
INVESTIGATING NEURAL MACHINE TRANSLATION FOR LOW-RESOURCE LANGUAGES: USING BAVARIAN AS A CASE STUDY*

Wan-Hua Her
 Information Science
 University of Regensburg
 Germany
 wan-hua.her@stud.uni-regensburg.de

Udo Kruschwitz
 Information Science
 University of Regensburg
 Germany
 udo.kruschwitz@ur.de

ABSTRACT

Machine Translation has made impressive progress in recent years offering close to human-level performance on many languages, but studies have primarily focused on high-resource languages with broad online presence and resources. With the help of growing Large Language Models, more and more low-resource languages achieve better results through the presence of other languages. However, studies have shown that not all low-resource languages can benefit from multilingual systems, especially those with insufficient training and evaluation data. In this paper, we revisit state-of-the-art Neural Machine Translation techniques to develop automatic translation systems between German and Bavarian. We investigate conditions of low-resource languages such as data scarcity and parameter sensitivity and focus on refined solutions that combat low-resource difficulties and creative solutions such as harnessing language similarity. Our experiment entails applying Back-translation and Transfer Learning to automatically generate more training data and achieve higher translation performance. We demonstrate noisiness in the data and present our approach to carry out text preprocessing extensively. Evaluation was conducted using combined metrics: BLEU, chrF and TER. Statistical significance results with Bonferroni correction show surprisingly high baseline systems, and that Back-translation leads to significant improvement. Furthermore, we present a qualitative analysis of translation errors and system limitations.

Keywords Neural Machine Translation, Low-resource Languages, Back-translation, Bavarian, German

1 Introduction

Neural Machine Translation (NMT) has progressed so far to reach human-level performance on some languages [1] and has become one of the most prominent approaches within the research area of Machine Translation (MT). Its easy-to-adapt architecture has achieved impressive performance and high accuracy. Promising methods that fall under NMT include Transfer Learning [2, 3], pre-trained language models [4, 5], and multilingual models [6, 7, 8, 9] etc.

However, existing NMT resources focus overwhelmingly on high-resource languages, which dominate a great portion of contents on the Internet and Social Media. Low-resource languages are often spoken by minorities with minimal online presence and insufficient amount of resources to achieve comparable NMT results [10, 11], but they might even have a very large population of speakers and still be under-resourced (such as Hindi, Bengali and Urdu). Growing interest in low-resource MT is evident through the annually held Conference on Machine Translation (WMT). In 2021, WMT featured tasks to promote MT in low-resource scenarios by exploring similarity and multilinguality [12]. Among all tasks, the objective of the Very Low Resource Supervised Machine Translation task [13] focused on Transfer Learning between German and Upper Sorbian. The task examined effects of utilizing similar languages and results

*Preprint accepted at SIGUL 2024

show that combining Transfer Learning and data augmentation can successfully exploit language similarity during training.

We introduce our experiment to develop bidirectional state-of-the-art NMT systems for German and Bavarian, a classic high-resource to/from low-resource language pair. Inspired by WMT21, our experiment explores the generalizability of Back-translation and Transfer Learning from the highest-ranking approach from [14]. Our approach covers the following: First, a simple Transformer [15] is trained as the baseline. Secondly, we use the base model for Back-translation and take the extended corpus to train our second model. Lastly, we experiment with Transfer Learning [3] by introducing German-French as the parent model. For evaluation we opt for a combination of three metrics: BLEU [16], chrF [17] and TER [18]. Recent studies have argued that using BLEU as a single metric neglects the complexity of different linguistic characteristics. Using combined metrics and having various penalization standards may be able to capture translation errors more diversely [19, 20].

By choosing the language pair Bavarian / German we offer one exemplar for a low-resource language (combined with a high-resource one) that can serve as a reference point for further experimental work applied to other low-resource MT. This will ultimately help addressing the imbalance that still prevails between a handful of well-resourced languages and the many others that are not. This paper makes the following contributions:

- We offer a systematic evaluation of state-of-the-art NMT approaches for a language pair involving a low-resource language that has attracted little attention so far. This investigation explores both translation from as well as into the low-resource language. We focus on a Transformer baseline against Back-translation and a Transfer Learning approach.
- To foster reproducibility and replicability (which is in the very spirit of SIGUL, LREC and COLING) we make all code available via a GitHub project repository².

2 Related Work

2.1 Low-Resource Languages

The challenges of low-resource languages can be very diverse, hence difficult to define in simple words.

For a start, even though large web-crawled data such as OPUS [21] has resulted in automatically generated parallel corpora for many minor languages, the quality of the data has been reported to be noisy. Examples include the Bantu (Niger-Congo) languages, where parallel data exists, but often too inconsistent to generate desirable MT performance and reproducible benchmarks [22]. Misalignments and mistranslations have also been reported while working with multilingual Indian languages [23]. The rise of Unsupervised NMT [24, 25, 26] alleviates the need for large amounts of labeled training data. Nonetheless, researchers have noted however strong the supervision during training is, there is an overall dependence on parallel data to support evaluation systems [27, 28]. We therefore see the problem of these less-studied languages as a problem caused by both the *quantity* and the *quality* of the resources. Without linguistically-trained speakers, parallel data is often curated in an unsupervised fashion and therefore noisy.

Furthermore, there are endangered languages [29], for example, the language Bri bri is an extremely low-resource indigenous language which is currently being displaced by English and Spanish [11]. Aside from suffering digital inequalities and having insufficient written data, it was more challenging to create standardized representations of Bri bri, since lexemes and rules vary from communities of speakers. Another similar study which focused on Alemannic dialects also highlights that dialects do not have uniform spelling rules, and that spelling reflect different regional pronunciations [30]. This raises a great challenge for MT to decide which variation should be given precedence. These under-resourced languages raise a string of challenges due to long years of absence of standardization, and that digital revitalization is not merely a question of gathering data and training models.

To optimize text processing and its size during training, the most common way is to create a joint vocabulary through Byte Pair Encoding (BPE) [31]. BPE is a highly effective subword segmentation algorithm. It iteratively merges frequent words and creates new subword units from infrequent words. A drawback of this approach is that the model learns patterns of smaller unit composition only by recognizing the infrequent words. To counter this, BPE dropout was introduced by [32] to stochastically corrupt the segmentation procedure within BPE.

²<https://github.com/wmher/nmt-de-bar>

2.2 Machine Translation

Nearest Neighbor Machine Translation Non- and semi-parametric methods have been successfully applied to MT tasks in recent years. [33] demonstrate a powerful combination of neural networks and non-parametric retrieval mechanisms to improve translation. *k*NN-MT follows the retrieval principle and proposes a more efficient non-parametric translation method, which augments the decoder of a pre-trained NMT model with a nearest neighbor retrieval mechanism, allowing direct access to data store of cached examples [34]. This approach scales the decoder to an arbitrary amount of examples at test time, particularly strengthening decoder’s translation capability. However, the big drawback is high computational cost and low decoding speed due to word-by-word generation. Chunk-based *k*NN-MT [35] solves this problem by processing translation in chunks of words instead of passing single tokens through the data store.

Transfer Learning in MT is often done by training a high-resource language pair and using this parent model to initialize parameters in a child model with low-resource languages. For example, [3] achieved translation improvements for Hansa, Turkish and Uzbek into English by using French-English as a parent model. Experiments from [36] showed improvements using Transformers [15] to train low-resource languages such as Estonian and Slovak. Their results pointed out key factors for a successful transfer include the size of the parent corpus and sharing the target or source language. For instance, Estonian-English as a child gained up to 2.44 BLEU with Finnish-English as a parent.

In Dual Transfer [2], two parent models are used to initialize one child. Monolingual and parallel parent data were trained separately so that inner layers and embeddings can be transferred separately. Another recent study extends conventional transfer learning by additionally transferring probability distributions from parent to child. The Consistency-based Transfer Learning [37] argues that parent prediction distribution is highly informative and can be useful to guide child translation. Their experiment showed that using German-English as a parent can achieve BLEU improvement up to 6.2 for Indonesian-English. Furthermore, the study from [6] investigated a technique to incrementally add new language pairs to a multilingual MT model based on knowledge transfer, without posing the original model at risk for catastrophic forgetting.

Pre-trained Language Models (PLMs) can be fine-tuned on low-resource languages. For instance, MT quality between Spanish and Quecha was shown to improve by leveraging Spanish-English and Spanish-Finnish PLMs [4], with the latter yielding better results. Furthermore, [38] combined a BERT [39] encoder with a vanilla NMT decoder. Evaluation on low-resource languages like English-Vietnamese show that their two-stage training improves performance significantly compared to simple fine-tuning. XLM extends the features of BERT by using Cross-Lingual Masked Language Modeling [40]. It has not only been reported to be beneficial for general unsupervised learning, but also for low-resource supervised MT such as English-Romanian. [41] acknowledged the success of PLMs and presented their granulated study of fine-tuning, which showed that cross-attention layers are crucial to continue training downstream tasks and that they are powerful when adapting to new languages.

2.3 Refined Solutions

Data Filtering and Normalization Translation data for low-resource languages are very difficult to come by and the primary source are often from the Web, making the data noisy and of poor quality [42]. Extra analysis and text normalization are often required to prevent overfitting. For instance, inaccurate translations, noisy data and a large amount of text-overlap was found in the parallel data for African languages collected from large crowd-sourced platforms [22]. Comparative results showed that an English-Zulu model trained with noisy data leads to unreliable results and a reduction of 7 BLEU. Research from [28] corroborated this and provided guidelines for removing low-quality translations. They presented translation filtering by way of n-gram models trained on monolingual data and sentence-level char-BLEU score [43] below 15 or over 90. Another novel filtering approach was proposed by [42], where cosine similarity is determined based on available parallel (good quality) data, which is then used as the threshold to filter out pseudo-parallel (noisy) sentences.

Multilinguality Previous findings have pointed out that one-to-many models with middle-sized parallel corpora have achieved better results than one-to-one models [44]. The multilingual model consisting of seven Asian languages developed by [9] using the Asian Language Treebank [45] is a great example. The presence of multiple in-domain aligned languages was argued to have contributed to better learn joint representations, hence leading to intra-language improvements. However, low-resource languages often face the risk of being overfitted in multilingual setups [46]. [7] investigated the extent of multilinguality for low-resource languages. Their corpus consists of Bible texts in 1,108 languages, all aligned by verse. Results show that BLEU increase/decrease with respect to the number of training languages is not uniform across languages. Although the 5-language models outperform bilingual baseline models for

Turkish and Xhosa, accuracy decrease can be found in Tagalog. The negative correlation between number of languages and translation quality is found to start at 10 languages, and maximal degeneration is observed at 100 languages, where addition of languages does not affect translation fluency anymore. This complication and pattern of degeneration can be explained by [47], where text repetition harms the likelihood function during decoding. Furthermore, the errors in sequence modeling are more obvious for multilingual corpora, indicating that increased number of languages leads to increased destructive interference.

Language Similarity Leveraging similarities between low-resource languages has been a growing interest in the MT community and is evident through the Similar Language Translation task (SLT) and Very Low Resource Supervised Machine Translation task at WMT21 [48]. Regardless of level of closeness and degree of mutual structures, similarity between languages has shown to have positive interactions with MT quality [49]. The goal of using language relatedness is similar to leveraging multilinguality. The major difference is they often do not use English as the pivot language, but translate between closely-related languages.

In the Very Low Resource Supervised Machine Translation task at WMT21 [13] between German and Upper Sorbian, the participants were encouraged to make use of Czech and Polish datasets (languages closely related to Sorbian). Results pointed out the importance of including related languages, and that carefully applying tricks can compensate for using smaller datasets substantially. For example, NoahNMT’s [50] approach entails a Dual Transfer [2] model that was initialized using German and Czech monolingual data as a parent model. The NRC-CNRC team’s [14] high-performance was attributed to the combination of minor tricks such as Back-translation [51], monolingual data selection by way of cosine similarity, Moore-Lewis filtering [52] and BPE dropout [32].

The technique Back-translation is further backed up by the study from [30]. They investigated the effect on Alemannic dialect translation and experienced significant improvement, suggesting that Back-translation is a highly promising method for low-resource languages.

3 Methodology

Motivated by the current findings, we present our experiment to develop bidirectional state-of-the-art NMT systems between German and Bavarian (ISO codes are de and bar respectively) - a language pair consisting of high- and low-resource languages. While Bavarian and Upper Sorbian are very different languages, they are both spoken by communities which are geographically located within or near Germany. We expect that applying the NMT methods that were found to be effective as part of WMT21 might result in similar findings for our setting.

We formulate the following three research questions (applied to the exemplar language pair Bavarian / German):

- **RQ1:** Does translating between similar languages achieve generally higher BLEU scores?
- **RQ2:** How well does Back-translation perform for (bidirectional) German-Bavarian?
- **RQ3:** Does cross-lingual transfer lead to improved results for German-Bavarian? More specifically, does the child model profit from related parent languages (i.e. German-French)?

3.1 Data Acquisition

The Tatoeba Challenge³ [53] is one of the most active projects advocating low-resource MT. It maintains a leader board to compare submitted MT system performance from the community. To our knowledge, we are the first to conduct MT for German-Bavarian systems. We discovered parallel and monolingual sources on OPUS⁴ [21], which we used for our experiments. More information about data sources can be found in our repository.

3.2 Framework

Inspired by the WMT21 Very Low Resource Supervised Machine Translation task [13], our experiment revisits solutions that have been proven to work effectively with low-resource languages.

- First, a simple Transformer [15] model using preprocessed parallel data is trained as the baseline model.
- Secondly, Back-translation is used to generate silver-paired parallel data to increase corpus size.
- Lastly, we experiment with Transfer Learning [3] by introducing German-French as the parent model.

³<https://github.com/Helsinki-nlp/tatoeba-challenge>

⁴<https://opus.nlpl.eu/>

For evaluation, we opt for an ensemble of automated MT metrics consisting of BLEU, chrF and TER for our systems. This is backed up by recent argumentation from [19] and [20], which states that multiple metrics instead of a single metric can diversify the evaluation based on different linguistic characteristics. This approach is a growing trend and has also been adopted by WMT21. Moreover, the study from [30] pointed out BLEU is insufficient in word matching due to ununified orthography.

4 Implementation

Data Preparation In total we found 99.7K parallel sentences between Bavarian and German on OPUS (details can be found in our repository). After extensive preprocessing, the corpus size was reduced to 42K. To conduct data augmentation for the second system, we downloaded an extra 258K of German and 295K Bavarian monolingual text, mainly from Wikipedia and Wikinews. For German-French, we collected a total size of 184K of parallel data from Tatoeba and WikiMedia, which was reduced to 165K after preprocessing. We argue that the amount of in-domain data could contribute positively to Transfer Learning. Text preprocessing removes special symbols and noisy annotation, as proposed in previous studies [14, 23].

In addition to conventional text preprocessing, we took two further measures to de-noise the data. The additional measures entail check and remove misaligned texts by way of cosine similarity between source and target languages and smart sentence truncation. Based on the knowledge that Bavarian and German share common script and that many morphemes are alike, cosine similarity is a great way to support misalignment removal. We assume that a low cosine correlation indicates a low relevance in context between source and target. Following exploratory experiments, we set the correlation threshold at 0.48 and treat anything that falls below 0.48 as misalignment and remove this. We leave a systematic investigation into this aspect as future work.

Our consideration for smart truncation comes from the long-tailed distribution of sentence lengths (outliers span up to 8000). Having long sentences in the corpus therefore poses potential threat that could damage MT performance [54]. However, if all longer sequences were simply removed, we might lose a significant amount of precious parallel data. Therefore, we implemented smart truncation to deal with longer sequences in the parallel corpus. The truncation is set at the sequence length of 90.

Cross Validation In low-resource MT training, it is important to implement Cross Validation (CV) to ensure robust predictive performance and address problems like overfitting. In this case, where the training corpus is small, CV can provide insights on the variability. We opt for 5-fold CV to compare training results. After text preprocessing, the cleaned text are randomly shuffled and split into 5 chunks. The subsets are then concatenated respectively before training. For our baseline systems, 4 of 5 iterations have the subset size of 33813 for training and 8453 for test. The last iteration has the size of 33812 and 8454 respectively.

System Implementation of all three systems is carried out as explained in Section 3.2. We utilized the MT development toolkit Sockeye [55] for BPE encoding, model training and evaluation.

Statistical Significance For statistical significance analysis, our experimental setup needs to take the multiple comparison problem into account. When testing multiple hypotheses simultaneously, the increased number of statistical inferences leads to increased probability of inexact inferences and Type I errors, making the conventional p threshold of 0.05 less reliable. This is a well-known problem, e.g. in the Genome- and Public Health-related research [56, 57].

Methods that counteract multiple testing generally adjust α so that the chance of observing inaccurate significant result is reduced. The Bonferroni correction is the simplest (and fairly conservative) approach to cut off the α value. Bonferroni corrects the α by considering the set of n comparisons, causing the α threshold to become α/n . With the Bonferroni correction, the p -value is set to 0.017 as opposed to 0.05.

5 Evaluation

5.1 Metrics

Despite the popularity of BLEU, recent studies from [19] and [58] questioned the phenomenon of using BLEU as a single metric, especially in low-resource scenarios, where language structures and scripts are complex and different from many high-resource languages. For example, the meta evaluation on Indian languages by [59] reported higher human judgement correlation using COMET [60] as opposed to BLEU. The limitation of BLEU also lies in the strong dependence on reference translation, whose quality can be highly unstable, especially when data is noisy. Issues such

	Model	BLEU	chrF	TER
bar-de	Baseline	66.0	78.1	32.7
	Back-translated	73.4	82.5	25.0
	Transferred	53.9	70.5	41.9
de-bar	Baseline	61.2	74.4	36.2
	Back-translated	63.4	76.3	31.9
	Transferred	48.2	63.9	44.4

Table 1: Overview of best performing models from each system

as translationese and poor reference diversity [20] might also jeopardize the entire evaluation. We therefore include chrF and TER for a more diverse evaluation. ChrF is language-independent and has been reported to better capture complex morpho-syntactic structures in MT evaluation [17]. TER (Translation Error Rate) quantifies the amount of edit operations it takes to change the system output to match the reference translation [18]. This intuitive technique avoids knowledge-intensive calculations and focuses on matching hypothesis with reference. The main advantage of TER as opposed to BLEU is the lower penalty for phrasal shifts. TER has also been reported to correlate highly with human judgement and has been implemented in recent WMT tasks [12, 61].

5.2 System 1: Baseline

Despite the lack of sufficient amount of parallel data, baseline models in both translation directions exceed 60 BLEU (see Table 1). For bar-de baseline, BLEU scores have an average of 66, chrF has an average of 78 and TER 33. We want to point out little variation between the folds - indicating that the results are robust. However, we observe relatively lower scores on the opposite direction, namely an average of 61 BLEU, 74 chrF and 36 TER. Variation are also small for the de-bar base systems.

5.3 System 2: Back-translation

Back-translation (BT) was applied to the best performing baseline folds with monolingual data. Significant improvements can be observed in all three metrics for bar-de, whereas de-bar systems show subtle increase. In contrast to baseline systems, we observe a systematic increase of standard deviation. Where SD was between 0.3 and 0.6 for base systems, 0.7 to 2.2 SD was found in back-translated systems.

5.4 System 3: Transfer Learning

In contrast to surprisingly high baselines, both parent models perform similarly moderate, the fr-de model scored 29 BLEU, 52 chrF and 65 TER, whereas the de-fr parent reached 30 BLEU, 53 chrF and 65 TER. Given the fact that the German-French corpus size is significantly bigger than the German-Bavarian corpus, we had expected better performance of the parent models. However, our results are comparable with available German-French models on Hugging Face, for instance the one from Helsinki-NLP⁵.

Despite the parents' BLEU scores are only a half of our baseline models, Transfer Learning improves children's performance considerably. For bar-de, the best system has 54 BLEU, 71 chrF and 42 TER, which is an increase of 25 BLEU and 19 chrF and decrease of 23 TER. For de-bar, the best model scored 51 BLEU, 65 chrF and 43 TER, which has a performance leap of 21 BLEU, 12 chrF and 22 TER from parent. We note that Transfer Learning improved translation capacity from parent to child with an enhancement of more than 20 BLEU. This corroborates with the recent studies on the use of Transfer Learning for low-resource languages. However, these improvement cannot compare with the very high baseline systems and their back-translated extensions.

5.5 Statistical Analysis

Two-tailed pairwise t-tests were conducted on all pairs with Bonferroni correction (p threshold is 0.017). Test statistics are shown in Tables 2 and 3. For bar-de models, the BLEU results from baseline ($M = 65.7$, $SD = 0.2$) and BT ($M = 70.5$, $SD = 2$) indicate that Back-translation leads to significant improvement, $t = -4.89$, $p = 0.0036$. BT also performs significantly better than transferred systems ($M = 52.8$, $SD = 0.7$), $t = 17.25$, $p < 0.0$. Further statistics from the metrics chrF and TER corroborate these findings.

⁵<https://huggingface.co/Helsinki-NLP/opus-mt-fr-de>

Metric	Group 1	Group 2	<i>t</i>	<i>p</i>	<i>p</i> (corr.)	Reject H_0
BLEU	Baseline	BT	-4.89	0.0012	0.0036	True
	Baseline	Transfer	37.86	0.0	0.0	True
	BT	Transfer	17.25	0.0	0.0	True
chrF	Baseline	BT	-5.83	0.0004	0.0012	True
	Baseline	Transfer	20.65	0.0	0.0	True
	BT	Transfer	19.82	0.0	0.0	True
TER	Baseline	BT	6.1	0.0003	0.0009	True
	Baseline	Transfer	-19.29	0.0	0.0	True
	BT	Transfer	-16.2	0.0	0.0	True

Table 2: Results of t-test with Bonferroni correction for bar-de systems.

Metric	Group 1	Group 2	<i>t</i>	<i>p</i>	<i>p</i> (corr.)	Reject H_0
BLEU	Baseline	BT	-2.85	0.0214	0.0641	False
	Baseline	Transfer	29.58	0.0	0.0	True
	BT	Transfer	22.04	0.0	0.0	True
chrF	Baseline	BT	-3.84	0.005	0.0149	True
	Baseline	Transfer	30.12	0.0	0.0	True
	BT	Transfer	26.28	0.0	0.0	True
TER	Baseline	BT	5.02	0.001	0.0031	True
	Baseline	Transfer	-23.74	0.0	0.0	True
	BT	Transfer	-15.91	0.0	0.0	True

Table 3: Results of t-test with Bonferroni correction for de-bar systems.

For de-bar models, the tendency is similar. ChrF results show a positive enhancement from baseline ($M = 74.1$, $SD = 0.4$) to BT ($M = 75.5$, $SD = 0.7$), $t = -3.84$, $p = 0.149$. The improvement of BT over transferred systems ($M = 64.2$, $SD = 0.6$) is significant as well. TER statistics also verify these findings. Interestingly, while chrF and TER successfully rejects the null hypothesis between baseline and BT performance, BLEU does the opposite. We argue that the results are nevertheless significant based on chrF and TER, and consider this disagreement between metrics as an occurrence derived from linguistically-different perspectives and computations.

5.6 Qualitative Analysis

We argue that the surprisingly high baseline results come from the similarity of the source and target languages. This corresponds to findings from [49] that language relatedness contributes positively to MT quality. The analysis of [23]’s multilingual NMT on Indo-Aryan languages lists linguistic characteristics such as word-order construction, degree of inflection, amount of similar word root, meaning and conjunct verbs as the key drivers for improving training. Our experiments corroborate these argumentation, thus answering **RQ1**.

The significant improvement from Back-translation, which can be seen with all metrics, aligns well with previous findings. Especially in the submitted systems for WMT21 Very Low Resource Supervised MT between Upper Sorbian and German by [14], Back-translation boosted the training corpus size and contributed to performance increase. However, we are aware of its limits. For instance, the augmented text includes many errors, which were inherited from the baseline systems. This issue of *Translationese* [62] is widely discussed, especially in the context of using silver-paired data for MT. In our case, we have opted for a smaller amount of augmented data, with the aim to reduce Translationese as much as possible while still allowing model improvement. We therefore answer **RQ2** that Back-translation contributes positively.

Regarding **RQ3**, we point out that while Transfer Learning did improve performance from parent to child, its final performance was not sufficient to exceed the other two systems.

We note that our results are similar to the ones from the German - Upper Sorbian translation task from WMT21. Our baseline and back-translated models have an accuracy range between 60 to 73 BLEU and 74 to 82 chrF, comparable with the final scores from the German - Upper Sorbian task. However, it is interesting to note that their chrF scores are substantially higher than ours (by 10), while our BLEU scores are similar. This brings us back to the notion that all metrics work linguistically different and these variations reflect through different languages.

German Input	System	Bavarian Output
sie hat heute abend im restaurant fisch bestellt.	Base	se hod heit abend im restaurant fisch bestöid.
	BT	se hod heid obend im restaurant fisch bestejd.

Table 4: Examples of German to Bavarian translation.

Furthermore, a common finding can be observed between our experimental results and the WMT21 experiments we compare against, namely the result discrepancy between high-to-low and low-to-high directions. In our study, de-bar is ca. 10 BLEU and 10 chrF behind bar-de. Similarly but not as extreme, Upper Sorbian - German also performs better than its high-to-low counter direction. This performance gap on the same corpus but different translation directions raises attention, with possible reasons due to the multiple orthographic standards and sub-dialects in our case.

Table 4 depicts two translation examples. We translate the German phrase “Sie hat heute Abend im Restaurant Fisch bestellt” (English meaning “she ordered fish in the restaurant tonight.”) into Bavarian using all of our systems. We observe that while Base and BT outputs look similar, their differences could come from various sub-dialects in the corpus. For instance, the term “heute” was translated into “heit” and “heid”, with only the last consonant different. However, in the Germanic linguistics, these consonants “t” and “d” differ themselves in voice. The linguistic notion of *Fortis and Lenis*⁶ differentiates oral pressure that is given to these consonants. Thus, we suspect these differences come from various dialects.

6 Conclusion

In this paper, we presented experimental work in Neural Machine Translation with the aim to push forward our understanding of how to best address the gap between a handful of well-resourced languages and the long tail of languages for which no sufficient resources are available. More specifically, we focused on methods and case studies that have shown promising results for languages with limited resources. We conceptualized the problems of noisy data and data shortage by way of recent studies. We revisited creative solutions designed to combat these challenges such as Back-translation, multilingual training and language relatedness. Our own low-resource implementation utilized data augmentation and cross-lingual transfer on German and Bavarian. We report our steps to preprocess the corpus and carry out training for three bidirectional systems. 5-fold cross validation was carried out on each system to compare robustness. We opted for a combined metric system using BLEU, chrF and TER to evaluate translation from different perspectives. For multiple hypothesis testing, pairwise t-tests with Bonferroni correction were conducted to test for statistical significance. Results show that translation between similar languages performs generally better and that augmented data contribute positively. However, even though cross-lingual transfer showed huge improvement from parent to child, it was not able to exceed baseline and back-translated models. We recognize that Transfer Learning is an effective approach for low-resource languages, but note that in our study language similarity played a more important role. To support reproducibility and replicability all code is made available via GitHub.

7 Limitations

The Bavarian orthography has been a known problem for decades, as it is mostly a spoken language and has not been properly standardized. For example, the word ‘Bavarian’ alone can be written in two ways: Boarisch or Bairisch. The investigation by [63] illustrates that there are multiple Bavarian orthographic conventions. From a computational perspective, the issue is “deciding which representation should be given precedence”, as stated in the Bri bri case study by [11]. Overcoming dialectal variations is also a problem of politics that can carry on for years. In light of the findings by [64], we would add that the automated translation of Bavarian should - like other under-sourced languages - be carefully planned with ethical considerations, and that purely using web-scraped data to deploy translation systems might neglect the concerns of speakers. Another challenge lies in multiple sub-dialects. This phenomenon can be observed in our corpus, which is mined from the Bavarian Wikipedia, where articles are written in different regional dialects. We argue that these sub-dialects in the parallel corpus lead to translation confusion, resulting in translation outputs which consist of mixed accents. Nevertheless, should there be a more refined and organized corpus of a particular sub-dialect, our systems can serve as baselines for fine-tuning. Another, more general limitation is the fact that throughout our work we conducted purely technical evaluations. The strength of such an experimental setup is that it can be reproduced and offers objective results. However, it is clearly necessary to involve native speakers to gain more insights into the quality of any translation process. We mitigated against the problem by choosing not just a

⁶https://en.wikipedia.org/wiki/Fortis_and_lenis

single evaluation metric (such as BLEU), but no matter how many different metrics are chosen they are no substitute for user studies.

8 Future Work

Following our findings and the limitations stated above, we propose further research directions to inspire future work: First, the curation of a more refined and organized parallel corpus for modern German-Bavarian to help establish a high quality benchmark for training and evaluation. An example to achieve this is through recruiting native speakers in both Bavarian and German who have an adequate amount of linguistic knowledge. This annotation could include not only translation of parallel sentences, but also the sub-dialects or Bavarian regional variations the speakers associate themselves with. This human-annotated dataset could furthermore be split into two parts, one for training and another for evaluation.

Additionally, identification of dialects would be an approach to counter translation confusion and mixed accents. This could help unify and isolate non-standardized languages or dialects. As mentioned in the previous section, a great way to start modelling sub-dialect detection is to automatically analyze the Wikipedia articles with their corresponding sub-dialects. This would greatly reduce the training corpus size, but additional measures to increase the corpus size could be taken, such as acquiring diverse datasets (i.e. open-source subtitles of Bavarian TV-programs or historical documents). More generally, we see our work as a reference benchmark for future work – be it to explore the same language pair further or other work into the general problem of low-resource language translation efforts.

9 Ethical Considerations

Ethical concerns arise whenever natural language is being sampled and used to train machine learning systems. For this experimental work we used existing test collections and other freely accessible data. All the experiments are conducted within the ethical framework imposed on us by our institution. In this context we did not identify a specific ethical issue.

However, it is clear that once any automated translation system is on its way to be deployed that care must be taken to (a) train it on *representative* samples, (b) mitigate against common biases, and (c) make sure no personal information is included in the training data. If trained on social media data there is also a risk that toxic content might surface. Care must be taken to take these issues seriously (rather than treating this as a box-ticking exercise), but we would argue that there are no ethical concerns arising from this work that have not already been identified previously.

10 Acknowledgment

We would like to thank the anonymous reviewers for their constructive feedback.

References

- [1] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. Phrase-Based & Neural Unsupervised Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium, 2018. Association for Computational Linguistics.
- [2] Meng Zhang, Liangyou Li, and Qun Liu. Two Parents, One Child: Dual Transfer for Low-Resource Neural Machine Translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2726–2738, Online, 2021. Association for Computational Linguistics.
- [3] Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas, 2016. Association for Computational Linguistics.
- [4] Nouman Ahmed, Natalia Flechas Manrique, and Antonije Petrović. Enhancing Spanish-Quechua machine translation with pre-trained models and diverse data sources: LCT-EHU at AmericasNLP shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 156–162, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [5] Stephane Clinchant, Kweon Woo Jung, and Vassilina Nikoulina. On the use of BERT for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 108–117, Hong Kong, November 2019. Association for Computational Linguistics.

- [6] Kaiyu Huang, Peng Li, Jin Ma, Ting Yao, and Yang Liu. Knowledge transfer in incremental learning for multilingual neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15286–15304, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [7] Aaron Mueller, Garrett Nicolai, Arya D McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. An Analysis of Massively Multilingual Neural Machine Translation for Low-Resource Languages. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, pages 3710–2718. European Language Resources Association (ELRA), 2020.
- [8] Röee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [9] Raj Dabre, Atsushi Fujita, and Chenhui Chu. Exploiting Multilingualism through Multistage Fine-Tuning for Low-Resource Neural Machine Translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1410–1416, Hong Kong, China, 2019. Association for Computational Linguistics.
- [10] Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [11] Isaac Feldman and Rolando Coto-Solano. Neural Machine Translation Models with Back-Translation for the Extremely Low-Resource Indigenous Language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online), 2020. International Committee on Computational Linguistics.
- [12] Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. Findings of the 2021 Conference on Machine Translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online, November 2021. Association for Computational Linguistics.
- [13] Jindřich Libovický and Alexander Fraser. Findings of the WMT 2021 Shared Tasks in Unsupervised MT and Very Low Resource Supervised MT. In *Proceedings of the Sixth Conference on Machine Translation*, pages 726–732, Online, November 2021. Association for Computational Linguistics.
- [14] Rebecca Knowles and Samuel Larkin. NRC-CNRC systems for Upper Sorbian-German and Lower Sorbian-German machine translation 2021. In *Proceedings of the Sixth Conference on Machine Translation*, pages 999–1008, Online, November 2021. Association for Computational Linguistics.
- [15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.
- [16] 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
- [17] Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics.
- [18] 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.
- [19] Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online, November 2021. Association for Computational Linguistics.

- [20] Markus Freitag, David Grangier, and Isaac Caswell. BLEU might be guilty but references are not innocent. *CoRR*, abs/2004.06063, 2020.
- [21] Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).
- [22] Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. AfroMT: Pretraining Strategies and Reproducible Benchmarks for Translation of 8 African Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1306–1320, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.
- [23] Vikrant Goyal, Sourav Kumar, and Dipti Misra Sharma. Efficient Neural Machine Translation for Low-Resource Languages via Exploiting Related Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 162–168, Online, 2020. Association for Computational Linguistics.
- [24] Alexandra Chronopoulou, Dario Stojanovski, and Alexander Fraser. Improving the Lexical Ability of Pretrained Language Models for Unsupervised Neural Machine Translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 173–180, Online, 2021. Association for Computational Linguistics.
- [25] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia, 2018. Association for Computational Linguistics.
- [26] Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations*, 2018.
- [27] Emily M. Bender. The #BenderRule: On Naming the Languages We Study and Why It Matters, 2019. <https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/>.
- [28] 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 *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 6097–6110, Hong Kong, China, 2019. Association for Computational Linguistics.
- [29] Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. Selection Criteria for Low Resource Language Programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).
- [30] Louisa Lambrecht, Felix Schneider, and Alexander Waibel. Machine translation from Standard German to alemannic dialects. In Maite Melero, Sakriani Sakti, and Claudia Soria, editors, *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 129–136, Marseille, France, June 2022. European Language Resources Association.
- [31] 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 2016. Association for Computational Linguistics.
- [32] Ivan Prosvilov, Dmitrii Emelianenko, and Elena Voita. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online, July 2020. Association for Computational Linguistics.
- [33] Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. Search engine guided non-parametric neural machine translation. *CoRR*, abs/1705.07267, 2017.
- [34] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Nearest neighbor machine translation. In *International Conference on Learning Representations*, 2021.
- [35] Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. Chunk-based nearest neighbor machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4228–4245, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [36] Tom Koci and Ondřej Bojar. Trivial Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Belgium, Brussels, 2018. Association for Computational Linguistics.

- [37] Zhaocong Li, Xuebo Liu, Derek F. Wong, Lidia S. Chao, and Min Zhang. ConsistTL: Modeling consistency in transfer learning for low-resource neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8383–8394, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [38] Kenji Imamura and Eiichiro Sumita. Recycling a pre-trained BERT encoder for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 23–31, Hong Kong, November 2019. Association for Computational Linguistics.
- [39] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.
- [40] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291, 2019.
- [41] Mozhdeh Gheini, Xiang Ren, and Jonathan May. Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1765, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics.
- [42] Akshay Batheja and Pushpak Bhattacharyya. Improving machine translation with phrase pair injection and corpus filtering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5395–5400, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
- [43] Etienne Denoual and Yves Lepage. BLEU in characters: Towards automatic MT evaluation in languages without word delimiters. In *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*, 2005.
- [44] Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. Multi-task learning for multiple language 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 1723–1732, Beijing, China, July 2015. Association for Computational Linguistics.
- [45] Ye Kyaw Thu, Win Pa Pa, Masao Utiyama, Andrew Finch, and Eiichiro Sumita. Introducing the Asian language treebank (ALT). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1574–1578, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).
- [46] Maha Elbayad, Anna Sun, and Shruti Bhosale. Fixing MoE over-fitting on low-resource languages in multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14237–14253, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [47] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *CoRR*, abs/1904.09751, 2019.
- [48] Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors. *Proceedings of the Sixth Conference on Machine Translation*. Association for Computational Linguistics, Online, November 2021.
- [49] Ife Adebara, El Moatez Billah Nagoudi, and Muhammad Abdul Mageed. Translating similar languages: Role of mutual intelligibility in multilingual transformers. In *Proceedings of the Fifth Conference on Machine Translation*, pages 381–386, Online, November 2020. Association for Computational Linguistics.
- [50] Meng Zhang, Minghao Wu, Pengfei Li, Liangyou Li, and Qun Liu. NoahNMT at WMT 2021: Dual transfer for very low resource supervised machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1009–1013, Online, November 2021. Association for Computational Linguistics.
- [51] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *CoRR*, abs/1511.06709, 2015.
- [52] Robert C. Moore and William Lewis. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

- [53] Jörg Tiedemann. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online, November 2020. Association for Computational Linguistics.
- [54] Philipp Koehn and Rebecca Knowles. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, 2017. Association for Computational Linguistics.
- [55] Tobias Domhan, Michael Denkowski, David Vilar, Xing Niu, Felix Hieber, and Kenneth Heafield. The sockeye 2 neural machine translation toolkit at AMTA 2020. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 110–115, Virtual, October 2020. Association for Machine Translation in the Americas.
- [56] M Aickin and H Gensler. Adjusting for multiple testing when reporting research results: The bonferroni vs holm methods, May 1996.
- [57] William S Noble. How does multiple testing correction work?, Dec 2009.
- [58] Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474, 12 2021.
- [59] Ananya Sai B, Tanay Dixit, Vignesh Nagarajan, Anoop Kunchukuttan, Pratyush Kumar, Mitesh M. Khapra, and Raj Dabre. IndicMT eval: A dataset to meta-evaluate machine translation metrics for Indian languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14210–14228, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [60] Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November 2020. Association for Computational Linguistics.
- [61] Nitika Mathur, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Online, November 2020. Association for Computational Linguistics.
- [62] Yvette Graham, Barry Haddow, and Philipp Koehn. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online, November 2020. Association for Computational Linguistics.
- [63] Ludwig Zehetner. Zur schreibung des bairischen. *Schmankerl*, 37:31–32, 1978.
- [64] Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada, July 2023. Association for Computational Linguistics.