Machine Translation Between High-resource Languages in a Language Documentation Setting

Katharina Kann and **Abteen Ebrahimi** and **Kristine Stenzel** and **Alexis Palmer**
University of Colorado Boulder
first.last@colorado.edu

Abstract

Language documentation encompasses translation, typically into the dominant high-resource language in the region where the target language is spoken. To make data accessible to a broader audience, additional translation into other high-resource languages might be needed. Working within a project documenting Kotiria, we explore the extent to which state-of-the-art machine translation (MT) systems can support this second translation – in our case from Portuguese to English. This translation task is challenging for multiple reasons: (1) the data is out-of-domain with respect to the MT system's training data, (2) much of the data is conversational, (3) existing translations include non-standard and uncommon expressions, often reflecting properties of the documented language, and (4) the data includes borrowings from other regional languages. Despite these challenges, existing MT systems perform at a usable level, though there is still room for improvement. We then conduct a qualitative analysis and suggest ways to improve MT between high-resource languages in a language documentation setting.

1 Introduction

We report on our investigations of whether and how existing machine translation (MT) systems can support the work of documenting and describing endangered languages. Rather than targeting low-resource MT, we look at translating between high-resource languages, aiming to save time for the language experts and language community members working on the language documentation project.

Specifically, we are working with a linguist documenting Kotiria (also known as *Wanano*), an East Tukano language spoken in the Brazil-Colombia borderlands in northwestern Amazonia. Documentation and description of Kotiria on the Brazilian side of the border has been ongoing since 2000, resulting in numerous publications, including a Reference Grammar (Stenzel, 2013), and a documentary archive of primarily monologic language data (approx. 10 hours of mythical, historical, and personal narratives, public addresses, and instructional speech). A second documentation project focusing on language use and interaction in daily life resulted in a much larger corpus – approximately 60 hours – of primarily conversational data. Both projects were carried out within the participatory research paradigm (Stenzel, 2014), with indigenous speakers involved in both recording and annotation of data in ELAN,[1] including translation of the indigenous language data into Portuguese.

Further grammatical analysis and annotation of these documentary materials, including translation from Portuguese into English, is ongoing but proceeds slowly. Researchers of endangered languages worldwide generally work alone or at best in small teams to deal with enormous amounts of data, further underscoring the gap between technological advances that facilitate production of large, high quality documentary corpora and researchers' ability to single-handedly process the resulting materials. The Kotiria case is no different, and even basic tasks, such as adding English translations to the two existing corpora, extend over years.

The corpus from the more recent Kotiria language documentation project presents additional challenges. First, language use in conversation is by its very nature more complex to annotate and analyze than monologic speech because it is rife with features such as reductions, cut-offs, overlaps, intonational contours, and other details of production, as well as grammatical structures whose meaning can only be understood in sequential context (Hepburn and Bolden, 2013). Additionally, due to the multilingual nature of social life in the region where Kotiria is spoken (Stenzel, 2005; Stenzel and Williams, 2021), recordings contain numerous instances of speech in other indigenous languages, such as Tukano. Though extremely rich, such data constitutes a lifetime (or perhaps several lifetimes)

[1] https://archive.mpi.nl/tla/elan

Proceedings of the first workshop on field linguistics, pages 26–33
October 16, 2022.

of processing work for a lone-wolf researcher.

Automatic translation – or *machine translation (MT)* – has made tremendous progress over the last few years (Sutskever et al., 2014; Bahdanau et al., 2014; Vaswani et al., 2017), and MT systems are used more and more in everyday life, e.g., in browser extensions, smartphone apps, or as a first translation pass in software for professional (human) translators. Initial translations in a language documentation project are often made into the dominant high-resource language in the documented language's region (Portuguese for Kotiria). As MT between high-resource languages is typically of high quality (Akhbardeh et al., 2021), we investigate if MT systems can assist with producing additional translations between the region's dominant high-resource language (Portuguese) and English, which can help make the created resources accessible to a broader community. Our goal is to produce first-pass translations automatically, such that the language experts in the language documentation project need not devote years to the process, but rather can do post-correction of the first-pass translations. This should yield significant time savings (Toral et al., 2018), freeing up the experts to work on other aspects of the project.

Importantly, such a translation in a documentation context constitutes multiple challenges not present in general MT: (1) the sentences that need to be translated are out-of-domain with respect to the system's training data, (2) the data is conversational, (3) the source-side data contains non-standard and uncommon expressions, often reflecting properties of the documented language, and (4) the text includes borrowings from other regional languages. While those challenges could be minimized by training on in-domain data from the concrete translation task, such data is generally either not available or too small for effective finetuning.

First, we employ 3 state-of-the-art MT systems to translate Portuguese sentences for which we have gold-standard translations into English. We evaluate the results both manually and with automatic metrics and find that Google Translate performs best. Second, we analyse the outputs of Google Translate, exploring what types of examples it fails and succeeds on. We observe that the conversational nature of the Kotiria data and particular properties of Kotiria-to-Portuguese translations cause many errors. We end by discussing how to improve MT for language documentation data.

2 Related Work

NLP for Language Documentation One goal of language documentation is to create permanent records of the linguistic and cultural practices of understudied speech communities and combat loss of linguistic diversity. It encompasses the audio and video recording of speech as well as the transcription, translation, and analysis of the recordings. This process is costly in terms of time and money, and, besides MT into additional high-resource languages, NLP has the potential to aid documentation via automatic speech recognition (Adams et al., 2018; Prud'hommeaux et al., 2021; Shi et al., 2021; Liu et al., 2022), improve access to legacy materials through OCR (Rijhwani et al., 2020), enrich text data with part-of-speech tags (Eskander et al., 2020) or word boundaries (Okabe et al., 2022) to eventually obtain interlinear glossed text, or to support the analysis of a language's morphology (Jin et al., 2020; Moeller et al., 2020), *inter alia*.

MT of Out-of-Domain Data Our setting requires MT models to generalize to out-of-domain data: available translations are too few for training or finetuning, and, in other language documentation settings, no translations into additional high-resource languages might be available at all. However, MT systems often struggle to perform well on data they have not been trained on – e.g., systems trained on 2019 news do not perform well on 2020 news, due to a topic shift towards the coronavirus (Anastasopoulos et al., 2020). Domain adaptation (DA), which has been studied extensively (Yang et al., 2018; Chu et al., 2018; Adams et al., 2022), though not in the context of a language documentation workflow, can yield improvements. Techniques include finetuning (Luong and Manning, 2015; Freitag and Al-Onaizan, 2016) or backtranslation (Sennrich et al., 2016). For surveys on DA for MT, we refer readers to Chu and Wang (2018); Saunders (2021). We investigate how well general state-of-the-art MT systems translate between high-resource languages in a language documentation setting. In future work, we will take inspiration from research on DA and investigate how to build better systems for our use case.

3 Experimental Setup

3.1 Data

Our dataset draws from the two Kotiria documentation projects described in Section 1, i.e., we have a

Meaning	5	Exactly the same meaning as gold (except for parts that appear in Portuguese but not in gold)
	4	About the same meaning as gold; maybe minor differences (like singular/plural or similar)
	3	Meaning can maybe be guessed but is not clear from the translation or something is misleading
	2	The meaning is different/partially misleading and only a few words are in common with gold
	1	The meaning of this translation has absolutely nothing to do with gold or is misleading
Rel. Fluency	5	Completely fluent in English; maybe more fluent than reference translation
	4	As fluent as reference translation or minor grammatical error that does not affect understanding
	3	Understandable, but not completely fluent
	2	Not a fluent sentence, understandable with lots of effort
	1	Not understandable because of lack of fluency

Table 1: The annotation instructions we provide to our annotators to assess translation quality in terms of *meaning* and *relative fluency*.

mix of monologic and conversational texts. Across the two projects, we have 2267 sentences with reference English translations, which we divide evenly into development and test sets. We report results on the development set to reserve the test set for future research on MT systems for this setting.[2]

3.2 MT Systems

M2M-100 M2M-100 (Fan et al., 2021) is a model trained to handle many-to-many translation between 100 languages. It is a transformer encoder–decoder, and for this work we use the version with 418M parameters. M2M-100 uses SentencePiece (Kudo and Richardson, 2018) tokenization and is trained on mined parallel data, extending prior work (El-Kishky et al., 2020; Schwenk et al., 2021). The model is not trained on data from all possible pairs – rather, languages are grouped, and only within-group language pairs are used for training. Bridge languages are chosen for each language group and trained against other bridge languages. In addition, all languages are trained against English. The training set has 7.5 billion examples.

mBART50 mBART (Liu et al., 2020) is a sequence-to-sequence autoencoder, pretrained with a denoising objective. The model is pretrained on 25 languages, with the goal of recovering the original input after it has been corrupted with a noise function, which involves sentence re-ordering and span masking. It is then finetuned for translation using parallel data. However, since Portuguese is not included in the original set of languages, for this work we use mBART50 (Tang et al., 2021), which builds upon the original mBART model and extends the number of languages from 25 to 50. We use the version trained with multilingual finetuning, allowing for many-to-many translation.

[2]Our data is publicly available at https://nala-cub.github.io/resources.

Google Translate We also compare to a state-of-the-art commercial MT system: Google Translate.[3] For our experiments we use Googletrans,[4] a python library accessing the Google Translate Ajax API.

3.3 Automatic Metrics

We use two automatic metrics for evaluation, which we calculate using SacreBLEU (Post, 2018).

BLEU First, we evaluate our outputs with BLEU (Papineni et al., 2002), the standard metric for MT. BLEU measures word overlap between the translation and the reference. We use SacreBLEU's default settings and tokenization.

ChrF We further compute ChrF (Popović, 2015). In contrast to BLEU, this metric measures the *character* overlap between a translation and a reference.

3.4 Human Evaluation

In addition to employing automatic metrics we also perform a manual/human evaluation of translations for a subset of 100 randomly sampled sentences from the development set. We show annotators the Portuguese source sentence, the English reference, and the system output and ask for an assessment along two axes: meaning (*does the translation's meaning correspond to the reference?*) and (relative) fluency (*is it as grammatical as the reference?*). Both meaning and fluency are assessed using a Likert scale from 1 to 5, with higher numbers indicating better quality. We give annotators the option to skip examples whose fluency and meaning they feel unable to judge, e.g., "Uhh". Each translation is rated by two annotators, and reported scores are averages over annotators. Table 1 shows the complete instructions given to annotators.

[3]https://translate.google.com

[4]https://py-googletrans.readthedocs.io/

System	BLEU	ChrF
Google Translate	**19.96**	**42.83**
mBART	9.40	31.39
M2M-100	10.25	30.50

Table 2: Automatic evaluation: BLEU and ChrF for all systems on the development set. Best scores in bold.

System	Meaning	Fluency
Google Translate	**3.82**	**4.07**
mBART	2.57	4.04
M2M-100	3.07	3.72

Table 3: Manual evaluation: meaning and fluency of all systems on 100 sentences from the development set. Scores are averaged over annotators. Best scores in bold.

4 Results and Discussion

4.1 Translation Performance

Automatic Evaluation Table 2 displays the performance of all systems on the development set according to automatic metrics. The best system is Google Translate with a BLEU (resp. ChrF) score of 19.96 (resp. 42.83). The other two systems obtain considerably lower and, surprisingly, quite similar scores: mBART achieves a BLEU and ChrF of 9.40 and, respectively, 31.39, while M2M-100's scores are 10.25 and 30.50.

In absolute terms, the score of Google Translate, the best system in our experiments, is reasonable, but not as good as for general in-domain MT, where BLEU scores higher than 40.00 were reported by Google already in 2017 (Johnson et al., 2017).

Human Evaluation Table 3 shows *meaning* (i.e., how well the translation represents the meaning of the gold translation) and *fluency* scores (i.e., how grammatical the sentence is, given the reference translation). They range from 2.57 to 3.82 for meaning and from 3.72 to 4.07 for fluency. As both scores are on a scale from 1 to 5 with higher being better, all systems perform reasonably well on our task. Thus, our first and main conclusion is that *MT systems can indeed help with language documentation*; specifically with translating from the dominant high-resource language in the region of the documented language into another high-resource language. However, *there is room for improvement*.

Comparing the 3 systems we get a picture similar to the one we get with automatic metrics: Google Translate performs best for both meaning and fluency. Surprisingly, mBART has with 4.04 a high fluency score, which nearly matches that of Google Translate, but a comparatively low meaning score with 2.57. M2M-100 is with 3.07 between the other two systems with regards to meaning, but lags behind the other two as far as fluency is concerned.

Comparing meaning with fluency scores, we observe that systems are similar with respect to the latter (max. delta: 0.35), but vary considerably for the former (max. delta: 1.25). This shows that all systems have been trained on enough English data to produce grammatical sentences. However, generating text that represents the meaning of the Portuguese sentence is more challenging.

4.2 Qualitative Analysis

We continue our analysis to investigate particular weaknesses and some unexpected strengths of MT by investigating the translations produced by Google Translate, the best performing system, according to both automatic and manual evaluations. We focus on issues relevant for data from a language documentation context.

Conversational/Dialog Speech Many fluency errors we see in the MT output can be at least partially attributed to the conversational nature of the original text. For example:

(1) *é, jogar, amanha vamos quebrar com chute*
(Ref) yeah, thrown away, and tomorrow we can kick them in
(GT) yeah, play, tomorrow we're going to break with kick

The utterance in (1) makes sense in its discourse context, with confirmation that an unspecified something has been thrown away: *jogar* means both "play" and "throw" and is used here as a shortened form of *jogar fora* ("throw out/away"). It is followed by a clause with a pronominal object. Absent that context, though, the MT system selects the wrong meaning, supplies no referents, and treats the verbs as infinitives. The result is a nearly incoherent English translation.

Transfer from Kotiria Some of the most interesting errors stem from L1 transfer, as nearly all of the Portuguese translations were written by speakers of Kotiria who had later learned Portuguese as one of their additional languages. In some translations, grammatical properties of Kotiria are transferred into Portuguese, resulting in non-standard

forms: e.g., serial verb constructions, in which multiple roots occur contiguously to form a single verb stem, are common in Kotiria but not in Portuguese. In (2), the Kotiria serialized verb construction indicating associated motion is rendered as a sequence of separately inflected verbs, resulting in understandable but odd-sounding Portuguese. Some differences reflect the different morphologi-

(2) *levaram arrastando e que ele estava sentindo mal (tristeza, raivoso)*
(Ref) they dragged him off and he was full of regret
(GT) led dragging and that he was feeling bad (sadness, angry)

cal inventories of the languages: Portuguese uses a range of different locative markers (indicating different spatial configurations, such as *in*, *on*, or *to*), but Kotiria has a single locative marker subsuming all of these functions. In cases like (3), we see *em* ("in") used as a generic locative marker rather than the context-appropriate *a* ("to") in Portuguese.

(3) *em são gabriel?*
(Ref) to São Gabriel?
(GT) in san gabriel?

Borrowings Another class of translation errors occurs when lexical borrowings from other regional languages appear in the Portuguese text. These are often not translated into English by the MT system.

Unexpected Strengths The translations found in our data often include clarifications/explanations (as seen in (2)) or reduced forms ((4), in which *pra* is a non-standard reduced form of *para*). Google Translate handles these issues surprisingly well.

(4) *pra bateria nao mexer*
(Ref) So the battery won't move again
(GT) so the battery doesn't move

4.3 How to MT for Language Documentation

Here, we investigate how general state-of-the-art models perform in a language documentation context. However, while existing MT models work surprisingly well for language documentation purposes, we believe that model adaptation to this specific domain (cf. Section 2) could further improve performance: English translations from documentation corpora of other languages could familiarize the model with conversational English and recurrent themes (e.g., *travel*, *food* or *ceremonies*). The more linguistically similar the documented languages are and the more topic overlap of collected text there is, the more this should help.

Another option – potentially combinable with the first one – would be a multilingual model that is trained (also) on parallel data between the documented language and the first high-resource language. This could teach the model about word choices and expressions, which, later on, would be beneficial for their translation into English.

Finally, the error types pointed out in Section 4.2 are frequent in our corpus, suggesting that MT models would benefit from incorporating explicitly-specified prior knowledge about key structural properties of the language being documented.

5 Conclusion

Using data from the documentation of Kotiria, we investigated how general state-of-the-art MT systems perform when translating from Portuguese to English in a language documentation setting. We found that, among 3 systems, Google Translate performs best and at a level that makes it a promising option for documentary linguists. We then performed a qualitative analysis of Google Translate and observed a number of systematic error patterns directly linked to properties of our language documentation project. Finally, we suggested multiple ways to improve systems for this setting, including model adaptation, targeted multilinguality, and the incorporation of linguistic features.

Acknowledgments

We would like to thank the Kotiria community for their permission to share data from their documentary corpora. We moreover acknowledge the invaluable work on transcription and translation by indigenous research team member Auxiliadora Ferreira Figueiredo. Stenzel's early research on Kotiria received funding through National Science Foundation dissertation (2002- 2004) and Endangered Languages Documentation Program MPD-155 (2007-2011) grants. Recent documentation and analysis were supported by the National Science Foundation under Grant No. BCS-1664348 (2017-2020) and the National Endowment for the Humanities fellowship FN-271117-20 (2020-2021). Any views, findings, conclusions, or recommendations expressed in this article do not necessarily reflect those of the National Endowment for the Humanities and the National Science Foundation.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluation phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Virginia Adams, Sandeep Subramanian, Mike Chrzanowski, Oleksii Hrinchuk, and Oleksii Kuchaiev. 2022. Finding the right recipe for low resource domain adaptation in neural machine translation. *arXiv preprint arXiv:2206.01137*.

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. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. 2020. TICO-19: the translation initiative for COvid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Chenhui Chu, Raj Dabre, and Sadao Kurohashi. 2018. A comprehensive empirical comparison of domain adaptation methods for neural machine translation. *Journal of Information Processing*, 26:529–538.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

Ramy Eskander, Smaranda Muresan, and Michael Collins. 2020. Unsupervised cross-lingual part-of-speech tagging for truly low-resource scenarios. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4820–4831, Online. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Çelebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22:107:1–107:48.

Markus Freitag and Yaser Al-Onaizan. 2016. Fast domain adaptation for neural machine translation. *arXiv preprint arXiv:1612.06897*.

Alexa Hepburn and Galina B. Bolden. 2013. The conversation analytic approach to transcription. In Jack Sidnell and Tanya Stivers, editors, *The Handbook of Conversation Analysis*, pages 57–76. Blackwell.

Huiming Jin, Liwei Cai, Yihui Peng, Chen Xia, Arya McCarthy, and Katharina Kann. 2020. Unsupervised morphological paradigm completion. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6696–6707, Online. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

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

Zoey Liu, Justin Spence, and Emily Tucker Prud'hommeaux. 2022. Enhancing documentation of Hupa with automatic speech

recognition. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 187–192, Dublin, Ireland. Association for Computational Linguistics.

Minh-Thang Luong and Christopher Manning. 2015. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the 12th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 76–79, Da Nang, Vietnam.

Sarah Moeller, Ling Liu, Changbing Yang, Katharina Kann, and Mans Hulden. 2020. IGT2P: From interlinear glossed texts to paradigms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5251–5262, Online. Association for Computational Linguistics.

Shu Okabe, Laurent Besacier, and François Yvon. 2022. Weakly supervised word segmentation for computational language documentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7385–7398, Dublin, Ireland. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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. Association for Computational Linguistics.

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, Belgium, Brussels. Association for Computational Linguistics.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation & Conservation*, 15:491–513.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

Danielle Saunders. 2021. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *arXiv preprint arXiv:2104.06951*.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging end-to-end ASR for endangered language documentation: An empirical study on yolóxochitl Mixtec. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1134–1145, Online. Association for Computational Linguistics.

Kristine Stenzel. 2005. Multilingualism: Northwest Amazonia Revisited. In *Proceedings of the Second Annual Congress CILLA (Congreso de Idiomas Indigenas de Latinoamérica)*.

Kristine Stenzel. 2013. *A Reference Grammar of Kotiria (Wanano)*. University of Nebraska Press.

Kristine Stenzel. 2014. The pleasures and pitfalls of a 'participatory' documentation project: an experience in northwestern Amazonia. *Language documentation & conservation*, 8:287–306.

Kristine Stenzel and Nicholas Williams. 2021. Toward an interactional approach to multilingualism: Ideologies and practices in the northwest Amazon. *Language & communication*, 80:136–164.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Antonio Toral, Martijn Wieling, and Andy Way. 2018. Post-editing effort of a novel with statistical and neural machine translation. *Frontiers in Digital Humanities*, page 9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Unsupervised domain adaptation for neural machine translation. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 338–343. IEEE.