Preparing the Vuk'uzenzele and ZA-gov-multilingual South African multilingual corpora

Richard Lastrucci[1], **Isheanesu Dzingirai**[1], **Jenalea Rajab**[2], **Andani Madodonga**[1], **Matimba Shingange**[1], **Daniel Njini**[1], **Vukosi Marivate**[1,3]
[1]Department of Computer Science, University of Pretoria
[2]School of Computer Science and Applied Mathematics, University of the Witwatersrand
[3]Lelapa AI
richard.lastrucci@tuks.co.za, ishe.dzingirai@gmail.com, jenalea.rajab@gmail.com, andanim412@gmail.com, mrosslyns@gmail.com, vukosi.marivate@cs.up.ac.za

Abstract

This paper introduces two multilingual government themed corpora in various South African languages. The corpora were collected by gathering the South African Government newspaper (Vuk'uzenzele), as well as South African government speeches (ZA-gov-multilingual), that are translated into all 11 South African official languages. The corpora can be used for a myriad of downstream NLP tasks. The corpora were created to allow researchers to study the language used in South African government publications, with a focus on understanding how South African government officials communicate with their constituents.

In this paper we highlight the process of gathering, cleaning and making available the corpora. We create parallel sentence corpora for Neural Machine Translation (NMT) tasks using Language-Agnostic Sentence Representations (LASER) embeddings. With these aligned sentences we then provide NMT benchmarks for 9 indigenous languages by fine-tuning a massively multilingual pre-trained language model.

1 Introduction

The advancement of Natural Language Processing (NLP) research in Africa is impeded due to the scarcity of data for training models for various NLP tasks (Nekoto et al., 2020) as well as availability of benchmarks and ways to reproduce them (Martinus and Abbott, 2019). For many South African languages there are still challenges finding easily available textual datasets (Marivate et al., 2020) even if there are many speakers for those languages (Ranathunga and de Silva, 2022). There is a need to focus on development of local language (Joshi et al., 2020) NLP resources.

This paper builds upon the work of Autshumato (Groenewald and Fourie, 2009; Groenewald and du Plooy, 2010) by creating automatically aligned parallel corpora from government textual data in the 11 official languages of South Africa. While the Autshumato project focused on creating Machine Translation tools for five indigenous languages, the resulting corpora lacked information about its origin or context, limiting its usefulness for other NLP tasks such as categorisation, topic modelling over time and other tasks that require contextual information of the content. Our approach provides more comprehensive data that can support a wider range of NLP applications.

Our belief is that there is a significant opportunity to create a more user-friendly data collection process that can be easily maintained and provide extraction tools for others. It is essential to preserve the data source and structure it in a way that enables extensions. Our goal is to enhance Neural Machine Translation (NMT) resources in the government data domain by including all indigenous languages and broadening the translation directions beyond English as the source language. Additionally, we recognise the importance of providing aligned data across all South African languages beyond English.

Further, this paper introduces parallel corpora datasets in the 11 official languages of South Africa, created from text data obtained from the government. These datasets are designed to facilitate the development of NMT models. The corpora are automatically aligned, and are expected to serve as a valuable resource for researchers and practitioners working in the field of machine learning.

The parallel corpora were generated using LASER encoders (Schwenk and Douze, 2017), facilitating the one-to-one alignment of tokenised sentence data. The data was sourced from credible sources such as newspapers and academic journals and covers diverse topics including health, finance, and politics.

We also provide NMT benchmarks for the parallel corpora by fine-tuning a massively multilingual model (M2M100 (Fan et al., 2021)) building on the work of Adelani et al. (2022).

This paper is structured as follows. In the following section, we detail the datasets that we have compiled, including their compilation methodology and the information they contain. We then describe how we have aligned and created parallel corpora using these datasets. The subsequent section presents our NMT experiments and provides an analysis of the results obtained. Finally, we conclude the paper with our findings and make recommendations for future research.

2 Main Datasets

2.1 The Vuk'uzenzele South African Multilingual Corpus

The Vuk'uzenzele dataset was constructed from editions of the South African government magazine Vuk'uzenzele[1]. Being a magazine, the text focuses mainly on current events, politics, entertainment, and other topics related to a magazine publication. The Vuk'uzenzele dataset provides a comprehensive view of the language and topics of discussion in South Africa during the respective period, giving researchers insight into the history and culture of South Africa. The Vuk'uzenzele dataset is thus a rich resource for any researcher wanting to analyse South African politics, current events, and popular culture.

2.1.1 Creation of Vuk'uzenzele

The raw Vuk'uzenzele data is scraped from PDF editions of the magazine. The main Vuk'uzenzele edition is in the English language. Only a few of these English articles are translated into the other 10 official South African languages (Afrikaans, isiNdebele, isiXhosa, IsiZulu, Sepedi, Sesotho, siSwati, Tshivenda, Xitsonga and Setswana). As such, we created a pipeline to identify which articles should be extracted from each language pdf from a specific edition. Individual articles were extracted and placed into text files. The extracted text files still have some challenges due to PDF extraction. To clean it, a team member goes through each extracted text file and formats it as follows:

- Line 1: Title of article (*in language*)
- Line 2: *empty line*
- Line 3: Author of article (*if available. If not, defaults to Vukuzenzele Unnamed*)
- Line 4: *empty line*
- Line 5-end: *body of article*

The data is easier to analyse and visualise after being manually reviewed. The labour-intensive effort was necessary to provide a comprehensive and meaningful analysis of the magazine's content. The time-consuming process of manual review and extraction was ultimately worth it, as it provides an opportunity to create a deeper understanding of the content within Vuk'uzenzele. As of writing we have *53* editions of the newspaper spanning *January 2020 to July 2022*. More additions will be added in time by the team. Automations have been built to download and archive the PDFs, however manual effort is still required to extract and identify translated articles. The dataset, code and automated scrapers are available at at https://github.com/dsfsi/vukuzenzele-nlp and Zenodo[2] (Marivate et al., 2023a). We make it available in a format that allows other researchers to extend, remix and add onto it (*CC-4.0-BY-SA licence for data and MIT License for code*).

2.2 The ZA-Gov Multilingual South African corpus

The ZA-Gov Multilingual corpus dataset was constructed from the speeches following cabinet meetings of the SA government. As such, the dataset carries a variety of topics including energy, labour, service delivery, crime, COVID, international relations, the environment, and government affairs such as government appointments, cabinet decisions, etc. This provides an eye into the workings of the South African government and how it has dealt with various challenges, both internal and external.

2.2.1 Creation of ZA-Gov-multilingual

The raw ZA-Gov Multilingual data is scraped from the the South African government website (https://www.gov.za/), where all cabinet statements, and their translations, are posted. The data was extracted and structured into a JSON format. The JSON payload for each speech records:

- Date,
- Datetime,
- Title (*in English*),
- Url (*top url for speech*),

[1] https://www.vukuzenzele.gov.za/

[2] https://doi.org/10.5281/zenodo.7598539

- Language payload for each language (*eng, afr, nbl, xho, zul, nso, sep, tsn, ssw, ven, tso*).
  - Title (*in language*),
  - Text (*in language*),
  - Url (*for the translation*).

This structure makes it convenient for researchers and analysts to perform various natural language processing, data mining and machine learning tasks such as sentiment analysis, topic modelling, categorisation, language modelling and more. For instance, through sentiment analysis and text mining, analysts can investigate opinions of cabinet members' statements and track the evolution of these topics over time. As of writing, the dataset contains *162* cabinet statements spanning *2 May 2013 to 1 December 2022*. The dataset will update automatically when new, *translated*, statements are available on the gov.za website. The dataset, code and automated scrapers are available at at https://github.com/dsfsi/gov-za-multilingual and Zenodo[3] (Marivate et al., 2023b). We make it available in such a way that other researchers can extend, remix and add onto it (*CC-4.0-BY-SA licence for data and MIT License for code*).

3 The corpora as a foundation for other NLP tasks and further study

In addition to supporting the creation of NMT models (discussed in the proceeding section), our datasets have the potential to serve as a foundation for many other NLP tasks beyond translation. We believe that these datasets will be a valuable resource for the study of South African government communication, and that it can be used for direct creation of multilingual document categorisation/classification (Schwenk and Li, 2018), simplification (Lu et al., 2021; Siddharthan, 2014; Martin et al., 2022), entity extraction (Tedeschi et al., 2021; Chen et al., 2018; Pappu et al., 2017; Emelyanov and Artemova, 2019), and other NLP tasks. To further extend the dataset's usefulness, we recommend looking at work such as the Parallel Meaning Bank (Abzianidze et al., 2017), which can act as an inspiration for transferring knowledge from one language to another and provide new benchmarks that may be helpful for Southern African languages beyond South Africa. We envision these datasets as a starting point for further research in the area of multilingual NLP for South African and African languages.

4 Methods for Processing and Compilation

The datasets are a two way parallel corpus of the 11 official languages of South Africa, which are listed in Table 1 with their corresponding ISO 639-2code. The datasets contain texts written in the official languages of South Africa, including Afrikaans, English, isiNdebele, isiXhosa, isiZulu, siSwati, Sepedi, Xitsonga, siSwati, Tshivenda, and Setswana. As such, there are 55 ways of combining these languages into pairs, producing 55 distinct corpora in each of the datasets. The dataset uses the ISO 639-2 language codes in its naming convention, i.e., 'aligned-afr-zul.csv'. By nature of compilation, some datasets have more observations than others, which could lead to varying results, i.e., if used for NMT, then a better model can be produced for two languages from a dataset with more observations as opposed to one with fewer observations. This compilation of data allows for further exploration into the complexities of South African language and discourse, creating a multi-dimensional representation of how language is used and interpreted in South Africa. Through these datasets, the range of language usage in South Africa can be explored, providing insights into how different languages interact.

Table 1: Language List with ISO 639-2 codes

Name	Code
isiZulu	zul
isiXhosa	xho
Afrikaans	afr
English	eng
Sepedi	nso
Setswana	tsn
Xitsonga	tso
Sesotho	sot
siSwati	ssw
Tshivenda	ven
isiNdebele	nbl

4.1 Preprocessing

Preprocessing was required to refine the raw scraped data prior to LASER encoding and alignment. The preprocessing steps are listed below

[3] https://doi.org/10.5281/zenodo.7635167

and differ slightly as the source data and method of scraping has an outcome on the data obtained. For example, ZA-Gov-Multilinguals involve a lot of nested points, i.e., *2.1.2*, which needed to be removed, while in contrast the Vuk'uzenzele data uses bullet points for listing.

4.1.1 Vuk'uzenzele

The raw text from the collected data was preprocessed in the following way:

- The text was set to lowercase.
- Hyphens and bullet points were removed.
- Double spaces, tabs, and newlines were replaced with a single space.
- The standard apostrophe, i.e., ", took the place of Unicode apostrophes.

4.1.2 ZA-Gov-multilingual

The raw text from the collected data was preprocessed in the following way:

- Removing the dots (or single- or multi-digit numbers) that began a line
- Inserting a period after a series of numbers in the format $x \cdot y$.
- Adding a period after a string of numbers in the format x.
- Replacing a sequence of punctuation marks, such as a period, colon, semi-colon, or a combination of these, followed by a letter with a single period.

4.2 Corpora Alignment

Once preprocessed, the text was passed to the NLTK tokeniser "punkt" which returns a vector of sentence tokens. The *n* tokenised sentences were sent to LASER encoder which encodes it into *n* sentence vectors, each of length 1024. The sentence vectors are compared and a cosine similarity algorithm was performed to produce a score from 0 to 1 on the similarity of the two vectors as described in section 4.2.1.

4.2.1 LASER Encodings

In order to compare the text for similarity, LASER encoders were utilised. LASER, which stands for Language-Agnostic Sentence Representations, is a research project by Facebook AI Research. LASER generates sentence representations by encoding sentences into a vector. The vectors serve as a machine representation of the sentence. The vectors can be compared using cosine similarity which outputs a score between zero and one. Cosine scores closer to one indicate high similarity. This score is recorded in the LASER datasets (available in the dataset repository). The observations present in each dataset with a score above 0.65, or 65% similarity, are listed in the following tables 2 and 3. Entire tables featuring the number of observations present in all datasets are featured on the READMEs of the dataset repos, https://github.com/dsfsi/gov-za-multilingual for ZA-Gov-Multilingual and https://github.com/dsfsi/vukuzenzele-nlp for Vuk'uzenzele.

Table 2: Top ten datasets with the most observations with a cosine score greater than or equal to 0.65 in Vuk'uzenzele.

Language pair	No. of observations in Vuk'uzenzele
ssw-xho	2,202
ssw-zul	2,183
xho-zul	2,102
nso-xho	2,081
nso-tso	2,071
ssw-tso	2,034
nso-ssw	2,021
tsn-tso	2,020
tsn-xho	2,009
tso-xho	2,009

Table 3: Top ten datasets with the most observations with a cosine score greater than or equal to 0.65

Language pair	No. of observations in ZAgov Multilingual
nbl-ven	18,984
nso-ssw	18,697
zul-ssw	18,563
xho-ssw	18,387
xho-zul	18,145
xho-nso	18,110
xho-tso	17,954
ssw-tso	17,880
zul-tso	17,789
zul-nso	17,630

4.3 Postprocessing

For the LASER datasets the source sentence, target sentence, and cosine score for the aligned data was written to a csv file with the naming convention 'aligned-{src_lang_code}-{tgt_lang_code}.csv', i.e. 'aligned-afr-zul.csv'. Refer to the language list in 1 for language codes used in naming the datasets.

For the simple aligned datasets the source sentence and the target sentence were written to a csv file with the same naming structure as the LASER datasets.

5 NMT Benchmarks

Minimal aligned sentence corpora, for low-resourced African languages, hinder the quality of NMT models trained from scratch (Martinus and Abbott, 2019; Nekoto et al., 2020; Adelani et al., 2022). Recently Adelani et al. (2022) approached this problem by fine-tuning massively multilingual models, including the M2M100 model (Fan et al., 2021), on a small number of aligned sentences. The M2M100 model is a Many-to-Many non-English centric language model trained to translate directly between 100 languages, including five South African official languages (Fan et al., 2021). Adelani et al. (2022) demonstrated how to effectively leverage this model for small quantities of data, to create NMT systems for languages and domains not included in pre-training.

Building on their work, we create baseline translation benchmarks for the Vuk'uzenzele and ZA-gov-multilingual datasets, in the government publication domain, by fine-tuning the M2M100 model. To provide our results in context and for comparison purposes we also fine-tune the M2M100 model on subsets of the existing Autshumato parallel corpora obtained from the South African Centre for Digital Language Resources (McKellar, (2021,2,2,2,0,2) (`https://sadilar.org`). We focus our efforts on providing NMT benchmarks for the low resource African languages in the datasets, as such Afrikaans translations are not included due to the relatively high availability of digital datasets in this language, and we leave this for future work.

5.1 Pre-processing

The aligned datasets were processed to remove duplicate and conflicting translations (in both the source and target sentences) then shuffled to remove any order bias before the train, test and dev set were created. The data splits are defined as 70% training, 20% test and 10% dev sets. For comparison, all models are fine-tuned using the 'xxx-eng' translation direction where 'xxx' represents the indigenous African source language and 'eng' is the English translation target.

The available Aushumato parallel corpora (extracted from various government resources and web-crawls (Groenewald and Fourie, 2009)) are comparable in domain to the ZA-gov-multilingual parallel corpora created, however the dataset sizes are currently much larger. We therefore extract the same number for pre-processed aligned sentences as the ZA-gov-multilingual corpora in the 'xxx-eng' translation direction, for direct NMT result comparison. The sentence and token counts of the corpora used for NMT benchmarking are provided in Appendix A.1 tables 5, 6 and 7.

5.2 Results

The M2M100 fine-tuning benchmark results for the Vuk'uzenzele, ZA-gov-multilingual and subsets of the available Autshumato parallel corpora are provided in table 4. The fine-tuning translation directions are provided, and any source languages which were not including in the original M2M100 pre-training are highlighted for references purposes. Additionally the highest BLEU score result achieved across the datasets is shown in bold. Cases where the Autshumato parallel corpora were not accessible or did not exist for a particular language are shown with a '-' symbol.

Table 4: BLEU scores for Massively Multilingual Transfer on xxx-eng translations using the Vuk'uzenzele (Vuk.), ZA-gov-multilingual (Gov.) and subsets of the available Autshumato datasets (Aut.)

Translation Direction	BLEU		
	Vuk.	Gov.	Aut.
nbl→eng	7.33	8.04	**12.24**
nso→eng	9.29	**26.50**	-
ssw→eng	4.80	**28.72**	-
sot→eng	4.55	10.21	**14.83**
tsn→eng	2.80	**29.68**	28.04
tso→eng	13.86	**35.40**	32.10
ven→eng	2.32	9.68	**17.24**
xho→eng	6.05	26.81	-
zul→eng	9.97	**30.03**	25.90

Fine-tuning on the Vuk'uzenzele datasets achieved the lowest overall BLEU scores, this is expected due to the small size of the aligned datasets in comparison to those of the ZA-gov-multilingual and Autshumato datasets. The highest BLEU scores are distributed inconsistently across the ZA-gov-multilingual and Autshumato NMT models, with the ZA-gov-multilingual models achieving a higher score for Setswana, Xitsonga and isiZulu. This could be due to the random subset selections from the Autshumato datasets, as well as a result of variations in source combinations and cleaning methods used by the Autshumato project when creating the aligned corpora. It is noted that variations in the subset selections will yield different results and an in-depth analysis is left for future work.

We achieve the highest benchmark result for Xitsonga ('tso-eng') across all datasets, which is a language that the M2M100 model has not been pre-trained on, demonstrating the effectiveness of transfer learning for new low-resource language datasets.

It is also noted that current NMT resources for the Autshumato datasets exist only in the 'eng-xxx' translation direction for Xitsonga, Setswana, Sepedi, Sesotho and isiZulu (Skosana and Mlambo, 2021). Our contributions therefore extend the benchmark translation resources (in the government data domain) to isiNdebele, isiXhosa, siSwati and Tshivenda; and broaden the translation direction beyond English as the source language.

6 Conclusion

Finally, this paper presented two multilingual corpora that are automatically aligned to facilitate the translation of texts between languages. These datasets contribute to an expanding collection of corpora for training African language NMT models that are left out or under-resourced in contemporary NLP research. It is the hope of the authors that these datasets will aid in creating NMT models for low-resource African languages and that the models can be used to facilitate access to translation services, empowering African speakers and writers to communicate more effectively in their native languages. It is also hoped that the datasets will further NLP research into the multilingualism of African languages and contribute to an understanding of the various dialects present in Africa.

7 Limitations

The NMT models discussed have been created for benchmarking the described datatsets and have not been exhaustively quality tested for production purposes. We also only tested the effectiveness of fine-tuning the M2M100 model with English as a target language and have have not extended the NMT systems to translations between indigenous South African languages, therefore further benchmarking still needs to be implemented. We would like to extend testing to include the evaluation set created by (McKellar and Puttkammer, 2020), which contains data (excluded from the Autshumato corpora) for all 11 official South African languages and could provide a more accurate comparison of the NMT models. It is noted that while the BLEU score results are promising, a qualitative linguistic analysis still needs to done on the translation models to determine if the BLEU scores for certain language translations (i.e. 'tso-eng') correlate to accurate translations within our domain. As future work we hope to collaborate with respective linguists to improve the quality and effectiveness of such NMT systems for South African Languages (Skosana and Mlambo, 2021).

8 Ethics Statement

The datasets created and used for the translation model benchmarks are taken solely from South African government resources. Therefore it is highlighted that if these models are used in production, they might ignore certain social/societal structures and will be representative of the dominant political party at the time the datasets were sourced (Bender et al., 2021). We also note that the benchmark models and datasets have not been curated to determine any biases that are present. As such, any existing biases in the system might have the potential to harm specific groups, when used in NLP downstream production tasks (Bender et al., 2021).

9 Acknowledgements

We would like to acknowledge funding from the ABSA Chair of Data Science, the Google Research scholar program, TensorFlow Award for Machine Learning Grant and the NVIDIA Corporation hardware Grant. Many thanks to all the anonymous RAIL reviewers for their valuable feedback.

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247.

David Adelani, Jesujoba 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 Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Muhao Chen, Yingtao Tian, Kai-Wei Chang, Steven Skiena, and Carlo Zaniolo. 2018. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. *arXiv preprint arXiv: Arxiv-1806.06478*.

Anton A. Emelyanov and E. Artemova. 2019. Multilingual named entity recognition using pretrained embeddings, attention mechanism and ncrf. *BSNLP@ACL*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Hendrik J Groenewald and Liza du Plooy. 2010. Processing parallel text corpora for three south african language pairs in the autshumato project. *AfLaT 2010*, page 27.

Hendrik Johannes Groenewald and Wildrich Fourie. 2009. Introducing the autshumato integrated translation environment. In *Proceedings of the 13th Annual conference of the European Association for Machine Translation*.

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

Xinyu Lu, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2021. An unsupervised method for building sentence simplification corpora in multiple languages. *Conference On Empirical Methods In Natural Language Processing*.

Vukosi Marivate, Daniel Njini, Andani Madodonga, Richard Lastrucci, Isheanesu Dzingirai, and Jenalea Rajab. 2023a. The Vuk'uzenzele South African multilingual corpus.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. In *Proceedings of the first workshop on Resources for African Indigenous Languages*, pages 15–20.

Vukosi Marivate, Matimba Shingange, Richard Lastrucci, Isheanesu Dzingirai, and Jenalea Rajab. 2023b. The South African Gov-za multilingual corpus.

Louis Martin, Angela Fan, Eric Villemonte de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. Muss: Multilingual unsupervised sentence simplification by mining paraphrases. *International Conference On Language Resources And Evaluation*.

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

Cindy McKellar. (2020). Autshumato english-tshivenda parallel corpora, version 1.0, [multilingual text corpora: Aligned]. Retrieved From `https://hdl.handle.net/20.500.12185/569`.

Cindy McKellar. (2021). Autshumato english-isindebele parallel corpora, version 1.0, [multilingual text corpora: Aligned]. Retrieved From `https://hdl.handle.net/20.500.12185/572`.

Cindy McKellar. (2022)a. Autshumato english-isizulu parallel corpora, version 2.0, [multilingual text corpora: Aligned]. Retrieved From `https://hdl.handle.net/20.500.12185/575`.

Cindy McKellar. (2022)b. Autshumato english-sesotho parallel corpora, version 1.0, [multilingual text corpora: Aligned]. Retrieved From `https://hdl.handle.net/20.500.12185/577`.

Cindy McKellar. (2022)c. Autshumato english-setswana parallel corpora, version 2.0, [multilingual text corpora: Aligned]. Retrieved From https://hdl.handle.net/20.500.12185/578.

Cindy McKellar. (2022)d. Autshumato english-xitsonga parallel corpora, version 2.0, [multilingual text corpora: Aligned]. Retrieved From https://hdl.handle.net/20.500.12185/579.

Cindy A McKellar and Martin J Puttkammer. 2020. Dataset for comparable evaluation of machine translation between 11 south african languages. *Data Brief*, 29:105146.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, et al. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160.

Aasish Pappu, Roi Blanco, Yashar Mehdad, Amanda Stent, and Kapil Thadani. 2017. Lightweight multilingual entity extraction and linking. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 365–374.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

Holger Schwenk and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. *ACL 2017*, page 157.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. *International Conference On Language Resources And Evaluation*.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Nomsa Skosana and Respect Mlambo. 2021. A brief study of the autshumato machine translation web service for south african languages. *Literator*, 42(1):7.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

A.1 Data Statistics

The data statistics for the datasets used for NMT bench-marking are provided in Tables 5, 6 and 7.

Table 5: Characteristics of the translation data for the Vuk'uzenzele (Vuk.) datasets

Translation Direction	**Size #sents (#src / #trg tokens)**
nbl→eng	136 (3.4k / 3.9k)
nso→eng	1715 (53.7k / 41.6k)
ssw→eng	1588 (29.9k / 37.6k)
sot→eng	260 (9.9k / 7.5k)
tsn→eng	1366 (49.8k / 31.7k)
tso→eng	1998 (58.9k/ 46.6k)
ven→eng	230 (9.1k / 7k)
xho→eng	1338 (25.8k / 31.5k)
zul→eng	1874 (34.1k / 43k)

Table 6: Characteristics of the translation data for the ZA-gov-multilingual (Gov.) datasets

Translation Direction	**Size #sents (#src / #trg tokens)**
nbl→eng	3513 (63.9k / 107k)
nso→eng	14742 (460.9k / 375k)
ssw→eng	15139 (291k/ 377.8k)
sot→eng	4995 (145.9k / 153.5k)
tsn→eng	14068 (493.1k / 362.2k)
tso→eng	15393 (466.4k / 381.2k)
ven→eng	3404 (68.2k / 96.6k)
xho→eng	15853 (318.2k / 389.5k)
zul→eng	15503 (327.5k / 384.1k)

Table 7: Characteristics of the translation data for the subsets of the available Autshumato datasets (Aut.)

Translation Direction	**Size #sents (#src / #trg tokens)**
nbl→eng	3513 (48.4k / 65k)
sot→eng	4995 (118.3k / 100.4k)
tsn→eng	14068 (335.1k / 274.7k)
tso→eng	15393 (193.7k/ 166k)
ven→eng	3404 (80.6k / 65.9k)
zul→eng	15503 (231.8k / 307.6k)