# Construction-Based Reduction of Translationese for Low-Resource Languages: A Pilot Study on Bavarian

**Peiqin Lin**[*1,2], **Marion Thaler**[*1], **Daniela Goschala**[1], **Amir Hossein Kargaran**[1,2], **Yihong Liu**[1,2], **André F. T. Martins**[3,4,5], **Hinrich Schütze**[1,2]

[1]Center for Information and Language Processing, LMU Munich
[2]Munich Center for Machine Learning [3]Instituto Superior Técnico (Lisbon ELLIS Unit)
[4]Instituto de Telecomunicações [5]Unbabel
linpq@cis.lmu.de, Marion.Thaler@campus.lmu.de

## Abstract

When translating into a low-resource language, a language model can have a tendency to produce translations that are close to the source (e.g., word-by-word translations) due to a lack of rich low-resource training data in pretraining. Thus, the output often is translationese that differs considerably from what native speakers would produce naturally. To remedy this, we synthetically create a training set in which the frequency of a construction unique to the low-resource language is artificially inflated. For the case of Bavarian, we show that, after training, the language model has learned the unique construction and that native speakers judge its output as more natural. Our pilot study suggests that construction-based mitigation of translationese is a promising approach. Code and artifacts are available at https://github.com/cisnlp/BayernGPT.

## 1 Introduction

The multilingual capabilities of large language models (LLMs) are impressive for medium- and high-resource languages, but they are still poor for low-resource languages for which the size of the available text corpus is small. While LLMs have recently improved their performance on low-resource comprehension tasks, little progress has been made on generation since the training demands for effective generation are much higher than those for comprehension. Bavarian is a low-resource language that instantiates this state of affair: some large state-of-the-art models' performance is decent for comprehension of Bavarian, but this does not carry over to generation.

Our hypothesis is that there are at least two different problems with limited generation capabilities of LLMs: lack of knowledge and translationese behavior.

Lack of knowledge mainly results in poor lexical choices. For example, our trained model (see below) translates German "Kuchen" 'cake' not as the correct Bavarian "Kuacha", but as "Kuchel" 'kitchen'. There is some promising work that addresses the lack of knowledge problem by prompting the LLM with relevant dictionary entries in in-context learning.

However, apart from the lack-of-knowledge problem, there is a second problem with the Bavarian generations of language models: translationese.

Translationese is a particular problem in machine translation with language models. The LMs tend to stick closely to the source sentence, especially when translating from a high-resource language to a closely related low-resource language as is the case for Standard German and Bavarian. Bavarian and Standard German are in a state of diglossia where Bavarian speakers produce forms of Bavarian that are closer to Standard German in more formal contexts and forms of Bavarian that can be completely incomprehensible to Standard German speakers in informal contexts.

This means that the Bavarian translationese generated by LMs is not necessarily incorrect: it may be appropriate Bavarian for certain contexts of language use. But clearly, the LMs do not have full competence of the Bavarian language if all they do is produce translationese.

In this paper, we take a small step towards addressing the translationese problem by training LMs to generate a Bavarian construction that does not occur in Standard German. This reduces the translationese property of what the LM generates because the output has clear indicators of being "genuine" Bavarian.

Specifically, we experiment with the article reduplication construction in Bavarian:

| | |
|---|---|
| Bavarian | Ea woa friara a recht a fidel Buam. |
| Gloss | He was formerly a RD-modifier a jolly boy. |
| translation | He used to be quite a jolly boy. |

[*]Equal contribution.

With certain reduplication modifiers (RD modifiers), in particular with “recht”, “so” and “ganz”, this Bavarian construction consists of the reduplication of the indefinite article, with the RD modifier occurring between the two indefinite articles.

We show that a model trained with data synthetically generated to contain article reduplication learns to produce the construction, reducing the translationese character of the language model translations.

To summarize, our method translates an originally Bavarian corpus to German using a state-of-the-art LM, resulting in an “unmodified” parallel corpus; generates a “modified” parallel corpus by semi-automatically editing parallel sentences (such that the Bavarian sentence contains a Bavarian construction and the German sentence is modified to reflect that change) and trains LMs on modified and unmodified corpora. We also create two evaluation datasets, one for sentences, one for noun phrases. We manually evaluate the performance of the two trained models. We find that the model trained on modified data successfully produces article reduplication and its output data is perceived as less “translationese” than the generations of the model trained on unmodified data.

## 2 Related Work

Multilingual language models have emerged as the dominant paradigm for supporting low-resource languages. These models range from smaller architectures such as multilingual BERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), mBART (Liu et al., 2020), mT5 (Xue et al., 2021), and Glot500 (Imani et al., 2023), to large-scale models including BLOOM (Scao et al., 2022), Aya (Üstün et al., 2024), MaLA500 (Lin et al., 2024a), EMMA500 (Ji et al., 2024), and Llama 3 (Dubey et al., 2024). Trained jointly on data from a wide range of languages, these models demonstrate strong cross-lingual transfer and generalization capabilities, offering a promising foundation for low-resource language applications.

Despite this progress, generative performance on low-resource languages remains limited – particularly for tasks such as machine translation (MT), which are highly sensitive to the quantity and quality of available training data.

MT has thus become a central benchmark for evaluating the generative abilities of multilingual models. In the context of large language models, recent efforts have explored two major directions: supervised fine-tuning on parallel corpora (Yang et al., 2023; Xu et al., 2023, 2024; Lin et al., 2024b; Alves et al., 2024; Rei et al., 2024), and in-context learning methods that incorporate external linguistic resources – such as grammar books and bilingual dictionaries – without modifying model weights (Lu et al., 2023; Tanzer et al., 2024; Zhang et al., 2024b,a; Pei et al., 2025).

These challenges are particularly pronounced for extremely low-resource languages such as Bavarian. Due to very limited annotated data, Bavarian remains largely excluded from multilingual pretraining. Her and Kruschwitz (2024) introduced one of the first Bavarian–German MT systems, demonstrating that translation between closely related language varieties can yield relatively strong BLEU scores. To further enhance translation quality while minimizing artifacts such as translationese, they employed back-translation (Sennrich et al., 2016) to generate a compact but effective set of synthetic training examples. However, their approach depends solely on WikiMatrix (Schwenk et al., 2021), a parallel corpus known to be noisy and dominated by simplistic sentence structures, which limits its ability to robustly capture more nuanced translation characteristics.

One such characteristic is translationese – a linguistic phenomenon that arises when translated text retains unnatural or non-native structures. This artifact is especially problematic for low-resource languages (Graham et al., 2020). To address it, Chowdhury et al. (2022) proposed removing translationese signals implicitly encoded in vector embeddings, leading to improved performance on natural language inference tasks. Similarly, Wein and Schneider (2023) employed Abstract Meaning Representation (AMR) to abstract away surface-level features and suppress translationese. While effective, these techniques do not explicitly assess whether the resulting text resembles naturally written language. More recently, Jalota et al. (2023) evaluated the success of style transfer techniques in mitigating translationese by analyzing classifier performance before and after post-editing. Kunilovskaya et al. (2024) further explored the use of GPT-4 to mitigate translationese by incorporating linguistic cues into the prompting context. Complementarily, Kuwanto et al. (2024) introduced a storyboard-based data collection method, in which native speakers generate descriptions from visual prompts without access to the source text—resulting in outputs that are more fluent and natural. However, these methods still fall short of enabling large language models to directly produce fluent, translationese-free output for truly low-resource languages, such as Bavarian.

## 3 Training

### 3.1 Data Preparation

Due to the scarcity of high-quality Bavarian–German parallel corpora, we use `GPT-4` to translate the Bavarian portion of the Wikipedia[1] into English and standard German. We use language identification (Kargaran et al., 2023) to filter out noise, such as when the model partially translates the source or directly copies it. From the resulting 22,564 Bavarian-English-German parallel documents, we reserve 1,000 for validation and another 1,000 for testing, with the remainder used for training. To create a sentence-level corpus, we segment the documents using line breaks and remove duplicate entries.

To reduce translationese effects and encourage native-sounding Bavarian output, we augment the original corpus using a rule-based algorithm grounded in syntactic analysis. We employ `spaCy` to parse the Standard German sentences and identify noun phrase structures of the form *indefinite article* + *adjective* + *noun*. These constructions serve as reliable anchors for inserting article reduplication in the aligned Bavarian sentence.

The algorithm first scans each tokenized German sentence for sequences where an indefinite article (e.g., *ein*, *eine*) is immediately followed by an adjective and a noun. To avoid semantically awkward or ungrammatical insertions, the algorithm filters out adjectives derived from nationalities (e.g., *deutsch*, *österreichisch*). For every such match, we check whether the corresponding Bavarian sentence has an equivalent syntactic pattern beginning with a Bavarian indefinite article (e.g., *a*, *oa*).

If the alignment is valid, we apply a transformation that inserts a reduplicated indefinite article separated by an RD modifier (randomly chosen from *recht*, *so*, *ganz*) between the original article and adjective. To maintain semantic alignment, the German counterpart is modified by inserting the intensifier *sehr* between the article and adjective.

This pipeline was run over sentence-aligned data and executed only where the token count matched between the Bavarian and German sides, ensuring high-precision transformations. Table 1 shows a representative example.

| **Before article reduplication transformation** |
|---|
| Bavarian: *A heiliga Lebnsbaam*<br>German: *Ein heiliger Lebensbaum* |
| **After article reduplication transformation** |
| Bavarian: *A recht a heiliga Lebnsbaam*<br>German: *Ein sehr heiliger Lebensbaum* |

Table 1: Example of article reduplication transformation in Bavarian–German parallel data.

### 3.2 Model Training

We develop a German-to-Bavarian machine translation system by instruction-tuning the `Llama 3.1 8B Chat` model (Dubey et al., 2024).

To accomplish this, we design a structured prompt format, as shown in Table 2. In this format, `[DEU_TEXT]` represents the input German sentence, and `[BAR_TEXT]` corresponds to the expected Bavarian translation. During training, both sentences are provided to the model, while at inference time, the model generates `[BAR_TEXT]` from the input German sentence.

To enable efficient fine-tuning, we use LoRA (Hu et al., 2022). The model is fine-tuned with a learning rate of $1 \times 10^{-4}$, weight decay set to 0.1, and the LoRA rank configured to 32.

We train two machine translation models:

- **m-base:** Trained on the original parallel dataset.
- **m-aug:** Trained on the dataset augmented with rule-based transformations.

## 4 Evaluation

To assess the effectiveness of our article reduplication augmentation strategy, we conducted both sentence-level and noun phrase–level (NP) evaluations using human judgments from two native Bavarian speakers (two of the authors of this paper).

### 4.1 Sentence-Level Evaluation

We used a test set of 141 Bavarian–Standard German sentence pairs which received the same augmentation as the training data of m-aug. Standard German inputs were translated into Bavarian by both m-base (baseline) and m-aug (augmented). The evaluation focused on three criteria:

[1] `dumps.wikimedia.org/barwiki`

```
<|start_header_id|>user<|end_header_id|>
Translate the following text from German to Bavarian.
German:  [DEU_TEXT]
Bavarian:  <|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
[BAR_TEXT]<|eot_id|>
```

Table 2: Prompt format used for instruction-tuned machine translation from German to Bavarian.

- **Correct application of article reduplication** – Is article reduplication used where grammatically appropriate?
- **Idiomatic and grammatical correctness** – Is the output of m-aug more idiomatic and grammatically natural and correct than m-base?
- **Pragmatic appropriateness** – Is article reduplication contextually suitable within the sentence?

The importance of pragmatic appropriateness can be illustrated by the following example.

> **Bavarian (with article reduplication):**
> *In da Eingobaauffordarung kennt ma an Untaschied zwischn Root und andan Nutzan duach* **a ganz a obschliaßends Rautzeichen** *(#) stott des Dollarzeichens ($).*
>
> **Standard German:**
> *In der Eingabeaufforderung erkennt man einen Unterschied zwischen Root und anderen Nutzern durch* ***ein sehr abschließendes Rautezeichen*** *(#) anstelle des Dollarzeichens ($).*
>
> **English translation:**
> *In the command prompt, one can recognize a difference between Root and other users by* ***a very final hash sign*** *(#) instead of the dollar sign ($).*

The emphatic use of article reduplication (*a ganz a obschliaßends Rautzeichen*) is unnatural and non-idiomatic in this technical context. As a result, it was evaluated as pragmatically inappropriate, even though the grammatical structure is correct.

The resulting outputs were evaluated by two native speakers. The evaluation regarding pragmatic appropriateness was conducted on the 70 sentences where reduplication was applied. The overall inter-annotator agreement was 100%, indicating high reliability of the judgments.

m-aug failed to apply article reduplication where grammatically possible in only 12 cases, and in just 19 cases the translation of m-base was assessed as more grammatically correct and idiomatic. However, regarding pragmatic appropriateness, there is a higher number of questionable cases. This is primarily due to the fact that the augmented training data was not filtered for pragmatic appropriateness, potentially including instances where article reduplication is not suitable. As such, m-aug provides a baseline that could be further improved with more appropriate training data.

| **Category** | **Count** | **Percentage** |
|---|---|---|
| Reduplication correctly applied | 70 | 49.65% |
| Not applicable (grammatically) | 59 | 41.84% |
| Reduplication missed (applicable) | 12 | 8.51% |
| **Total** | 141 | 100.00% |

Table 3: Sentence-level evaluation: Article reduplication accuracy.

| **Comparison Result** | **Count** | **Percentage** |
|---|---|---|
| m-aug sentence is better | 103 | 73.05% |
| Sentences are equivalent | 19 | 13.48% |
| m-aug sentence is worse | 19 | 13.48% |
| **Total** | 141 | 100.00% |

Table 4: Sentence-level evaluation: Idiomatic and grammatical correctness comparison (m-base vs m-aug).

| **Evaluation Result** | **Count** | **Percentage** |
|---|---|---|
| Reduplication is pragmatically correct | 42 | 60.00% |
| Reduplication is questionable | 28 | 40.00% |
| **Total** | 70 | 100.00% |

Table 5: Sentence-level evaluation: Pragmatic appropriateness of reduplication.

### 4.2 Noun Phrase–Level Evaluation

Given that article reduplication targets noun phrases, we conducted a focused evaluation. A test set of 200 Standard German NPs in the structure *indefinite article* + *intensifier* + *adjective* + *noun* was generated using random combinations of *adjective* + *noun* from the Wikipedia corpus. The translations of both models were evaluated by a native speaker. This evaluation focused on whether the article reduplication was applied accurately and whether the idiomatic and grammatical correctness was improved compared to the NPs produced by the original Model A.

| Category | Count | Percentage |
|---|---|---|
| Reduplication applied | 200 | 100% |
| Reduplication not applied | 0 | 0% |
| **Total** | 200 | 100% |

Table 6: NP-level evaluation: Article reduplication accuracy.

| Comparison Result | Count | Percentage |
|---|---|---|
| NP of Model B is better | 189 | 94.5% |
| NP of Model B is worse | 11 | 5.5% |
| **Total** | 200 | 100% |

Table 7: NP-level evaluation: Idiomatic and grammatical correctness (Model A vs. B).

These results indicate that Model B systematically learned the reduplication pattern within the structure *indefinite article* + *intensifier* + *adjective* + *noun*, producing outputs that are both idiomatic and grammatically well formed.

To assess whether the model overgeneralizes article reduplication, we conducted a complementary evaluation using 200 Standard German noun phrases of the form *indefinite article* + *adjective* + *noun*, i.e., without an intensifier. This test aimed to determine if the model incorrectly applies reduplication to structures where it is not licensed. Only 4 out of 200 outputs contained article reduplication without an intensifier present. Interestingly, these instances were all triggered by the word *ganz* used adjectivally, as in the Standard German phrase *ein ganzer Ortsteil*, translated in Bavarian as *a ganza **a** Orstei* (gloss: *a whole subdistrict*). In these cases, *ganz*, previously encountered as an RD-modifier in the training data, was likely misinterpreted by the model as licensing reduplication, even when used adjectivally.

| Category | Count | Percentage |
|---|---|---|
| Reduplication falsely applied | 4 | 2% |
| Reduplication not applied | 196 | 98% |
| **Total** | 200 | 100% |

Table 8: NP-level evaluation: Reduplication overgeneralization in NPs without intensifiers.

These findings suggest that the model applies article reduplication in a targeted and controlled manner, largely avoiding false positives.

## 5 Conclusion

We propose a method to remedy the problem of translationese when translating to low-resource languages and apply it to Bavarian. Our approach synthetically creates a training set in which the frequency of a construction unique to the low-resource language is artificially inflated. We show that a model trained with this synthetic data produces output with this construction and that it is perceived as being more natural than the baseline.

## Limitations

Our pilot study has numerous limitations.

- We know of no linguistic studies that quantify the impact of constructions on the “naturalness” of linguistic production. Other factors may also have to be addressed to produce fully natural output.
- For a given pair of high resource and low resource languages, there may be no constructions that meet our selection criterion: that is, they are relatively frequent in the low-resource language and do not occur at all in the high-resource language.
- The construction we chose is easy to match and to generate. For more complex constructions, there is a risk that the modified low-resource sentences would not be correct, thereby introducing a new source of errors.
- We only implemented a baseline method for changing source and target languages. There are several ways this baseline method can be improved, e.g., by trying to eliminate the pragmatically inappropriate language we detected in the experiments.
- Our evaluation is very basic. Due to the difficulty of finding native speakers of Bavarian, two of the authors (who are native speakers of Bavarian) performed the annotation.

## Acknowledgements

This work was funded by DFG (SCHU 2246/14-1), the European Research Council (DECOLLAGE, ERC-2022-CoG #101088763), EU’s Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the

Federal Ministry of Education and Research, and by FCT/MECI through national funds and when applicable co-funded EU funds under UID/50008: Instituto de Telecomunicações.

## References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro Henrique Martins, João Alves, M. Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *CoRR*, abs/2402.17733.

Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, and Josef van Genabith. 2022. Towards debiasing translation artifacts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3983–3991. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 72–81. Association for Computational Linguistics.

Wan-Hua Her and Udo Kruschwitz. 2024. Investigating neural machine translation for low-resource languages: Using bavarian as a case study. *CoRR*, abs/2404.08259.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André F. T. Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1082–1117. Association for Computational Linguistics.

Rricha Jalota, Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2023. Translating away translationese without parallel data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7086–7100. Association for Computational Linguistics.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. 2024. EMMA-500: enhancing massively multilingual adaptation of large language models. *CoRR*, abs/2409.17892.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages. In *Findings* *of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 6155–6218. Association for Computational Linguistics.

Maria Kunilovskaya, Koel Dutta Chowdhury, Heike Przybyl, Cristina España-Bonet, and Josef van Genabith. 2024. Mitigating translationese with GPT-4: strategies and performance. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1), EAMT 2024, Sheffield, UK, June 24-27, 2024*, pages 411–430. European Association for Machine Translation (EAMT).

Garry Kuwanto, Eno-Abasi Urua, Priscilla Amondi Amuok, Shamsuddeen Hassan Muhammad, Aremu Anuoluwapo, Verrah Otiende, Loice Emma Nanyanga, Teresiah W. Nyoike, Aniefon D. Akpan, Nsima Ab Udouboh, Idongesit Udeme Archibong, Idara Effiong Moses, Ifeoluwatayo A. Ige, Benjamin Ajibade, Olumide Benjamin Awokoya, Idris Abdulmumin, Saminu Mohammad Aliyu, Ruqayya Nasir Iro, Ibrahim Said Ahmad, Deontae Smith, Praise-EL Michaels, David Ifeoluwa Adelani, Derry Tanti Wijaya, and Anietie Andy. 2024. Mitigating translationese in low-resource languages: The storyboard approach. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 11349–11360. ELRA and ICCL.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024a. Mala-500: Massive language adaptation of large language models. *CoRR*, abs/2401.13303.

Peiqin Lin, André F. T. Martins, and Hinrich Schütze. 2024b. A recipe of parallel corpora exploitation for multilingual large language models. *CoRR*, abs/2407.00436.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. *CoRR*, abs/2305.06575.

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schütze. 2025. Understanding in-context machine translation for low-resource languages: A case study on manchu. *CoRR*, abs/2502.11862.

Ricardo Rei, José Pombal, Nuno Miguel Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tânia Vaz, Duarte M. Alves, M. Amin Farajian, Sweta Agrawal, António Farinhas, José Guilherme Camargo de Souza, and André F. T. Martins. 2024. Tower v2: Unbabel-ist 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation, WMT 2024, Miami, FL, USA, November 15-16, 2024*, pages 185–204. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 1351–1361. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ahmet Üstün, Viraat Aryabumi, Zheng Xin 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. 2024. Aya model: An instruction finetuned open-access multilingual language model. *CoRR*, abs/2402.07827.

Shira Wein and Nathan Schneider. 2023. Translationese reduction using abstract meaning representation. *CoRR*, abs/2304.11501.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *CoRR*, abs/2309.11674.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: Pushing the boundaries of LLM performance in machine translation. *CoRR*, abs/2401.08417.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. *CoRR*, abs/2305.18098.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024a. Teaching large language models an unseen language on the fly. *CoRR*, abs/2402.19167.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024b. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions. *CoRR*, abs/2402.18025.