covering domains such as religion, news, tales, and legal texts. The collected data was partitioned into training, validation, and test sets for machine translation system development, with specific splits created for the FLORES+ evaluation benchmark (*dev* and *devtest* sets). The chapter also highlights the challenges encountered, such as domain imbalances from web-scraped religious texts and significant inter-annotator disagreement on standard Emakhuwa orthography during the evaluation phase. Overall, this indicates a need for personalized orthographic outputs in MT systems for such languages.

Chapter 5

Benchmarks

This chapter aims to establish the benchmark performance for machine translation between Portuguese and Emakhuwa. Building on the dataset introduced in Chapter 4, we describe the development and training of a baseline neural machine translation system based on the Transformer architecture. We begin by outlining the model architecture, preprocessing pipeline, and training procedures for both translation directions: Portuguese to Emakhuwa and Emakhuwa to Portuguese. Next, we present a detailed evaluation of translation quality using standard metrics across a range of test sets, including both in-domain and out-of-domain data. Finally, we conduct a qualitative error analysis to uncover common translation errors and recurring patterns, to shed light on the specific challenges raised by this low-resource language pair. These benchmarks provide a solid foundation for future research aimed at improving translation performance in this context.

5.1 Baseline System

Architecture We adopted the transformer architecture (Vaswani et al., 2017) for training MT systems to translate in both directions, from Portuguese to Emakhuwa and from Emakhuwa to Portuguese – the baseline system. We used the OpenNMT toolkit (Klein et al., 2017), applying the vanilla transformer architecture, which comprises an encoder and a decoder, each with 6 layers, 8 attention heads, and 512 hidden units in the feed-forward layers. Both source and target word embeddings are represented in 512 dimensions. We used layer normalization and added dropout with a 0.1 probability to all transformer block.

Preprocessing We used the data splits shown in Table 4.4 for training and validation. To segment words into subword units, we applied Byte Pair Encoding (BPE) with a vocabulary size of 40,000 tokens. We used OpenNMT's pyonmttok[1] module to learn the BPE model. The tokenizer was configured with an aggressive mode to help with better granularity, given the fact that Emakhuwa is morphologically complex. We enabled joiner annotations and soft case region handling. The

[1] https://pypi.org/project/pyonmttok/

training of the BPE model was conducted by ingesting all corpus files in both languages, and the learned subword model was then applied to tokenize each file before training and testing.

Training Training was performed using OpenNMT's token-based batching, setting the batch size to 4,096 tokens and accumulating gradients over 4 batches to stabilize updates. The training was run for a maximum of 100,000 steps, with validation performed every 1,000 steps. To prevent overfitting, we used early stopping with a patience of 4 validation steps. Model checkpoints were saved every 1,000 updates. Optimization was performed using the Adam optimizer (Kingma and Ba, 2014) with in conjunction with the Noam learning rate schedule, employing a base factor of 2, a warm-up period of 8,000 steps, and a β_2 parameter set to 0.998. Label smoothing of 0.1 was applied to improve generalization, and the model was trained using 16-bit floating-point precision.

We train two models, Portuguese-Emakhuwa (*pt-vmw*) and Emakhuwa-Portuguese (*vmw-pt*), in three runs using different seeds (123, 42, 456). The plots in Figure 5.1 illustrate the decrease in cross-entropy loss over training steps. We observe a consistent pattern across both training runs and language directions. Notably, the models generally converge after around 7,000 steps, with the exception of the run using seed 42, which converges at approximately 8,000 steps.

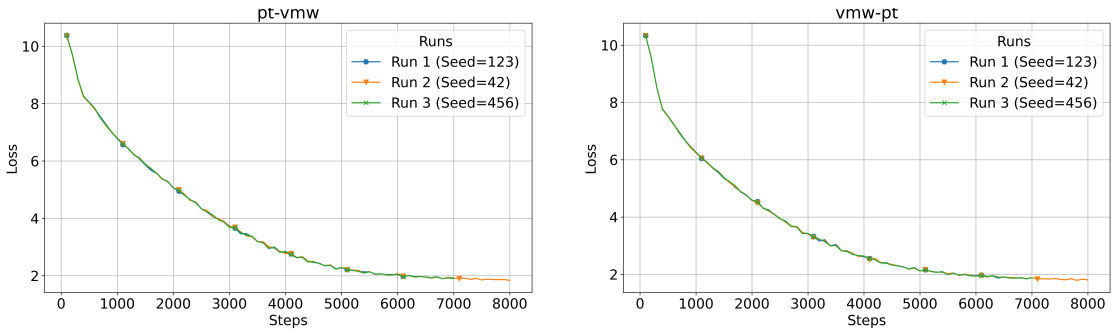


Figure 5.1: Cross-entropy loss over training steps between *pt-vmw* and *vmw-pt* directions.

Similarly, Figure 5.2 shows that accuracy increases steadily as training progresses, reaching a plateau around the same convergence points (7,000 steps, or 8,000 steps for the run with seed 42).

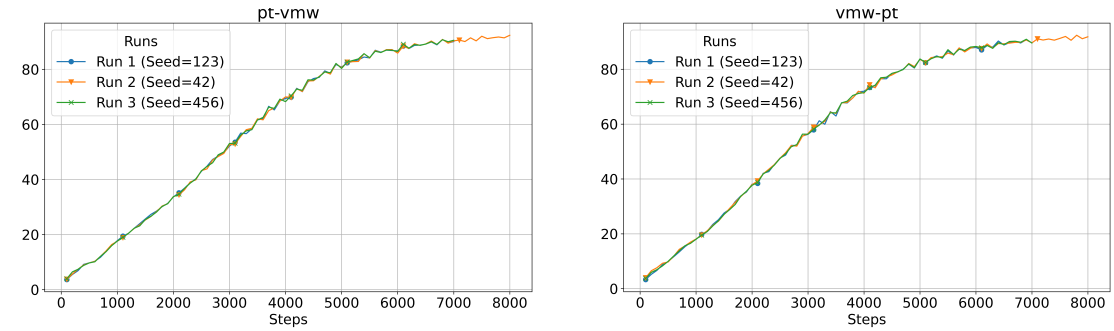


Figure 5.2: Accuracy over training steps between *pt-vmw* and *vmw-pt* directions.

Evaluation We evaluated the performance of the MT systems using two standard metrics: BLEU (Papineni et al., 2002) and ChrF (Popović, 2015). The evaluation was conducted on two datasets: the test split in Table 4.4, here referred to as *Multi-source*, and the *FLORES+* evaluation set, which includes the *Dev* and *DEVTEST* subsets (see Section 4.5 for details).

Results Table 5.1 reports the average performance of the translation models across three runs using different random seeds. Overall, both translation directions (*pt-vmw* and *vmw-pt*) exhibit relatively low scores, reflecting the difficulty of the translation task in this language pair. As anticipated, both models perform better on the *Multi-source* test set compared to the *FLORES+* benchmark. This difference can be attributed to domain similarity: the *Multi-source* test set contains sentences drawn from the same or similar distribution as the training data and can therefore be considered in-domain. In contrast, the *FLORES+* dataset, which is based on Wikipedia text, represents a challenging out-of-domain setting. The performance gap between the two test sets highlights the limited generalization capabilities of the models when evaluated on unseen and topically diverse data.

Notably, the BLEU score of the *vmw-pt* model on the *Multi-source* test set reaches an average of 10.40, compared to 7.14 for the *pt-vmw* model. A similar trend is observed across both the *FLORES+ dev* and *devtest* sets.

These results underscore not only the challenges of building robust translation systems for under-resourced languages like Emakhuwa, but also the importance of domain diversity and directionality in model performance and generalization.

Table 5.1: Evaluation scores on the test set for the MT models

		Multi-source		*FLORES+* *DEV*		*FLORES+* *DEVTEST*	
	Seed	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF
pt-vmw	123	6.88	34.48	2.84	27.76	2.64	27.16
	42	7.37	35.77	3.29	29.63	2.72	28.58
	456	7.18	35.55	2.85	28.93	2.41	27.68
AVG.		7.14 ± 0.20	35.27 ± 0.56	2.99 ± 0.21	28.77 ± 0.77	2.59 ± 0.13	27.81 ± 0.59
vmw-pt	123	10.27	32.44	5.51	27.51	3.52	25.76
	42	10.67	32.87	5.58	27.74	3.95	26.12
	456	10.27	32.50	5.55	27.58	3.36	25.79
AVG.		10.40 ± 0.19	32.60 ± 0.19	5.55 ± 0.03	27.61 ± 0.10	3.61 ± 0.25	25.89 ± 0.17

5.2 Error Analysis

To better understand the limitations of our models, we performed a qualitative error analysis by manually examining translation outputs. We randomly selected four examples from the *FLORES+ devtest* set and analyzed the translations produced by the best-performing model (i.e., seed = 42). We used the Multidimensional Quality Metrics (MQM) framework Lommel et al. (2013), focusing

specifically on accuracy, i.e., errors where the meaning of the source text is distorted, omitted, or augmented in the target translation. In other words, we classified the errors into four main categories: Mistranslation, Omission, Addition, and Untranslated spans.

5.2.1 Emakhuwa Translations

The *pt-vmw* model consistently showed the weakest performance across evaluation metrics. Table 5.2 presents representative examples and our analysis.

- **Mistranslation**: Mistranslations were the most frequent and often the most severe error type. In some cases, the meaning of the entire sentence was altered or lost. For instance, in Example 1, the translation does not bear semantic resemblance to the source sentence, which suggests misalignment during training or a failure in lexical mapping. Similarly, in Example 3, the named entity "Springboks" was incorrectly rendered as "Sinkapura" (Singapore), likely due to a faulty approximation or confusion during tokenization and vocabulary mapping.

- **Omission**: Omission errors were also common and possibly affected by prepositional phrases or low-frequency terms. In Example 2, the phrase "*em latim*" (in Latin) is completely removed from the translation, leading to a partial loss of meaning. Similarly, Example 4 omits two important content elements: "*causando as marés*" and the proper noun "*Sagitário*", which reduces the coherence and informativeness of the output.

- **Addition**: In certain cases, the translation introduced content that is not present in the source sentence. In Example 3, the phrase "*etthukinwe mwaha wa ophurela musurukhu 5*" is shown in the output, but has no corresponding string in the source text. This is a type of hallucination that is common in a poorly performing MT system.

- **Untranslated spans**: We also observed instances of untranslated spans, where terms from the source sentence remained unchanged in the target output. For instance, in Example 4, the term "*Via-Láctea*" (Milky Way) was retained in Portuguese, rather than being adapted to Emakhuwa (e.g., via lexical borrowing). This is inappropriate, as such terms should be localized to align with the phonetic structure and rules of Emakhuwa, as demonstrated in the reference translation, "*Eviiya Laateya*".

Overall, the *pt-vmw* model struggles with both fluency and adequacy. The prevalence of hallucinated and mistranslated spans underscores the need for improved vocabulary coverage, enhanced alignment mechanisms, and more targeted training on domain-specific or culturally salient content.

Table 5.2: Error Analysis Evaluation on the *pt-vmw* model

Example 1	
Source	*O Aerosmith cancelou os shows restantes da sua turnê.*
Reference	*Aerosmith ohiyererya miteko soowooniherya saahanle wa ikwaha satthokihiwe.*
Translation	*Mameediku manceene aahiiriha opaka soothanla mwa enamuna ekina.*
Errors	**Mistranslation** - The whole sentence is mistranslated

Example 2	
Source	*Naturalistas e filósofos se concentraram em textos clássicos e, em particular, na Bíblia em latim.*
Reference	*Axineene ni anamiruku atthipanyerya malepo akhalayi ni, otepaxa weene, Piipiliya wa latim.*
Translation	*Omwiita ni musuweli a elimaliherya soolempwa sa khalai, vanceenexa Biibiliya.*
Errors	**Mistranslation** – Naturalistas ≠ Omwiita; **Omission** – "*em latim*" is omitted in the translation;

Example 3	
Source	*Para o Springboks, acabou uma sequência de cinco derrotas consecutivas.*
Reference	*Wa nikhuro na Springboks, wa mmalela wa ikhwaha sooxinttiwa thanu sottharelana.*
Translation	*Wa Sinkapura, etthukinwe mwaha wa ophurela musurukhu 5.*
Errors	**Mistranslation** – Sinkapura ≠ Springboks ; **Addition** – "etthukinwe mwaha wa ophurela musurukhu 5" is not present in the source;

Example 4	
Source	*Assim como a Lua exerce uma atração sobre a Terra, causando as marés, a Via Láctea exerce uma força sobre a galáxia de Sagitário.*
Reference	*Siisaale mweeri okumihenryaaya exookhi ni etthaya, Eviiya Laateya ennikumiherya tho nave ikuro ni makhuru a itheneeri sa opuro.*
Translation	*Nave-tho, khula mweeri anneemela itthu sa vathi, mwaattelo wa axilalaxia ya Via-Láctea.*
Errors	**Mistranslation** – "exerce uma atração sobre a" ≠ "anneemela itthu sa", "exerce uma força sobre a galáxia de" ≠ "mwaattelo wa axilalaxia"; **Addition** – "khula" is not present in the source; **Omission** – "causando as marés" and "Sagitário" is omitted; **Untranslated** – "Via-Láctea" is left untranslated in the translation

5.2.2 Portuguese Translations

While the *vmw-pt* model performs comparatively better, it still exhibits significant translation quality issues (see Table 5.3). All error categories identified in the reverse direction are also present in the Portuguese translation.

- **Mistranslation**: This remains the most common error type. Example 1 reveals a complete semantic drift, where the translated sentence shares no recognizable meaning with the Emakhuwa source. In Examples 2 and 3, lexical mismatches such as “Springboks” being translated as “exército de Adore” point to misrecognition of named entities.

- **Omission**: Content omission is another recurring issue. In Example 4, the entire clause "*ennikumiherya tho nave ikuro ni makhuru a itheneeri sa opuro*" is missing from the translation, leading to a significant loss of intended meaning.
- **Addition**: Several translations include hallucinated content not present in the source. For instance, in Example 2, the mention "os líderes da Renamo" is presented unexpectedly, suggesting that the model may be drawing from unrelated or domain-inappropriate training data.
- **Untranslated Spans**: The output also contains untranslated strings and poorly adapted transliterations.

Table 5.3: Error Analysis Evaluation on the *vmw-pt* model

Example 1	
Source	*Aerosmith ohiyererya miteko soowooniherya saahanle wa ikwaha satthokihiwe.*
Reference	*O Aerosmith cancelou os shows restantes da sua turnê.*
Translation	Acumular o regresso às suas actividades é também garantido.
Errors	**Mistranslation** - The whole sentence is mistranslated

Example 2	
Source	*Axineene ni anamiruku atthipanyerya malepo akhalayi ni, otepaxa weene, Piipiliya wa latim.*
Reference	*Naturalistas e filósofos se concentraram em textos clássicos e, em particular, na Bíblia em latim.*
Translation	*As autoridades e os chamados abrir os olhos dos antigos e deleitas, particularmente os líderes da Renamo.*
Errors	**Mistranslation** – "*Axineene ni anamiruku atthipanyerya malepo akhalayi*" ≠ "*As autoridades e os chamados abrir os olhos dos antigos e deleitas*"; **Addition** – "*os líderes da Renamo*" is not present in the source; **Omission** – "*Piipiliya wa latim*" is omitted

Example 3	
Source	*Wa nikhuro na Springboks, wa mmalela wa ikhwaha sooxinttiwa thanu sottharelana.*
Reference	*Para o Springboks, acabou uma sequência de cinco derrotas consecutivas.*
Translation	*Para o exército de Adore, a favor de umas seis meses.*
Errors	**Mistranslation** – "*Springboks*" ≠ "*exército de Adore*"; **Addition** – "*a favor de umas seis meses*" is not present in the source; **Omission** – "*wa mmalela wa ikhwaha sooxinttiwa thanu sottharelana*" is omitted

Example 4	
Source	*Siisaale mweeri okumihenryaaya exookhi ni etthaya, Eviiya Laateya ennikumiherya tho nave ikuro ni makhuru a itheneeri sa opuro.*
Reference	*Assim como a Lua exerce uma atração sobre a Terra, causando as marés, a Via Láctea exerce uma força sobre a galáxia de Sagitário.*
Translation	*No Website da Lua cheia, a Vila Lava vale a pena ler:*
Errors	**Mistranslation** – "*Siisaale*" ≠ "*No Website da*", "*okumihenryaaya exookhi ni etthaya*" ≠ "*cheia*", and "*Eviiya Laateya*" ≠ "*a Vila Lava*"; **Addition** – "*vale a pena ler:*" is not present in the source; **Omission** – "*ennikumiherya tho nave ikuro ni makhuru a itheneeri sa opuro.*" is omitted

In both translation directions, the models exhibit systematic errors that compromise translation faithfulness. Mistranslation, omissions, additions, untranslations. Addressing these challenges will typically require expanding the coverage of training data, both in quantity and in domain diversity, and improving the alignment between Emakhuwa and Portuguese. However, given the resource constraints typical in low-resource settings, leveraging pre-trained multilingual models and adopting techniques tailored for low-resource machine translation (e.g., transfer learning, back-translation, or lexicon-guided training) offers more time- and cost-efficient, yet effective pathways to enhance translation quality and robustness.

5.3 Summary

This chapter presented a benchmark evaluation of NMT systems for the Portuguese–Emakhuwa language pair. A baseline Transformer model was implemented using the OpenNMT toolkit, adopting a standard architecture. We preprocessed the input using BPE with a shared vocabulary of 40,000 tokens. Models were trained in both translation directions, *pt-vmw* and *vmw–pt*, using three different random seeds, with convergence typically occurring between 7,000 and 8,000 training steps. Quantitative evaluation using BLEU and ChrF metrics revealed generally low translation quality, reflecting the inherent challenges of this low-resource language pair. Performance was consistently higher on the in-domain *Multi-source* test set compared to the out-of-domain *FLORES+* benchmark, highlighting difficulties in generalization. Notably, the *vmw–pt* model outperformed the *pt–vmw* model across all test sets, achieving an average BLEU score of 10.40 on the *Multi-source* set, versus 7.14 for the *pt–vmw* direction. A detailed qualitative error analysis, grounded in the MQM framework, was carried out using outputs from the *FLORES+ devtest* set. Across both translation directions, the error is due to the MQM categories, including mistranslation, omission, addition, and untranslated content.

In summary, these benchmark results and error analyses underscore the significant challenges involved in building effective NMT systems for Portuguese and Emakhuwa. The low scores and recurrent error types point to both data scarcity and linguistic divergence as key barriers. These findings reinforce the need for more extensive and diverse training data, as well as the adoption of more advanced techniques tailored to low-resource settings. Both these topics will be explored in the following chapter.

The error analysis conducted reinforces the quantitative findings and highlights specific linguistic challenges in Portuguese–Emakhuwa machine translation. In the next chapter, we consider improving the quality of the translation by exploring proven strategies applied to low-resource scenarios. Moreover, incorporating human-in-the-loop evaluation and targeted data augmentation could help mitigate the translation issues.

Chapter 6

Transfer Learning

Transfer learning is a widely adopted approach that has demonstrated effectiveness across various NLP tasks, particularly in scenarios where resources are limited. As outlined in Chapter 2, it is a popular strategy adopted to mitigate the challenges posed by data scarcity. By leveraging the shared representation space of pretrained models, transfer learning enables the adaptation of knowledge from high-resource settings to low-resource ones. Cross-lingual transfer learning is a form of transfer learning where insights learned from one language or domain are transferred to another, thereby enhancing performance even when task-specific data is limited (Ruder, 2019). In the previous chapter, we trained from scratch a transformer-based model for MT in Portuguese-Emakhuwa language pairs, but the results produced poor translations, both in terms of adequacy and fluency. Thus, in this chapter, we explore whether transfer learning can also be applied effectively to our task and assess the extent to which it can improve upon the results presented in the previous chapter.

6.1 Related Work

Pre-training and fine-tuning have emerged as state-of-the-art strategies in NLP, consistently demonstrating effectiveness across a wide range of tasks. The **pre-training** phase involves training a large, general-purpose model on massive amounts of unlabeled text data. This phase allows the model to learn broad linguistic patterns, world knowledge, and contextual representations. In contrast, the **fine-tuning** phase adapts this generalist model to a specific downstream task (e.g., sentiment analysis, question answering, or translation) using a smaller, task-specific dataset (Ruder, 2019). Fine-tuning is often much less resource-intensive, when compared to pre-training, both in terms of data and computational power.

The pre-training data is typically diverse in both domain and language. Many modern models are trained on multilingual corpora, allowing them to generalize across multiple languages. This multilingual exposure is particularly beneficial for low-resource languages, which may lack sufficient data for training a monolingual model from scratch. One key factor enabling this is shared tokenization; most multilingual models use a common vocabulary across languages. As a result,

low-resource languages can leverage the representational strengths of high-resource ones through shared subword units and embeddings (Kocmi and Bojar, 2018).

In recent years, the NLP field has seen a surge in the release of pretrained models, sometimes at the pace of several per month. While these models vary in architecture, size, training objective, and language coverage, they can generally be categorized roughly into encoder-only models, encoder-decoder models, and decoder-only models, each serving different purposes and task requirements.

6.1.1 MT as a Downstream Task

Encoder-decoder and decoder-only pretrained models are particularly well-suited for MT downstream tasks, as both architectures are designed for text generation.

Encoder-decoder models can be further divided into general-purpose text-to-text multilingual models – such as mT5 (Xue et al., 2021), ByT5 (Xue et al., 2022), Aya-101 (Üstün et al., 2024), and MT0 (Muennighoff et al., 2023) – and translation-optimized multilingual models – such as mBART (Tang et al., 2020), OPUS (Tiedemann et al., 2023), M2M-100 (Fan et al., 2021), NLLB (NLLBTeam et al., 2024), and MADLAD (Kudugunta et al., 2023). The key difference between them is that translation-optimized models are specifically designed for MT and are typically trained using explicit language pair information (e.g., source and target language tags). Additionally, translation-optimized models are capable of many-to-many translation, enabling a single model to handle multiple translation directions.

Decoder-only models, while primarily designed for instruction following generation, have more recently been explored for translation tasks (Kocmi et al., 2023). Although they are still relatively new in this context, early results suggest that they are becoming competitive in comparison to encoder-decoder models, especially in high-resource settings. For instance, For example, *Claude-3.5-Sonnet* ranked first among 105 systems in the WMT 2024 General Machine Translation Task (Kocmi et al., 2024). In the subsequent WMT 2025 edition, *Gemini 2.5 Pro* achieved the highest score, followed closely by OpenAI's *GPT-4.1o*.

In addition, decoder-only models specifically optimized for translation are beginning to emerge as a growing area of research. The instruction-following paradigm applied to decoder-only architectures shows promise in addressing several open challenges in machine translation, such as handling non-literal or domain-specific translations and leveraging external information sources (e.g., online dictionaries or domain-specific databases) to resolve ambiguities (Ataman et al., 2025). Notable examples include ALMA (Xu et al., 2024) and Tower70B (Rei et al., 2024), which are open-source models trained with multilingual translation objectives.

6.1.2 Pre-trained Models for MT and African Languages

The effectiveness of pre-trained models for machine translation is strongly influenced by the availability and quality of the training data used during the pre-training phase. Unfortunately, African

languages remain severely underrepresented in most large-scale multilingual models. This underrepresentation is mainly due to data scarcity, limited digital resources, and a lack of high-quality corpora, all of which hinder the performance of these models in African languages.

Table 6.1 provides an overview of the language coverage of several popular pre-trained models. Most of these models cover only a small subset of African languages. One notable exception is Cheetah (Adebara et al., 2024), a relatively recent model that significantly expands language coverage to include 517 African languages. Its development was motivated by the need for a greater representation and support of low-resource African languages in large language models.

In addition to Cheetah, there have been other multilingual Afro-centric efforts aimed at improving language representation. These efforts include pre-training models from scratch, such as InkubaLM (Tonja et al., 2024), as well as adapting existing multilingual models through fine-tuning, for instance, Tucan (Elmadany et al., 2024b), AfriVeTa (Oladipo et al., 2023), Lugha-Llama (Buzaaba et al., 2025), AfriInstruct (Uemura et al., 2024), AfrimT5 and Afribty5 (Adelani et al., 2022a).

In Table 6.1, only the OPUS model includes explicit pre-training support for Emakhuwa. However, we hypothesize that this support is domain-specific and potentially limited, given that its training data consists predominantly of Bible translations (Tiedemann et al., 2023). While OPUS offers a potential starting point, a more comprehensive study is necessary to determine the most effective pre-trained model for our language pair. Therefore, we will conduct a comparison to assess the impact of transfer learning. This study will involve fine-tuning several prominent multilingual models on the Portuguese-Emakhuwa translation task and evaluating their performance against a baseline model (see Section 5.1).

6.2 Research Objectives

Our experiment is structured around the following objectives:

- To investigate the impact of transfer learning from large multilingual language models on improving machine translation quality for the low-resource Portuguese–Emakhuwa language pair.
- To compare the performance and error characteristics of pre-trained encoder-decoder and decoder-only architectures when applied to Portuguese–Emakhuwa machine translation.
- To evaluate whether fine-tuning African-centric pre-trained language models produces significantly better translation outcomes for the Portuguese–Emakhuwa language pair than general multilingual models or large language models (LLMs).

6.3 Methodology

We perform supervised fine-tuning on models listed in Table 6.2. Below, we describe the training procedures adopted for each model category.

Table 6.1: Overview of pretrained language models and their coverage of African languages. Models optimized for machine translation are indicated with *.

LM	Type	Lang/Total	Backbone	African Languages
Foundation Multilingual Models (Pre-trained from scratch)				
mBART*	Enc-Dec	3/50	-	afr, swh, yor.
mT5	Enc-Dec	12/101	-	afr, amh, nya, hau, ibo, mlg, sna, som, swh, xho, yor, and zul
OPUS*	Enc-Dec	84/290+	-	bas, bem, bnt, bss, cce, cjk, cwe, dig, dug, gog, gwr, hay, heh, her, jmc, kam, kdc, kdn, kik, kin, kki, kkj, kmb, kng, kon, ksb, kua, ldi, lem, lin, lon, lsm, lua, lug, luy, mcp, myx, nbl, nde, ndo, nim, nnb, nso, nuj, nya, nyf, nyn, nyo, nyy, old, ozm, pkb, rim, run, seh, sna, sot, ssw, suk, swa, swc, swh, sxb, thk, tlj, toh, toi, tsn, tso, tum, umb, ven, **vmw**, vun, wmw, xho, xog, zul
MADLAD-400*	Enc-Dec	34/400	-	af, am, arz, ary, bem, bm, din, ee, ff, gaa, ha, hz, ig, kg, kj, kr, ln, mg, mos, nr, nso, nus, ny, rn, rw, sn, so, ss, sw, ti, tn, ts, tzm, ve, wo, xh, yo, zu
M2M-100*	Enc-Dec	16/100	-	amh, far, ff, hau, ibo, lg, ln, nso, som, ss, sw, tn, wo, xh, your, zu
NLLB*	Enc-Dec	56/202	-	acq, aeb, afr, aka, amh, ary, arz, bam, bem, cjk, dik, dyu, ewe, fon, fuv, hau, ibo, kab, kam, kbp, kea, kik, kin, kmb, knc, lin, lua, lug, luo, mos, nso, nus, nya, orm, plt, run, sag, sna, som, sot, ssw, swh, taq, tir, tsn, tso, tum, twi, tzm, umb, wol, xho, yor, zul
Cheetah	Enc-Dec	517/527	-	517 African languages, but vmw not supported
InkubaLM	Dec-only	5/7	-	hau, sw, xho, your, zu
Multilingual Models Fine-Tuned with African Languages				
MT0	Enc-Dec	14/101	mT5	afr, amh, hau, ibo, lin, mlg, nyj, orm, sot, sna, som, swh, xho, your, and zul
AfriVeTa	Enc-Dec	16/20	mT5	afr, amh, arz, eng, fra, hau, ibo, kin, mlg, nya, orm, por, sna, som, sot, swa, tir, xho, yor, zul
AfriMT5	Enc-Dec	17/20	mT5	bam, bbj, ewe, fon, hau, ibo, lug, luo, pcm, mos, swa, tsn, twi, wol, yor, zul
AfriByT5	Enc-Dec	17/20	ByT5	bam, bbj, ewe, fon, hau, ibo, lug, luo, pcm, mos, swa, tsn, twi, wol, yor, zul
Aya-101	Enc-Dec	15/101	mT5	afr, amh, hau, ibo, mlg, mlt, nso, nya, sna, som, sot, swa, xho, yor, zul
Lugha-Llama	Dec-only	16/20	Llama	afr, amh, nya, hau, ibo, mlg, sot, sna, som, swa, xho, yor, zul, orm, kin, tir
AfroLlama-v1	Dec-only	5/6	Llama	Swahili, Xhosa, Zulu, Yoruba, Hausa
AfriInstruct	Dec-only	16/20	Llama	afr, amh, nya, hau, ibo, mlg, sot, sna, som, swa, xho, yor, zul, orm, kin, tir
Tucan*	Enc-Dec	47/49	Cheetah	aar, ach, afr, aka, amh, amh, ara, bam, bam, bas, bem, bfe, btg, ewe, fan, fon, gez, hau, ibo, kau, kbp, kin, kon, lgg, lin, lug, mlg, nnb, nya, nyn, orm, pom, som, sot, nso, swa, swe, tiv, tin, tir, tso, twi, wal, wol, xho, yor, zul

6.3.1 Training

Sequence-to-Sequence *text-to-text* multilingual pretrained models We perform full supervised fine-tuning (SFT) on *text-to-text* multilingual pretrained models using the Hugging Face Transformers library[1]. As part of the preprocessing step, we prepend a task-specific prefix to each input sentence to indicate the desired translation direction (e.g., `translate Emakhuwa to Portuguese:`). This helps guide the model during both training and inference.

[1] https://github.com/huggingface/transformers/tree/main/examples/pytorch/translation

Multilingual MT encoder-decoder pre-trained models This category includes two main groups of models:

- **Models requiring explicit source and target language:** This group includes mBART50, M2M-100, and NLLB. These models require both source and target language codes to be specified during training and inference. For instance, mBART50 uses `pt_XX` for Portuguese, M2M-100 uses `pt`, and NLLB uses `por_Latn`. However, none of these models supports Emakhuwa natively. To address this limitation, we extended the tokenizer of the selected multilingual model by registering Emakhuwa with a custom language code. For example, `vmw_Latn`, `vmw_XX`, and `vmw` for NLLB, M2M, and mBART50 respectively. The tokenizer extension involved updating the tokenizer's mappings, followed by adding the new token and resizing the model's embedding layer to incorporate the added vocabulary.
- **Target language Prefixed models:** These models only require the target language code to be specified as a prefix. This includes Tucan and OPUS (MarianMT-based model). Here, during preprocessing, we prepend the corresponding target language token (e.g., »`vmw`« or »`pt`«) to each source sentence to indicate the desired output language.

Multilingual Decoder-only Pre-trained Models For the smallest model in this category, the InkubaLM model, we performed full supervised fine-tuning using the Hugging Face Transformers library. The remaining models in this group are substantially larger, ranging from 3B to 9B parameters. Due to computational constraints, we used parameter-efficient fine-tuning techniques instead of full fine-tuning. Specifically, we applied LoRA (Hu et al., 2022), which reduces trainable parameters drastically. For Qwen2.5, Llama 3.1, Llama 3.2, Phi3.5-mini, Gemma-2, and Gemma-3, we used the Unsloth framework (Daniel Han and team, 2023), which enables us to significantly reduce VRAM requirements. For the TowerInstruct-v0.1 model, which is not supported by Unsloth, we used the axolotl tool[2] with LoRA.

6.3.2 Evaluation

Automatic Evaluations We extend our evaluation beyond traditional metrics such as BLEU and CHRF by also considering AfriCOMET (Wang et al., 2024a), a neural-based evaluation metric specifically tuned for African languages. AfriCOMET builds on the COMET framework (Rei et al., 2020) and currently supports 76 African languages. Although Emakhuwa is not among the supported languages, AfriCOMET has shown strong correlation with human judgments even for languages it was not explicitly trained on (Tapo et al., 2025). This makes it a promising candidate for assessing translation quality in the broader African multilingual context.

We compared our results with a baseline model using CHRF as the primary reference point. CHRF was chosen based on the findings of Li et al. (2025c), who identified it as the metric that best aligns with human evaluation in the Portuguese-Emakhuwa language pair.

[2] `https://axolotl.ai/`

Table 6.2: Pre-trained Models used in our experiments. Prefixes and prompts followed the standard format typically used with each model. M2M-100, NLLB, and mBART models do not require a prefix; instead, the source and target languages are specified directly through the tokenizer.

	LM	Size	Prefix/Prompt	Hparams.
Encoder-Decoder	*Full Supervised Fine-Tuning*			
	mT5/AfriMT5	580M	Figure 6.1, Template 1	Table A.2
	MT0	582M	Figure 6.1, Template 1	Table A.2
	AfriTeva	428M	Figure 6.1, Template 1	Table A.2
	byT5/AfriByT5	580M	Figure 6.1, Template 1	Table A.2
	Cheetah	582M	Figure 6.1, Template 1	Table A.2
	Tucan	582M	Figure 6.1, Template 2	Table A.2
	mBART50	610M	-	Table A.5
	M2M-100	418M	-	Table A.4
	NLLB	600M	-	Table A.3
	OPUS	240M	Figure 6.1, Template 3	Table A.6
Decoder-only	*Full Supervised Fine-Tuning*			
	InkubaLM	0.4B	Figure 6.1, Template 4	Table A.7
	Supervised Fine-Tuning w/ LORA			
	Qwen2.5	7B	Figure 6.1, Template 4	Table A.8
	Llama 3.1	8B	Figure 6.1, Template 4	Table A.8
	Llama 3.2	3B	Figure 6.1, Template 4	Table A.8
	Phi3.5-mini	3.8B	Figure 6.1, Template 4	Table A.8
	Gemma-2	9B	Figure 6.1, Template 4	Table A.8
	Gemma-3	4B	Figure 6.1, Template 4	Table A.8
	TowerInstruct-v0.1	7B	Figure 6.1, Template 5	Table A.9

6.4 Results and Discussion

The detailed results are presented in Tables 6.3, 6.4, and 6.5, which compare the performance of each model in the three evaluation sets. The key findings and insights are discussed below.

Transfer learning: Our experiments demonstrate that fine-tuning multilingual pre-trained language models generally leads to substantial improvements in translation quality across all evaluation metrics (BLEU, CHRF, and AfriCOMET) when compared to baseline systems. However, the degree of improvement is not uniform across models; it depends heavily on both the underlying architecture and the pre-trained model.

For instance, models built on the mT5 backbone, namely Afriteva, mT5, AfriMT5, and MT0, consistently struggled to surpass the baseline on the MULTI-SOURCE test set for the Portuguese-to-Emakhuwa translation direction. This could point to limitations in the mT5 architecture when

Template 1

translate `{source_language}` to `{target_language}`:

Template 2

`{target_language_code}: {source_sentence}`

Template 3

`>>{target_language_code}<< {source_sentence}`

Template 4

Translate the following sentence from `{source_language}` into `{target_language}`:
`{source_sentence}`

Template 5

Translate the following text from `{source_language}` into `{target_language}`.
`{source_language}: {source_sentence}`
`{target_language}:`

Figure 6.1: Prefix and Prompt templates for MT

dealing with languages characterized by spelling variation, such as Emakhuwa. By contrast, the byT5 model, a byte-level variant of T5, performed competitively with the top-tier models (i.e., multilingual MT and decoder-only LLMs) in our evaluation. This is a noteworthy outcome, given that byT5 was designed to be more robust to noisy input and better suited for tasks involving non-standard or highly variable orthographies. Its strong performance suggests that byte-level processing may offer a significant advantage for under-resourced languages with free or inconsistent spelling systems.

A particularly surprising case is the InkubaLM model. Although specifically designed with African languages in mind, it consistently underperformed relative to the baseline in all evaluation metrics and test sets. This poor performance points to potential shortcomings in the model's multilingual coverage and capacity to generalize. One possible explanation is that InkubaLM's pre-training data may not have adequately represented the language pair, resulting in limited transferability to the supervised fine-tuning task.

Encoder-decoder vs. Decoder-only: Both encoder-decoder and decoder-only architectures yield promising results with our language pair when compared to the baseline. Among the **encoder-decoder** models, fine-tuning multilingual MT models generally led to improved performance. In particular, NLLB and M2M-100 consistently achieved competitive results across all evaluation sets. These two models were especially effective in the more challenging translation direction, i.e., Portuguese to Emakhuwa, by outperforming other models. While mBART50 also consistently outperformed the baseline, its performance lagged behind NLLB and M2M-100. We hypothesize that this is due to limitations in the pre-training data coverage, which may have affected its ability to generalize across low-resource language pairs, thus constraining its multilingual transferability.

Table 6.3: Evaluation results on the **DEV** set. CHRF underlined if significantly better than baseline ($p < 0.05$). COMET refers to the AfriCOMET metric.

	pt→vmw			**vmw→pt**		
Model	BLEU	CHRF	AfriCOMET	BLEU	CHRF	AfriCOMET
OPUS	4.40	37.39 ↑8.62	0.50	16.82	43.49 ↑15.88	0.58
Afriteva	4.83	29.47 ↑0.70	0.38	12.72	34.24 ↑6.63	0.47
Afri-byT5	8.89	38.73 ↑9.96	0.49	22.63	47.63 ↑20.02	0.61
byT5	9.51	39.54 ↑10.77	0.50	22.36	46.95 ↑19.34	0.60
mT5	5.41	29.38 ↑0.61	0.41	13.51	34.83 ↑7.22	0.50
Afri-mT5	5.52	30.45 ↑1.68	0.42	14.37	36.36 ↑8.75	0.51
Cheetah	6.47	32.38 ↑3.61	0.45	13.17	34.89 ↑7.28	0.50
MT0	3.35	27.66 ↓1.11	0.31	12.52	37.02 ↑9.41	0.48
Toucan	6.88	34.28 ↑5.51	0.47	14.46	36.66 ↑9.05	0.51
M2M-100	9.07	**40.09** ↑11.32	0.50	20.24	44.75 ↑17.14	0.59
mBart	6.72	34.80 ↑6.03	0.47	13.05	35.21 ↑7.60	0.50
NLLB	7.75	40.06 ↑11.29	0.50	22.07	46.75 ↑19.14	0.61
InkubaLM	0.93	23.56 ↓5.21	0.28	1.14	19.36 ↓8.25	0.24
Llama 3.2	6.54	35.36 ↑6.59	0.47	20.31	44.71 ↑17.10	0.57
Qwen2.5	5.38	32.94 ↑4.17	0.47	14.52	38.44 ↑10.83	0.54
Phi3.5-mini	5.44	31.52 ↑2.75	0.46	16.62	39.22 ↑11.61	0.53
Gemma-3	7.19	35.57 ↑6.80	0.49	21.20	43.99 ↑16.38	0.57
Tower	2.89	27.60 ↓1.17	0.32	7.31	34.25 ↑6.64	0.44
Qwen2.5	5.96	33.87 ↑5.10	0.48	17.54	42.38 ↑14.77	0.58
Llama 3.1	8.25	37.79 ↑9.02	0.50	24.40	47.92 ↑20.31	0.60
Gemma-2	8.09	37.18 ↑8.41	0.49	25.43	**48.16** ↑20.55	0.62

Decoder-only models also delivered competitive results, with LLaMA and Gemma models standing out in this category. Notably, Gemma-2 (8B) emerged as the top-performing system for the Emakhuwa-to-Portuguese translation direction, achieving substantial improvements of +20.55 (DEV), +13.50 (DEVTEST), and +14.82 (MULTI-SOURCE) in CHRF scores. These findings suggest that LLMs are particularly well-suited for translating from low-resource to high-resource languages, likely due to their extensive pretraining on massive amounts of text in high-resource languages. This pretraining exposure allows the models to develop strong representations and fluent generation capabilities when translating from a low-resource language into a more represented one.

Tables 6.6 and 6.7 present some qualitative improvements in translation quality achieved by the fine-tuned models. For error analysis, we revisit the example discussed in Section 5.1, where baseline systems are shown to have failed to provide appropriate translations. To gain deeper insight into the nature of the improvements, we analyze outputs from the top-performing models within each architecture category: **text-to-text** (byT5, AfriByT5), **multilingual MT models** (M2M-100 and NLLB), and **decoder-only LLMs** (LLaMA and Gemma).

In both translation directions, substantial improvements were observed. While lexical choices sometimes deviated from the reference, the translated meanings often approximated the intended message. For instance, in Portuguese to Emakhuwa, NLLB mistranslated the word "*tournê*" ("tour") as "*championship*," while Gemma rendered the phrase "*os shows restantes da sua tournê*"

Table 6.4: Evaluation results on the **DEVTEST** set. CHRF underlined if significantly better than baseline ($p < 0.05$). COMET refers to the AfriCOMET metric.

	pt→vmw			**vmw→pt**		
Model	BLEU	CHRF	AfriCOMET	BLEU	CHRF	AfriCOMET
OPUS	3.79	34.63 ↑6.82	0.48	11.50	36.53 ↑10.64	0.53
Afriteva	4.30	28.63 ↑0.82	0.36	8.50	31.01 ↑5.12	0.43
Afri-byT5	6.47	34.42 ↑6.61	0.46	14.10	37.92 ↑12.03	0.54
byT5	6.72	34.59 ↑6.78	0.46	13.67	37.37 ↑11.48	0.54
mT5	4.78	27.84 ↑0.03	0.40	8.53	29.98 ↑4.09	0.45
Afri-mT5	4.71	28.65 ↑0.84	0.40	9.05	31.47 ↑5.58	0.47
Cheetah	5.38	30.27 ↑2.46	0.42	8.92	31.49 ↑5.60	0.47
MT0	3.07	26.42 ↓1.39	0.30	7.40	31.11 ↑5.22	0.42
Toucan	5.75	32.02 ↑4.21	0.44	9.64	33.07 ↑7.18	0.48
M2M-100	7.01	36.24 ↑8.43	0.47	13.19	37.16 ↑11.27	0.53
mBart	5.66	32.95 ↑5.14	0.45	8.95	31.50 ↑5.61	0.46
NLLB	6.68	**37.21** ↑9.40	0.48	13.88	38.61 ↑12.72	0.55
InkubaLM	0.88	24.23 ↓3.58	0.27	0.90	19.08 ↓6.81	0.24
Llama 3.2	5.17	32.54 ↑4.73	0.45	12.37	37.52 ↑11.63	0.51
Qwen2.5	4.72	31.36 ↑3.55	0.46	8.96	33.17 ↑7.28	0.49
Phi3.5-mini	4.65	29.91 ↑2.10	0.44	10.32	33.96 ↑8.07	0.48
Gemma-3	5.63	32.21 ↑4.40	0.45	13.32	36.77 ↑10.88	0.51
Tower	2.64	27.10 ↓0.71	0.32	4.99	30.35 ↑4.46	0.40
Qwen2.5	5.34	32.28 ↑4.47	0.46	11.38	35.95 ↑10.06	0.52
Llama 3.1	6.65	34.71 ↑6.90	0.47	14.76	39.70 ↑13.81	0.54
Gemma-2	7.14	34.51 ↑6.70	0.47	15.39	**39.39** ↑13.50	0.55

as something similar to “they stopped singing forever,” reflecting a more severe loss of meaning.

Conversely, in the Emakhuwa-to-Portuguese direction, although all models showed improved adequacy, the encoder-decoder models struggled with adequacy, particularly in rendering the Emakhuwa equivalent of "show." Interestingly, decoder-only models like LLaMA and Gemma successfully produced fluent and accurate Portuguese translations in these cases, likely benefiting from their strong language generation capabilities in high-resource languages.

African-centric models: To address the third research objective, we conducted a pairwise bootstrap resampling analysis comparing models fine-tuned from Afrocentric multilingual language models with those fine-tuned from general-purpose multilingual models. In this context, Afrocentric models refer to multilingual models with a higher proportional representation of African languages during pre-training or fine-tuning. These include models such as Cheetah, Toucan, InkubaLM, AfriByT5, AfriMT5, and AfriTeva. For evaluation, we merged all test sets into a single evaluation set, which therefore comprises 3,001 sentences. Pairwise bootstrap resampling with 1,000 samples was performed to estimate the statistical significance of performance differences, measured using the CHRF score.

Figure 6.2 and Figure 6.3 present a detailed analysis of the confusion matrices generated during evaluation. When compared with general-purpose multilingual encoder-decoder models such

Table 6.5: Evaluation results on the **MULTI-SOURCE** set. CHRF underlined if significantly better than baseline ($p < 0.05$). COMET refers to the AfriCOMET metric.

	pt→vmw			**vmw→pt**		
Model	BLEU	CHRF	AfriCOMET	BLEU	CHRF	AfriCOMET
OPUS	5.03	41.19 ↑5.92	0.52	17.26	42.98 ↑10.38	0.60
Afriteva	6.58	34.69 ↓0.58	0.44	15.29	37.13 ↑4.53	0.52
Afri-byT5	9.64	41.30 ↑6.03	0.51	19.52	43.14 ↑10.54	0.60
byT5	10.31	41.90 ↑6.63	0.52	18.90	42.27 ↑9.67	0.59
mT5	7.07	34.28 ↓0.99	0.46	14.25	35.54 ↑2.94	0.53
Afri-mT5	7.04	34.87 ↓0.40	0.46	14.83	37.09 ↑4.49	0.54
Cheetah	7.30	36.11 ↑0.84	0.47	14.55	36.82 ↑4.22	0.54
MT0	4.76	32.15 ↓3.12	0.37	12.27	35.81 ↑3.21	0.51
Toucan	8.35	38.58 ↑3.31	0.49	16.14	38.88 ↑6.28	0.56
M2M-100	10.97	43.89 ↑8.62	0.52	20.86	44.43 ↑11.83	0.61
mBart	8.56	39.86 ↑4.59	0.51	14.78	36.70 ↑4.10	0.54
NLLB	11.87	**46.09** ↑10.82	0.53	23.24	46.82 ↑14.22	0.62
InkubaLM	1.52	27.48 ↓7.79	0.29	1.79	22.41 ↓10.19	0.25
Llama 3.2	8.30	39.66 ↑4.39	0.50	19.35	43.46 ↑10.86	0.59
Qwen2.5	7.08	37.76 ↑2.49	0.49	15.63	39.33 ↑6.73	0.57
Phi3.5-mini	7.14	37.36 ↑2.09	0.49	17.63	40.17 ↑7.57	0.56
Gemma-3	9.58	41.14 ↑5.87	0.52	21.08	44.05 ↑11.45	0.59
Tower	3.83	31.84 ↓3.43	0.37	8.09	35.68 ↑3.08	0.47
Qwen2.5	7.67	39.23 ↑3.96	0.50	18.47	42.24 ↑9.64	0.59
Llama 3.1	10.54	43.30 ↑8.03	0.53	23.58	46.63 ↑14.03	0.61
Gemma-2	11.60	43.34 ↑8.07	0.52	25.19	**47.42** ↑14.82	0.62

as mT5 and MT0, Afrocentric models tend to perform comparably well or slightly better. This suggests that, for this language pair, models explicitly fine-tuned or designed for African languages can match, and sometimes exceed, the capabilities of broad multilingual systems. In particular, byT5 consistently stands out among these, probably due to its byte-level tokenization approach. This feature proves advantageous for morphologically rich languages like Emakhuwa, as it captures subtle variations in word formation and inflection that are often missed by standard subword tokenization. However, this advantage decreases when Afrocentric models are compared to dedicated multilingual MT systems such as OPUS-MT, mBART50, M2M-100, and particularly NLLB. While OPUS-MT occasionally yields lower CHRF scores and mBART50 sometimes performs on par with Afrocentric models, M2M-100 and NLLB consistently and significantly outperform them. These results underscore the benefit of training on massive, multilingual corpora with broad language coverage, where exposure to diverse linguistic patterns enhances the model's ability to handle low-resource languages like Emakhuwa.

The most significant performance gaps emerge when Afrocentric models are evaluated against decoder-only LLMs such as Qwen2.5, Llama, and Gemma. These models, when adapted to translation tasks, exhibit a better performance, likely due to their larger scale and richer pre-training data. However, not all decoder-only models excel. The Tower model performs poorly on this language pair despite being fine-tuned on a diverse set of translation-related tasks. We believe that limited language coverage and focus on high-resource languages likely hinders its ability to

Table 6.6: Error Analysis Evaluation on the *pt-vmw* model. Note, only *byT5* is shown, as it outperformed *AfriByT5* in this translation direction.

Example 1		
	Source	*O Aerosmith cancelou os shows restantes da sua turnê.*
	Reference	*Aerosmith ohiyererya miteko soowooniherya saahanle wa ikwaha satthokihiwe.*
Translation	byT5	*Aerosmith ookhootta ishow sikina sa eturnee awe.*
	M2M-100	*Aerosmith ohiyererya ishow soohala sa eturneeti awe.*
	NLLB	*Nikhuru na Aerosmith nohiyererya mixinttano sotheene sa nikhuru nenla naasimwa.*
	Llama	*Aerosmith oohiya ishow sotheene sa mukwaha awe.*
	Gemma	*Aerosmith ohiyererya wiipa wa mahiku otheene.*
Errors	byT5	-
	M2M-100	-
	NLLB	**Mistranslation** – "*mixinttano*" ≠ "*turnê*";
	Llama	-
	Gemma	**Mistranslation** – "*wiipa wa mahiku otheene.*" ≠ "*os shows restantes da sua turnê*";

Table 6.7: Error Analysis Evaluation on the *vmw-pt* model. Note, only *AfriByT5* is shown, as it outperformed *byT5* in this translation direction.

Example 2		
	Source	*Aerosmith ohiyererya miteko soowooniherya saahanle wa ikwaha satthokihiwe.*
	Reference	*O Aerosmith cancelou os shows restantes da sua turnê*
Translation	Afri-byT5	*A Aerosmith abandonou as exposições que estavam preparadas.*
	M2M-100	*Aerosmith desistiu das demonstrações que ficaram para viagens organizadas.*
	NLLB	*A Aerosmith cancelou os trabalhos de apresentação que haviam ficado na agenda.*
	Llama	*Aerosmith cancelou os restantes shows da digressão*
	Gemma	*Os Aerosmith cancelaram as restantes datas da digressão.*
Errors	Afri-byT5	**Mistranslation** – "*A*" ≠ "*O*", "*abandonou as exposições*" ≠ "*cancelou os shows*";
	M2M-100	**Mistranslation** –"*demonstrações*" ≠ "*shows*";
	NLLB	**Mistranslation** – "*os trabalhos de apresentação*" ≠ "*shows*";
	Llama	-
	Gemma	-

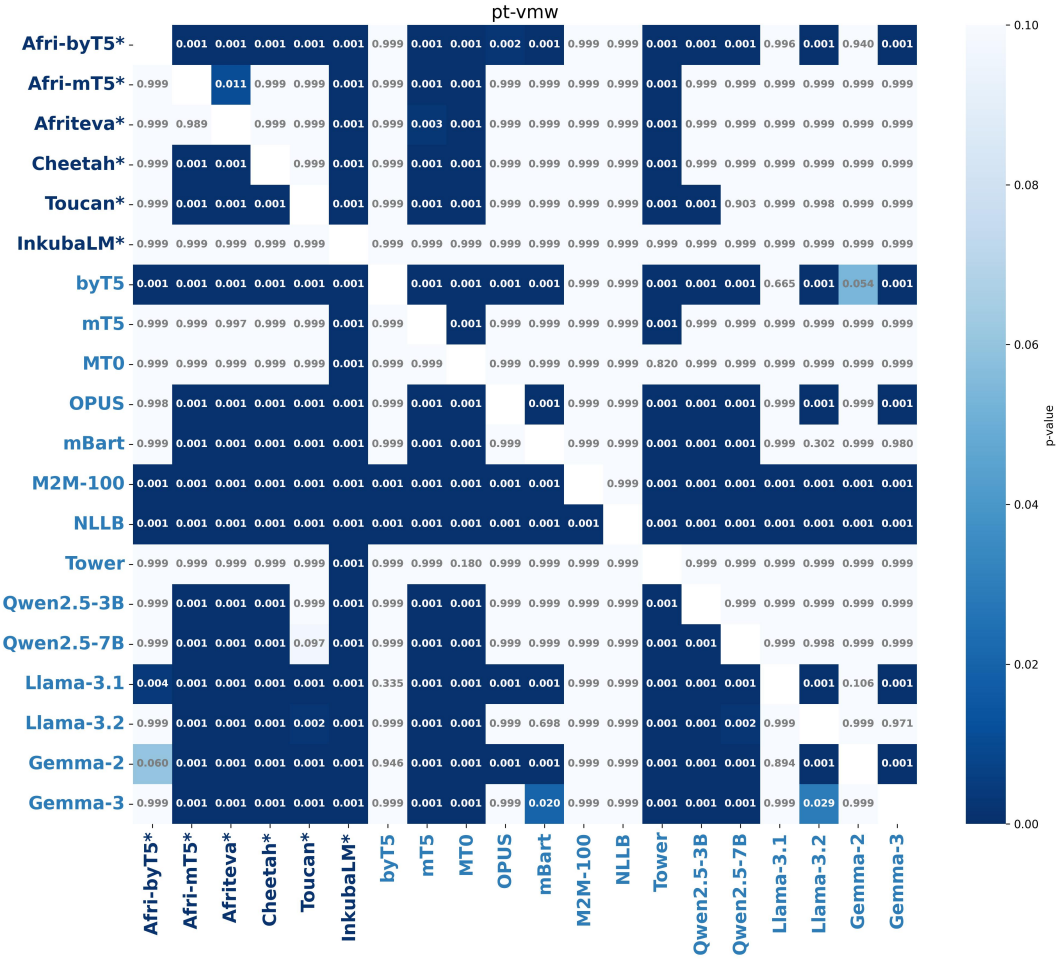


Figure 6.2: Pairwise bootstrap resampling significance test results. The heat map cells highlight where the performance of the system in the row is significantly higher and different than the system in the column ($p < 0.05$). The value inside each cell corresponds to the p-value. Afrocentric systems are indicated with *

generalize to our language pair.

6.5 Discussion

Our findings confirm that transfer learning, through fine-tuning, offers a substantial improvement over baseline systems, which restates its value in data-scarce scenarios. However, the extent of this improvement is highly dependent on the model's architecture and pre-training data. When comparing architectures, both encoder-decoder and decoder-only models proved effective. Translation-optimized encoder-decoder models, particularly NLLB and M2M-100, excelled at translating into the low-resource language (pt-vmw). Conversely, the strongest decoder-only models, Gemma-2 and Llama 3.1, achieved top performance when translating into the high-resource language (vmw-pt), capitalizing on their superior target-side language modeling capabilities. Regarding Afrocentric models, our analysis found limited evidence that they outperform general multilingual models,

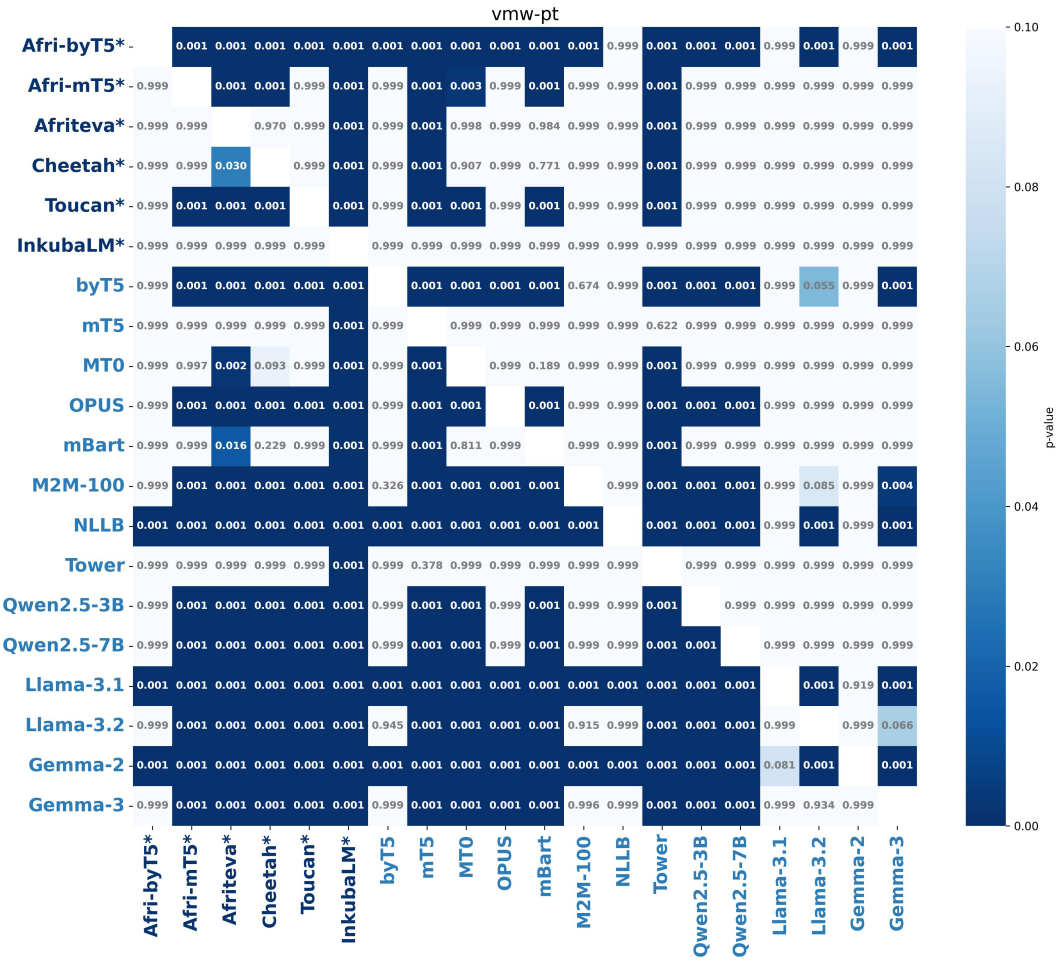


Figure 6.3: Pairwise bootstrap resampling significance test results. The heat map cells highlight where the performance of the system in the row is significantly higher and different than the system in the column ($p < 0.05$). The value inside each cell corresponds to the p-value. Afrocentric systems are indicated with *

at least for the Portuguese and Emakhuwa language pair. Although some Afrocentric models performed competitively, particularly against general-purpose text-to-text encoder-decoder models, they were consistently outperformed by larger, translation-specialized or decoder-only models. This suggests that while Afrocentric pretraining holds promise, current models have limited support for Lusophone African languages like Emakhuwa, reflecting a broader imbalance in language coverage that appears to favor Anglophone and Francophone African languages. Addressing this gap is essential for building more inclusive and effective multilingual African language models.

6.6 Summary

This chapter investigated the application of transfer learning to enhance machine translation for the low-resource Portuguese-Emakhuwa language pair. We systematically evaluated a diverse range of pretrained models, spanning general-purpose, translation-optimized, and Afrocentric variants of both encoder-decoder and decoder-only architectures, to address three central research questions.

Building on the results obtained, the next chapter will explore strategies to further push these performance gains, specifically by combining transfer learning with data augmentation techniques.

Chapter 7

Data Augmentation

In the previous chapter, we demonstrated that transfer learning significantly improves MT quality for the Portuguese–Emakhuwa language pair. However, we observed an asymmetry in translation performance; the Emakhuwa-to-Portuguese direction achieved comparatively stronger results, while the reverse, Portuguese-to-Emakhuwa, remained considerably more challenging.

Motivated by these findings, this chapter explores strategies to further improve translation quality. Specifically, we investigate the use of data augmentation, a widely adopted technique in low-resource MT settings aimed at increasing both the quantity and diversity of training data. We focus on backtranslation using monolingual corpora to generate synthetic parallel sentences. Experiments are conducted to assess both the benefits and potential trade-offs of incorporating synthetic data into the training pipeline.

7.1 Related Work

Data augmentation has become a key technique in improving MT, particularly for low-resource language pairs. It involves generating additional synthetic data to expose models to a wider range of training examples (Pellicer et al., 2023). In the context of MT, commonly used data augmentation strategies include backtranslation, paraphrasing, and synthetic corpus generation using pretrained models (Ranathunga et al., 2023). These methods aim to enrich the training data beyond the original parallel corpora, thereby improving translation robustness and generalization.

Among these techniques, backtranslation has emerged as one of the most effective and widely used approaches for low-resource MT (Sennrich et al., 2016a; Edunov et al., 2018). The method involves translating monolingual target-language data back into the source language to create synthetic parallel sentence pairs. This allows the model to leverage the relatively abundant monolingual data available for many target languages.

Several studies have demonstrated the effectiveness of backtranslation for low-resource languages, including African languages. For example, McNamee and Duh (2023) explored its application in 60 languages, including 13 African ones: Swahili, Afrikaans, Xhosa, Zulu, Shona,

Hausa, Nyanja, Amharic, Somali, Lingala, Oromo, Yoruba, and Igbo. Their results showed consistent improvements when translating into English, with BLEU score gains ranging from +1.0 to +9.7 on the FLORES-101 benchmark. Other works, such as Akera et al. (2022), Emezue and Dossou (2021b), NLLB Team et al. (2022), Kimera et al. (2025), and Ali et al. (2024c), have also reported translation improvements for African languages using backtranslation and similar augmentation techniques. These studies underscore the potential of leveraging synthetic data to boost MT performance in data-scarce scenarios.

Despite its promise, data augmentation is not without limitations. The quality of synthetic data plays a critical role in its effectiveness. When generated using low-quality or domain-mismatched models, synthetic sentences can introduce noise and inconsistencies (NLLB Team et al., 2022), potentially degrading the overall performance of the model, rather than improving it. Therefore, careful selection of monolingual data and translation systems is essential when applying backtranslation in low-resource settings.

7.2 Research Question

We aim to empirically evaluate the impact of backtranslation as a data augmentation technique in the Portuguese–Emakhuwa MT setting. We frame our experiment to answer the following research question:

- Does augmenting data with back-translations from monolingual Emakhuwa sources lead to measurable improvements in MT performance?

Essentially, our goal is to use the best-performing model to translate Emakhuwa text into Portuguese. This process produces synthetic source sentences (see Figure 7.1), which are then paired with the original Emakhuwa sentences as targets. Finally, we fine-tune new models and evaluate the extent to which this approach yields performance gains in both translation directions.



Figure 7.1: Synthetic source sentence generation via backtranslation as in Sennrich et al. (2015)

7.3 Methodology

To explore the effectiveness of data augmentation in Portuguese–Emakhuwa MT, we apply backtranslation to generate synthetic parallel data (see Figure 7.2). The resulting synthetic dataset is subsequently integrated into the training pipeline to develop and evaluate new models.

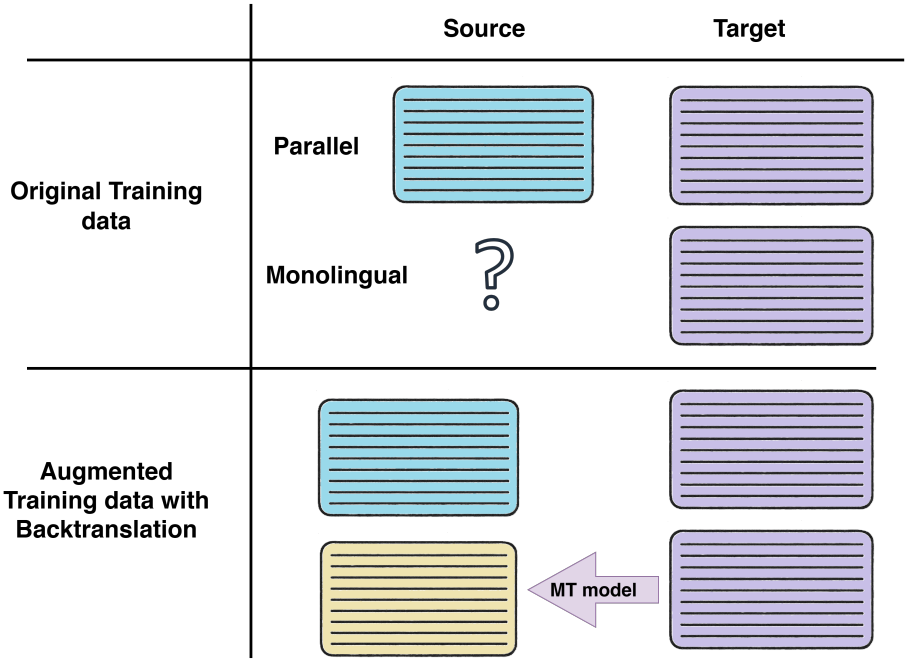


Figure 7.2: Augmenting parallel data with backtranslation

7.3.1 Monolingual Data

We use a set of 14,463 Emakhuwa monolingual sentences collected as explained in Chapter 4 (see Table 4.3) as the source for backtranslation, which corresponds to approximately 22% increase from our previous experiment training data (Table 4.4 in chapter 6).

7.3.2 Experimental Setup

Backtranslation System Backtranslation is performed in the Emakhuwa-to-Portuguese direction, as this direction showed stronger performance in our earlier experiments. For this task, we use the best-performing model from Chapter 6, Gemma-2 (9B), which consistently outperformed all other models across evaluation sets.

Data Augmentation Each Emakhuwa sentence is translated into Portuguese using the Gemma-2 (9B) model. These translated sentences are then paired with their original Emakhuwa counterparts to form synthetic parallel data. This synthetic dataset is combined with the original Portuguese–Emakhuwa parallel corpus (Table 4.4) to create an augmented training set.

Model Selection and Training To evaluate the impact of data augmentation across different model architectures, we retrain six models, the best two from each of the three main types evaluated in Chapter 6: (1) **Text-to-text models**: byT5 and AfriByT5; (2) **Multilingual MT models**: M2M-100 and NLLB; (3) **Decoder-only LLMs**: LLaMA and Gemma.

All models are trained on the augmented training data using the same hyperparameters and training settings described in Section 6.3, ensuring fair and consistent comparisons.

Evaluation Each model is evaluated in both translation directions, Portuguese-to-Emakhuwa and Emakhuwa-to-Portuguese, using the same test sets and metrics as in previous chapters. For comparison, we use as baselines the model results reported in Tables 6.3, 6.4, and 6.5.

7.4 Results

The detailed results are presented in Tables 7.1, 7.2, and 7.3. Across all three evaluation sets, we observe a consistent trend: backtranslation generally yields improvements in the Portuguese-to-Emakhuwa translation direction, except for the Gemma model, which was the model used for backtranslation. For instance, on the *DEV* set (Table 7.1), Afri-byT5 shows an increase of +2.98 in CHRF for pt→vmw, while NLLB improves by +2.26. Similar improvements are observed on the *DEVTEST* and *MULTI-SOURCE* sets. This suggests that the synthetic data generated through backtranslation from Emakhuwa monolingual sentences helps the models to better learn the intricacies of translating into Emakhuwa.

However, the impact of backtranslation on the Emakhuwa-to-Portuguese direction is less uniformly positive. While some text-to-text models show slight increases in CHRF and BLEU on certain datasets, others, particularly the decoder-only LLMs, exhibit decreases in performance. For example, Llama 3.1 shows a decrease of -0.92 CHRF on the DEVTEST set for vmw→pt. Moreover, Gemma-2, which served as the backtranslation engine, experienced a performance drop in both directions and across most evaluation sets. This suggests a form of self-contamination or overfitting, where exposure to its own output during training degrades generalization.

Interestingly, the improvements from backtranslation are more pronounced on the FLORES datasets (DEV and DEVTEST). We attribute this to the overlap in domain between the FLORES benchmark and the monolingual Emakhuwa data, which partially originates from the Wikipedia. This improvements highlights the importance of domain alignment in backtranslation: when the synthetic data is closer to the evaluation set in style and content, it becomes more effective. This finding suggests that incorporating domain-specific monolingual data can substantially enhance translation quality.

The varying impact across different model architectures is also noteworthy. Text-to-text and multilingual models have shown to be more responsive to synthetic data.

In general, performance gains for encoder-decoder models tend to be minimal or modest, indicating that while these architectures can benefit from additional synthetic data, their improvements are not substantial. Future work could explore the use of more diverse and domain-rich monolingual corpora to further increase the effectiveness of backtranslation.

Table 7.1: Evaluation results on the *DEV* set. CHRF underlined if significantly better than baseline (see Table 6.3) ($p < 0.05$).

	pt→vmw			**vmw→pt**		
Model	BLEU	CHRF	AfriCOMET	BLEU	CHRF	AfriCOMET
Afri-byT5	10.32	41.71 ↑2.98	0.51	23.15	47.74 ↑0.11	0.61
byT5	10.50	42.11 ↑2.57	0.51	22.95	47.41 ↑0.46	0.61
M2M-100	8.98	40.39 ↑0.30	0.50	20.02	44.33 ↓0.42	0.59
NLLB	9.34	42.32 ↑2.26	0.50	23.71	47.86 ↑1.11	0.62
Llama 3.1	7.85	37.91 ↑0.12	0.49	24.24	47.81 ↓0.11	0.60
Gemma-2	7.93	36.83 ↓0.35	0.48	23.97	47.36 ↓0.80	0.61

Table 7.2: Evaluation results on the *DEVTEST* set. CHRF underlined if significantly better than baseline (see Table 6.4) ($p < 0.05$).

	pt→vmw			**vmw→pt**		
Model	BLEU	CHRF	AfriCOMET	BLEU	CHRF	AfriCOMET
Afri-byT5	7.56	36.53 ↑2.11	0.47	14.67	38.50 ↑0.58	0.55
byT5	7.58	36.72 ↑2.13	0.48	14.44	38.21 ↑0.84	0.54
M2M-100	7.25	37.09 ↑0.85	0.48	13.88	37.28 ↑0.12	0.54
NLLB	7.41	38.57 ↑1.36	0.48	14.61	39.06 ↑0.45	0.55
Llama 3.1	6.50	34.86 ↑0.15	0.47	13.85	38.78 ↓0.92	0.54
Gemma-2	6.74	34.15 ↓0.36	0.46	15.03	38.91 ↓0.48	0.55

Table 7.3: Evaluation results on the **MULTI-SOURCE** set. CHRF underlined if significantly better than baseline (see Table 6.5) ($p < 0.05$).

	pt→vmw			**vmw→pt**		
Model	BLEU	CHRF	AfriCOMET	BLEU	CHRF	AfriCOMET
Afri-byT5	10.80	43.20 ↑1.90	0.52	20.38	43.69 ↑0.55	0.60
byT5	10.98	43.15 ↑1.25	0.52	20.07	43.17 ↑0.90	0.60
M2M-100	11.60	44.99 ↑1.10	0.53	21.76	45.16 ↑0.73	0.61
NLLB	12.32	46.29 ↑0.20	0.53	24.31	47.46 ↑0.64	0.63
Llama 3.1	10.69	43.67 ↑0.37	0.52	23.89	46.54 ↓0.09	0.61
Gemma-2	11.19	43.52 ↑0.18	0.52	24.61	47.15 ↓0.27	0.62

7.5 Summary

This chapter explored data augmentation, specifically backtranslation, as a strategy to improve MT for the Portuguese–Emakhuwa language pair. While previous results in Chapter 6 showed stronger performance for Emakhuwa-to-Portuguese translation, the reverse direction remained challenging.

To approach this problem, we empirically evaluated the application of backtranslation; thus, we backtranslated 14,463 monolingual Emakhuwa sentences using the Gemma-2 model to create synthetic Portuguese–Emakhuwa pairs. These are added to the original dataset and used to retrain six MT models across the following categories: text-to-text (byT5 and AfriByT5), multilingual MT (M2M-100 and NLLB), and decoder-only LLms (Gemma-2 and Llama 3.1).

Results show consistent improvements in the Portuguese-to-Emakhuwa direction, particularly for text-to-text and multilingual models. However, results in the reverse direction are mixed, and some models, especially Gemma, exhibit performance drops, likely due to overfitting on their own synthetic outputs. Nevertheless, these improvements remain modest, and the performance asymmetry persists even when back-translation is applied. This trend is especially evident in the BLEU metric. This points to the orthographic choices of the MT systems, which often diverge from the references, largely due to the inherent variability in Emakhuwa spelling. In the next chapter, we aim to address these issues by incorporating external resources, such as loanword glossaries and dictionaries.

Chapter 8

Leveraging External Information

In the previous chapter, we used monolingual data and backtranslation to expand the training corpus with synthetic examples. This approach yielded modest improvements; we also observed that it introduced noise into the data, which can negatively affect model performance.

In this chapter, we explore an alternative data augmentation strategy grounded in the linguistic features of Emakhuwa. This approach is motivated by the following empirical observations:

- **Borrowings from Portuguese:** Manual translation practices between Portuguese and Emakhuwa often rely heavily on lexical borrowing, where spelling and morphological adaptation vary depending on the translator. However, MT systems struggle to capture the subtle morphological nuances that human translators apply when integrating borrowed words into the target language. We hypothesize that explicitly modeling and managing these borrowings can help the system better align source and target representations.
- **Over-borrowing:** In some cases, MT models tend to overuse borrowed terms or even generate non-existent lexical forms, despite the availability of native equivalents. This poses a linguistic and cultural risk, as such artificial words are likely to be rejected by native speakers. Moreover, since modern MT systems are increasingly used to produce synthetic data, these invented forms could gradually contaminate future datasets with unnatural or nonstandard vocabulary.
- **Availability of Lexical Resources:** Existing bilingual dictionaries, glossaries, and other lexical resources can serve as valuable aids in translation. Incorporating them into the training process can help guide the model toward more accurate, consistent, and culturally appropriate outputs.

To this end, we explore two complementary strategies. First, we constrain the translation model using information about Portuguese loanwords in Emakhuwa. Second, we expand our approach by incorporating entries from a bilingual dictionary to guide the model during translation. These methods allow us to integrate linguistic knowledge and external lexical resources into the training and inference pipeline, with the goal of improving translation adequacy and fluency in the Portuguese-Emakhuwa context.

8.1 Related Work

Guiding MT systems to use external knowledge has emerged as a promising research direction (Zhang and Zong, 2016), particularly for improving translation quality in specialized or low-resource contexts. One prominent approach in MT, which leverages external knowledge, is Constrained Neural Machine Translation (CNMT) (Sennrich et al., 2016c), which involves guiding the output generation of the model through the use of external constraints based on predefined terms, dictionaries, or glossaries. For example, when translating the sentence "I spent the whole night chasing a bug," the correct translation depends heavily on the domain (e.g., software engineering vs. biology) since the term bug carries different meanings. By guiding the MT system with domain-appropriate terminology, we can ensure that the model produces accurate and contextually suitable translations.

One major application of this technique is in domain-specific translation, where ensuring the correct use of terminology is crucial. By providing the model with a set of predefined terms, CNMT helps ensure consistency and accuracy, especially in fields like medicine, law, or technical documentation.

8.1.1 Lexical handling in MT

Several studies have explored the use of dictionaries to enhance supervised machine translation. Traditional research has primarily focused on integrating dictionaries into encoder-decoder architectures. For instance, Zhang and Zong (2016) improved MT by incorporating a bilingual dictionary containing rare or unseen words from the training data, while Arthur et al. (2016) addressed the challenge of translating infrequent words through discrete translation lexicons and attention mechanisms that select the most appropriate lexical probabilities. Similarly, Hämäläinen and Alnajjar (2020) leveraged dictionaries to generate synthetic parallel data, resulting in better model training. More recently, Moslem et al. (2023), Ghazvininejad et al. (2023), and Lu et al. (2024) demonstrated the value of bilingual dictionaries in MT with LLM prompting, aiming to further enhance translation quality by leveraging external lexical resources.

In the context of low-resource languages, constrained translation becomes a valuable technique. Due to the limited availability of training data, particularly parallel corpora, MT systems often suffer from poor vocabulary coverage and incorrect lexical choices. Incorporating external lexical resources, such as bilingual dictionaries or glossaries, can help guide the model toward producing more accurate and fluent translations, especially when translating from high-resource to low-resource languages. Recent studies have explored various methods to implement CNMT, broadly falling into three categories:

- **Decoding-based** These approaches focus on modifying the decoding process to enforce terminological constraints. These methods typically extend the search space during decoding to ensure that specific terms are correctly translated (Hokamp and Liu, 2017; Post and Vilar, 2018; Hu et al., 2019; Hasler et al., 2018). While effective, these approaches often incur

significant computational overhead and may yield only marginal improvements in overall translation quality (Guanhua et al., 2021; Zhang et al., 2023a).

- **Architecture-based** These approaches, conversely, explore modifications to the neural network architecture to integrate external information. Examples include incorporating alignment information (Song et al., 2020; Guanhua et al., 2021), vectorized terminology representations (Wang et al., 2022; Conia et al., 2024), non-autoregressive translation models (Song et al., 2020), and retrieval-augmented techniques combined with knowledge graphs (Conia et al., 2024). These methods aim to enhance the model's ability to handle constrained translation by leveraging additional contextual or structural information.

- **Input augmentation** Another widely used strategy is input augmentation, a simple yet effective method for constraining translation outputs (Crego et al., 2016; Song et al., 2019; Dinu et al., 2019; Michon et al., 2020; Niehues, 2021; Chen et al., 2021; Ailem et al., 2021; Zhang et al., 2023a; Conia et al., 2024). Input augmentation methods can be further divided into three main categories: placeholder-based, code-switching, and post-editing methods. In placeholder-based methods, the raw bitext is pre-processed by replacing source and target constraints with ordered labels during training. At inference time, the source constraints are marked with these labels, and the model predicts the corresponding target terms autonomously (Zhang et al., 2023a). In contrast, code-switching methods directly replace source constraints with their corresponding target terms in the input sentence. This allows the model to learn a copy behavior, where the decoder generates the target text step-by-step while preserving the pre-specified constraints (Zhang et al., 2023a). Placeholder-based and code-switching methods aim to signal to the model to prioritize the correct use of constraints during translation. Lastly, post-processing methods, or post-editing, include directly incorporating the required terminology into the translated output. Recently, Bogoychev and Chen (2023) demonstrated the potential of leveraging Large LLMs for refining translation output with specific terminology.

These approaches share the common goal of integrating human-collected or automatically extracted linguistic knowledge into the translation process, ultimately aiming to bridge the performance gap between high- and low-resource language pairs.

In this chapter, we adopt this paradigm by experimenting with constraints derived from Portuguese and Emakhuwa lexical resources and observing patterns of borrowings from Portuguese. Our approach builds on this body of work while adapting it to the unique linguistic features and data limitations of Emakhuwa.

8.1.2 Loanwords Challenge in MT

In multilingual contexts, lexical borrowing is a common phenomenon that facilitates communication by addressing linguistic diversity and filling vocabulary gaps. This is particularly evident in societies where multiple languages coexist, often with a dominant one (e.g., Portuguese in

Mozambique) serving as an official or widely used medium alongside regional languages. In such settings, translation practices frequently rely on lexical borrowing as a strategy to compensate for the lack of equivalent vocabulary in the target language, ensuring that the intended meaning is effectively conveyed (Gauton et al., 2008).

For example, when finding translation equivalents of technical or modern concepts—such as "radio" or "computer"—from Portuguese into Emakhuwa, human translators may adopt phonetic adaptations or construct sound-meaning equivalents— *raatio*, *kumuputatore*. These adaptations may vary based on the translator's orthographic style preference (e.g., radio — *ratio*, *ratiyo*, *raatiyo*, etc.), creating word innovations that help fill lexical gaps. Although such borrowed or newly coined words are generally accepted and understood by speakers, they can introduce lexical sparsity, presenting significant challenges for MT systems.

A key challenge in MT is handling loanword adaptation effectively. Unlike human translators, who can make context-sensitive decisions about adaptation strategies—such as phonetic transliterations, calques (literal translations), or culturally appropriate substitutions—MT models often struggle with these nuances. Moreover, the absence of standardized orthography further complicates the process, as variations in spelling must be accounted for while maintaining consistency in translation outputs.

To illustrate this issue, consider Example 4 in Table 5.2, where we evaluated our initial baseline model, trained from scratch using the Transformer architecture. The Portuguese source contains the term *Via Láctea*, which translates to *Milky Way* in English. This term is particularly challenging to render in Emakhuwa, as there is no direct native equivalent. As a result, the human reference opted for a phonetic adaptation aligned with Emakhuwa's sound system: *Eviiya Laateya*.

However, when examining the performance of MT systems fine-tuned on our dataset (see Table 8.1), we observe that all models struggled to adapt this loanword appropriately. Some failed to match the reference entirely, while others introduced mistranslations. For instance, the Gemma-2 model rendered the phrase as *ephiro ya oGalileya*, which incorrectly translates to "the way to Galilee."

Another critical aspect of lexical borrowing is its integration into the target language's grammatical structure. This involves applying appropriate pluralization rules, verb conjugations, and other morphological adjustments to ensure that borrowed terms conform to the target language's linguistic norms. For instance, a borrowed noun may require modifications to match native pluralization patterns, or a verb may need conjugation to fit syntactic constraints. While human translators apply linguistic rules to produce acceptable adaptations, MT systems must be tuned to handle such challenges.

8.1.3 Loanword Handling in MT

Loanwords present a unique challenge in MT. Nath et al. (2022) highlight the potential of effectively managing loanwords to improve translation systems, particularly for handling out-of-vocabulary (OOV) terms, co-referents, and named entities. Despite this potential, practical implementations of loanword handling techniques remain limited, with only a few studies exploring this

Table 8.1: Loanword adaptations by fine-tuned translation models for Portuguese into Emakhuwa

	Source	*Assim como a Lua exerce uma atração sobre a Terra, causando as marés, a Via Láctea exerce uma força sobre a galáxia de Sagitário.*
	Reference	*Siisaale mweeri okumihenryaaya exookhi ni etthaya, Eviiya Laateya ennikumiherya tho nave ikuro ni makhuru a itheneeri sa opuro.*
Translation	byT5	*Ntoko mweeri onikumiherya awe ephareya, onikumiherya iphareya, Ephiro ya Láctea ennikumiherya ikuru sa egalaxia ya Sagitário.*
	Afri-byT5	*Ntoko mweri onettiha awe elapo ya vathi, onettiha iphareya, Erowa Lactea ennivara ikuru sa egalaxia ya Sagitario.*
	M2M-100	*Ntoko mweeri oneettiha awe valaponi vaavo oneettiha awe ephareya, Via-Láctea oneettiha galáxia ya Sagitário.*
	NLLB	*Ntoko siisaale mweeri onikuruxa awe elapo, nto khiyakumiherya maasi, Via-Láctea onniwerya okuruxa egalaxia ya Sagitário.*
	Llama 3.1	*Ntoko siisaale mweeri oniiriha awe ephareya, nave-tho ephiro ya La-Tea enniiriha ephiro ya Sagitário okhala yoolipa.*
	Gemma-2	*Ntoko mweeri onikuxa awe maasi a mphareya, wira sipakiweke mawimpi, nave-tho ephiro ya oGalileya enniwerya ohela muhina egalaxia ya Sagitário.*

area in depth.

A notable contribution in this field comes from Mi et al. (2018), who investigated the role of loanwords in Neural MT (NMT) systems, focusing on their ability to mitigate the OOV problem. Building on Luong et al. (2015), their approach enriches the training data with explicit information, allowing the NMT system to generate special tokens for OOV words that match the source sentence. These tokens are subsequently replaced with the correct translations after they have been processed using a bilingual dictionary. This method shares similarities with placeholder-based CNMT approaches, as described earlier. However, a key limitation of placeholder-based augmentation is the difficulty in producing fluent translations (Zhang et al., 2023a; Chen et al., 2025).

While constrained translation has shown promise in improving translation quality, the intersection of constrained translation and loanword handling remains underexplored. To the best of our knowledge, there is limited research on leveraging data augmentation techniques to enforce loanword terminologies in translation systems. Our study addresses this gap by exploring input augmentation approaches with a particular emphasis on enforcing loanword constraints, thus enhancing the translation of loanwords while maintaining translation fluency.

8.2 Research Questions

In this chapter, we seek to answer the following questions:

- **RQ1**: To what extent does incorporating Portuguese loanword constraints improve translation quality in Portuguese–Emakhuwa MT systems?
- **RQ2**: How does the integration of bilingual dictionary information further enhance constrained MT performance beyond the use of loanword constraints alone?

8.3 Methodology

Our methodology consists of a two-fold approach. In the first phase, we investigate constraining MT systems using loanword information. In the second phase, we extend this approach by incorporating additional contextual information derived from bilingual dictionaries.

Figure 8.1 illustrates the overall framework. We begin by highlighting common translation challenges observed in baseline MT systems fine-tuned using standard parallel data. For instance, the system omits or mistranslates words that are either OOV or rare in the training corpus. As shown, the translation omits the word "plane" (*avião*), likely due to absence or low frequency in the training data. Similarly, "hyena" (*hiena*) is mistranslated as "lion" (*kharamu*), and "boar" (*javalis*) as "rabbit" (*namarokolo*).

Our first approach is based on the assumption that many OOV errors arise from loanwords, i.e., terms borrowed from Portuguese that have been adapted into Emakhuwa but occur infrequently in text corpora. We hypothesize that by injecting explicit mappings between source tokens and their corresponding loanword translations, the model can better learn when to copy or align these correctly during decoding.

However, we assume that loanword mappings alone are insufficient. The model may still struggle to translate words that are familiar within the Emakhuwa community but underrepresented in the training data, such as "hyena" and "boar". Therefore, in the second approach, we extend the first approach by also integrating dictionary entries.

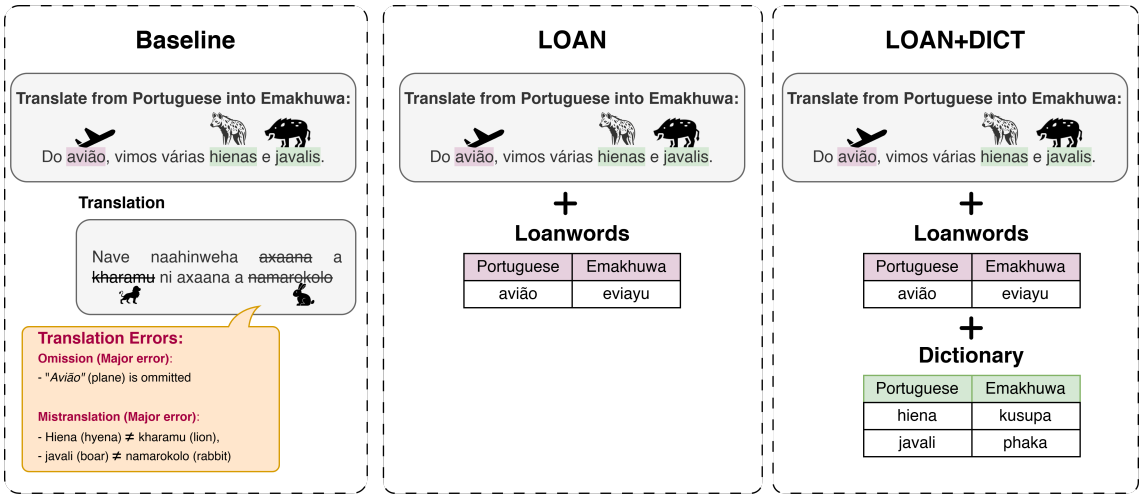


Figure 8.1: Methodology

Both approaches use a CNMT strategy, which we describe in detail in the next section.

8.4 Approach 1: Constrained Neural Machine Translation with loanwords

As illustrated in Figure 8.2, our first approach involves training a MT system that leverages mappings of loanwords to guide the model during translation. The aim is to help the model learn to copy the correct loanword into the target translation, improving accuracy and consistency in handling borrowed terms.

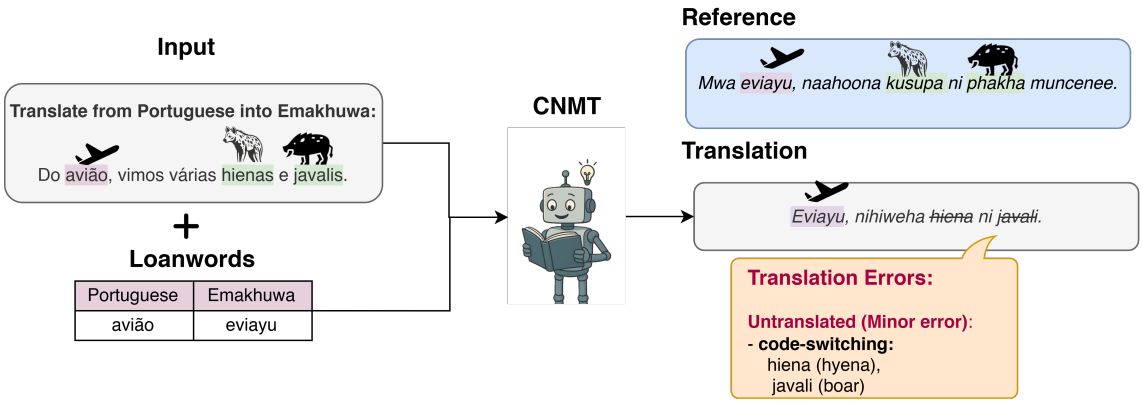


Figure 8.2: Constrained Neural Machine Translation with loanwords

Loanword definition We define loanwords as words that originate in a donor language and are borrowed into a recipient language. This borrowing process may sometimes involve transliteration, where the original word is adapted to fit the phonological or orthographic system of the recipient language. Cognates are excluded from this definition because, unlike loanwords, they are not the result of a borrowing process. Instead, cognates share a common etymological origin across languages and have evolved independently over time.

Problem statement Let $\mathbf{x} = \{x_1, \ldots, x_M\}$ represent the source sentence and $\mathbf{y} = \{y_1, \ldots, y_N\}$ represent the target sentence. Let $\mathscr{C} = \{\langle s_1, t_1 \rangle, \ldots, \langle s_K, t_K \rangle\}$ denote the set of loanword constraints between $\mathbf{x}$ and $\mathbf{y}$, where s_i and t_i correspond to the i-th source and target constraint, respectively. Each constraint can be a single word or a multi-word span, i.e., $|s_i| \geq 1$ and $|t_i| \geq 1$. Then, the problem requires that the translation systems enforce that s_i is translated into t_i.

Loanword-constrained Machine Translation Given $\mathbf{x}$, $\mathbf{y}$ and constraints $\mathscr{C} = \{\langle s_1, t_1 \rangle, \langle s_2, t_2 \rangle, \ldots, \langle s_K, t_K \rangle\}$, we augment the source text $\mathbf{x}$ into $\widehat{\mathbf{x}}$ using a replacement and prompting strategy.

8.4.1 Replacement

In this approach, loanwords identified in the source sentence s_i are replaced with an extended schema that explicitly links s_i to its target counterpart t_i. The schema follows a structured format:

$$\texttt{<|start|>}s_i\texttt{<|translate-as|>}t_i\texttt{<|end|>}$$

This method is inspired by data augmentation techniques for guiding NMT systems, commonly used in prior work on encoder-decoder architectures (Crego et al., 2016; Song et al., 2019; Dinu et al., 2019; Michon et al., 2020; Niehues, 2021; Chen et al., 2021; Ailem et al., 2021; Zhang et al., 2023a; Conia et al., 2024). In particular, we adopt the schema proposed by Conia et al. (2024) (see Table 8.2). This format serves as an explicit signal to the model, guiding it to correctly incorporate the target term t_i in the output. While the "replacement" strategy is primarily designed for encoder-decoder models, alternatives for decoder-only models are discussed in the following section.

Table 8.2: Standard input and CNMT input with replacement strategy.

Standard Input	*Assim como a Lua exerce uma atração sobre a Terra, causando as marés, a Via Láctea exerce uma força sobre a galáxia de Sagitário.*
CNMT Input	*Assim como a Lua exerce uma atração sobre a Terra, causando as marés, a* `<|start|>` ***Via Láctea*** `<|translate-as|>` ***Eviiya Laateya*** `<|end|>` *exerce uma força sobre a galáxia de Sagitário.*

8.4.2 Prompting

Similar to the replacement approach, this strategy involves augmenting the source text with constraints but adopts a prompt-based format. The prompt template (see Figure 8.3b) provides detailed instructions for the model, incorporating the constraints of the loanword alongside the translation task. Our prompt design builds on prior work in terminology-constrained prompting for large language models, particularly the studies by Moslem et al. (2023), Ghazvininejad et al. (2023), and Lu et al. (2024). This method is designed for decoder-only instruction-tuned models, capitalizing on their capabilities in following natural language directives.

Both "Replacement" and "Prompting" aim to enforce a copy behavior based on the provided constraints to encourage the model to map s_i and t_i. However, while this approach provides strong guidance for loanword adaptation, it does not fully guarantee that the model will always adhere to the constraints. This is due to the inherent unpredictability of language models' decoding behavior. To thoroughly assess the effectiveness of the proposed methods, we provide a comprehensive evaluation in the following sections.

Standard Input prompt tamplate

Translate the following sentence from `{source_language}` into `{target_language}`:
`{source_sentence}`

(a) Standard Prompt template for translation task.

CNMT Input prompt template

Translate the following sentence from `{source_language}` into `{target_language}`:

Sentence: `{source_sentence}`

Guidelines:
You may adapt loanwords as necessary.
For instance:
$\{s_0\}$ should be translated as $\{t_0\}$
$\{s_1\}$ should be translated as $\{t_1\}$
...
$\{s_k\}$ should be translated as $\{t_k\}$

(b) Prompt template for translation task. The placeholders s_i and t_i are replaced with corresponding loanword constraint pairs $\mathscr{C} = \{\langle s_1, t_1\rangle, \langle s_2, t_2\rangle, \ldots, \langle s_K, t_K\rangle\}$.

Figure 8.3: Prompt template for standard input (a) and with CNMT (b)

8.4.3 Experimental Setup

In this section, we begin by outlining the datasets used in the experiments and data preparation. Lastly, we detail the models used for fine-tuning in the machine translation task.

We used a Portuguese-Emakhuwa bitext dataset augmented with loanword constraint annotations to train our models. The dataset includes considerable manually annotated constraints, but a significant proportion of the data lacks these annotations. To address this, we used weak supervision techniques to annotate the remaining data. Further details on the dataset and our annotation methodology are provided below.

Given source sentences $\mathbf{x}$, target sentences $\mathbf{y}$, and $\mathscr{C}$, we augment all source sentences into $\widehat{\mathbf{x}}$ using $\mathscr{C}$ as described in Section 8.4. This preprocessing step was applied to the training, validation, and test sets.

8.4.3.1 Training and Validation Data

We used the same training, validation, and test sets described in the previous chapters. However, for this experiment, we applied additional preprocessing by incorporating loanword annotations. As discussed in Chapter 4—Section 4.5, the dataset combines biblical parallel texts, manually translated news articles, and other web-sourced materials (see Table 8.3). Among these, only the news article subset includes manually annotated loanwords, collected during the translation process.

To extend these annotations to the remaining unannotated portions of the data, we applied a weak annotation approach using LLM as an annotator. Specifically, we reformatted the existing gold-standard annotations into instruction-tuning prompts and used them to perform supervised fine-tuning of the GPT4o-mini model. The fine-tuned model was then used to automatically annotate the remaining unlabeled subset. Full details of this procedure are provided in the following subsection.

Table 8.3: Training and validation Portuguese-Emakhuwa bitext data (Table 4.4). Note: Some strings in the news category did not contain loanword constraints, meaning translations were performed entirely using native words in the target language.

Category	**TRAIN**	**DEV**	**Manual Loanword Annotations**
religious	45,386	290	×
news-politics	1,950	67	✓
news-economy	1,950	65	✓
news-culture	3,611	108	✓
news-sports	3,030	63	✓
news-health	2,256	83	✓
news-society	1,922	90	✓
news-world news	2,515	89	✓
tales	2,059	20	×
legal	893	24	×
Wikipedia	27	28	×
history	9	37	×
Total	65,608	964	

Generating Loanword Constraint Pairs For the model M, we leveraged the capabilities of *gpt-4o-mini-2024-07-18*[1], a state-of-the-art LLM by OpenAI. We fine-tuned on manually annotated sentence-level loanword pairs. We divided the gold-standard instances into an 80/20 split for training and validation, respectively. Then, for fine-tuning, we used an instruction prompt to guide the model in predicting loanword pair alignments $\mathscr{C}$ from the parallel sentences $\mathbf{x}$ and $\mathbf{y}$ (see Figure 8.4 and Figure 8.5). Finally, for inference, we set the temperature hyperparameter to 0 to ensure deterministic outputs.

We evaluated the model's performance using the dataset introduced by Ali et al. (2024b), which extends *FLORES+* for Portuguese–Emakhuwa machine translation evaluation. A key advantage of this evaluation set is the inclusion of manually annotated loanwords in each pair of parallel sentences. Below, we present a detailed analysis of the model's performance.

Evaluation of LLM as annotator To assess the effectiveness of the model described above, we compared its output against the gold standard annotations of our test dataset. For that, for each

[1]We fine-tuned this model during OpenAI's free 1M-token offer period held from August through September 2024.

Prompt template

Given a source sentence in `{source_language}` and its corresponding translation in `{target_language}`, identify the `{source_language}` loanwords used in the `{source_language}` sentence.
Present the results in the format:
"donor sequence" => "recipient sequence" without any further explanation.

Sentences:
`{source_language}`: {source_sentence}
`{target_language}`: {target_sentence}

Output:
"donor sequence" => "recipient sequence"

Figure 8.4: Prompt template. The placeholders are substituted with source and target languages and corresponding sentences.



Figure 8.5: Generating Loanword Constraint Pairs Using the Fine-Tuned *gpt-4o-mini-2024-07-18* Model

sentence pair we concatenated source and target text, then for each token we assigned a tag from one of two categories: LOAN–indicating that the token is part of a loanword–or O–indicating that the token is not part of a loanword. We adopted the BILOU scheme, which is primarily used in Named Entity Recognition (NER) tasks. In our case, the BILOU format was adapted to label tokens based on their position within or outside a loanword category (see Figure 8.6).

The evaluation results, presented in Table 8.4, show that the model achieves an F1-score of 95% on the gold standard test set. This high accuracy provided us with some confidence on the reliability of the automatic annotations using the model described above.



Figure 8.6: Labeling scheme for tagging source and target sentence tokens with "LOAN" or "O" category.

Table 8.4: Evaluation results of loanword constraints generation model

Tag	Precision	Recall	F1-Score	Accuracy
LOAN	0.87	0.95	0.91	–
O	0.99	0.98	0.99	–
Macro Avg.	0.93	0.97	0.95	–
Accuracy				0.98

8.4.3.2 Test Data

To evaluate the performance, we also used the FLORES+ *DEV* and *DEVTEST* datasets (Ali et al., 2024b), which contain 997 and 1012 pairs of sentences, respectively. Each sentence pair includes manually annotated loanword constraint information.

To provide further insight into the characteristics of each test set, Table 8.5 reports the average sentence length, average number of loanwords per sentence, and the overall proportion of loanwords. Notably, compared to the DEVTEST, the DEV set contains a higher ratio of loanwords per sentence.

Table 8.5: Details of the evaluation sets

Evaluation	AVG. Sent. length	AVG. Lonwords	Ratio
DEV	22.48	2.79	12%
DEVTEST	23.14	1.44	6%

8.4.3.3 Models

We fine-tuned, evaluated, and compared the following models to assess the effectiveness of our approach. Specifically, we retrained the same models used in Chapter 6, namely: (1) **Text-to-text models**: byT5 and AfriByT5; (2) **Multilingual machine translation models**: M2M-100 and NLLB; (3) **Decoder-only models**: LLaMA and Gemma.

Training Training is performed using pairs of source sentences $\mathbf{x}$, their augmented counterparts $\widehat{\mathbf{x}}$, and the corresponding translations $\mathbf{y}$. This setup is applied to both translation directions: Portuguese–Emakhuwa and Emakhuwa–Portuguese. The model is trained on a mixed setup: (1) translating unconstrained source sentences $\mathbf{x} \rightarrow \mathbf{y}$, and (2) translating source sentences with constraints explicitly injected $\widehat{\mathbf{x}} \rightarrow \mathbf{y}$.

8.4.3.4 Evaluation

For our machine translation evaluation, we used three key metrics: BLEU (Papineni et al., 2002), CHRF (Popović, 2015), and AfriCOMET (Wang et al., 2024a). AfriCOMET, an extended version of the COMET framework (Rei et al., 2020), was initially introduced to support a limited set of African languages. It has since been expanded by Wang et al. (2024a) to encompass a broader range of 76 African languages.

In addition, we assess loanword accuracy using sentence-level constraint accuracy (SCA), a metric introduced by Zhang et al. (2023a) which measures how accurately loanwords are translated within sentences. Under this metric, a translation is considered correct only if it accurately renders all required loanwords within a sentence.

Human Evaluation To assess the quality of translations produced by the different models, we conducted a human evaluation focusing on adequacy. Two independent professional translators volunteered to rate the translations based on a direct assessment score. This score ranged from 0 to 100, with specific guidelines: a score of 0 indicates that no meaning is preserved in the translation; a score between]0 - 34] means that the translation retains some of the source meaning but loses significant parts; a score between]34 - 67] indicates that most of the meaning is preserved; a score between]67 - 99] reflects a translation that is consistent with the source text, while a score of 100 represents a perfect translation.

Inter-annotator agreement was measured to ensure the reliability of the human judgments. We calculated Pearson's correlation coefficient (0.620), Spearman's rank correlation coefficient (0.580), and the Intraclass Correlation Coefficient ICC(3,2) (0.764). These values collectively suggest a moderate level of agreement between the two evaluators.

8.4.4 Results

We evaluated translation performance in both directions: Portuguese to Emakhuwa (pt→vmw) and Emakhuwa to Portuguese (vmw→pt), as detailed in Table 8.6 and in Table 8.7. Across all evaluation metrics, models trained with loanword constraints consistently outperformed the baseline. The most substantial gains in BLEU and CHRF scores were observed on the FLORES+ DEV set in the Portuguese-to-Emakhuwa direction. This split contains a higher ratio of loanwords (see Table 8.5), suggesting that constraint-aware training is particularly beneficial with lexical contextual clues. In the pt→vmw direction, the SCA results further reinforce this observation, showing substantial improvements in loanwords adherence when constraints are applied, ranging from 60.88 to 84.55 on the DEV set, and from 79.15 to 89.23 on the DEVTEST set. These scores highlight the effectiveness of constraint-aware training in preserving loanword usage. Additionally, AfriCOMET scores also improved under constrained conditions, indicating that lexical fidelity gains were accompanied by enhancements in overall translation quality.

In the vmw→pt direction, although the improvements were more modest, the constrained model still achieved significant gains, particularly in BLEU and CHRF scores. However, SCA did not show a meaningful increase, which aligns with our expectations. In this translation direction, it is generally easier for the model to recover the original Portuguese loanwords from Emakhuwa, making explicit constraints less critical.

Both encoder-decoder models and decoder-only models benefit from constraint-aware training. This indicates that the effectiveness of loanwords constraints generalizes across architectural paradigms, reinforcing their practical utility in low-resource translation settings.

Robustness We assessed the robustness of CNMT models when provided with standard inputs, as shown in Table 8.8 and Table 8.9. The models exhibited varied behavior under these conditions. Notably, NLLB and LLaMA 3.1 consistently outperformed the baseline, showing statistically significant improvements in CHRF scores. M2M-100 maintained stable performance, performing slightly below the baseline but without a significant difference. In contrast, Gemma-2, AfriByT5, and ByT5 showed consistent performance drops compared to the baseline. But, overall, the findings highlight NLLB and LLaMA 3.1 as more robust and reliable choices for our approach.

Human Judgments The overall rankings derived from these human judgments are presented in Figure 8.7. A key finding is that constraint-trained models (in blue) generally achieved higher adequacy scores compared to their counterparts trained solely on unconstrained inputs. Furthermore, the results indicate that both encoder-decoder and decoder-only architectures demonstrated competitive performance in this evaluation, with specific NLLB and Llama variants standing out within their respective categories. This suggests that training with awareness of constraints contributes positively to translation adequacy as perceived by human evaluators. Interestingly, human evaluation findings align with the trends observed in our automatic evaluation metrics and corroborate the quantitative improvements reported in metrics such as BLEU, CHRF, and AfriCOMET.

Table 8.6: Evaluation results on the **DEV** set. CHRF underlined if significantly better than baseline (Table 6.3) ($p < 0.05$).

	pt→vmw				**vmw→pt**			
Model	BLEU	CHRF	AfriCOMET	SCA	BLEU	CHRF	AfriCOMET	SCA
Afri-byT5	18.57	49.93 ↑11.20	0.57	80.94	26.74	49.58 ↑1.95	0.63	12.44
byT5	19.49	50.58 ↑11.04	0.57	81.04	25.88	48.56 ↑1.61	0.62	12.44
M2M-100	17.46	49.73 ↑9.64	0.54	70.61	25.25	48.60 ↑3.85	0.62	12.34
NLLB	16.83	50.84 ↑10.78	0.55	60.88	24.09	48.79 ↑2.04	0.63	12.34
Llama 3.1	19.53	**52.08** ↑14.29	0.57	82.65	29.19	52.14 ↑4.22	0.63	12.44
Gemma-2	19.12	50.46 ↑13.28	0.56	84.55	31.40	52.91 ↑4.75	0.64	12.44

Table 8.7: Evaluation results on the **DEVTEST** set. CHRF underlined if significantly better than baseline (Table 6.4) ($p < 0.05$).

	pt→vmw				**vmw→pt**			
Model	BLEU	CHRF	AfriCOMET	SCA	BLEU	CHRF	AfriCOMET	SCA
Afri-byT5	10.36	38.18 ↑3.76	0.49	89.23	15.87	38.86 ↑0.94	0.54	40.32
byT5	10.42	38.60 ↑4.01	0.50	89.53	15.46	38.24 ↑0.87	0.54	40.22
M2M-100	10.06	39.54 ↑3.30	0.49	85.97	14.86	38.75 ↑1.59	0.55	39.92
NLLB	10.50	**40.84** ↑3.63	0.50	79.15	15.14	39.73 ↑1.12	0.56	39.82
Llama 3.1	9.99	39.42 ↑4.71	0.50	87.65	15.63	40.97 ↑1.27	0.54	40.12
Gemma-2	10.11	37.78 ↑3.27	0.49	87.45	18.10	41.56 ↑2.17	0.56	39.92

Table 8.8: Robustness evaluation on the **DEV** set.

	pt→vmw				**vmw→pt**			
Model	BLEU	CHRF	AfriCOMET	SCA	BLEU	CHRF	AfriCOMET	SCA
Afri-byT5	8.20	36.24 ↓2.49	0.47	13.14	19.44	44.21 ↓3.42	0.58	12.14
byT5	9.14	37.34 ↓2.20	0.48	13.14	18.49	42.95 ↓4.00	0.57	12.34
M2M-100	9.27	39.48 ↓0.61	0.50	13.94	20.33	44.55 ↓0.20	0.58	11.94
NLLB	9.47	41.82 ↑1.76	0.51	14.94	22.66	46.95 ↑0.20	0.61	12.04
Llama 3.1	9.27	40.11 ↑2.32	0.51	14.44	24.66	48.37 ↑0.45	0.60	12.24
Gemma-2	6.64	33.69 ↓3.49	0.47	13.04	23.30	46.13 ↓2.03	0.60	12.34

Table 8.9: Robustness evaluation on the **DEVTEST** set.

	pt→vmw				**vmw→pt**			
Model	BLEU	CHRF	AfriCOMET	SCA	BLEU	CHRF	AfriCOMET	SCA
Afri-byT5	6.76	33.91 ↓0.51	0.45	42.00	12.93	36.37 ↓1.55	0.53	41.21
byT5	6.92	34.67 ↑0.08	0.46	41.40	12.54	36.00 ↓1.37	0.52	41.01
M2M-100	7.05	36.30 ↑0.06	0.47	43.48	13.20	37.06 ↓0.10	0.53	40.22
NLLB	7.56	38.18 ↑0.97	0.49	43.28	13.95	38.74 ↑0.13	0.55	40.02
Llama 3.1	7.20	36.20 ↑1.49	0.48	44.66	14.81	39.71 ↑0.01	0.54	40.22
Gemma-2	6.11	32.53 ↓1.98	0.45	42.00	15.22	39.07 ↓0.32	0.55	39.92

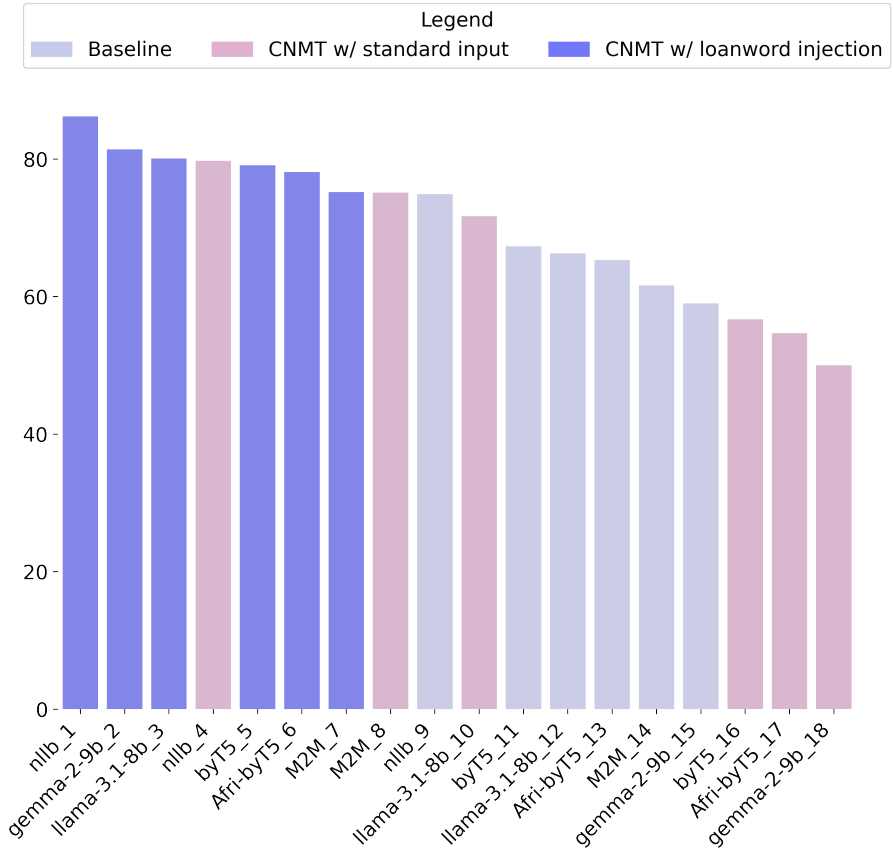


Figure 8.7: Model rankings based on human evaluations. Each bar plot shows the average direct assessment across the evaluated 200 sample from test set.

8.4.5 Case Study

To demonstrate the impact of loanword injection, we present two translation examples in Table 8.10 and Table 8.11. Table 8.10 illustrates how our approach improves upon the baseline results reported in Table 8.1. Across both translation directions—**pt**→**vmw** and **vmw**→**pt**—models trained with loanword injection exhibit a stronger alignment with reference translations, particularly in preserving loanwords. This leads to improvements not only in loanword faithfulness but also in overall translation quality.

Table 8.11 presents an additional example from the DEVTEST set. In the case of the **pt**→**vmw** direction, all baseline models—trained without loanword injection—struggled to accurately translate the Portuguese loanwords *murais* and *rabiscos*, often omitting or mistranslating them, which resulted in lower adequacy. While most models attempted to adapt the term *grafite* into Emakhuwa, these adaptations frequently violated phonological norms. For instance, letter *g* does not exist in the orthography of Emakhuwa and should ideally be replaced with *k* Ali et al. (2024a).

A similar pattern emerges in the reverse direction, **vmw**→**pt**. Loanwords such as *imuraaxi*, *irapiixiku*, and *ekarafiiti* were typically identified as phonetically adapted borrowings from Portuguese. However, most models failed to correctly map them back to their original forms. The no-

Table 8.10: A sample from DEV set showcasing the improvement over the baseline translation in Table 8.1

		Example 1
	pt	Assim como a Lua exerce uma atração sobre a Terra, causando as marés, a Via Láctea exerce uma força sobre a galáxia de Sagitário
	vmw	Siisaale mweeri okumihenryaaya exookhi ni etthaya, Eviiya Laateya ennikumiherya tho nave ikuro ni makhuru a itheneeri sa opuro.
CNMT w/ loaword injection		
Afri-byT5	pt-vmw	Ntoko siisaale mweeri onikumiherya awe valaponi, onikumiherya iphareya, Eviiya Laateya ennikumiherya ikuru sa elapo ya Sagitaario.
	vmw-pt	Assim como o mês causou a fome e a terra, a Via Láctea também aproveita também a força e as estrelas locais.
byT5	pt-vmw	Ntoko mweeri onikhumelela awe valaponi, onikhumelela iphareya, Eviiya Laateya ennikhaliherya ikuru sa elapo ya Sagitaario.
	vmw-pt	Assim como o mês que produz fogo e terra, a Via Láctea também produz forças e grupos de estrelas locais.
M2M	pt-vmw	Ntoko siisaale mweeri onikumiha awe mphareya valaponi, nto Eviiya Laateya ennikumiha ikuru valaponi.
	vmw-pt	Assim como no mês passado, a Via Láctea produziu também energia e grupos de estrelas no espaço.
NLLB	pt-vmw	Ntoko siisaale mweeri onaanaanaanaana awe elapo, nto khiyakumiherya ephareya, nto Eviiya Laateya ennaanaanaanaanaana ni egalaxia eniitthaniwa Sagitário.
	vmw-pt	Assim como a lua produz o pó e o solo, a Via Láctea produz também órgãos e glóbulos celestiais.
Llama 3.1	pt-vmw	Ntoko siisaale mweeri siniira aya waavarela atthu ni iphareya, Eviya Laateya enniira waavarela atthu a esuphi Sagitário.
	vmw-pt	Assim como o mês criou o sol e a terra, a Via Láctea criou também as estrelas do seu espaço
Gemma-2	pt-vmw	Ntoko siisaale mweeri onimukuxa awe maasi a mphareya, eviiya laateya enniwiiriha atthu a elapo ya Sagitário okhala oowiiraana.
	vmw-pt	Assim como a Lua faz com a Terra, a Via Láctea também faz com que as estrelas pareçam brilhar.

Table 8.11: A sample from DEVTEST set, showcasing the improvement over the baseline translation with CNMT with loaword injection

		Example 2
	pt	Murais ou rabiscos indesejados são conhecidos como grafite.
	vmw	Imuraaxi wala irapiixiku soohitthuneya ti soosuwelaxiwa okhala ekarafiiti.
Baseline		
Afri-byT5	pt-vmw	Ikaruma wala irapisku sohiphavela, sinnisuweliwa ntoko egrafiti.
	vmw-pt	Os murais ou os rapiscos inseguros são conhecidos como grafite .
byT5	pt-vmw	Okhala wira imatta wala imatta soohiloka sinisuweliwa ntoko grafite.
	vmw-pt	A água ou os plásticos necessários são conhecidos como ecarbonetos.
M2M-100	pt-vmw	Makhalelo oovirikana a atthu anisuweliwa okhala grafite.
	vmw-pt	MURAGES ou rapiosos desnecessários são conhecidos como grafiços
NLLB	pt-vmw	Ithoonyeryo soohiloka sa exiri sinnisuweliwa okhala karafiiti.
	vmw-pt	Buracos desnecessários são conhecidos como grafite.
Llama 3.1	pt-vmw	Murais wala soolempwa soohiloka siniitthaniwa grafite.
	vmw-pt	Imunidades ou isentimentos injustificados são conhecidos como grafitis.
Gemma-2	pt-vmw	Murais wala rabisko woohiphavela onisuweliwa okhala grafite.
	vmw-pt	As muralhas ou grafites são conhecidos como arte urbana.
CNMT w/ loaword injection		
Afri-byT5	pt-vmw	Imuraaxi wala irapiixiku soohiphavela sinisuweliwa ntoko ekarafiiti.
	vmw-pt	Murais ou rabiscos inesperados são conhecidos como grafite.
byT5	pt-vmw	Imuraaxi wala irapiixiku soohiphavela sinnisuweliwa ntoko ekarafiiti.
	vmw-pt	Murais ou rabiscos inesperados são conhecidos como grafite.
M2M-100	pt-vmw	Makhalelo oovirikana a atthu anisuweliwa okhala grafite.
	vmw-pt	Murais ou rabiscos desnecessários são conhecidos como grafite
NLLB	pt-vmw	Imuraaxi wala irapiixiku soohitthuneya sinisuweliwa ntoko ekarafiiti
	vmw-pt	Murais ou debulhos desnecessários são conhecidos como grafite.
Llama 3.1 8B	pt-vmw	Imuraaxi wala irapiixiku soohiloka siniitthaniwa ekarafiiti.
	vmw-pt	Muranjos ou rabiscos são conhecidos como grafite.
Gemma-2 9B	pt-vmw	Irapiixiku wala ekarafiiti soohiloka sinisuweliwa okhala muraaxi wala irapiixiku soohiloka.
	vmw-pt	Murais ou rabiscos não autorizados são conhecidos como grafite.

table exception was the Gemma-3 4B model, which successfully retrieved the correct Portuguese terms, suggesting a higher capacity for loanword adaptation. These challenges likely stem from the fact that such loanwords are rare or absent in the training data, making it harder for models to recognize and translate them accurately.

In summary, the results indicate that injecting loanwords into the source sentences substantially improves translation quality in both directions. These improvements are consistently reflected in automatic evaluation metrics and human assessments, confirming the effectiveness of our approach. Essentially, this method enables the model to better handle rare words that arise from the adaptation of foreign terms into the target language. However, rare words may also appear due to the scarcity of native word examples in the training data. Recalling the example discussed earlier in this section (see Figure 8.2), we observe that CNMT with loanword injection enables the model to produce more appropriate translations of loanwords. Nevertheless, the system still struggles with words like "hyena" and "boar," which have native equivalents and should not be treated as loanwords. Therefore, the next section explores an extension of the CNMT approach that aims to incorporate native or non-loanword mappings by leveraging dictionary entries.

8.5 Approach 2: Incorporating Dictionary Information

To address RQ2, we extend Approach 1 by incorporating additional information from bilingual dictionaries (see Figure 8.8).

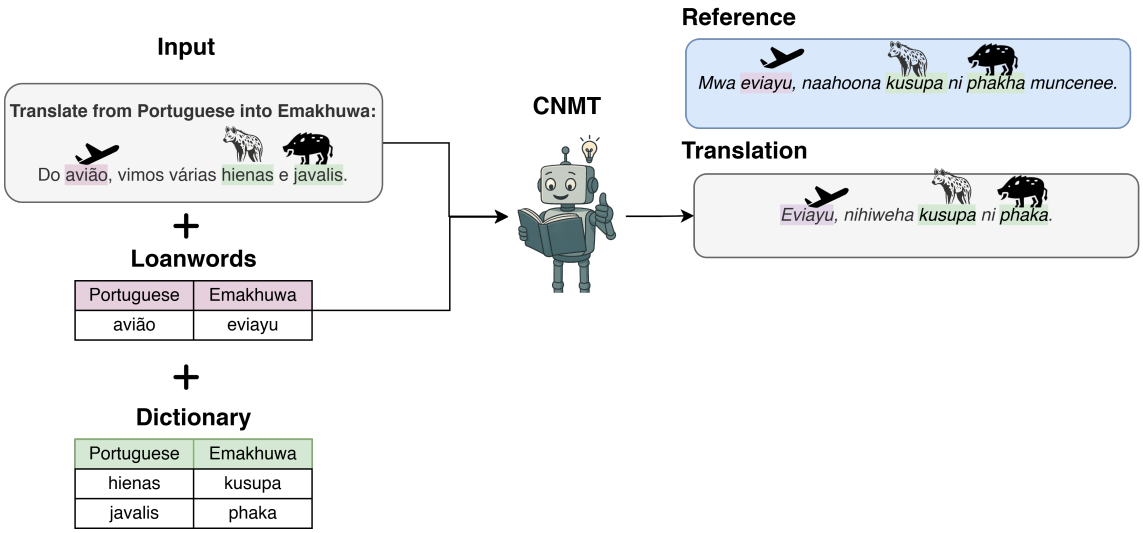


Figure 8.8: Constrained Neural Machine Translation with loanwords and dictionary

We enriched each source sentence with glossary information that contains candidate source–target word pairs relevant to that sentence. These pairs are sourced from direct matches from a bilingual dictionary. The following sections provide a detailed description of how dictionary resources were constructed.

8.5.1 Dictionary

We collected different Portuguese to Emakhuwa dictionaries and glossaries, scanned those available only in printed format, used optical character recognition (OCR) to extract information, and compiled all dictionary entries into a single bilingual structured dictionary. The process of extracting lexical information from scanned dictionaries involved the following steps:

- **Dictionary entry segmentation**: Scanned dictionary was split into pages, then each page was further divided into individual dictionary entries. We used Algorithm 1 (see also Figure 8.9) to crop individual entries in dictionary pages. This ensured that each dictionary entry was isolated for further processing.
- **OCR**: After dividing the pages into individual dictionary entries, we used GPT-4o for OCR to convert the images into text. We then applied Regular Expressions (RegEx) to detect Portuguese words and their corresponding Emakhuwa translations. Finally, we organized the extracted information into a JSON file, where each entry includes a Portuguese term and one or more associated Emakhuwa translations.

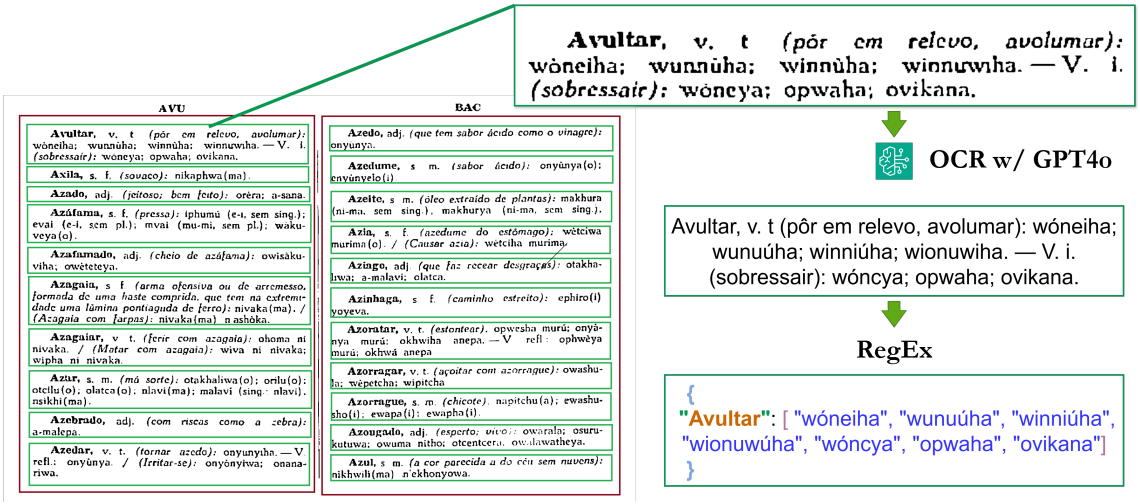


Figure 8.9: Visualization of the process for extracting dictionary entries. Step-by-step operation of Algorithm 1, followed by OCR and the subsequent JSON formatting.

We compiled bilingual entries with approximately 8,084 Portuguese terms, each paired with one or more Emakhuwa translations. This dictionary was then used to generate sentence-specific glossaries by matching relevant entries to terms in the source sentences.

8.5.2 Matching Dictionary

As described in Section 8.5.1, the dictionary d maps each term t to a set of corresponding translations:

$$d(t) = \{\text{translation}_1, \text{translation}_2, \ldots, \text{translation}_M\}$$

Algorithm 1 Cropping a Scanned Dictionary Page

```
procedure CROPPDFPAGE(scanned_page)
    // Step 1: Split the page vertically into columns
    Draw a vertical midpoint of the page as initial cut line
    // Adjust the cut line
    for each small shift left or right from midpoint do
        if no text is intersected by the line (i.e., blank vertical gap found) then
            Save this position as a clean vertical split
            Break loop
        end if
    end for
    Split the page into two columns along the clean vertical line
    // Step 2: Dictionary entry Crop
    for each column in {left, right} do
        // Find top crop line
        Draw two horizontal lines at the top of the column
        for each small shift upwards from the 1st horizontal line do
            if no text is intersected by the line (i.e., blank horizontal gap found) then
                Use this as top crop boundary
                Break
            end if
        end for
        //Find bottom crop line
        for each small shift downwards from the 2nd horizontal line do
            if no text is intersected by the line (i.e., blank horizontal gap found) then
                Use this as bottom crop boundary
                Break
            end if
        end for
        Crop the column using top and bottom boundaries
        Repeat top/bottom boundary cropping until no more crops found
    end for
    return Individual dictionary entry crops
end procedure
```

Each dictionary entry may consist of a single word or a multiword expression. To identify matching terms in a source sentence $x = \langle x_1, \ldots, x_n \rangle$, we adopt the following strategy. First, we generate n-grams of up to N tokens (e.g., $N = 5$) from the source sentence to capture both single words and multiword expressions as potential matches:

$$x' = \bigcup_{k=1}^{\min(n,N)} \{\langle x_i, x_{i+1}, \ldots, x_{i+k-1} \rangle \mid 1 \leq i \leq n-k+1\}$$

Next, we expand the candidate set by including lemmatized forms of unigrams, obtained through a language-specific lemmatization function that reduces words to their base or dictionary form (e.g., "running" to "run"). These lemmatized unigrams, along with the original n-grams, form a unified candidate set C. For each candidate $c_i \in C$, we compute a similarity score using a fuzzy matching algorithm using Levenshtein distance-based strategy, against each term t in the dictionary d. The similarity score quantifies the degree of lexical overlap, accounting for minor

Table 8.12: Standard input and CNMT input with appending strategy.

Standard Input	*Assim como a Lua exerce uma atração sobre a Terra, causando as marés, a Via Láctea exerce uma força sobre a galáxia de Sagitário.*
CNMT Input w/ LOAN	*Assim como a Lua exerce uma atração sobre a Terra, causando as marés, a* <|start|> ***Via Láctea*** <|translate-as|> ***Eviiya Laateya*** <|end|> *exerce uma força sobre a galáxia de Sagitário.*
CNMT Input w/ LOAN+DICT appending	*Assim como a Lua exerce uma atração sobre a Terra, causando as marés, a* <|start|> ***Via Láctea*** <|translate-as|> ***Eviiya Laateya*** <|end|> *exerce uma força sobre a galáxia de Sagitário.* </sep> <dict> <term> <src> ***lua*** </src> <tgt> ***mweeri*** </tgt> </term> <term> <src> ***assim*** </src> <tgt> ***siisa*** </tgt> </term> <term> <src> ***terra*** </src> <tgt> ***etthaya*** </tgt> </term> </dict>

variations such as typos or morphological differences. Only candidates c_i with a similarity score w exceeding a predefined threshold θ (e.g., $\theta = 0.85$) are retained as C'. For each retained match $c'_i \in C'$, we apply fuzzy matching between the tokens in the target sentence $y = \langle y_1, \ldots, y_m \rangle$ and the translations in $d(c'_i)$ to rank the translations based on lexical similarity, selecting the translation m with the highest similarity score. The glossary entry for a specific sentence pair is thus defined as:

$$G = \left\{ (s, m, w) \mid s \in C', m \in d(s), w \geq \theta \in [0,1] \right\}$$

8.5.3 Integration of Glossary

We examined two strategies for incorporating the glossary into the translation workflow: (i) appending the glossary to the input and (ii) prompting the model to use glossary terms explicitly.

Appending Given a source–target sentence pair and glossary G, we augment the input with relevant and matched dictionary entries and loanword mappings, as illustrated in Table 8.12. We explore appending dictionary entries by post-appending, i.e, placing entries after the source sentence.

Prompting Given a source–target sentence pair and glossary G, we construct the input as a prompt that augments the translation instruction with the relevant dictionary entries and loanword mappings, as illustrated in Figure 8.10.

Glossary items are incorporated into the prompt in a structured format. By explicitly integrating these elements into the prompt, we aim to guide the model toward producing translations that are better aligned with the terminologies injected. It is worth noting that not all sentence pairs contain both resource types. In some cases, only dictionary entries are available, while in others only loanword mappings apply. For such cases, we adopt the corresponding prompt formats.

Prompt augmented with glossary entries

Translate the following sentence from Portuguese into Emakhuwa:

$\{x\}$

Guidelines:
You may use glossary terms or adapt loanwords as needed to ensure clarity, cultural relevance, and linguistic consistency.

Dictionary:
$\{s_i\} : \{m_i\}, \{s_{i+1}\} : \{m_{i+1}\}, \ldots, \{s_{i+k}\} : \{m_{i+k}\}$

Loanwords:
$\{s_j\}$ should be adapted as $\{m_j\}$
$\{s_{j+1}\}$ should be adapted as $\{m_{j+1}\}$
...
$\{s_{j+g}\}$ should be adapted as $\{m_{j+g}\}$

Figure 8.10: Prompt template with glossary integration. The translation task is augmented by incorporating structured glossary information, which includes dictionary terms and loanword adaptations.

8.5.4 Experimental Setup

In this section, we describe the data preparation process used in our experiments, followed by the models used for fine-tuning on the machine translation task.

Data Preparation We integrated sentence-specific glossaries using two approaches: appending and prompting, as described in Section 8.5.3. For encoder-decoder models, we applied the "appending" method during preprocessing, while prompting was used for LLMs. We preprocess on the same training, validation, and test set described in Section 8.4.3. Table 8.13 summarizes the statistics of the dataset, including the average sentence length, the average number of loanwords, and the average number of glossary terms.

Table 8.13: Statistics of the training, validation, and test sets. Note that the average number of injected loanwords from glossaries is computed using a threshold of $\theta = 85$.

		Average		
	# Sent.	**Length**	**Loanwords**	**Glossary**
TRAIN	65,574	21.26	1.63	7.5
DEV	997	22.48	2.79	6.56
DEVTEST	1012	23.14	1.44	6.62

Models We fine-tuned and evaluated the top-performing models identified through human assessment in our first approach (Figure 8.7). These models include: (1) **Multilingual machine translation model**: NLLB 200(600M); (2) **Decoder-only models**: LLaMA-3.1(8B) and Gemma-2(9B)

The models were fine-tuned using mixed inputs: standard translation data and data augmented with sentence-specific glossaries (LOAN+DICT). This mixed-input approach was designed to improve robustness, allowing the models to handle inputs both with and without additional sentence-specific glossaries.

8.5.5 Results

Table 8.14 presents the results of adding more context with dictionary information. Incorporating dictionary-based context led to notable performance differences, particularly for the LLaMA 3.1 model. On the DEV set, LLaMA 3.1 demonstrated improved performance across all evaluation metrics, indicating that the model effectively leveraged the enriched contextual input. However, on the DEVTEST set, its performance declined slightly compared to its version without dictionary context. Despite this drop, the differences are not statistically significant, suggesting that LLaMA can still benefit from additional contextual information without incurring major trade-offs.

These findings suggest that the architecture of LLaMA, probably influenced by its instruction following design, is well-suited to integrate external information sources, such as dictionaries. Conversely, NLLB exhibited substantial performance degradation when dictionary context was added, particularly in BLEU and CHRF scores. This underperformance may be attributed to limitations in how contextual embeddings were concatenated, potentially preventing the model from effectively capturing and utilizing the additional semantic cues. Further research is required to develop more efficient methods of incorporating external context into encoder-decoder architectures like NLLB.

Overall, our results indicate that instruction-tuned LLMs, such as LLaMA and Gemma, are more suitable for the task, making them promising candidates for CNMT with dictionary entries.

Table 8.14: Results of Portuguese-to-Emakhuwa translation when incorporating additional context using a dictionary ($\theta = 85$) on *FLORES+ DEV* and *DEVTEST* sets. Variations refer to the results in Table 8.6 and Table 8.7.

	DEV			***DEVTEST***		
Model	BLEU	CHRF	AfriCOMET	BLEU	CHRF	AfriCOMET
NLLB	14.72 ↓2.11	49.44 ↓1.40	0.58 ↑0.03	9.03 ↓1.47	40.00 ↓0.84	0.53 ↑0.03
Llama 3.1	21.51 ↑1.98	<u>53.63</u> ↑1.55	0.59 ↑0.02	10.21 ↑0.22	40.74 ↑1.32	0.51 ↑0.01
Gemma-2	19.16 ↑0.04	50.62 ↑0.16	0.57 ↑0.01	10.26 ↑0.15	38.14 ↑0.36	61.67 ↑0.49

8.5.5.1 Error Analysis

To illustrate the impact of our approach, we examine the example shown in Table 8.15, focusing on LLaMA 3.1, which overall yielded the strongest performance.

In the **Baseline** setup, the translation diverged significantly from the reference. LLaMA 3.1 produced a sentence with both mistranslation and addition errors, introducing concepts not present

in the source (e.g., "lion cubs" and "rabbit cubs"). This behavior reflects a common issue in low-resource machine translation: when encountering unfamiliar vocabulary, the model struggles to maintain semantic fidelity and often hallucinates content. This underscores the limitations and risks of relying solely on supervised fine-tuning under data-scarce conditions.

In the **LOAN** setup, translation quality improved substantially. First, the model correctly adopted the expected borrowing for "plane" (*eviayu*). Second, the output aligned more closely with the reference in meaning. However, challenges persisted with terms absent from the training data, such as "hyenas" and "wild boars." Instead of using the native Emakhuwa equivalents (*kusupa* and *phaka*), the model resorted to borrowed forms (*hiena*, *javali*). While this resembles code-switching, which is natural in informal speech, it raises critical concerns in machine translation. Artificially introduced forms can fragment the lexicon, displace culturally significant vocabulary, and, if deployed in real-world applications, risk normalizing inconsistent or non-native terms, potentially weakening the integrity of the language over time.

Finally, in the **LOAN+DICT** setup, LLaMA 3.1 successfully produced translations that respected both borrowings and native terms, yielding outputs much closer to the expected Emakhuwa rendering. This effect was consistent across other models as well, suggesting that dictionary-based lexical augmentation is broadly applicable and effective in enhancing translation quality beyond loanword handling alone.

8.5.5.2 Ablation study

Our main results show a consistent trend: translation quality improves as we move from **Baseline** → **LOAN** → **LOAN+DICT** training setups. In particular, loanword augmentation drives the largest gains, while dictionary entries provide complementary improvements by covering additional vocabulary.

These results motivate a closer examination of the **LOAN+DICT** setup. In particular, we consider two aspects: (i) the extent to which performance on standard prompts inference differs from the baseline, and (ii) the sensitivity of the approach to glossary design, especially the choice of matching thresholds.

To explore these questions, we conduct a series of ablation studies described below. We restrict this analysis to the LLM-based models, as the NLLB model underperformed on the LOAN+DICT approach.

Providing standard prompts in LOAN+DICT setup To assess whether the improvements of the **LOAN+DICT** configuration come at the expense of performance on standard prompts, we evaluate translations generated under this condition (Table 8.16). Results show that LLaMA 3.1 8B achieves statistically significant gains in CHRF above the baseline, with improvements of +2.42 and +2.59 on DEV, and +1.39 and +1.52 on DEVTEST, respectively. In contrast, Gemma 2 9B shows a small but consistent decrease compared to the baseline.

[1] **Note**: We designed this example to showcase the strengths of our approach, particularly its ability to manage loanwords (e.g., plane) as well as vocabulary absent from the training data (e.g., native names of wild animals)

Table 8.15: Example

EN:	*From the plane, we saw several hyenas and wild boars.*
PT:	Do avião, vimos várias hienas e javalis.
VMW:	Mwa eviayu, naahoona kusupa ni phakha muncenee.
Glossary	**LOAN:** { avião → eviayu }, **DICT:** { hiena → kusupa, javalis → phakha }
Baseline	
NLLB	Mwaha wa eviyau, naahinweha ehiena ni ihopa sinceene. *EN: When we were on the plane, we saw several ~~ehiena~~ and ~~fish~~.*
LLaMA 3.1	Nave naahinweha axaana a kharamu ni axaana a namarokolo. *EN: We also saw the lion cubs and the rabbit cubs.*
Gemma 2	Okathi naavira ahu ni ntteeke, naahoona inama sinceene ntoko kharamu, ni inama sikina. *EN: While flying, we saw many animals like lions and other animals.*
LOAN	
NLLB	Neemenle ene eviayu, naahinweha kusupa ni iwali *EN: When we flew over, we saw hyenas and ~~iwali~~.*
LLaMA 3.1	Eviayu, nihiweha ~~hiena~~ ni ~~javali~~. *EN: On the plane, we saw hyena and wild boar.*
Gemma 2	Okathi naakhuma ahu eviayu, naahoona axinama anceene ntoko kharamu, mwalapwa, ni ikuluwe *EN: When we got off the plane and saw many lions, dogs, and pigs.*
LOAN+DICT	
NLLB	Okathi naavira ahu ni eviayu, naahiwoona kusupa ~~Javali ni Kusupa Javali.~~ *EN: On the plane, we saw the hyena ~~javali and hyena javali.~~*
LLaMA 3.1	Okhuma eviayu, nooweha kusupa ni phaka. *EN: Getting off the plane, we saw the hyena and the warthog.*
Gemma 2	Okhumaka eviayu, nahoona kusupa ni phaka. *EN: When we got off the plane, we saw the hyena and the warthog.*

These findings indicate that incorporating loanword and dictionary resources does not reduce model performance overall when prompted in the standard format. Instead, for most models, the additional training signals yield measurable benefits even without specialized prompting. The exception of Gemma 2 9B suggests that the effect may depend on the architecture or pretraining characteristics of the underlying model.

Matching thresholds effect Beyond prompt robustness, another key factor in the **LOAN+DICT** setup is the construction of the glossary itself. In particular, dictionary entries must be matched with the training data, and the choice of threshold directly determines which terms are incorporated. A low threshold may increase coverage but risks introducing noise resulting from incorrect matches, while a high threshold ensures precision but may exclude potentially useful entries. To

Table 8.16: Portuguese(pt) →Emakhuwa(vmw) performance of results of standard prompts on LOAN+DICT setup. For CHRF values underlined indicate significant evidence that they outperform the baseline ($p < 0.05$), based on pairwise randomized significance tests (Koehn, 2004)

	DEV		DEVTEST	
Model	CHRF	AfriCOMET	CHRF	AfriCOMET
LLaMA 3.1 8B	40.38$_{\uparrow 2.59}$	62.07	36.23$_{\uparrow 1.52}$	61.09
Gemma 2 9B	36.76$_{\downarrow 0.42}$	61.08	34.05$_{\downarrow 0.46}$	60.70

better understand this trade-off, we analyze the effect of varying matching thresholds by lowering θ to 80, by training a model augmented with data with lower θ. Table 8.17 reports CHRF scores for $\theta = 80$ across the DEV and DEVTEST splits. The results reveal consistent patterns: LLaMA 3.1 8B is stable across thresholds. On both *DEV* and *DEVTEST*, the change from $\theta = 85$ (see Table 8.14) to $\theta = 80$ results in no significant variation in CHRF, which suggests that this model can tolerate the additional noise introduced at a lower threshold. Gemma 2 9B, in contrast, is much more sensitive to threshold changes. At $\theta = 80$, performance drops drastically (to 24.23 CHRF on *DEV* and 21.39 CHRF on *DEVTEST*). This suggests that Gemma 2 relies heavily on precise dictionary matches and is disproportionately affected by noisy entries.

Overall, these findings indicate that the optimal threshold may depend on the underlying model. While some architectures are robust to looser matching, others require stricter alignment to fully benefit from dictionary integration. This observation underscores the importance of tuning θ not only globally but also with respect to the characteristics of the target model.

Table 8.17: Portuguese(pt) →Emakhuwa(vmw) performance of results, threshold of $\theta = 80$.

	DEV		DEVTEST	
Model	CHRF	AfriCOMET	CHRF	AfriCOMET
LLaMA 3.1 8B	53.18$_{\downarrow 0.43}$	65.31	40.67$_{\downarrow 0.07}$	62.26
Gemma 2 9B	26.39$_{\downarrow 24.23}$	29.96	16.75$_{\downarrow 21.39}$	29.53

8.6 Summary

In low-resource MT, the lack of parallel data severely limits quality. This chapter presented a targeted approach to improve Portuguese-Emakhuwa MT, by leveraging specific linguistic knowledge: the frequent use of Portuguese loanwords in Emakhuwa and the availability of bilingual dictionaries. Our method involved Constrained Neural Machine Translation, where we explicitly guided MT models to handle loanwords correctly. We augmented the training data with loanword pairs, using special tokens for encoder-decoder models and direct instructions for large language models. We also experimented with incorporating a bilingual dictionary to provide additional lexical context.

The results were statistically significant. Constraining models with loanwords consistently and substantially improved translation quality across all models and metrics, a finding confirmed by human evaluators. The models became much better at correctly translating loanwords (as reflected in higher SCA scores), which in turn enhanced overall sentence adequacy and fluency. Adding a full bilingual dictionary yielded mixed results, by improving the performance of the instruction-tuned LLaMA model, but degrading the performance of the NLLB model. This finding highlights that the integration method is critical.

In conclusion, this research presents a practical and effective strategy for enhancing low-resource MT. By creating and using curated glossaries of known linguistic phenomena like loanwords, we can directly infuse linguistic knowledge into MT systems, and subsequently lead to statistically significant improvements in translation quality.

Chapter 9

Conclusion and Future Work

This research addressed the challenge of developing a MT system for Portuguese-Emakhuwa, a language that is spoken by a substantial portion of the Mozambican population, yet is systematically marginalized. We began by characterizing Emakhuwa and outlining the unique obstacles inherent in building MT systems for such a low-resource context. The primary challenge identified was the scarcity of parallel data, a critical resource for current MT approaches. Furthermore, Emakahuwa presents complex linguistic features, orthographic variation, and a reliance on lexical borrowing from Portuguese to fill vocabulary gaps, which pose distinct challenges for developing translation models.

Our main contribution lies in laying a foundational dataset for the Portuguese-Emakhuwa language pair. This was a multifaceted endeavor that involved aggregating dispersed open-source texts, digitizing printed materials, and, most importantly, undertaking a large-scale manual translation project to generate the first significant parallel corpus in the news domain. This effort extended to the translation of established MT evaluation benchmarks, thereby creating a standardized framework for training and evaluating MT systems in this language pair.

With this in place, we systematically investigated several state-of-the-art methodologies to determine the most effective path forward for Portuguese-Emakhuwa MT. Our experimental analysis centered on three core approaches. Based on this foundation, we explored and evaluated several approaches to machine translation for Emakhuwa:

- **Transfer Learning:** In line with current research emphasizing the value of transfer learning in low-resource scenarios, we experimented with fine-tuning multilingual pretrained models. Our evaluation included both encoder–decoder and decoder-only architectures, and we compared general-purpose models with Africa-centric ones. The results confirmed the effectiveness of fine-tuning in significantly improving translation quality for Portuguese and Emakhuwa. Africa-centric models showed promise, suggesting that regionally contextualized pretraining can offer measurable benefits. An interesting finding was the potential of tokenization-free models for this language pair. The byte-level tokenization of models like byT5 proved highly effective at handling Emakhuwa's orthographic variation. Surprisingly, a relatively small ByT5-base model outperformed several larger, token-based models

such as Llama 3.2-3B, Gemma-3-4B, Phi3.5-mini, Qwen2.5-3B, and Qwen2.5-7B. Furthermore, our results underscore the parameter efficiency of encoder-decoder architectures. Encoder–decoder models, especially those optimized for translation such as NLLB and M2M-100, achieve a performamce comparable to or even higher than decoder-only models, while being more computationally efficient. This finding is critical for low-resource contexts, where computational power is often as scarce as data, making smaller models more practical for real-world adoption by the community. Finally, we observed the expected performance asymmetry: translation from Portuguese into the morphologically complex Emakhuwa language was significantly more challenging. This insight directly motivated follow-up experiments designed to improve these results.

- **Data Augmentation:** To mitigate the persistent issue of data scarcity, we investigated back-translation as a strategy for generating synthetic data. We employed the best-performing model from our transfer learning experiments, Gemma-2-8B, to translate monolingual Emakhuwa data into Portuguese, thereby creating a synthetic parallel corpus. This augmented dataset was then used to retrain our models. The results revealed consistent but modest performance gains across both translation directions, with the most noticeable improvements occurring in encoder-decoder architectures. In contrast, decoder-only models showed no significant positive improvements. This suggests that while back-translation can be beneficial, researchers must carefully consider the trade-off, as the process can introduce noise that may undermine the modest performance gains.

- **Leveraging External Linguistic Resources:** We explored the integration of external resources, including a glossary of loanword adaptations and a bilingual dictionary. Our methodology was twofold. First, to address the challenge of loanwords, we pre-processed the source text to incorporate glossary loanwords information by augmenting input sentences for encoder-decoder models and the translation prompt for decoder-only models. Second, focusing on the more challenging Portuguese-to-Emakhuwa direction, we enriched the input with contextual information from our bilingual dictionary. For encoder-decoder models, this was achieved by concatenating the input embeddings with embeddings of matched dictionary terms. For decoder-only models, we extended the prompt to include dictionary definitions for relevant source words. Our results demonstrated that integrating loanword information significantly improves performance in both directions, with a more pronounced effect when translating into Emakhuwa. This held for both model architectures. Furthermore, enriching the input with dictionary information pushed performance even higher for LLMs. Conversely, this strategy did not yield similar gains for encoder-decoder models. We hypothesize that this is attributable to our specific method of embedding concatenation, and we leave the exploration of more effective integration techniques for encoder-decoder models as an avenue for future work.

Future Work

While this thesis has made meaningful progress toward building MT systems for Portuguese-Emakhuwa, several avenues remain open for future research:

- **Scaling Data Creation:** There is a need to expand the dataset to cover additional domains such as health, education, and legal language. Because MT systems trained only on high-low data become biased and brittle. This also involves collecting source texts originally written in Emakhuwa and translating them into Portuguese. Low-high data introduces authentic sentence structures, idioms, and cultural references. Collaborations with local institutions and communities can help scale high-quality data collection through participatory methods.

- **Leveraging Speech for Data Expansion:** Another promising avenue for collecting natural and contemporary language data is the transcription of existing audio materials in Emakhuwa—such as oral narratives, community radio broadcasts, interviews, and other spoken recordings—followed by their translation into Portuguese. Since Emakhuwa is primarily a spoken language, leveraging spoken artifacts not only allows the dataset to capture genuine speech patterns and discourse structures, but also helps preserve rich linguistic features that are often missing from written sources, particularly those influenced by Portuguese. This approach therefore contributes both to improved language resources and to the documentation of authentic Emakhuwa usage.

- **Multilingual and Multitask Learning:** Future work could explore the development of multilingual models that include additional Mozambican and related Bantu languages. Such an approach would build on existing strategies that have sought to leverage cross-lingual transfer through transfer learning from large pretrained models in African languages—a direction already explored within broader community-driven initiatives like Masakhane. Integrating more languages with shared linguistic features has the potential to unlock even stronger cross-lingual transfer, thereby improving performance for all languages involved. In parallel, implementing multitask learning frameworks that jointly train for translation and complementary linguistic tasks (such as morphological analysis or named entity recognition) could lead to better performance.

- **Tokenization-Free Multilingual Models:** Building on our finding that byte-level models excel at handling the orthography of Emakhuwa, a logical next step is to train a tokenization-free multilingual model from scratch on a corpus of related Bantu languages. Such a model could learn to embed information directly from raw bytes, potentially offering superior handling of morphological complexities and code-switching across multiple related languages.

- **Investigating and Mitigating Gender Bias:** Future work should include a dedicated investigation into gender bias, particularly in Emakhuwa-Portuguese translations. This would

involve assessing possible bias arising from the fact that Emakhuwa is a gender-neutral language, and subsequently creating a challenging set to evaluate gender-specific phenomena and exploring debiasing techniques that are effective in a morphologically rich setting.

- **Efficient Integration of External Knowledge in Encoder-Decoder Models:** Our initial attempt to incorporate dictionary information by simply concatenating it with the input in encoder-decoder architectures left room for improvement. Future work should focus on exploring more effective integration techniques.

- **Evaluation Methodologies:** While automatic metrics are indispensable for model development, a more holistic evaluation framework is needed. Future research should incorporate taxonomy-based human evaluations to assess translation adequacy and fluency. Furthermore, developing culturally informed quality metrics that capture nuances specific to Emakhuwa is essential for ensuring that translated outputs are not just correct but also pragmatically and culturally appropriate.

In summary, this work represents a foundational step toward democratizing language technologies in Mozambique. By building resources, evaluating practical approaches, and identifying key challenges, we contribute not only to Emakhuwa MT, but also to broader efforts in equitable and inclusive NLP research.

References

Idris Abdulmumin, Michael Beukman, Jesujoba O. Alabi, Chris Emezue, Everlyn Asiko, Tosin Adewumi, Shamsuddeen Hassan Muhammad, Mofetoluwa Adeyemi, Oreen Yousuf, Sahib Singh, and Tajuddeen Rabiu Gwadabe. Separating grains from the chaff: Using data filtering to improve multilingual translation for low-resourced african languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, page 1001 – 1014. Association for Computational Linguistics, 2022. ISBN 978-195942929-6. URL https://aclanthology.org/2022.wmt-1.98/.

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. Correcting FLORES evaluation dataset for four African languages. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 570–578, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.44. URL https://aclanthology.org/2024.wmt-1.44/.

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse S. Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. Correcting flores evaluation dataset for four african languages. In Haddow B., Kocmi T., Koehn P., and Monz C., editors, *Conference on Machine Translation - Proceedings*, volume 2024-November, page 570 – 578. Association for Computational Linguistics, 2024b. ISBN 979-889176179-7. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85216122083&partnerID=40&md5=fbc2f8fe41ffb47934fdc83a00bd89f3.

Ife Adebara and Muhammad Abdul-Mageed. Towards afrocentric NLP for African languages: Where we are and where we can go. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.265. URL https://aclanthology.org/2022.acl-long.265/.

Ife Adebara, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. Cheetah: Natural language generation for 517 African languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12798–12823, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.691.

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen H. Muhammad, Guyo D. Jarso, Oreen Yousuf, Andre N. Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin A. Ajibade, Tunde Oluwaseyi Ajayi, Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire M. Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States, July 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.223. URL https://aclanthology.org/2022.naacl-main.223/.

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen R. Gwadabe, Freshia Sackey, Bonaventure F.P. Dossou, Chris Chinenye Emezue, Colin Leong, Michael Beukman, Shamsuddeen H. Muhammad, Guyo D. Jarso, Oreen Yousuf, Andre N. Rubungo, Gilles Hacheme, Eric P. Wairagala, Muhammad U. Nasir, Benjamin A. Ajibade, Tunde Oluwaseyi Ajayi, Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire M. Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for african news translation. In *NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*, page 3053 – 3070. Association for Computational Linguistics (ACL), 2022b. ISBN 978-195591771-1. doi: 10.18653/v1/2022.naacl-main.223. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85135861784&doi=10.18653%2fv1%2f2022.naacl-main.223&partnerID=40&md5=9cad5eff08ded268fb2b55017f142ad0.

David Ifeoluwa Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta R. Costa-Jussà, Jesse Dodge, Fahim Faisal, Natalia Fedorova, Christian Federmann, Francisco Guzmán, Sergey Koshelev, Jean Maillard, Vukosi Marivate, Jonathan Mbuya, Alexandre Mourachko, Safiyyah Saleem, Holger Schwenk, and Guillaume Wenzek. Findings of the wmt'22 shared task on large-scale machine translation evaluation for african languages. In *Conference on Machine Translation - Proceedings*, page 773 – 800. Association for Computational Linguistics, 2022c. ISBN 978-195942929-6. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85140869276&partnerID=40&md5=8068b731fc95fe70ea65e9dce81f6da7.

Lawrence B. Adewole, Adebayo O. Adetunmbi, Boniface K. Alese, Samuel A. Oluwadare, Oluwatoyin B. Abiola, and Olaiya Folorunsho. Automatic vowel elision resolution in yorùbá language. In *Conference of the South African Institute of Computer Scientists and Information Technologists 2020*, page 126 – 133. Association for Computing Machinery, 2020. ISBN

978-145038847-4. doi: 10.1145/3410886.3410917. URL https://doi.org/10.1145/3410886.3410917.

Željko Agić and Ivan Vulić. JW300: A wide-coverage parallel corpus for low-resource languages. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1310. URL https://aclanthology.org/P19-1310/.

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1388. URL https://aclanthology.org/N19-1388/.

Orevaoghene Ahia, Anuoluwapo Aremu, Diana Abagyan, Hila Gonen, David Ifeoluwa Adelani, Daud Abolade, Noah A. Smith, and Yulia Tsvetkov. Voices unheard: Nlp resources and models for yorùbá regional dialects. In Al-Onaizan Y., Bansal M., and Chen Y.-N., editors, *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, page 4392 – 4409. Association for Computational Linguistics (ACL), 2024. ISBN 979-889176164-3. doi: 10.18653/v1/2024.emnlp-main.251. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85214724560&doi=10.18653%2fv1%2f2024.emnlp-main.251&partnerID=40&md5=a25ff4dd7bf8e52b254065265e5f16cf.

Melissa Ailem, Jingshu Liu, and Raheel Qader. Encouraging neural machine translation to satisfy terminology constraints. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1450–1455, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.125. URL https://aclanthology.org/2021.findings-acl.125/.

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. Machine translation for african languages: Community creation of datasets and models in uganda. In *3rd Workshop on African Natural Language Processing*, 2022. URL https://openreview.net/forum?id=BK-z5qzEU-9.

Md Mahfuz Ibn Alam and Antonios Anastasopoulos. Language adapters for large-scale mt: The gmu system for the wmt 2022 large-scale machine translation evaluation for african languages shared task. In *Conference on Machine Translation - Proceedings*, page 1015 – 1033. Association for Computational Linguistics, 2022. ISBN 978-195942929-6. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85159725711&partnerID=40&md5=9c12fdd2e4c727dcd64f74f042510adf.

Felermino D. M. A. Ali, Andrew Caines, and Jaimito L. A. Malavi. Towards a parallel corpus of portuguese and the bantu language emakhuwa of mozambique, 2021. URL https://arxiv.org/abs/2104.05753.

Felermino Dario Mario Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. Detecting loanwords in Emakhuwa: An extremely low-resource Bantu language exhibiting significant borrowing

from Portuguese. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4750–4759, Torino, Italia, May 2024a. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.425.

Felermino Dario Mario Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. Expanding FLORES+ benchmark for more low-resource settings: Portuguese-emakhuwa machine translation evaluation. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 579–592, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.45. URL https://aclanthology.org/2024.wmt-1.45/.

Felermino D.M.A. Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. Building resources for Emakhuwa: Machine translation and news classification benchmarks. In Al-Onaizan Y., Bansal M., and Chen Y.-N., editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, page 14842 – 14857. Association for Computational Linguistics (ACL), 2024c. ISBN 979-889176164-3. doi: 10.18653/v1/2024.emnlp-main.824. URL https://aclanthology.org/2024.emnlp-main.824/.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*, 2019. URL https://arxiv.org/abs/1907.05019.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=Sy2ogebAW.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. Incorporating discrete translation lexicons into neural machine translation. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1162. URL https://aclanthology.org/D16-1162/.

Duygu Ataman, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. Machine translation in the era of large language models:a survey of historical and emerging problems. *Information*, 16(9), 2025. ISSN 2078-2489. doi: 10.3390/info16090723. URL https://www.mdpi.com/2078-2489/16/9/723.

Akinbowale Nathaniel Babatunde, Christiana Oluwakemi Abikoye, Abdulkarim Ayopo Oloyede, Roseline Oluwaseun Ogundokun, Afeez Adeshina Oke, and Hafsat Omolola Olawuyi. English to yoruba short message service speech and text translator for android phones. *International Journal of Speech Technology*, 24(4):979 – 991, 2021. ISSN 13812416. doi: 10.1007/s10772-021-09852-w. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85107263039&doi=10.1007%2fs10772-021-09852-w&partnerID=40&md5=fd0545d810367d3099dc9324b9482386.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1409.0473.

Nikolay Bogoychev and Pinzhen Chen. Terminology-aware translation with constrained decoding and large language model prompting. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.80. URL https://aclanthology.org/2023.wmt-1.80/.

Matthias Brenzinger, Arienne M. Dwyer, Tjeerd de Graaf, Colette Grinevald, Michael Krauss, Osahito Miyaoka, Nicholas Ostler, Osamu Sakiyama, María E. Villalón, Akira Y. Yamamoto, and Ofelia Zepeda. Unesco ad hoc expert group on endangered languages, 2003. URL https://ich.unesco.org/doc/src/00120-EN.pdf. Report by the Ad Hoc Expert Group on Endangered Languages.

Happy Buzaaba, Alexander Wettig, David Ifeoluwa Adelani, and Christiane Fellbaum. Lugha-llama: Adapting large language models for african languages. *arXiv preprint arXiv:2504.06536*, 2025.

Guanhua Chen, Yun Chen, Yong Wang, and Victor O. K. Li. Lexical-constraint-aware neural machine translation via data augmentation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 2021. ISBN 9780999241165.

Shangfeng Chen, Xiayang Shi, Pu Li, Jingjing Liu, and Yinlin Li. Refining translations with large language models: A constraint-aware iterative prompting approach. *Data Intelligence*, pages –, 2025. doi: https://doi.org/10.3724/2096-7004.di.2025.0130. URL http://www.sciengine.com/publisher/BeijingZhongkeJournalPublisingCo.Ltd./journal/DataIntelligence///10.3724/2096-7004.di.2025.0130.

Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. Joint training for pivot-based neural machine translation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3974–3980, 2017. doi: 10.24963/ijcai.2017/555. URL https://doi.org/10.24963/ijcai.2017/555.

Paddington Chiguvare and Christopher W. Cleghorn. Improving transformer model translation for low resource south african languages using bert. In *2021 IEEE Symposium Series on Computational Intelligence, SSCI 2021 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., 2021. ISBN 978-172819048-8. doi: 10.1109/SSCI50451.2021.9659923. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85125818143&doi=10.1109%2fSSCI50451.2021.9659923&partnerID=40&md5=ee3ed8b61071f095bf2bf49854fe6ade.

Everlyn Asiko Chimoto and Bruce A. Bassett. Very low resource sentence alignment: Luhya and swahili. In Ojha A.K., Ojha A.K., Liu C.-H., Vylomova E., Abbott J., Washington J., Oco N., Pirinen T.A., Malykh V., Malykh V., Logacheva V., and Zhao X., editors, *Proceedings - International Conference on Computational Linguistics, COLING*, page 1 – 8. Association for Computational Linguistics (ACL), 2022. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85159616718&partnerID=40&md5=f2a4591fd9558713b7e28b9a87f9d67d.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.914. URL https://aclanthology.org/2024.emnlp-main.914/.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

Marta R. Costa-Jussà and José A. R. Fonollosa. Character-based neural machine translation. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 357–361, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2058. URL https://aclanthology.org/P16-2058/.

Josep Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, Satoshi Enoue, Chiyo Geiss, Joshua Johanson, Ardas Khalsa, Raoum Khiari, Byeongil Ko, Catherine Kobus, Jean Lorieux, Leidiana Martins, Dang-Chuan Nguyen, Alexandra Priori, Thomas Riccardi, Natalia Segal, Christophe Servan, Cyril Tiquet, Bo Wang, Jin Yang, Dakun Zhang, Jing Zhou, and Peter Zoldan. Systran's pure neural machine translation systems, 2016. URL https://arxiv.org/abs/1610.05540.

Michael Han Daniel Han and Unsloth team. Unsloth, 2023. URL http://github.com/unslothai/unsloth.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. Training neural machine translation to apply terminology constraints. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1294. URL https://aclanthology.org/P19-1294/.

Shubhashis Roy Dipta and Sai Vallurupalli. Umbclu at semeval-2024 task 1: Semantic textual relatedness with and without machine translation. In Ojha A.K., Dohruoz A.S., Madabushi H.T., Da San Martino G., Rosenthal S., and Rosa A., editors, *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, page 1351 – 1357. Association for Computational Linguistics (ACL), 2024. ISBN 979-889176107-0. URL https://aclanthology.org/2024.semeval-1.195/.

Sibonelo Dlamini, Edgar Jembere, Anban Pillay, and Brett van Niekerk. Isizulu word embeddings. In *2021 Conference on Information Communications Technology and Society,*

ICTAS 2021 - Proceedings, page 121 – 126. Institute of Electrical and Electronics Engineers Inc., 2021. ISBN 978-172818080-9. doi: 10.1109/ICTAS50802.2021.9395011. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104579174&doi=10.1109%2fICTAS50802.2021.9395011&partnerID=40&md5=473db1f314a92ea4f1f69e49835f69ac.

Moussa Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory Conde, Kalo Mory Diané, Chris Piech, and Christopher Manning. Machine translation for nko: Tools, corpora, and baseline results. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 312–343, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.34. URL https://aclanthology.org/2023.wmt-1.34.

Kevin Duh, Paul McNamee, Matt Post, and Brian Thompson. Benchmarking neural and statistical machine translation on low-resource african languages. In Calzolari N., Bechet F., Blache P., Choukri K., Cieri C., Declerck T., Goggi S., Isahara H., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J., and Piperidis S., editors, *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, page 2667 – 2675. European Language Resources Association (ELRA), 2020. ISBN 979-109554634-4. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096584447&partnerID=40&md5=1b5352f878e586fafec28f4cbba1bd19.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig. Ethnologue: Languages of the world., 2022. URL http://www.ethnologue.com.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1045. URL https://aclanthology.org/D18-1045/.

Khalid N. Elmadani and Jan Buys. Neural machine translation between low-resource languages with synthetic pivoting. In Calzolari N., Kan M.-Y., Hoste V., Lenci A., Sakti S., and Xue N., editors, *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings*, page 12144 – 12158. European Language Resources Association (ELRA), 2024. ISBN 978-249381410-4. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195960580&partnerID=40&md5=a38102560f5ffe6e2aca9fc731b6923a.

AbdelRahim Elmadany, Ife Adebara, and Muhammad Abdul-Mageed. Toucan: Many-to-many translation for 150 african language pairs. In Ku L.-W., Martins A., and Srikumar V., editors, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, page 13189 – 13206. Association for Computational Linguistics (ACL), 2024a. ISBN 979-889176099-8. doi: 10.18653/v1/2024.findings-acl.781. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85205295194&doi=10.18653%2fv1%2f2024.findings-acl.781&partnerID=40&md5=a4248b3fead6f07b54926d534cfefcb5.

Abdelrahim Elmadany, Ife Adebara, and Muhammad Abdul-Mageed. Toucan: Many-to-many translation for 150 african language pairs. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13189–13206, 2024b.

Chris C. Emezue and Bonaventure F.P. Dossou. Mmtafrica: Multilingual machine translation for african languages. In *WMT 2021 - 6th Conference on Machine Translation, Proceedings*, page 398 – 411. Association for Computational Linguistics (ACL), 2021a. ISBN 978-195408594-7. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85127163605&partnerID=40&md5=a32613aed125d605081254fa715bad57.

Chris Chinenye Emezue and Bonaventure F. P. Dossou. MMTAfrica: Multilingual machine translation for African languages. In *Proceedings of the Sixth Conference on Machine Translation*, pages 398–411, Online, November 2021b. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.48.

Goodwill Erasmo Ndomba, Medard Edmund Mswahili, and Young-Seob Jeong. Tokenizers for african languages. *IEEE Access*, 13:1046–1054, 2025. doi: 10.1109/ACCESS.2024.3522285.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48, 2021. URL http://jmlr.org/papers/v22/20-1307.html.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.2. URL https://aclanthology.org/2024.wmt-1.2/.

{Jay P.} Gala, {Pranjal A.} Chitale, Raghavan AK, Varun Gumma, Sumanth Doddapaneni, {Kumar M.} Aswanth, {Janki Atul} Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, {Mitesh M.} Khapra, Raj Dabre, and Anoop Kunchukuttan. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*, 2023, 2023. ISSN 2835-8856. DBLP's bibliographic metadata records provided through http://dblp.org/search/publ/api are distributed under a Creative Commons CC0 1.0 Universal Public Domain Dedication. Although the bibliographic metadata records are provided consistent with CC0 1.0 Dedication, the content described by the metadata records is not. Content may be subject to copyright, rights of privacy, rights of publicity and other restrictions.

Tanja Gaustad, Cindy A. McKellar, and Martin J. Puttkammer. Dataset for siswati: Parallel textual data for english and siswati and monolingual textual data for siswati. *Data in Brief*, 54, 2024. ISSN 23523409. doi: 10.1016/j.dib.2024.110325. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85189931454&doi=10.1016%2fj.dib.2024.110325&partnerID=40&md5=48deed3861e96b60ae01bd0ae1e9383e.

Rachelle Gauton, Elsabe Taljard, Tirhani Mabasa, and Lufuno Netshitomboni. Translating technical (lsp) texts into the official south african languages: A corpus-based

investigation of translators' strategies. *Language Matters*, 39(2):148 – 180, 2008. doi: 10.1080/10228190802579569. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-77649111251&doi=10.1080%2f10228190802579569&partnerID=40&md5=d3c201a55f0b507c92d181c0db9185ea. Cited by: 2.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. Dictionary-based phrase-level prompting of large language models for machine translation, 2023. URL https://arxiv.org/abs/2302.07856.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538, 2022. doi: 10.1162/tacl_a_00474. URL https://aclanthology.org/2022.tacl-1.30/.

Chen Guanhua, Chen Yun, and Li Victor O.K. Lexically constrained neural machine translation with explicit alignment guidance. In *Proceedings of AAAI*, volume 35, pages 12630–12638, 2021.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024. doi: 10.1162/tacl_a_00683. URL https://aclanthology.org/2024.tacl-1.54/.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1632. URL https://aclanthology.org/D19-1632.

Frederick Gyasi and Tim Schlippe. Twi machine translation. *Big Data and Cognitive Computing*, 7(2), 2023. ISSN 25042289. doi: 10.3390/bdcc7020114. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85163613962&doi=10.3390%2fbdcc7020114&partnerID=40&md5=7a265c38dc5cd2da0f8f731c15ed1866.

Mika Hämäläinen and Khalid Alnajjar. A template based approach for training nmt for low-resource uralic languages - a pilot with finnish. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI '19, page 520–525, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450372619. doi: 10.1145/3377713.3377801. URL https://doi.org/10.1145/3377713.3377801.

Eva Hasler, Adrià de Gispert, Gonzalo Iglesias, and Bill Byrne. Neural machine translation decoding with terminology constraints. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 506–512, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2081. URL https://aclanthology.org/N18-2081/.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. Bitext mining using distilled sentence representations for low-resource languages. In Goldberg Y., Kozareva Z., and Zhang Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, page 2101 – 2112. Association for Computational Linguistics (ACL), 2022. doi: 10.18653/v1/2022.findings-emnlp.420. URL https://aclanthology.org/2022.findings-emnlp.154/. Cited by: 33; All Open Access, Green Open Access, Hybrid Gold Open Access.

Chris Hokamp and Qun Liu. Lexically constrained decoding for sequence generation using grid beam search. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1141. URL https://aclanthology.org/P17-1141/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR 2022*, April 2022. URL https://www.microsoft.com/en-us/research/publication/lora-low-rank-adaptation-of-large-language-models/.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. Improved lexically constrained decoding for translation and monolingual rewriting. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 839–850, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1090. URL https://aclanthology.org/N19-1090/.

John Hutchins. From first conception to first demonstration: the nascent years of machine translation, 1947–1954. a chronology. *Machine Translation*, 12(3):195–252, 1997. ISSN 1573-0573. doi: 10.1023/A:1007969630568. URL https://doi.org/10.1023/A:1007969630568.

Wenxiang Jiao, Zhaopeng Tu, Jiarui Li, Wenxuan Wang, Jen-Tse Huang, and Shuming Shi. Tencent's multilingual machine translation system for wmt22 large-scale african languages. In *Conference on Machine Translation - Proceedings*, page 1049 – 1056. Association for Computational Linguistics, 2022. ISBN 978-195942929-6. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85146908926&partnerID=40&md5=ccbe722fa0f91b05119d3b09d8515578.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.35. URL https://aclanthology.org/2024.wmt-1.35/.

Samta Kamboj, Sunil Kumar Sahu, and Neha Sengupta. Dentra: Denoising and translation pre-training for multilingual machine translation. In *Conference on Machine Translation - Proceedings*, page 1057 – 1067. Association for Computational Linguistics, 2022. ISBN 978-195942929-6. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85160576079&partnerID=40&md5=2d4ee50682f06c40867f4cb7ca61d31d.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. GlotLID: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=dl4e3EBz5j.

Mamadou K. Keita, Elysabhete Amadou Ibrahim, Habibatou Abdoulaye Alfari, and Christopher Homan. Feriji: A french-zarma parallel corpus, glossary & translator. In Fu X. and Fleisig E., editors, *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 4, page 1 – 9. Association for Computational Linguistics (ACL), 2024. ISBN 979-889176097-4. doi: 10.18653/v1/2024.acl-srw.1. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85204061797&doi=10.18653%2fv1%2f2024.acl-srw.1&partnerID=40&md5=9da79ffd96bb6139c9d52b75b8e67fd7.

Richard Kimera, DongNyeong Heo, Daniela N. Rim, and Heeyoul Choi. Data augmentation with back translation for low resource languages: A case of english and luganda. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '24, page 142–148, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400717383. doi: 10.1145/3711542.3711594. URL https://doi.org/10.1145/3711542.3711594.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://api.semanticscholar.org/CorpusID:6628106.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P17-4012.

Tom Kocmi and Ondřej Bojar. Trivial transfer learning for low-resource neural machine translation. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6325. URL https://aclanthology.org/W18-6325/.

Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada,

Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland, June 2023. European Association for Machine Translation. URL https://aclanthology.org/2023.eamt-1.19/.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, Mariya Shmatova, and Jun Suzuki. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.1. URL https://aclanthology.org/2023.wmt-1.1/.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.1. URL https://aclanthology.org/2024.wmt-1.1/.

Philipp Koehn. Statistical significance tests for machine translation evaluation. In Dekang Lin and Dekai Wu, editors, *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-3250/.

Philipp Koehn. *Neural Machine Translation*. Cambridge University Press, 2020. ISBN 1108497322.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Eduardo Blanco and Wei Lu, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012/.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. Madlad-400: A multilingual and document-level large audited dataset. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 67284–67296. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d49042a5d49818711c401d34172f9900-Paper-Datasets_and_Benchmarks.pdf.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. Madlad-400: a multilingual and document-level large

audited dataset. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

Guillaume Lample, Alexis Conneau, Laurent Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1711.00043.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378, 2017. doi: 10.1162/tacl_a_00067. URL https://aclanthology.org/Q17-1026/.

Senyu Li, Felermino D. M. A. Ali, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, Colin Cherry, and David Ifeoluwa Adelani. Evaluating wmt 2025 metrics shared task submissions on the ssa-mte african challenge set. In *Proceedings of the WMT 2025 Workshop at the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2025a.

Senyu Li, Felermino Dario Mario Ali, Jiayi Wang, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, Colin Cherry, and David Ifeoluwa Adelani. Evaluating WMT 2025 metrics shared task submissions on the SSA-MTE African challenge set. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 913–919, Suzhou, China, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-341-8. doi: 10.18653/v1/2025.wmt-1.65. URL https://aclanthology.org/2025.wmt-1.65/.

Senyu Li, Jiayi Wang, Felermino D. M. A. Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. Ssa-comet: Do llms outperform learned metrics in evaluating mt for under-resourced african languages?, 2025c. URL https://arxiv.org/abs/2506.04557.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*, 2020. URL https://arxiv.org/abs/2001.08210.

Chi-kiu Lo. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5358. URL https://aclanthology.org/W19-5358/.

Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. The multi-range theory of translation quality measurement: Mqm scoring models and statistical quality control, 2024. URL https://arxiv.org/abs/2405.16969.

Arle Richard Lommel, Aljoscha Burchardt, and Hans Uszkoreit. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK, November 28-29 2013. Aslib. URL https://aclanthology.org/2013.tc-1.6/.

Ângela Filipe Lopes and Maria da Graça L. Castro Pinto. Do ensino bilingue em moçambique: Elementos em jogo na sua implementação e desenvolvimento, 2017. URL https://ojs.letras.up.pt/index.php/LinguarumArena/article/download/3249/2933/8858.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. Chain-of-dictionary prompting elicits translation in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.55. URL https://aclanthology.org/2024.emnlp-main.55/.

Minh-Thang Luong and Christopher D. Manning. Achieving open vocabulary neural machine translation with hybrid word-character models. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1100. URL https://aclanthology.org/P16-1100/.

Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1002. URL https://aclanthology.org/P15-1002.

Hlaudi Daniel Masethe, Lawrence M Mothapo, Sunday O. Ojo, Pius A. Owolawi, Mosima Anna Masethe, and Fausto Giunchigilia. Machine translation for morphologically rich low-resourced south african languages. In Zuva T., Brown A., and Rikhotso M., editors, *Proceedings of 2024 4th International Multidisciplinary Information Technology and Engineering Conference, IMITEC 2024*, page 71 – 78. Institute of Electrical and Electronics Engineers Inc., 2024. ISBN 979-835038798-8. doi: 10.1109/IMITEC60221.2024.10850972. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85218357192&doi=10.1109%2fIMITEC60221.2024.10850972&partnerID=40&md5=82514887c6ac6d95c64d1aaafddbaa59.

Derguene Mbaye, Moussa Diallo, and Thierno Ibrahima Diop. Low-resourced machine translation for senegalese wolof language. *Lecture Notes in Networks and Systems*, 696 LNNS:243 – 255, 2024. ISSN 23673370. doi: 10.1007/978-981-99-3236-8_19. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85174734354&doi=10.1007%2f978-981-99-3236-8_19&partnerID=40&md5=182b87bd020e5be1f99660317b609b74.

Cindy A. McKellar and Martin J. Puttkammer. Dataset for comparable evaluation of machine translation between 11 south african languages. *Data in Brief*, 29, 2020. ISSN 23523409. doi: 10.1016/j.dib.2020.105146. URL https://www.scopus.com/inward/record.

uri?eid=2-s2.0-85078589371&doi=10.1016%2fj.dib.2020.105146&partnerID=40&md5=f5b1aea4db5750d7142517bdd15871d6.

Paul McNamee and Kevin Duh. An extensive exploration of back-translation in 60 languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8166–8183, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.518. URL https://aclanthology.org/2023.findings-acl.518/.

Chenggang Mi, Yating Yang, Lei Wang, Xi Zhou, and Tonghai Jiang. Toward better loanword identification in Uyghur using cross-lingual word embeddings. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle, editors, *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3027–3037, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1256/.

Elise Michon, Josep Crego, and Jean Senellart. Integrating domain terminology into neural machine translation. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.348. URL https://aclanthology.org/2020.coling-main.348/.

Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty. *Translation Quality Assessment: From Principles to Practice*. Springer, 2018. doi: 10.1007/978-3-319-91241-7. URL https://link.springer.com/book/10.1007/978-3-319-91241-7.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. Domain terminology integration into machine translation: Leveraging large language models. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.82. URL https://aclanthology.org/2023.wmt-1.82/.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.891. URL https://aclanthology.org/2023.acl-long.891/.

Ngessimo M. Mutaka and Pius Ngwa Tamanji. *An introduction to African linguistics*. Number 16 in Handbooks in Linguistics. Lincom Europa, Munich, 2000.

Pooja Nakamura, Kaitlin Carson, Dustin Davis, Nisha Rai, and Amy Todd. Language mapping study in mozambique, final report, 2017.

Abhijnan Nath, Sina Mahdipour Saravani, Ibrahim Khebour, Sheikh Mannan, Zihui Li, and Nikhil Krishnaswamy. A generalized method for automated multilingual loanword detection. In

Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4996–5013, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.442/.

António Miguel Ndapassoa, editor. *Glossários de Conceitos Políticos, Desportivos e Sociais*. Rádio Moçambique – E.P., 2015. Tiragem: 350 exemplares. Data de Impressão: Março de 2016. Registo nº 8593/RLIND/2015.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddee Hassan Muhammad, Salomon Kabongo, Salomey Osei, Sackey Freshia, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa, Mofe Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Jane Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkabir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Espoir Murhabazi, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Emezue, Bonaventure Dossou, Blessing Sibanda, Blessing Itoro Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, page 2144 – 2160. Association for Computational Linguistics (ACL), 2020. ISBN 978-195214890-3. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85101816736&partnerID=40&md5=18ea535b25a3b331b0d646e991db2c61.

Armindo Ngunga and Osvaldo Faquir. *Padronização da Ortografia de Línguas Moçambicanas: Relatório do III Seminário*. Centro de Estudos das Línguas Moçambicanas, 2012.

Armindo Ngunga, Carlos Manuel, David Langa, Inês Machungo, and Crisófia Langa da Câmara. *Padronização da Ortografia de Línguas Moçambicanas: Relatório do IV Seminário*. Centro de Estudos Africanos, 2022.

Jan Niehues. Continuous learning in neural machine translation using bilingual dictionaries. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 830–840, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.70. URL https://aclanthology.org/2021.eacl-main.70/.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. The zeno's paradox of 'low-resource' languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.983. URL https://aclanthology.org/2024.emnlp-main.983/.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler

Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia-Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No language left behind: Scaling human-centered machine translation. 2022.

NLLBTeam, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. Scaling neural machine translation to 200 languages. *Nature*, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07335-x. URL https://doi.org/10.1038/s41586-024-07335-x.

Odunayo Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. Afriteva: Extending "small data" pretraining approaches to sequence-to-sequence models. In Cherry C., Fan A., Foster G., Haffari G., Khadivi S., Peng N., Ren X., Shareghi E., and Swayamdipta S., editors, *DeepLo 2022 - 3rd Workshop on Deep Learning Approaches for Low-Resource NLP, Proceedings of the DeepLo Workshop*, page 126 – 135. Association for Computational Linguistics (ACL), 2022. ISBN 978-195591797-1. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85137557845&partnerID=40&md5=28a9eff7b256ae4ff02076f7aabf667e.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. AfroBench: How good are large language models on African languages? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.976. URL https://aclanthology.org/2025.findings-acl.976/.

Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. Better quality pre-training data and t5 models for African languages. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.11.

Abigail Oppong. Building a participatory data design approach to examine gender bias in english-twi machine translation. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394222. doi: 10.1145/3544549.3583942. URL https://doi.org/10.1145/3544549.3583942.

Miles Osborne. *Statistical Machine Translation*, pages 912–915. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_783. URL https://doi.org/10.1007/978-0-387-30164-8_783.

Hamed Joseph Ouily, Aminata Sabané, Delwende Eliane Birba, Rodrique Kafando, Abdoul-Kader Kabore, and Tégawendé F. Bissyandé. A low-resource language translation: French to mooré. *Communications in Computer and Information Science*, 2462 CCIS:405 – 412, 2025. ISSN 18650929. doi: 10.1007/978-3-031-88226-5_30. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-105003861960&doi=10.1007%2f978-3-031-88226-5_30&partnerID=40&md5=9f9802ef6456a13311df1162bc197c0d.

Chester Palen-Michel, June Kim, and Constantine Lignos. Multilingual open text release 1: Public domain news in 44 languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2080–2089, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.224.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Lucas Francisco Amaral Orosco Pellicer, Taynan Maier Ferreira, and Anna Helena Reali Costa. Data augmentation techniques in natural language processing. *Applied Soft Computing*, 132:109803, 2023. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2022.109803. URL https://www.sciencedirect.com/science/article/pii/S1568494622008523.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049/.

Matt Post and David Vilar. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1119. URL https://aclanthology.org/N18-1119/.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11), February 2023. ISSN 0360-0300. doi: 10.1145/3567592. URL https://doi.org/10.1145/3567592.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL https://aclanthology.org/2020.emnlp-main.213/.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.52/.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022b. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.60/.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.12. URL https://aclanthology.org/2024.wmt-1.12/.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. Afromt: Pretraining strategies and reproducible benchmarks for translation of 8 african languages. In *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, page 1306 – 1320. Association for Computational Linguistics (ACL), 2021. ISBN 978-195591709-4. doi: 10.18653/v1/2021.emnlp-main.99. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85127452016&doi=10.18653%2fv1%2f2021.emnlp-main.99&partnerID=40&md5=d2d3b1cfd0922f58cdd1c8b1eb7e5e43.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.478. URL https://aclanthology.org/2020.emnlp-main.478.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. Chatgpt mt: Competitive for high- (but not low-) resource languages. In *Conference on*

Machine Translation - Proceedings, page 439 – 465. Association for Computational Linguistics, 2023. ISBN 979-889176041-7. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85179128125&partnerID=40&md5=c20b48eeee59e59137e542874f8c3a78.

Fábio Bonfim Duarte Ronaldo Rodrigues de Paula. Diversidade linguística em moçambique. *Kadila: culturas e ambientes - Diálogos Brasil-Angola: culturas e ambientes - Diálogos Brasil-Angola*, 1(1):343 – 362, 2016. ISSN 9788580392111. doi: 10.5151/9788580392111-19. URL openaccess.blucher.com.br/article-details/diversidade-lingustica-em-mocambique-20168.

Sebastian Ruder. *Neural Transfer Learning for Natural Language Processing*. PhD thesis, National University of Ireland, Galway, 2019.

Sakayo Toadoum Sari, Angela Fan, and Lema Logamou Seknewna. Ngambay-french neural machine translation (sba-fr). In Gutierrez R.L., Escribe M., Mihajlov T., Kunilovskaya M., and Mitkov R., editors, *NLP4TIA 2023 - 1st Workshop on Natural Language Processing Tools and Resources for Translation and Interpreting Applications, associated with the 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023 - Proceedings*, page 32 – 38. Incoma Ltd, 2023. ISBN 978-954452091-5. doi: 10.26615/978-954-452-091-5_006. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85185004898&doi=10.26615%2f978-954-452-091-5_006&partnerID=40&md5=86d304a6b45005f8187c4f049b6e49d6.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704/.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *CoRR*, abs/1511.06709, 2015. URL http://arxiv.org/abs/1511.06709.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016a. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL https://aclanthology.org/P16-1009/.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016b. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 35–40, 2016c.

Iyanuoluwa Shode, David Ifeoluwa Adelani, Jing Peng, and Anna Feldman. Nollysenti: Leveraging transfer learning and machine translation for nigerian movie sentiment classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 2, page 986 – 998. Association for Computational Linguistics (ACL), 2023. ISBN 978-195942971-5. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-86000189144&partnerID=40&md5=f3d2220a0587226c794204e8e11a2cee.

Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. Assessing digital language support on a global scale. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.379/.

Nomsa J. Skosana and Respect Mlambo. A brief study of the autshumato machine translation web service for south african languages. *Literator*, 42(1), 2021. ISSN 02582279. doi: 10.4102/lit.v42i1.1766. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85118660550&doi=10.4102%2flit.v42i1.1766&partnerID=40&md5=d69b77b31362009a64db294849f923cc.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12 2006. Association for Machine Translation in the Americas. URL https://aclanthology.org/2006.amta-papers.25/.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. Code-switching for enhancing NMT with pre-specified translation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 449–459, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1044. URL https://aclanthology.org/N19-1044/.

Kai Song, Kun Wang, Heng Yu, Yue Zhang, Zhongqiang Huang, Weihua Luo, Xiangyu Duan, and Min Zhang. Alignment-enhanced transformer for constraining nmt with pre-specified translations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8886–8893, Apr. 2020. doi: 10.1609/aaai.v34i05.6418. URL https://ojs.aaai.org/index.php/AAAI/article/view/6418.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press.

Taofik O. Tafa, Siti Zaiton Mohd Hashim, Mohd Shahizan Othman, Hitham Alhussian, Maged Nasser, Said Jadid Abdulkadir, Sharin Hazlin Huspi, Sarafa O. Adeyemo, and Yunusa Adamu Bena. Machine translation performance for low-resource languages: A systematic literature review. *IEEE Access*, 13:72486–72505, 2025. doi: 10.1109/ACCESS.2025.3562918.

Shaomu Tan and Christof Monz. ReMedy: Learning machine translation evaluation from human preferences with reward modeling. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4370–4387, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.217. URL https://aclanthology.org/2025.emnlp-main.217/.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning, 2020. URL https://arxiv.org/abs/2008.00401.

Gabriel Tanzer, Mirac Suzgun, Emiel Visser, Dan Jurafsky, and Luke Melas-Kyriazi. A benchmark for learning to translate a new language from one grammar book. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*, Kigali, Rwanda, May 2023. URL https://arxiv.org/abs/2309.16575.

Allahsera Auguste Tapo, Kevin Assogba, Christopher M Homan, M. Mustafa Rafique, and Marcos Zampieri. Bayelemabaga: Creating resources for Bambara NLP. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12060–12070, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.602. URL https://aclanthology.org/2025.naacl-long.602/.

Brian Thompson and Matt Post. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.8. URL https://aclanthology.org/2020.emnlp-main.8/.

Brian Thompson and Matt Post. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online, November 2020b. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.67/.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raul Vazquez, and Sami Virpioja. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755, 2023. ISSN 1574-0218. doi: 10.1007/s10579-023-09704-w.

Atnafu Lambebo Tonja, Bonaventure FP Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Aremu Anuoluwapo, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, et al. Inkubalm: A small language model for low-resource african languages. *arXiv preprint arXiv:2408.17024*, 2024.

Kosei Uemura, Mahe Chen, Alex Pejovic, Chika Maduabuchi, Yifei Sun, and En-Shiun Annie Lee. AfriInstruct: Instruction tuning of African languages for diverse tasks. In Yaser Al-Onaizan,

Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13571–13585, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.793. URL https://aclanthology.org/2024.findings-emnlp.793/.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya model: An instruction finetuned open-access multilingual language model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.845. URL https://aclanthology.org/2024.acl-long.845/.

J. Van der Wal. *Word Order and Information Structure in Makhuwa-Enahara*. LOT international series. LOT, 2009. ISBN 9789078328902. URL https://books.google.pt/books?id=_BkqAQAAIAAJ.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. URL https://arxiv.org/abs/1706.03762.

Pavanpankaj Vegi, J. Sivabhavani, Biswajit Paul, K.R. Prasanna Kumar, and Chitra Viswanathan. Anvita-african: A multilingual neural machine translation system for african languages. In *Conference on Machine Translation - Proceedings*, page 1090 – 1097. Association for Computational Linguistics, 2022. ISBN 978-195942929-6. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85160538315&partnerID=40&md5=71a640a4eee4c1784aff1454a0d9025f.

Cécile B. Vigouroux and Salikoko S. Mufwene. *Globalization and Language Vitality: Perspectives from Africa*. Continuum, 2008.

Jiayi Wang, David Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Mohamed, Hassan Ayinde, Oluwabusayo Awoyomi, Lama Alkhaled, Sana Al-azzawi, Naome Etori, Millicent Ochieng, Clemencia Siro, Njoroge Kiragu, Eric Muchiri, Wangari Kimotho, Toadoum Sari Sakayo, Lyse Naomi Wamba, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Iro, Saheed Abdullahi, Stephen Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Ogbu, Sam Ochieng', Verrah Otiende, Chinedu Mbonu, Yao Lu, and Pontus Stenetorp. AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico, June 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.334. URL https://aclanthology.org/2024.naacl-long.334.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed,

Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed, Ayinde Hassan, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-Azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Samuel Njoroge, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdullahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Raphael Ogbu, Sam Brian, Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoum Sari, Yao Lu, and Pontus Stenetorp. Afrimte and africomet: Enhancing comet to embrace under-resourced african languages. In Duh K., Gomez H., and Bethard S., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, volume 1, page 5997 – 6023. Association for Computational Linguistics (ACL), 2024b. ISBN 979-889176114-8. doi: 10.18653/v1/2024.naacl-long.334. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85199581086&doi=10.18653%2fv1%2f2024.naacl-long.334&partnerID=40&md5=81d947ae5e3b9acc07c709d69a8c2b3b. Cited by: 7; All Open Access, Green Open Access.

Jiayi Wang, David Ifeoluwa Adelani, and Pontus Stenetorp. Evaluating wmt 2024 metrics shared task submissions on afrimte (the african challenge set). In Haddow B., Kocmi T., Koehn P., and Monz C., editors, *Conference on Machine Translation - Proceedings*, volume 2024-November, page 505 – 516. Association for Computational Linguistics, 2024c. ISBN 979-889176179-7. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85216076559&partnerID=40&md5=c3353f01830b1c5c04125e50cf577f67.

Mingxuan Wang and Lei Li. Pre-training methods for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics: Tutorials*, 2021. URL https://aclanthology.org/2021.acl-tutorials.4/.

Shuo Wang, Zhixing Tan, and Yang Liu. Integrating vectorized lexical constraints for neural machine translation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7063–7073, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.487. URL https://aclanthology.org/2022.acl-long.487/.

H. Ekkehard Wolff. Language ideologies and the politics of language in post-colonial africa. *Stellenbosch Papers in Linguistics Plus*, 51(1):1–22, 2017. doi: 10.5842/51-0-701. URL https://journals.co.za/doi/abs/10.5842/51-0-701.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=farT6XXntP.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy,

Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41/.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022. doi: 10.1162/tacl_a_00461. URL https://aclanthology.org/2022.tacl-1.17/.

Hans M. Zell. Publishing in african languages: a review of the literature. *African Research and Documentation*, 132:3–90, 2018. doi: 10.1017/S0305862X00024031.

Huaao Zhang, Qiang Wang, Bo Qin, Zelin Shi, Haibo Wang, and Ming Chen. Understanding and improving the robustness of terminology constraints in neural machine translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6029–6042, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.332. URL https://aclanthology.org/2023.acl-long.332/.

Jiajun Zhang and Chengqing Zong. Bridging neural machine translation and bilingual dictionaries, 2016. URL https://arxiv.org/abs/1610.07272.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.925. URL https://aclanthology.org/2024.findings-acl.925/.

Miaoran Zhang, Mingyang Wang, Jesujoba O. Alabi, and Dietrich Klakow. Aadam at semeval-2024 task 1: Augmentation and adaptation for multilingual semantic textual relatedness. In Ojha A.K., Dohruoz A.S., Madabushi H.T., Da San Martino G., Rosenthal S., and Rosa A., editors, *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, page 800 – 810. Association for Computational Linguistics (ACL), 2024b. ISBN 979-889176107-0. URL https://aclanthology.org/2024.semeval-1.114/.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.43. URL https://aclanthology.org/2023.wmt-1.43/.

Appendix A

Appendix

A.1 Quality Assessment



Figure A.1: Annotation Tool User Interface.

A.2 Base Models

Table A.1: Base models hugging-face checkpoints.

Models	Size	Model name
mT5	580M	`google/mt5-base`
AfriMT5	580M	`masakhane/afri-mt5-base`
MT0	582M	`bigscience/mt0-base`
AfriVeTa	428M	`castorini/afriteva_v2_base`
byT5	580M	`google/byt5-base`
AfriByT5	580M	`masakhane/afri-byt5-base`
Cheetah	582M	`UBC-NLP/cheetah-base`
Tucan	582M	`UBC-NLP/toucan-base`
mBART50	610M	`facebook/mbart-large-50-many-to-many-mmt`
M2M-100	418M	`facebook/m2m100_418M`
NLLB	600M	`facebook/nllb-200-distilled-600M`
OPUS	240M	`Helsinki-NLP/opus-mt-tc-bible-big-mul-deu_eng_fra_por_spa`
InkubaLM	0.4B	`lelapa/InkubaLM-0.4B`
Tower	7B	`Unbabel/TowerInstruct-7B-v0.1`
Qwen2.5	3B	`unsloth/Qwen2.5-3B-Instruct`
Qwen2.5	7B	`unsloth/Qwen2.5-7B-Instruct`
Llama 3.1	8B	`unsloth/Llama-3.2-3B-Instruct`
Llama 3.2	3B	`unsloth/llama-3.1-8b-instruc`
Phi3.5-mini	3.8B	`unsloth/Phi-3.5-mini-instruct`
Gemma-3	4B	`unsloth/gemma-3-4b-it`
Gemma-2	9B	`unsloth/gemma-2-9b-it`

A.3 Hyperparameters

Table A.2: Text-to-Text models Hyperparameters for fine-tuning.

Hyperparameter	Value
Epochs	3
Learning Rate	5e-05
Optimizer	`Adamw with betas=(0.9,0.999) and epsilon=1e-08`

Table A.3: NLLB Hyperparameters for fine-tuning.

Hyperparameter	**Value**
Epochs	3
Learning Rate	1e-4
Optimizer	`Adafactor, relative_step=False, scale_parameter=False, clip_threshold=1.0 weight_decay=1.3`

Table A.4: M2M Hyperparameters for fine-tuning.

Hyperparameter	**Value**
Epochs	3
Learning Rate	3e-05
Optimizer	`Adamw with betas=(0.9,0.98) and epsilon=1e-06`

Table A.5: mBART50 Hyperparameters for fine-tuning.

Hyperparameter	**Value**
Epochs	3
Learning Rate	3e-05
Optimizer	`Adamw with betas=(0.9,0.98) and epsilon=1e-06, Scheduler=polynomial`

Table A.6: OPUS Hyperparameters for fine-tuning.

Hyperparameter	**Value**
Epochs	3
Learning Rate	2e-05
Optimizer	`Adam with betas=(0.9,0.98) and epsilon=1e-08, Scheduler=linear`

Table A.7: InkubaLM Hyperparameters for fine-tuning.

Hyperparameter	**Value**
Epochs	3
Learning Rate	2e-05
Optimizer	`Adam`

Table A.8: LLMs Hyperparameters for fine-tuning (with Unsloth).

Hyperparameter	**Value**
Epochs	1
Learning Rate	2e-4
Lora Alpha	16
Lora Rank	16
Optimizer	`Adamw, Scheduler=linear`

Table A.9: TowerInstruct-v0.1 Hyperparameters for fine-tuning.

Hyperparameter	**Value**
Epochs	1
Learning Rate	3e-05
Lora Alpha	32
Lora Rank	16
lora dropout	0.05
Optimizer	`Adam with betas=(0.9, 0.999) and epsilon=1e-08, Scheduler=cosine`