Quantifying the Ethical Dilemma of Using Culturally Toxic Training Data in AI Tools for Indigenous Languages

Pedro Henrique Domingues, Claudio Pinhanez, Paulo Cavalin, Julio Nogima
PUC-Rio (first author), IBM Research (remaining authors)
phd.engmec@gmail, {csantosp,pcavalin,jnogima}@br.ibm.com

Abstract

This paper tries to quantify the ethical dilemma of using culturally toxic training data to improve the performance of AI tools for ultra low-resource languages such as Indigenous languages. Our case study explores the use of Bible data which is both a commonly available source of training pairs for translators of Indigenous languages and a text which has a trail of physical and cultural violence for many Indigenous communities. In the context of fine-tuning a WMT19 German-to-English model into a Guarani Mbya-to-English translator, we first show, with two commonly-used Machine Translation metrics, that using only Bible data is not enough to create successful translators for everyday sentences gathered from a dictionary. Indeed, even fine-tuning with only 3,000 pairs of data from the dictionary produces significant increases in accuracy compared to Bible-only models. We then show that simultaneously fine-tuning with dictionary and Bible data achieves a substantial increase over the accuracy of a dictionary-only trained translator, and similarly happens when using two-step methods of fine-tuning. However, we also observed some, measurable, contaminated text from the Bible into the outputs of the best translator, creating concerns about its release to an Indigenous community. We end by discussing mechanisms to mitigate the negative impacts of this contamination.

Keywords: Machine Translation, Indigenous Languages, Domain Contamination

1. Introduction

One of the most common ethical concerns in the development of *Artificial Intelligence (AI)* and *Machine Learning (ML)* systems are the presence of toxic content in the training data which can sometimes spill over to the final systems (Abbasi et al., 2022; Van Aken et al., 2018). The most advocated solutions to the problem involve the removal of the toxic elements from the training sets (Mehrabi et al., 2021) or its detection and removal from the outputs of the system (Garg et al., 2023). However, in the case of *ultra-low resource (ULR)* languages, i.e. languages with so low resources that religious texts such as the Bible comprise the largest source of data (such as many Indigenous languages), the exiguity of training data available creates an ethical-technical dilemma since the removal of toxic training content may render the final system unfeasible due to the lack of sufficient training data.

In this paper we address this dilemma in the context of creating a *Guarani Mbya*-to-*English* machine translation (MT) system for Indigenous communities in Brazil. The *Guarani Mbya* language is spoken by approximately 8,000 people, mostly in the South-Southeast area of Brazil, and, although being a language still actively spoken and well-studied, it has very few sources of translated texts which can be used to mine bilingual pairs of sentences essential for the training of today's ML translators.

State-of-the-art translators, such as the `WMT19` German-to-English translator used in this work (Ng et al., 2019), are trained with hundreds of millions of sentence pairs, including original sources such as translations of known books, web data, and synthetically generated data based on linguistic knowledge. In contrast, for most ULR languages, even finding tens of thousands of bilingual pairs is difficult, often having to rely on dictionaries, tales and other cultural narratives, and translations of religious materials such as the Bible and the Qur'an. Moreover, given this lack of training data for ULR languages, a popular technique to create AI tools for those languages is to *fine-tune* a *large language model (LLM)* with small amounts of data from the targeted final language.

However, for Indigenous peoples in the Americas, translations of the Bible are connected to a history of violence to convert Indigenous peoples to those religions (Franchetto, 2008) and to colonialist practices (Stoll, 1982) and therefore negatively viewed by many communities. As argued by Nunpa (2020), a Dakota author, *"the Bible was a tool for the colonization process [...working] hand-in-hand in the exploitation, subjugation, and continued oppression of the Indigenous Peoples of the U.S."*. Similarly, Ogden (2005), a California Indian writer, points out that *"at the beginning of the colonization process two tools of genocide were forced upon Native people: the bottle and the bible."* Therefore, we consider here the Bible, in the Indigenous context, as a potential toxic source of training material, that is, training data which can potentially pushes ML systems and translators to produce undesirable or offensive text.

At the same time, and also considering that the Bible is a text sacred to millions of people in the world, including to members of Indigenous com-

SIGUL2024 Workshop, pages 283–293
21-22 May, 2024.

munities, we use in this work the term *culturally toxic* to strengthen the cultural context of the toxicity of the data. In fact, as noted by Sheth et al. (2022), *"... culture provides essential context to the determination of any toxic content"*, and therefore we consider that it is more appropriate to refer to the Bible as *culturally toxic* content in the specific context of Indigenous languages of this paper.

We explore here different methods to use Guarani Mbya data to fine-tune a `WMT19` translator, including about 3,300 sentences from a comprehensive dictionary and from a compilation of traditional tales (culturally non-toxic data) and 4,000 pairs from a translation of the New Testament of the Bible (which is, in our view, culturally toxic data). We also study whether *multilingual* approaches such as fine-tuning with data from translation of the Bible to related Indigenous languages, which can provide more training data, help or hinder the development of a Guarani Mbya translator able to handle everyday sentences.

Ideally, AI systems for Indigenous languages should not be biased by content from the Bible, to not perpetuate even further the memory and impact of past abuses. Therefore, avoiding biblical data is the safest solution for this problem, an ethical decision which may cause diminished accuracy. This work fills a gap of the research in this area by studying the counter-balancing effects of Bible data with additional commonly-available data such as dictionaries, quantifying the impacts both in accuracy and output contamination, and discussing the ethical impacts of the results.

Considering commonly-used metrics to measure the quality of MT systems, our study found that fine-tuning only with Bible data produces poor translators, significantly worse than fine-tuning only with the non-toxic dictionary and tales data. However, we also found that using a two-step fine-tuning process, first with the culturally toxic and then with non-toxic data, or simultaneously fine-tuning with non-toxic and culturally toxic data, produces translators with the same quality for dictionary and tales, which are, at the same time, also significantly better for Bible content. We then did a detailed qualitative analysis of 300 outputs of the mixed input translator, finding 2 clear cases and 12 other with content potentially linked to the Bible (4.7%). We finish the paper by discussing ways to mitigate the negative effects of culturally toxic data.

This paper explores, in a quantitative way, an important ethical issue present in many scenarios of ML tools for ULR languages and contributes by providing actionable data about the advantages and disadvantages of the use of culturally toxic data. This work also contributes to a deeper understanding of fine-tuning methods by suggesting that diverse sources of fine-tuning data, even in very small amounts, seem to have a large positive impact in the performance of fine-tuned systems and at the same time, are detectable in the outputs. The most important contribution of this paper is the quantification of the levels of performance improvement and contamination which, although suggested in other works, were never actually measured, especially for very small fine-tuning datasets.

2. Related Work

Large Language Models (LLMs) are currently a big trend in *Natural Language Processing (NLP)* and one of the biggest promises of AI technology. Such models have proved to be useful to speed up the development of increasingly better applications for problems such as text classification (Devlin et al., 2019) and machine translation (Raffel et al., 2020). More recently, the potential of LLMs was delivered to the masses with the release of LLM-based personal assistants such as ChatGPT (OpenAI, 2022).

The main approach behind LLMs consists of training a *Transformer* neural network (Vaswani et al., 2017), or only a part of it, on large amounts of self-supervised data, relying on auto-regressive and masked language modelling learning objectives (Devlin et al., 2019; Liu et al., 2020a; Raffel et al., 2020; Chowdhery et al., 2022). Then, an LLM can be used directly for a downstream application either in a zero-shot manner or by passing instructions in the input, what is usually called *prompt engineering/tuning* (Liu et al., 2023).

Another way to employ LLMs is *fine-tuning* its parameters to more specific downstream datasets, so that the knowledge of the base, general-purpose language model is transferred to a more specific problem, usually involving a more restricted domain (Zhou and Srikumar, 2022; Arase and Tsujii, 2019). In comparison, fine-tuning is usually more costly than prompt tuning since it requires adjusting parameters of the model and that can be a computationally-intensive job. On the other hand, fine-tuning might be the only option for some cases, for instance teaching a new language to an LLM, or getting the best out of very small training datasets.

2.1. Fine-Tuning LLMs

Since the goal of fine-tuning is to transfer knowledge from a general-purpose model to a more specific task, the fine-tuning process normally involves two steps (Wei et al., 2022). The first step consists of pre-training a neural network with self-supervised data (Devlin et al., 2019; Brown et al., 2020). Next, in the second step, its parameters are fine-tuned on a downstream dataset with annotated data for applications such as classification, question answering, machine translation (Raffel et al., 2020).

Another approach that is gaining popularity is to conduct intermediate steps of fine-tuning before generating a final model, an approach usually referred to as *intermediate training* or *intertraining* (Ein-Dor et al., 2022). Intermediate training can be done by using additional pre-training steps with self-supervised domain specific data (Pruksachatkun et al., 2020) or by fine-tuning a model on a larger dataset, usually related to the downstream dataset, before the final downstream fine-tuning (Phang et al., 2019; Gururangan et al., 2020).

2.2. Multilingual Training

Aiming at improving translation quality for low-resource languages, multilingual training emerged as a sought-after solution. This method consists of using corpora of multiple languages at once, to leverage shared linguistic features among diverse but related languages (Aharoni et al., 2019; Dabre et al., 2020).

The way multilingual training is implemented depends on the stage at which it is used, and the final task. Multilingual datasets can be used during *pre-training* often by mixing data from several languages in a single training set (Liu et al., 2020b; Xue et al., 2021). When handling downstream datasets, such as machine translation corpora, one can rely on creating multi-way translations where the source or target language is usually specified (Dabre et al., 2019; Mueller et al., 2020).

The Bible is a document which have translations for several languages in the world, including many Indigenous languages. For this reason, the Bible has been used to test the feasibility of current NLP tools for such languages, and multiple works with low-resource languages have shown that such content can help the construction of MTs, particularly multilingual ones, and often as an additional source (Mayer and Cysouw, 2014; Bollmann et al., 2021; Vázquez et al., 2021; Nagoudi et al., 2021; Adelani et al., 2022). In the case of Indigenous languages, exploring such a source of data is an important option given the scarceness of data and the common availability of translations of the Bible. However, the use of the Bible in the context of Indigenous languages is problematic, not only due to its association to a history of abuse and colonialism but also because the translation process is often marred with poor quality and a Western-centred view (Franchetto, 2008; Stoll, 1982).

This paper contributes in quantifying to which extent using the Bible as training data is beneficial and harmful in terms of generating texts at inference time, considering cultural issues of Indigenous communities with this document.

3. Working Ethically with Indigenous Languages and Communities

Working with Indigenous communities and languages is the subject of specific guidelines and legal issues. Mihesuah (1993) gives a comprehensive set of guidelines for research with US American Indigenous communities. Straits et al. (2012) is an example of research guidelines on how to engage in research with Native US American communities, both in more traditional research and cases where technology development and deployment is involved. Besides the ethical considerations, there are specific legal and regulatory procedures which have to be followed in different countries and when working with specific Indigenous communities (Harding et al., 2012). Specific provisions are needed related to data ownership and sovereign rights since those concepts may be understood differently by the community (Harding et al., 2012; Sahota, 2007).

The use of technology for documentation and vitalization is discussed as part of the UNESCO engagement framework known as the *Los Pinos Declaration*[1]. For AI-related work, a good proposal is the *The Indigenous Protocol and Artificial Intelligence (A.I.) Working Group* (Lewis et al., 2020), the result of two workshops with Indigenous leaderships, linguistic professionals, and computer researchers.

We follow here the methods proposed by Pinhanez et al. (2023) to mitigate and control the negative effects of using religious texts in Indigenous contexts by creating a "containment process" where the team was made aware of the potential harmful aspects of the culturally toxic data for Indigenous communities. Also, we do not plan to make available this data or the created prototypes and tools publicly, as a way to avoid unwanted releases. Researchers interested in checking or duplicating our results can contact us to access the data and code under strict conditions.

This work is related to a collaboration with the *Tenondé Porã* Indigenous community in the South of São Paulo City, comprising about 3,000 etnical *Guaranis* who use the *Guarani Mbya* as their primary language. The collaboration has focused on the creation of writing-support tools for high-school native students fluent both in Guarani Mbya and Portuguese.This collaboration informs the use of Guarani Mbya as the language in this study.

[1] https://www.worldindigenousforum.com/products/los-pinos-declaration-chapoltepek-outcome-document

4. The Datasets Used in the Study

In this study we considered two datasets which cover two extremes of the toxicity versus performance dilemma. The first one, *Dictionary*, consists of limited non-toxic data, with a small number of sentences with a large proportion of short sentences. The second one, *Bibles*, contains translations of the Bible and, as mentioned, can be considered a culturally toxic dataset in the Indigenous languages context, but it is larger and contains longer and more elaborated sentences than the former.

4.1. The *Dictionary* Dataset

Sentences from three different sources were used in the construction of *Dictionary* dataset. The first source was a set of Guarani Mbya short stories with 1,022 sentences, also available in Portuguese and English (Dooley, 1988a,b). The second comprises 245 texts extracted from PDF files with a pedagogical character (Dooley, 1985). The third source was Robert A. Dooley's *Lexical Guarani Mbya dictionary* (Dooley, 2016), a reference work for the language, from which we extracted 2,230 sentence pairs, and the reason why the dataset was named Dictionary. The last two sources contained sentence pairs in Guarani Mbya and Portuguese only. We converted them to English using a Portuguese-to-English commercial translation service. We have permission from the author to use this data.

After concatenating the data from the three sources, we cleaned it, removing some non-alphanumeric characters (e.g. *, ≫, •) and normalizing Unicode values. Then, the Dictionary dataset was split into training and test sets and finalized by removing repeated sentences in each set and cross-contamination between sets, totaling 3,155 and 300 sentences pairs, respectively.

4.2. The Culturally Toxic *Bibles* Dataset

We use in this work translations of the *New Testament* of the Bible, a book which comprises about 7,000 sentences in its English versions, to 39 Indigenous languages spoken in Brazil. Brazil has been home to about 270 Indigenous languages according to the Census of 2010, the last comprehensive assessment of linguistic diversity in Brazil (IBGE, 2010). These languages are spoken by approximately 800 thousand people (IBGE, 2010), half of them living in Indigenous lands. Storto (2019) provides a good overview of the history, structure, and characteristics of *Brazilian Indigenous Languages (BILs)*. Almost all of those languages are considered endangered (Moseley, 2010). We adopted here the Indigenous language classification, nomenclature, and data from the 2010 Brazilian Census by IBGE (IBGE, 2010) and language acronyms according to ISO 639-3.

Indigenous Languages					# Aligned Sentences		
Name	Acron	Branch	Family	Speakers	Train	Test	Total
Bororó	bor	Macro-Jê	Bororó	1035	1861	202	2063
Apinayé	apn	Macro-Jê	Jê	1386	877	75	952
Kaingáng	kgp	Macro-Jê	Jê	19905	5695	917	6612
Kayapó	txu	Macro-Jê	Jê	5520	2669	510	3179
Xavánte	xav	Macro-Jê	Jê	11733	1275	342	1617
Karajá	kpj	Macro-Jê	Karajá	3119	2828	333	3161
Maxakali	mbl	Macro-Jê	Maxakali	1024	5566	905	6471
Rikbaktsa	rkb	Macro-Jê	Rikbaktsa	10	3560	710	4270
Mawé	maw	Tupi	Mawé	8103	6381	970	7351
Mundurukú	myu	Tupi	Mundurukú	3563	3110	190	3300
Guajajára	gub	Tupi	Tupi-Guarani	8269	4956	934	5890
Guaraní (West Bolivia)	gnw	Tupi	Tupi-Guarani	NA	5263	970	6233
Guaraní (East Bolivia)	gui	Tupi	Tupi-Guarani	NA	5263	924	6187
Guaraní Kaiowá	kgk	Tupi	Tupi-Guarani	24368	3034	479	3513
Guaraní Mbyá	gun	Tupi	Tupi-Guarani	3248	6340	970	7310
Guaraní (Paraguay)	gug	Tupi	Tupi-Guarani	NA	5196	970	6166
Ka'apor	urb	Tupi	Tupi-Guarani	1241	3380	436	3816
Kaiabi	kyz	Tupi	Tupi-Guarani	673	2187	280	2467
Nheengatu (LGA)	yrl	Tupi	Tupi-Guarani	3771	5035	691	5726
Tenharim	pah	Tupi	Tupi-Guarani	32	3215	844	4059
Jamamadí-Kanamanti	jaa	no branch	Arawá	217	4759	715	5474
Kulina Madijá	cul	no branch	Arawá	3043	4319	697	5016
Paumarí	pad	no branch	Arawá	166	3653	372	4025
Apurinã	apu	no branch	Aruak	824	6329	970	7299
Palíkur	plu	no branch	Aruak	925	6137	904	7041
Paresí	pab	no branch	Aruak	122	6381	970	7351
Teréna	ter	no branch	Aruak	6314	6381	970	7351
Wapixána	wap	no branch	Aruak	3154	5081	853	5934
Kadiwéu	kbc	no branch	Guaikurú	649	4523	793	5316
Apalaí	apy	no branch	Karib	252	5548	970	6518
Bakairí	bkq	no branch	Karib	173	4000	317	4317
Hixkaryána	hix	no branch	Karib	52	4270	472	4742
Makuxi	mbc	no branch	Karib	4675	4900	940	5840
Nadëb	mbj	no branch	Makú	326	5213	811	6024
Nambikwára	nab	no branch	Nambikwára	951	2774	844	3618
Kashinawá (Peru)	cbs	no branch	Pano-Tacanan	3588	2136	130	2266
Tukano	tuo	no branch	Tukano	4412	3750	846	4596
Yanomámi	guu	no branch	Yanomámi	12301	1283	196	1479
Tikúna	tca	no branch	no family	30057	3097	386	3483
TOTAL	**39**	**3**	**16**	**169201**	**162225**	**25808**	**188033**

Table 1: Indigenous languages and corresponding size of the datasets used in the study. Language name, branch, family, and number of speakers (considering only who speak the language at home in an Indigenous land in Brazil) according to the table 1.13 of the Indigenous data of the Brazilian census of 2010 (IBGE, 2010).

Table 1 lists the 39 Indigenous languages used in this work which includes 36 languages spoken primarily in Brazil and 3 other Guarani languages used mostly in Paraguay and Bolivia but also spoken in some areas in Brazil.

The *Bibles* dataset consists of 188,033 parallel verses from the New Testament in English and their translations into these 39 Indigenous languages. The parallelism among translations of the same verse were done by the authors. We are aware that some of those translations were performed by non-specialists and have linguistic problems (Franchetto, 2008, 2020; Stoll, 1982). Also, since some of those translations were created as part of efforts to convert Indigenous peoples to Western religions, in particular to different forms of Christianism, such translations of the Bible are often not only associated to different forms of cultural abuse and violence to Indigenous communities but also, in many ways, are connected to *orthographies of domination* (Franchetto, 2008) and to questionable practices of indoctrination (Stoll, 1982). That is the main reason for referring to this dataset as culturally toxic in this work, since the use of this

data can result in MT systems which reproduce the language that is associated to cultural violence for Indigenous communities.

The Bibles dataset was split into training and test sets, considering the *Matthew* chapter for testing and the remaining content for training. As Guarani Mbya is the language under study in this work, all translators were evaluated in the test set of this language which comprises 970 sentences.

5. The Fine-Tuned Models

The models used in this study were obtained by performing different fine-tunings of the `WMT19` model (Ng et al., 2019), which is a 315M-parameter German-to-English machine translator pre-trained with about 28M pairs of translated sentences and more than 500M back-translated sentences. We have also evaluated other LLMs for this task, such as mBART and mT5, but `WMT19` presented the best results in terms of translation quality with these very small datasets. We suspect that, given that Guarani Mbya and most of the Indigenous languages related to this work were not included in the pre-training of either LLM, a smaller model is more suitable for this scenario with ULR languages involved.

As a baseline, we rely on the `zeroshot` model, consisting of the original German-English `WMT19` model without any fine-tuning. This model enables us to evaluate any intrinsic bias which the pre-training process may have introduced. Next, we describe the different fine-tuned models.

5.1. The Bibles-Tuned Models

Using only the Bibles training set, we generated three different models based on directly fine-tuning `WMT19`: `mbya`, the `WMT19` model fine-tuned with only the Guarani Mbya data from the Bibles training set;`TGf`, the `WMT19` model fine-tuned with Bibles data from 10 languages of the *Tupi-Guarani* linguistic family, *(Guarani* of Paraguay and Bolivia (2); *Guarani Kayowá*, *Guarani Mbya*; *Ka'apor*, *Kaiabi*, *Nheengatu*, *Guajajára*, and *Tenharim*, aiming to take advantage of the geo-linguistically similarity of those languages; and `all`, the `WMT19` model fine-tuned with data from all the 39 Indigenous languages of the Bibles training set.

These models help evaluating the impact of multilingual fine-tuning of language models with the use of culturally toxic data only. In this case, `mbya` is the simpler bilingual model and `TGf` and `all` are multilingual models with different number of languages. Although the former rely on less languages than the latter, i.e. only 10 languages versus 39, the use of linguistically similar languages is expected to optimize the gains with multilingual training. Thus, one goal is to show the improvements, if there is any, of using more languages. But another goal is to understand if the use of such data magnifies the contamination of this type of data.

The three models considered different subsets of the Bibles dataset for training. The `mbya` model performed the `WMT19` fine-tuning using only the Guarani Mbya sentences, 6,340 pairs. The `TGf` model is fine-tuned with 43,869 pairs of sentences from 10 Tupi-Guarani family languages. Finally, the `all` model is generated based on a multilingual fine-tuning approach which considers all Indigenous languages available, totaling 162,225 training pairs. All models were fine-tuned considering a batch size of 32 and learning rate of 2.10^{-5} decaying to 2.10^{-6} according to a cosine function. Number of epochs from 2 to 100 were evaluated. 50, 5 and 20 epochs were selected for `mbya`, `TGf` and `all` models, respectively.

5.2. The *Dictionary*-Tuned Models

Using the data from the Dictionary training set, we generated four additional models: `dict`, the `WMT19` model fine-tuned with Dictionary data; `mbya>dict`, the `mbya` model fine-tuned a second time with Dictionary data; `TGf>dict`: the `TGf` model fine-tuned a second time with Dictionary data; and `all>dict`: the `all` model fine-tuned a second time with Dictionary data.

Notice that while `dict` was obtained by a direct fine-tuning process on top of `WMT19` with no Bibles data, the other three models use a two-step process where Bibles data was employed in a first training step and the resulting model was then fine-tuned on Dictionary data. The goal was to evaluate how the introduction of culturally toxic data in intermediate training steps affects the quality of the translator and how much contamination of problematic data is still present after the fine-tuning with Dictionary data, considering three different levels of multilingualism. Fine-tuning hyper-parameters were adjusted considering 32-sized batches and a learning rate of 2.10^{-5} which decays to 2.10^{-6} in 50 fine-tuning epochs according to a cosine function.

5.3. Both-at-Once Model

Finally, we trained `mbya+dict`, consisting of the `WMT19` model fine-tuned with Guarani Mbya data from the Bibles training set and Dictionary data at the same time, simultaneously. The goal is to understand the gains and perils of using culturally toxic together with non-toxic data compared to the use of culturally toxic data in two-step training fine-tuning processs. The same fine-tuning hyper-parameters of the Dictionary-tuned models were considered here but for 100 epochs.

WMT19		original	only Bibles data			dictionary	Bibles then Dictionary			both-at-once
METRIC	TEST SET	0shot	mbya	TGf	all	dict	mbya>dict	TGf>dict	all>dict	mbya+dict
BLEU	Dictionary	1 ± 2	7 ± 7	4 ± 3	6 ± 5	11 ± 12	**15 ± 15**	14 ± 14	**15 ± 16**	**15 ± 17**
	Bibles	1 ± 1	24 ± 22	11 ± 12	16 ± 14	3 ± 2	11 ± 10	7 ± 7	8 ± 8	**28 ± 24**
chrF	Dictionary	12 ± 4	20 ± 11	17 ± 9	19 ± 9	25 ± 16	**32 ± 18**	29 ± 18	31 ± 20	**32 ± 20**
	Bibles	15 ± 3	46 ± 18	32 ± 13	39 ± 14	21 ± 5	35 ± 11	29 ± 9	32 ± 10	**50 ± 20**

Table 2: Performance in the Dictionary and Bibles test sets of the original WMT19 model and its fine-tuning into 8 models using different training data sets.

6. Performance Evaluation

We relied on standard *machine translation (MT)* evaluation methods to compare the different models. That is, we evaluated MT metrics on both Bibles and Dictionary datasets and we quantitatively measured the impact of each fine-tuning method on both culturally toxic and non-toxic test data with automated MT evaluation metrics.

We used two metrics to evaluate the results: the **BLEU** metric which is the BLEU score computed with the SacreBLEU Python package (Post, 2018); and the **chrF** metric (Popović, 2015) which, although being a metric for poly-synthetic languages, has been widely applied in recent works with low-resource languages. For the two metrics, we computed the average and standard deviation over the score of each sentence in the two test sets created from the Dictionary and Bibles datasets.

6.1. Results

For the 9 models used in this study, we performed an evaluation with the BLEU and chrF metrics of the outputs of both the Dictionary and the Guarani Mbya Bibles test sets (referred, for simplicity, as the Bibles test set throughout the end of the paper). Table 2 shows the average and standard deviations of the `zeroshot`, the Bibles-tuned models (`mbya`, `TGf`, and `all`), with only Dictionary data (`dict`), intermediately fine-tuned with Bibles data and then fine-tuned with Dictionary data (`mbya>dict`, `TGf>dict`, and `all>dict`) and with Bibles and Dictionary simultaneously (`mbya+dict`) when evaluated with the Dictionary and Bibles test sets for the two metrics.

6.2. Findings and Discussion

We focus first here on the results when using the Dictionary test set which correspond to the generic use of the translator for everyday activities as shown in table 2.

For the two metrics, the `zeroshot` has an extremely low performance and it is the worst model, especially when compared to the models fine-tuned with Dictionary data. This was expected since this is basically a German-to-English translator. The poor results are, however, an evidence that the original `WMT19` translator was not exposed to the Guarani Mbya language in its training process.

Also, the performance of the three models fine-tuned with Bibles data is poor, as expected since they were trained with the very specialized vocabulary and style of biblical verses. This becomes clearer when we compare the `dict` model to them: the average accuracy is considerably improved. Although `dict` has a large standard deviation, it is significantly better than the other four models ($p < 0.001$) for all 2 metrics, using standard *one-tailed Student t-tests*.

When we consider the three two-step models (marked with `>dict`), gains of about 16% to 36% in accuracy are seen over `dict`. The t-tests confirm that each of those models are significantly better than `dict` ($p < 0.001$). The best nominal performance is achieved with the both-at-once model, `mbya+dict`, in all metrics and test sets, although there is no statistically significant difference to the two-step models.

The results with the Dictionary test set seem to show, with high confidence and for all metrics, that the best results were achieved by the fine-tuning of the `WMT19` model with the two types of data. We discuss in the next sections both the quality of the outputs generated by those models, the level of contamination from the culturally toxic data, and the ethical and practical implications of it.

But before doing so, we would like to point out that the results for Bibles test set are very similar, except that the performance of the `dict` is not as good, as expected, and that simultaneous fine-tuning with Dictionary data (`mbya+dict`) significantly improves the performance (7-16%) over the best Bibles model (`mbya`), with a similar standard deviation. Fine-tuning simultaneously seems to be a good generic strategy.

The results also indicate that multilingual strategies (`TGf`, `all`) do not pay off, first as it requires more effort both to obtain the data and to convince different Indigenous communities, which may be historically distant, to use their language in the same model, while it produces worse results than Bilingual (`mbya` and `mbya+dict`.

Finally, the fine-tuning in the second domain (Dictionary) reduces the performance in the Bibles test set of the first domain: in all evaluations with the

models	mbya			dict			mbya+dict		
EXPECTED OUTPUT FROM TEST SET	mbya OUTPUT	BLEU	chrF	dict OUTPUT	BLEU	chrF	mbya+dict OUTPUT	BLEU	chrF
long ago there lived a giant.	there were two men in the crowd.	13	30	long ago there lived a giant.	100	100	long ago there lived a giant.	100	100
one day he went to the woods again.	then one of them went to the other side of the lake.	13	32	one day he went again to the woods.	46	76	and so one day he went again to the woods.	36	73
when he arrived at his house, he said to his wife, "can there be anyone who can hunt like me?"	when he came to the tomb, he said to his mother, how can i not know where i am.	17	30	when he got home, he said to his wife, "could it be that i'm from here?"	37	38	when he got home, he said to his wife, "couldn 't i find the ring?"	32	39
when he fell, he hit his back on the ground and died then and there.	so then, how much more will the earth bear down on him than the earth will bear down on him.	3	21	as he fell, he hit his forehead on the rock.	32	38	and as soon as he touched the ground, he died too.	7	30
years ago when i was a child, i didn't know the language of non-indians.	i have not been able to speak the word of the one who sent me into heaven.	3	13	years ago when i was a lot younger, i didn 't know what to do with the books.	29	48	years ago when i was a child, i did not understand the meaning of portuguese.	52	54
when my brother went, saw a snake.	when he came to my house, he saw me.	6	18	my brother went out to see the snake.	22	56	my brother went and saw the snake.	24	61
one day, one of them said to his younger brother, now then, i'm going to the woods.	then one of them said to him, look, i am going to die.	22	40	one day he said to his brother-in-law, "now i 'll go to the woods."	19	26	then one day he said to his brother, "now i 'll go to the woods."	21	45
there comes an inhabitant of the hare village.	you are one of the twelve living creatures.	10	20	there comes the hare from the hare.	15	38	there comes the tapixi village.	15	43
each time the giant went to the woods, he would kill two or three peccaries.	but the one who comes after him will eat the bread, and the bread will come out of his mouth.	2	20	he went very early to the woods to kill two coatis, one of whom was a shotgun.	11	32	this giant will go every day to the woods and kill two or three people.	22	48
is your father at home?	but what do you want me to do for you	0	15	have you come yet?	8	16	your father is?	23	52
he grabbed him by his arm	so he went up to heaven with his brother.	5	12	he took his brother -in -law there.	7	12	then he took hold of the indian in the sky.	4	9
when evening came, the birds were singing and singing, but the indian was still stuck.	but the spirit of the spirit is in the spirit, and the spirit is in the spirit.	5	17	and then it was the turn to eat the birds, both of which were indians.	6	35	and the one who drinks the spirit remains in it, though the spirit remains.	3	17
you changed arbitratriously what you were	if i am a believer, i will be a believer in you	3	9	if you guys believe me, i will believe you.	5	11	you will defraud me even more.	6	12
even though his face got completely bloodied, he smiled.	now the world was divided into three parts.	4	12	that type of wound has already healed lit., it has already healed lit., it already has peel.	2	14	he had bruising on his face.	9	17
who come with lower and higher people;	and all who are in the world and all who are in the world	4	17	has a lot of faith in him.	0	10	low-cost and high-cost carriers also must go;	7	25

Table 3: Examples of outputs of the mbya, dict, and mbya+dict models with BLEU and chrF scores and the expected output from the test set; segments which are associated with biblical texts and expressions are marked in red.

Bibles test set, the performance of the models only trained with Bible data significantly decreased after they are fine-tuned with Dictionary data.

7. Output Quality Evaluation

Our previous experiences with fine-tuning translators for Indigenous languages has taught us the importance of qualitatively checking the outputs generated by such systems (anonymous). In the study of this paper, we focused the qualitative evaluation of the results mainly on the issue of *contamination* of the outputs with elements from the culturally toxic data used in the fine-tuning which is, in this case, verses from the the New Testament.

The question is whether, when tested with the approximately 300 sentences from the Dictionary test set, the different translators we created would produce output which contained, explicitly or not, typical words or language from the Bible. In particular, we were interested to determine whether the best translator, `mbya+dict` suffered from this problem. We performed this analysis manually, reading every generated translation, comparing it with the expected translation, and marking cases where there were possible issues. We also looked for typical biblical words in the generated sentences such as *"Jesus"*, *"God"*, *"cross"*, *"disciples"*, etc. In some cases, we also performed a search in the Internet using suspicious parts of the outputs, looking for possible matches with biblical texts.

Table 3 shows 15 examples from this evaluation process. As a reference, in table 3 we also include, for each of the 15 examples, the output of the Bibles-trained `mbya` translator, where we expected lots of contaminations, and of the `dict` translator, where we would expect no contamination. The examples shown cover an ample range of the two metrics.

All examples from `mbya` in table 3 seem to have contamination (marked in red) and none in the `dict` outputs. It also shows one example in the `mbya+dict` outputs which has been considered as a possible case of contamination, since the output *"and the one who drinks the spirit remains in it, though the spirit remains."* has a resemblance to the John 6:56 verse *"The one who eats my flesh and drinks my blood resides in me, and I in him."*.

The qualitative evaluation of all of the 300 outputs of the `mbya+dict` for the Dictionary test set yielded that 14 (4.7%) of the 300 outputs may have some level of contamination, including only 2 obvious cases where the word *"Jesus"* appeared. The 14 outputs are shown in table 4. We considered as contaminated outputs examples where words such as *"prayer"*, *"dove"*, *"bagel"* were produced; and expressions such as *"washed his feet"*, *"blind eye to evil"*, and *"if I die, I die"*, whose degree of connection to the Bible is questionable.

8. Ablation Studies

We also performed an ablation study to determine how critical is the size of the training set in the performance of the `dict` model. To do that, we fine-tuned the `WMT19` model with 500, 1,000, and 1,500 pairs extracted randomly from the Dictionary

input in Guarani Mbya	expected output	generated by mbya+dict	BLEU	chrF
ha'e vy je ojuka jevy mboapy tapi'i.	he killed three tapirs	he killed three of the jesus,	24	61
yvy ku'i rataxĩ	dust.	a bagel of dust.	21	64
jaye'o 'rã kova'e poryko	let's take out the stomach of this pig.	i will pronounce this pig.	18	24
ha'e gui ma je ita ova oĩ va'e gui je peteĩ ava ramigua okẽ oipe'a jekuaa .	then he saw something like a man open a door in the rock cliff.	then just as the stone was coming out of the tomb, something like a man opened the door.	14	49
ha'e va'e ava poapẽ va'e ma je ogueraa nhande'i va'e omerã .	the "claw-man" took the indian home to be her husband.	this man took the indian and brought him to life.	13	36
ha'e rã je ipy jevy oja .	but his foot stuck too.	once again he washed his feet.	8	13
enhoẽ eme yaku tyrã pa remoakỹmbaxe vi tupa?	don't spill the tea or do you want to wet the bed completely?	do you not untie the strap of your sandals or sandals?	7	17
nhaxi' ũ oixu'u oke va'e	the pernilongs bit the one who was sleeping.	the dove dove.	5	6
xapy'a-py'a py rive nda'u mba'emo vai oiko nhandere?	will it be by chance that bad things happen to us?	have we not turned a blind eye to evil?	4	13
ha'e ndaxyi va'e rire ndoui va'erãgue	if he wasn't sick, it wouldn't have come.	if i hadn 't been born, i would have never been born.	4	23
xepoano va'e rire	if you had treated me I would have been cured.	if i die, i die;	4	4
ha'e gui ka'aru ma guyra'i onhe'ẽ-nhe'ẽ ma teĩ nhande'i va'e oĩ tema ha'e py .	when evening came, the birds were singing and singing, but the indian was still stuck.	and the one who drinks the spirit remains in it, though the spirit remains.	3	17
ha'e vy tu porami rei xeayvu: "você que ta veio já que risada guaraní," ha'e rive.	i said something like, "you what came already what laugh at guaraní."	he was very pleased with the way he talked about it: "we could have bought a hat that would belong to jesus."	3	17
yvyra kyxĩa 'y	sawing cable.	small tree with adjective subordinate prayer;	0	11

Table 4: Outputs of the mbya+dict model which were identified as possible cases of contamination; segments which are possibly associated with biblical texts and expressions are marked in red.

Ablation tests		dictionary only (`dict`)			
metric	test set	500	1000	1500	3000
BLEU	Dictionary	6 ± 6	7 ± 7	8 ± 10	$\mathbf{11 \pm 12}$
	Bibles	2 ± 1	2 ± 1	3 ± 2	$\mathbf{3 \pm 2}$
chrF	Dictionary	16 ± 8	18 ± 11	20 ± 13	$\mathbf{25 \pm 16}$
	Bibles	16 ± 4	18 ± 4	20 ± 4	$\mathbf{21 \pm 5}$

Table 5: Ablation results: performance in the Dictionary and Bibles test sets of the `WMT19` model when fine-tuned with 500, 1000, and 1500 pairs and the full Dictionary training set.

training set and compared to the performance of the `dict` model. The results are shown in table 5. The `dict` significantly outperformed the other three models, in a quasi-linear improvement in accuracy as the number of training pairs increased. That suggests not only that the amount of data is key to improve performance but also that there is room for improvement in the current models if more pairs like the ones in the Dictionary dataset are available.

9. Final Discussion

This paper presents a study of the trade-offs of using non-toxic (dictionary and tales) and culturally toxic (biblical texts) data in the fine-tuning of LLM-based translators of ULR languages. The results in the development of a Guarani Mbya-to-English translator showed that the use of data from the Bible can generate significant improvements over the use of only dictionary-based data in a context with similar amounts of both. In particular, training simultaneously with the two types of data achieved best results, about 30% better than using dictionary data only but similar to two-step processes. A qualitative analysis of the results of the best translator showed, however, 2 cases and other 12 of possible contamination, or about 4.7% of 300 test outputs.

From the results described, it is clear that there is some level of potentially culturally toxic contamination in the best translator we could build for the Guarani Mbya language, due to the use of data from the the Bible for fine-tuning. In many ways, identifying and quantifying the extent of this problem is our main role as technologists and the next steps are to communicate clearly to the communities involved in our findings, provide ideas on how to mitigate the issues, and wait and respect their decision about using the contaminated translator.

Based on those findings we would advise against its release in broader contexts and would recommend its use only in tightly controlled situations where negative effects can be mitigated. Of course, following the ethical guidelines also discussed in the paper, we leave the final decision to the Indigenous communities involved. We abide to the belief that the decision of whether to use a translator for an Indigenous language has to be done by the people who speak the language, fully informed and, whenever possible, as participants in the process (Mihesuah, 1993; Sahota, 2007; Straits et al., 2012), as outlined in the *Los Pinos Declaration*[2].

The results also suggest that more training data is needed. However, as it is the case of most ULR languages, there are few other sources available. We intend to explore the use of those other sources such as academic works and to work with the community to create with them more data. Another possible direction is to explore the use of *synthetic data* which can be generated by working with linguists and language experts from the community to create reliable synthetic language generators.

We finish by acknowledging how honored we are to be working with the extensive cultural and linguistic heritage of the Indigenous peoples of Brazil.

[2]https://en.unesco.org/sites/default/files/los_pinos_declaration_170720_en.pdf.

10. Ethics Statement

In this work we have found that the translator fine-tuned simultaneously with dictionary and bible data is significantly better than the one only tuned with sentences from the dictionary. At the same time, the manual evaluation of the results showed that about 4.7% of the outputs had possibly some contamination, including two clear cases.

Some of those contaminated outputs may be avoided by a filtering system which looks for words often associated with biblical texts and exclude those translations. This would probably take care of the obvious cases but certainly not all (Van Aken et al., 2018; Abbasi et al., 2022).

These results should inform the decision of deploying or not the better but contaminated translator. Ultimately, this decision belongs to the communities interested in the tool. In situations where translators are immediately and highly needed, our advice would be to deploy it but to restrict its use to members which clearly understand the risks involved and establish, possibly with our help, a monitoring system to measure the translator behavior over time. As a more generic tool, available for a larger population, especially of non-Indigenous people, we would not advice its use, since it may occasionally misrepresent the culture and possibly be considered offensive. In this latter case, it seems safer to deploy the translator based only on dictionary data and, with the permission of the community and its users, gradually collect more data and improve its performance.

Also, for similar reasons we cannot share publicly neither the datasets nor the models created in this study without the knowledge and clear acceptance of the Guarani Mbya-speaking people.

Ahmed Abbasi, Abdul Rehman Javed, Farkhund Iqbal, Natalia Kryvinska, and Zunera Jalil. 2022. Deep learning for religious and continent-based toxic content detection and classification. *Scientific Reports*, 12(1):17478.

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.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. 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. Association for Computational Linguistics.

Yuki Arase and Jun'ichi Tsujii. 2019. Transfer fine-tuning: A BERT case study. 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 5393–5404, Hong Kong, China. Association for Computational Linguistics.

Marcel Bollmann, Rahul Aralikatte, Héctor Murrieta Bello, Daniel Hershcovich, Miryam de Lhoneux, and Anders Søgaard. 2021. Moses and the character-based random babbling baseline: CoAStaL at AmericasNLP 2021 shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 248–254, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, et al. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, et al. 2022. Palm: Scaling language modeling with pathways.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A survey of multilingual neural machine translation. *ACM Comput. Surv.*, 53(5).

Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. 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. Association for Computational Linguistics.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of ACL'19*.

Robert Dooley. 1985. Nhanhembo'e aguã nhandeayvu py [1-5].

Robert A. Dooley. 1988a. Arquivo de textos indígenas – guaraní (dialeto mbyá) [1].

Robert A. Dooley. 1988b. Arquivo de textos indígenas – guaraní (dialeto mbyá) [2].

Robert A. Dooley. 2016. Léxico guarani, dialeto mbyá: Guarani-português.

Liat Ein-Dor, Ilya Shnayderman, Artem Spector, Lena Dankin, Ranit Aharonov, and Noam Slonim. 2022. Fortunately, discourse markers can enhance language models for sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10608–10617.

Bruna Franchetto. 2008. The war of the alphabets: indigenous peoples between the oral and the written. *Mana*, 14(SE):31–59.

Bruna Franchetto. 2020. Língua (s): cosmopolíticas, micropolíticas, macropolíticas. *Campos-Revista de Antropologia*, 21(1):21–36.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Computing Surveys*, 55(13s):1–32.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

A. Harding, B. Harper, D. Stone, C. O'Neill, et al. 2012. Conducting research with tribal communities: Sovereignty, ethics, and data-sharing issues. *Environmental health perspectives*, 120(1):6–10.

IBGE. 2010. Censo demográfico 2010. Accessed = 2022-12-30.

J. Lewis, A. Abdilla, N. Arista, K. Baker, et al. 2020. *Indigenous protocol and artificial intelligence position paper*. Indigenous Protocol and Artificial Intelligence Working Group.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

D. Mihesuah. 1993. Suggested guidelines for institutions with scholars who conduct research on american indians. *American Indian Culture and Research Journal*, 17(3):131–139.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. 2020. An analysis of massively multilingual neural machine translation for low-resource languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3710–3718, Marseille, France. European Language Resources Association.

El Moatez Billah Nagoudi, Wei-Rui Chen, Muhammad Abdul-Mageed, and Hasan Cavusoglu. 2021. IndT5: A text-to-text transformer for 10 indigenous languages. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 265–271, Online. Association for Computational Linguistics.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Chris Mato Nunpa. 2020. *The great evil: Christianity, the bible, and the Native American genocide*. See Sharp Press.

Stormy Ogden. 2005. The prison-industrial complex in indigenous california. In *Global lockdown: Race, gender, and the prison-industrial complex*, pages 57–65. Routledge New York.

OpenAI. 2022. Introducing chatgpt. https://openai.com/blog/chatgpt. Accessed on August 14, 2023.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2019. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks.

C. Pinhanez, P. Cavalin, M. Vasconcelos, and J. Nogima. 2023. Balancing social impact, opportunities, and ethical constraints of using ai in the documentation and vitalization of indigenous languages. In *Proc. of IJCAI'23*, Macau, China.

Maja Popović. 2015. 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. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

P. Sahota. 2007. Research regulation in American Indian/Alaska native communities: Policy and practice considerations. In *NCAI*.

Amit Sheth, Valerie L Shalin, and Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318.

David Stoll. 1982. The summer institute of linguistics and indigenous movements. *Latin American Perspectives*, 9(2):84–99.

Luciana Raccanello Storto. 2019. *Línguas indígenas: tradição, universais e diversidade*. Mercado de Letras.

K. Straits, D Bird, E. Tsinajinnie, J. Espinoza, et al. 2012. Guiding principles for engaging in research with Native American communities. *UNM Center for Rural and Community Behavioral Health*.

Betty Van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *Proc. of the 2nd Workshop on Abusive Language Online (ALW) at EMNLP'18*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264, Online. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

L. Xue, N. Constant, A. Roberts, M. Kale, et al. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proc. of NAACL'21*.

Yichu Zhou and Vivek Srikumar. 2022. A closer look at how fine-tuning changes BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1046–1061, Dublin, Ireland. Association for Computational Linguistics.