On the Ethics and Linguistic Impacts of Using the Bible as Training Data for Yucatec Maya-to-Spanish Machine Translation

Jade A. Phoreman

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2026

Committee:

Dr. Emily M. Bender

Dr. Shane Steinert-Threlkeld

Program Authorized to Offer Degree:
Linguistics

University of Washington

Abstract

On the Ethics and Linguistic Impacts of Using the Bible as Training Data for Yucatec Maya-to-Spanish Machine Translation

Jade A. Phoreman

Chair of the Supervisory Committee:
Dr. Emily M. Bender
Linguistics

Religious texts, primarily the Christian Bible, are commonly used as training data for low-resource machine translation (MT) systems because they constitute some of the most extensive and systematically digitized parallel corpora available for many languages. However, this practice raises both linguistic and ethical concerns, particularly for Indigenous language communities for whom Bible translation has historically been intertwined with colonialism and cultural erasure. This thesis investigates the trade-offs associated with using Bible-derived parallel data to fine-tune machine translation models for the Yucatec Maya-to-Spanish translation task. I fine-tuned two models – TowerInstruct-7B-v.02 and T5S – across seven experimental conditions varying the proportion and quantity of Bible training data, ranging from 0% to 100% Bible data. Translation quality was evaluated using BLEU, chrF, METEOR, and COMET. Bible-related content drift in model outputs was assessed through two complementary methods: a semantic similarity analysis using BETO sentence embeddings, and a Bible n-gram contamination analysis using log-likelihood ratio statistics. Results show that increasing the proportion of Bible training data consistently degraded translation quality across both models. For TowerInstruct-7B-v.02, this degradation was strictly monotonic. For T5S, the relationship was broadly similar but not strictly monotonic. Neither model benefited from increased quantities of Bible-dominated training data. Semantic drift toward biblical Spanish was negligible across all conditions for both

models, with a single exception of T5S trained only on a subset of Bible data. These findings are contextualized by a community survey of 84 Yucatec Maya speakers, who broadly supported machine translation development while expressing concern about data sovereignty, colonial training data, and the risk of epistemic extractivism. Together, the computational and community findings argue that domain-matched, community-generated data should be prioritized over Bible corpora in low-resource MT development, even when data scarcity creates pressure to use all available resources.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables iv

Chapter 1: Introduction 1

Chapter 2: Background 4
2.1 Bible Translation Usage History 4
2.2 The Yucatec Maya Language Community 7
2.3 Models 8
2.4 Summary 9

Chapter 3: Related Work 10
3.1 Recent Ethical Critiques of Religious Data Use 10
3.2 Low-Resource MT Methodology 13
3.3 MT for Yucatec Maya 14
3.4 Corpus Comparison and Semantic Similarity Methods 15
3.5 Summary 16

Chapter 4: Survey 17
4.1 Survey Methodology 17
4.2 Survey Results 25

Chapter 5: Experimental Methodology 41
5.1 Model Selection 41
5.2 Data 42
5.3 Experimental Design 44
5.4 Summary 48

Chapter 6: Experimental Results 50
6.1 Translation Quality 50
6.2 Semantic Drift 60
6.3 Summary 67

Chapter 7: Discussion 68
7.1 Translation Quality 68
7.2 Semantic Drift 75
7.3 Limitations and Future Work 77
7.4 Summary 78

Chapter 8: Conclusion 80
8.1 Summary of Findings 80
8.2 Broader Significance 81

LIST OF FIGURES

Figure Number — Page

4.1 Distribution of Participant Roles and Language Background (n=79). 25
4.2 Yucatec Maya Usage across Contexts 28
4.3 Degree of Support for Yucatec Maya Machine Translation 29
4.4 Training Data Sources of Concern
Question: Are you concerned that some of these texts might be used to train machine translators? 36
6.1 BLEU Scores Across Experimental Conditions for TowerInstruct and T5S . . 53
6.2 chrF Scores Across Experimental Conditions for TowerInstruct and T5S . . . 54
6.3 METEOR Scores Across Experimental Conditions for TowerInstruct and T5S 54
6.4 COMET Scores Across Experimental Conditions for TowerInstruct and T5S . 55
6.5 Proportion of Outputs Flagged as Semantically Bible-Influenced 62
6.6 Distribution of Semantic Similarity Deltas by Condition 63

LIST OF TABLES

Table Number Page

5.1 Training data composition for each experimental condition. All conditions manipulating the proportion of Bible vs non-Bible training data use 14,634 sentence pairs and share the same validation and test sets. 44

5.2 Condition F uses all available training data from both the Bible dataset and the non-Bible dataset. Condition G uses all available Bible data. 44

6.1 Translation quality across ordered experimental conditions with varying Bible-to–non-Bible training data ratios (A–E). Condition F uses all available training data from both the Bible dataset and the non-Bible dataset and is therefore approximately 68% Bible data. Condition G uses all available Bible data and is therefore 100% Bible data. Higher scores indicate better performance. *baseline* refers to the respective unfine-tuned model performing the Yucatec Maya-to-Spanish translation task with the same testing set that was used for experimental conditions A-G. The T5S and the TowerInstruct model baselines perform similarly, further highlighting the significance of the performance gains achieved by the T5S model after fine-tuning. 52

6.2 *gold2gold* refers to the gold standard Spanish translation being evaluated against itself. *jpn2gold* refers to a set of Japanese sentences evaluated against the gold standard Spanish translation. *rand2gold* refers to a random reordering of sentences from the gold standard Spanish translation being compared to the original ordering of sentences from the gold standard Spanish translation. *src2targ* refers to the testing set in Yucatec Maya being compared against the gold standard Spanish translation. *fr–en* and *fr–es* refer to the French pivot experiments. 57

6.3 Selected examples of intelligible model outputs across experimental conditions. 58

6.4 Outputs Flagged as Semantically Similar to Bible Data 65

ACKNOWLEDGMENTS

I would like to thank my faculty advisor, Dr. Emily M. Bender, for her guidance at every step of this process, for helping me shape the ethical framing of this thesis, and for helping me appreciate the contribution I was making by coupling computational experiments with a community survey. Her encouragement, flexibility, and deep empathy made it possible for me to complete this work while navigating personal challenges and running a tutoring business. I could not have asked for a more understanding advisor.

I would also like to thank my second reader, Dr. Shane Steinert-Threlkeld, whose LING 575 course first gave me the space to begin formulating the research questions and methodology that became this thesis. I am grateful for his willingness to serve as my reader while on sabbatical, and for his patience through the many timeline changes along the way.

This research would not have been possible without the cooperation and support of Silvia Fernandez Sabido and César Can of the Yucatec Maya language community. Silvia provided the UNAM corpus and César provided the T'aantsil corpus; both contributed their expertise to refining the n-gram lists, helped distribute the community survey, and offered invaluable guidance in formulating the survey questions. I am deeply grateful for their trust in allowing me to work with their community and their language data, and for their patience as this project took longer than any of us had planned.

I would also like to thank my dad, Jim Phoreman Jr., who has championed me and believed in me at every step of my academic journey. Early in this process, when the enormity of writing a thesis felt impossibly unmanageable, he told me: "How do you eat the elephant? one bite at a time." I repeated that quote to myself countless times through many periods of overwhelm in this process. My dad has always been one of my biggest sources of intellectual inspiration and motivation.

My mom, Melissa Phoreman, has always supported me with so much selflessness through

every season, every challenge, and every endeavor. She and my dad made graduate school financially possible for me, and I am deeply grateful for the opportunities that I have had because of their hard work and generosity.

Finally, my best friend Kent Cosme, whose love for language and heartfelt response to reading my unfinished first draft reminded me why this work matters.

DEDICATION

For my students, who motivate me to keep learning.

Chapter 1

INTRODUCTION

Religious texts, primarily the texts of the Christian Bible, are commonly used as training data in low-resource machine translation (MT) because, in many cases, they constitute the most extensive and systematically digitized parallel corpora available, reflecting historical and institutional patterns of language documentation.The Christian Bible in particular also contains a chapter-verse structure that allows for easy alignment of parallel datasets, and it belongs to the public domain, adding further motivation for use as MT training data. However, this use of the Christian Bible in the training of MT systems for low-resource languages raises potential linguistic concerns because biblical texts exhibit domain-specific lexical and syntactic patterns that diverge from those found in contemporary everyday language use. More importantly, it raises a host of ethical and cultural concerns related to colonialism, proselytism, oppression of minority language communities' culture and religion, and the possible production of offensive content.

Until recent work, the task of quantifying the linguistic impacts that give rise to these ethical concerns had gone largely unexplored. In this thesis, I build upon the work of Domingues et al. (2024) by extending machine translation evaluation to a new language pair, employing more robust evaluation metrics, and developing computational methods for identifying the frequency of erroneously Bible-related output. Identifying the frequency with which Bible-related content deviates from gold-standard translations provides one way to operationalize the ethical risks associated with using religious texts as training data. Given the historical entanglement of Christian missionary activity with colonial projects characterized by physical and cultural violence, erroneous Bible-related output in MT systems has the potential to produce cultural harm and thus constitutes an ethical concern. However, judgments regarding whether specific content is culturally harmful must ultimately remain with members of the affected communities, rather than external researchers.

It is usually expected that additional parallel data in the target language pair should improve translation quality, as widely demonstrated in the MT literature (Gordon et al., 2021; Fadaee et al., 2017). Similarly, when the total quantity of training data is held constant, substituting one parallel corpus for another from the same language pair might be expected to yield similar translation quality. My research was designed to test both of these assumptions by asking: What is the trade-off, if any, between translation quality (measured by BLEU, chrF, COMET, and METEOR) and frequency of erroneous Bible-related outputs when fine-tuning existing MT models with different amounts of Bible data vs non-Bible data for the Yucatec Maya-to-Spanish ultra-low-resource translation scenario? I investigated this question across two model architectures: TowerInstruct-7B-v.02, a decoder-only instruction-tuned LLM, and T5S, an encoder-decoder model used in prior work on this language pair. My results showed that there was no trade-off between translation quality and the frequency of semantic drift towards Bible content across either model. Instead, increasing proportions of Bible fine-tuning data consistently degraded translation quality for both models, though the relationship was strictly monotonic for TowerInstruct-7B-v.02 and broadly but not strictly monotonic for T5S. Neither fine-tuning on all available Bible data alone nor combining all available Bible and non-Bible data outperformed the non-Bible-only condition, despite both of these conditions involving a larger quantity of training data overall.

In addition to computational evaluation, this thesis includes a survey study to contextualize these technical findings within community perspectives. This survey component responds to recent calls by researchers such as Hutchinson (2024) and Pinhanez et al. (2023) for culturally grounded evaluation and community consultation in NLP system design. Because the cultural and ethical significance of religious language cannot be determined solely through quantitative metrics, I surveyed Yucatec Maya speakers to examine attitudes toward religious texts and other sources being used to train language models, attitudes toward machine translation generally, and the community's challenges with technological accessibility. The survey also collected information on the frequency of usage of Yucatec Maya across contexts. This survey provides essential interpretive grounding for the computational results, ensuring that technical measurements of contamination are situated within the lived

linguistic and cultural realities of the communities most directly affected.

This thesis is organized as follows. Chapter 2 provides historical, computational, and sociolinguistic background on the role of Christian Bible translation in linguistic documentation and its subsequent use as training data in low-resource machine translation, alongside background on the Yucatec Maya language and community. It also includes background information about the two models fine-tuned for the experiments: TowerInstruct-7B-v.02 and T5S. Chapter 3 reviews related work in low-resource MT, surveying Indigenous language communities, and computational approaches to domain contamination, establishing the research gap that this thesis addresses. Chapter 4 presents a survey study investigating community perspectives on religious language, translation, and cultural appropriateness in Yucatec Maya contexts. Chapter 5 describes my experimental methodology, including dataset construction, model training, and evaluation procedures. Chapter 6 reports quantitative results on translation quality and Bible-related contamination across experimental conditions. Chapter 7 discusses these findings in relation to prior work, ethical considerations, and broader implications for low-resource MT. Finally, Chapter 8 concludes with a summary of contributions, limitations, and directions for future research.

Chapter 2

BACKGROUND

This chapter provides the background necessary to situate the experiments and findings of this thesis within their broader historical, linguistic, and technical contexts. It begins with a history of Bible translation as a linguistic and computational resource, tracing the origins of religious text data in missionary linguistics and examining the ethical concerns that arise when such data is used in machine translation systems. It then describes the Yucatec Maya language community, whose sociolinguistic situation motivates the specific ethical stakes of this research. Finally, it introduces the two models used as the foundation for the fine-tuning experiments: TowerInstruct-7B-v.02 and T5S.

2.1 Bible Translation Usage History

The Bible has frequently been used as training data for multilingual computational language models throughout the history of NLP due to its public domain status, its verse-level alignment, and its availability in hundreds of languages (Resnik et al., 1999; Christodouloupoulos and Steedman, 2015). However, the ethical implications of this usage had gone largely unexplored throughout the history of NLP until relatively recently.

2.1.1 Missionary Linguistics and the Origins of Written Language Data

For many languages that became targets of Christian missionary activity, written language itself arrived as a missionary technology. European missionaries introduced writing systems to communities without pre-existing orthographic traditions specifically in order to produce Bible translations (Wonderly and Nida, 1963; Errington, 2001), meaning that written documentation and religious content did not merely co-occur – the latter was the explicit purpose of the former. As a result, for such languages, the earliest surviving written records are religious texts not by coincidence but by design.

Beginning in late antiquity and continuing through the medieval and early modern periods, Bible translation efforts by Christian missionaries produced some of the first grammars, orthographies, and lexicons for numerous languages, including Gothic, Armenian, Ethiopic, Old Church Slavonic, and many others (Wonderly and Nida, 1963). In many cases, these translations constitute the earliest or only extensive written documentation of these languages.

By the twentieth century, missionary linguistics had become increasingly institutionalized. Organizations such as the Summer Institute of Linguistics (SIL), Wycliffe Bible Translators, and the American Bible Society used community development projects and humanitarian aid to legitimize their proselytizing goals in the eyes of language communities and governments. They also employed modern descriptive linguistic methods to aid their Bible translation efforts, producing grammars and dictionaries. While these methods benefited the cause of linguistic documentation, the fact that the linguistic documentation was happening specifically to further the cause of Bible translation led to religious content being the dominant written records in many languages.

As a result, for many low-resource languages today, religious texts are not merely a subset of available data but form the core linguistic record upon which later computational resources are built.

2.1.2 The Bible as a Parallel Corpus in Computational Linguistics

The Bible being the most widely translated document in history has directly impacted practices in machine translation. As early statistical approaches to machine translation were developed, researchers required parallel corpora that were multilingual, aligned, and publicly available. The Bible uniquely met these requirements.

Resnik et al. (1999) formalized the use of the Bible as a computational resource under the banner of the American Bible Society's proposed "Book of 2000 Tongues" initiative (p.132), identifying several features that made it particularly attractive for NLP research: its availability in hundreds of languages, its consistent verse-level alignment across translations, the careful translation practices applied to its production, and its relatively large size

compared to other single-text corpora available at the time. Resnik et al. (1999) also noted that because the Bible spans narrative, poetry, and epistolary genres written by at least 30–40 authors, it offers broad stylistic diversity within a single aligned resource (p.129). They further claimed that the corpus covers approximately 80% of the most useful vocabulary in modern English, arguing that despite its specialized subject matter, it remains relevant for research on contemporary language (p.148).

This framing positioned the Bible as a practical solution to the data scarcity problem in multilingual NLP and led to its widespread adoption as training data. As machine translation paradigms shifted from rule-based to statistical and later neural approaches, Bible-based corpora continued to be reused, often without reevaluating the assumptions that initially justified their adoption.

2.1.3 Ethical and Postcolonial Critiques of Religious Training Data

Scholars in linguistic anthropology and history have long argued that missionary linguistic work cannot be separated from colonial power structures (Errington, 2001). Linguistic documentation was frequently intertwined with efforts to convert Indigenous populations to Christianity, control Indigenous territory, and impose external categories of language and identity on communities that had not sought such classification.

From this perspective, religious texts function not only as linguistic artifacts but as vehicles of ideological transmission. When such texts dominate the available data for a language, they shape both how that language is represented computationally and how it is reproduced in downstream technologies – encoding not just vocabulary and grammar but the worldview and communicative purposes of the missionaries who produced them.

Domingues et al. (2024), working in collaboration with the Tenondé Porã Guarani Mbya community in Brazil, characterize Bible data as culturally toxic in Indigenous language contexts, arguing that its use in MT training is not merely a technical limitation but an ethical one: outputs that carry religious content risk perpetuating the same colonial dynamics that originally produced the data, in communities where Bible translation has historically been inseparable from violence and cultural erasure.

Despite growing recognition of these concerns, there remains a lack of systematic analysis of how Bible training data shapes model behavior across varying proportions of Bible and non-Bible data. Addressing this gap requires both controlled experiments varying training data composition and a detection framework capable of identifying religious content drift in model outputs – the dual approach this thesis develops.

2.2 The Yucatec Maya Language Community

Yucatec Maya is a Mayan language spoken primarily in the Mexican states of Yucatán, Quintana Roo, and Campeche, as well as in parts of Belize. According to the 2020 Mexican census, it is spoken by 774,755 people aged three and older, making it the second most widely spoken indigenous language in Mexico after Náhuatl (Instituto Nacional de Estadística y Geografía, 2020). Despite this relatively large population of speakers, the language is under sustained pressure from Spanish, driven by urbanization, migration, mass media exposure, and the structural dominance of Spanish in education, government, and economic life.

Bilingualism with Spanish is now widespread among Yucatec Maya speakers: the 2020 census records a monolingual rate of just 4.3%, and 65.18% of Yucatán's population self-identify as Indigenous culturally, the second highest rate of any Mexican state (Instituto Nacional de Estadística y Geografía, 2020). Even so, this strong Indigenous identity has not prevented language shift: Yamasaki (2019) documents the declining rate of Yucatec Maya's intergenerational transmission, resulting in a community where Mayan identity persists while active language use weakens across generations.

The dominance of religious content in available Yucatec Maya parallel data is not a coincidence of the language's low-resource status, but a consequence of who has historically been motivated to produce written translations. During the colonial period, Spanish missionaries produced the earliest known translations of Catholic texts into Yucatec Maya, using the language as a vehicle for proselytism and cultural governance (Hanks, 2012). This pattern continued into the twentieth century: Protestant missionary organizations, most notably SIL International (operating in Mexico under government contract from the 1930s through 1979), pursued Bible translation across more than one hundred Mexican Indigenous languages as part of a broader evangelical agenda (Stoll, 1982). The parallel data available

today reflects this cumulative history — while non-religious Yucatec Maya-to-Spanish corpora do exist, such as the Corpus Paralelo de Lenguas Mexicanas (CPLM) [The Mexican Languages Parallel Corpus] (Sierra Martínez et al., 2020), they remain substantially smaller in scale compared to the digitally accessible, verse-aligned Bible translations that are among the largest available parallel resources for the language (Hutchinson, 2024).

2.3 Models

2.3.1 TowerInstruct-7B-v.02

For the purposes of this thesis, I needed a baseline model with strong Spanish-language competence and no documented exposure to Yucatec Maya, to ensure that any translation capability the model develops for Yucatec Maya in the course of the experiments comes from fine-tuning rather than pretraining, allowing the effect of training data composition to be isolated and measured.

TowerInstruct (Alves et al., 2024) met these requirements. TowerInstruct is an open-source large language model designed specifically for translation-related tasks. It was developed by first continuing the pretraining of LLaMA-2 on multilingual monolingual and parallel data across ten languages – including Spanish – to produce TowerBase, which was then instruction-tuned on translation tasks using the TowerBlocks dataset. This translation-specific training distinguishes TowerInstruct from general-purpose multilingual models such as mBART or NLLB, which are trained on broader objectives such as multilingual masked language modeling or general translation across hundreds of language pairs.

2.3.2 T5S

In addition to TowerInstruct-7B-v.02, this thesis also uses T5S as a second model for experiments with fine-tuning. T5S refers to the `vgaraujov/t5-base-spanish` checkpoint, a T5-base model (Raffel et al., 2020) that has been further pretrained on Spanish text. Unlike TowerInstruct-7B-v.02, T5S uses an encoder-decoder sequence-to-sequence architecture and was not instruction-tuned. It was selected for this thesis because it was used by Rangel and Kobayashi (2024) for the same language pair (but in the opposite direction), enabling a

direct architectural comparison with their results, and because its smaller size made full-parameter fine-tuning feasible under the same hardware constraints, without requiring the LoRA adaptation used for TowerInstruct-7B-v.02.

2.4 Summary

This chapter has traced three interconnected contexts that together motivate this thesis. The history of Bible translation as a computational resource shows how practical data scarcity led machine translation researchers to adopt religious corpora without adequately examining the assumptions underlying that choice—assumptions that postcolonial and ethical critiques have since called into question. The sociolinguistic situation of the Yucatec Maya community illustrates how this history plays out in a specific language. TowerInstruct-7B-v.02 and T5S provide two complementary model architectures for isolating the effect of training data composition: TowerInstruct for its strong Spanish-language foundation and translation-specific training, and T5S for its architectural comparability with prior work on this language pair. Together, these contexts frame the central question this thesis investigates: What happens to translation quality and output semantics when a model is fine-tuned on increasing proportions and amounts of Bible-derived parallel data?

Chapter 3

RELATED WORK

This chapter reviews the literature most directly relevant to this thesis, organized around three themes. The first is the growing body of ethical scholarship on the use of religious texts in NLP and machine translation for Indigenous languages, which frames Bible-derived training data not as a neutral resource but as a carrier of colonial history and potential cultural harm. The second is the methodological literature on low-resource machine translation, including parameter-efficient fine-tuning approaches and the domain mismatch challenges that arise when available parallel data is drawn from narrow or specialized sources. The third is prior machine translation work specifically involving Yucatec Maya, which provides a performance reference point and establishes the architectural context for the models used in this thesis. Together, these three areas of prior work define the research gap that this thesis addresses: the absence of a systematic, quantitative framework for measuring how Bible training data composition affects both translation quality and the frequency of Bible-related content in model outputs, grounded in the expressed perspectives of the affected community.

3.1 Recent Ethical Critiques of Religious Data Use

Domingues et al. (2024) examined the ethical dilemma of using potentially culturally toxic training data (Bible text) for Indigenous languages, specifically in a Guarani Mbya-to-English translation context. Guarani Mbya is spoken in Brazil, Argentina, and Paraguay. The researchers frame Bible-related content as culturally toxic in Indigenous contexts due to the historical entanglement of Bible translation with colonialism, violence, and cultural erasure. They fine-tuned a WMT19 neural machine translation model (Ng et al., 2019) under multiple training configurations using Guarani Mbya-to-English data from the Bible and a non-Bible dictionary corpus. They compared Bible-only, dictionary-only, and mixed-domain

models, varying both the linguistic scope of Bible data (single language, language-family, and multilingual Indigenous sets) and the training strategy (sequential versus simultaneous fine-tuning). Their experimental design aimed to assess whether exposure to Bible data introduces culturally toxic contamination in MT outputs and whether such contamination persists after subsequent fine-tuning on non-religious data. The findings of Domingues et al. (2024) showed that MT models fine-tuned exclusively on Bible data perform poorly on everyday translation tasks, as measured by BLEU and chrF, and that introducing even a small amount of dictionary data yields substantial gains in translation quality. They further found that training with both Bible and dictionary data – whether sequentially or simultaneously – outperformed training on dictionary data alone, with the simultaneous model achieving the best overall performance. Despite these quality improvements, their best-performing model still exhibited measurable Bible-derived contamination in its outputs, identified through manual inspection of 300 test outputs and estimated at roughly 4.7%, raising ethical concerns about deployment in Indigenous community contexts.

Domingues et al. (2024) did not computationally operationalize the identification of culturally toxic outputs in the behavior of their MT systems beyond keyword identification or manual human qualitative analysis. They also did not systematically vary the proportion of Bible versus non-Bible data. Instead, they compared Bible-only and non-Bible-only models, as well as sequential and simultaneous fine-tuning configurations, to assess whether contamination introduced by Bible data persists after subsequent fine-tuning on non-Bible data. This design did not quantify how contamination frequency or translation quality change as a function of increasing Bible data proportions, nor did it examine potential trade-offs between translation quality and Bible-related contamination. Thus, in this thesis, I extend the work of Domingues et al. (2024) by building a machine translation system for Yucatec Maya-to-Spanish. Like Guarani Mbya, Yucatec Maya is a low-resource Indigenous language, and like English, Spanish is a high-resource colonial language with a history of use in Indigenous contexts. To extend translation quality evaluation while being able to compare my results to those of Domingues et al. (2024), I employ COMET and METEOR alongside BLEU and chrF. To quantify the frequency of erroneously Bible-related content in system outputs, I implement a two-part detection framework involving stylistic n-gram

distribution analysis and semantic similarity comparisons with Bible sentence embeddings, as described in detail in Chapter 5.

Hutchinson (2024) examined the ethical implications of using religious texts in NLP, arguing that such data cannot be treated as neutral training material divorced from its cultural, historical, and spiritual contexts. Hutchinson (2024) critiqued the widespread use of translated religious texts and highlighted concerns that extend beyond model bias to include data origins, context collapse, proselytism, and power asymmetries between researchers and marginalized communities. Hutchinson also emphasized researcher positionality and the need to account for the perspectives of linguistic and religious communities for whom these texts hold sacred significance. While this work establishes a strong ethical framework and calls for more reflexive and community-aware data practices, it does not provide computational methods for identifying or measuring harm in machine translation outputs, nor does it examine how training data composition influences model behavior. This thesis addresses these gaps by pairing controlled machine translation experiments with a community survey, quantifying contamination in model outputs while asking Yucatec Maya speakers themselves whether and how machine translation should be used.

Mager et al. (2023) conducted an interview study with Indigenous language community members from the Aymara, Chatino, Maya, Mazatec, Mixe, Nahua, Otomí, Quechua, Tenek, Tepehuano, Kichwa of Otavalo, and Zapotec communities to examine ethical questions surrounding machine translation for Indigenous languages, including how community members wish to be involved in the MT process, which domains may be inappropriate to translate without explicit permission, and how linguistic data can be collected ethically. Similar to the survey results presented in Chapter 4 of this thesis, they find that many community members support the development of MT for their languages provided that there is close collaboration and community governance. Notably, several participants identified Western religious content as a sensitive domain, citing concerns about the historical use of Bible translation as a tool of cultural oppression and colonial violence. Accordingly, Mager et al. (2023) caution against the use of Bible translations in MT development unless such use is explicitly approved by the community.

Pinhanez et al. (2023) reflect on the tensions they encountered while developing LLMs for Brazilian Indigenous languages, acknowledging that their prototype violated several ethical guidelines – including training on Bible translations without community consent – while arguing that the potential for social impact justified proceeding under strict damage containment procedures. They frame this as a deliberate balancing act requiring what they call damage containment procedures – explicit awareness of harm, restricted deployment, and community disclosure. This thesis addresses the same tension by experimentally evaluating translation quality alongside Bible-related contamination frequency, making the trade-off between data availability and ethical risk empirically measurable rather than assumed.

3.2 Low-Resource MT Methodology

Work by Haddow et al. (2022) surveys strategies for low-resource machine translation, including transfer learning, synthetic data generation, careful domain balancing, and fine-tuning pretrained models on small parallel corpora, which is the approach taken in my experiments. They note that in many low-resource settings, available parallel data is often drawn from narrow domains such as religious texts or dictionaries, which can lead to domain mismatch between training and evaluation data – a challenge directly relevant to this thesis, which investigates the effects of Bible-derived parallel data on translation quality and semantic drift.

To adapt a pretrained machine translation model under constrained computational resources, I employ Low-Rank Adaptation (LoRA), introduced by Hu et al. (2022). LoRA fine-tunes models by injecting trainable low-rank matrices into attention layers while keeping the original model parameters frozen, substantially reducing memory usage and training cost compared to full fine-tuning. Hu et al. demonstrate that LoRA achieves performance comparable to full fine-tuning across a range of NLP tasks; I adopt it here because it makes fine-tuning a 7B parameter model feasible under GPU memory constraints, and because holding the underlying model architecture constant across all experimental conditions enables cleaner comparison of data composition effects.

3.3 MT for Yucatec Maya

Rangel and Kobayashi (2024) demonstrated the feasibility of producing usable translations involving Yucatec Maya despite severe data scarcity. They developed NMT systems for Yucatec Maya and Chol using large language models pretrained on Spanish and further fine-tuned on 28,135 parallel sentence pairs extracted from the Corpus Paralelo de Lenguas Mexicanas (CPLM) (Sierra Martínez et al., 2020). They evaluated both one-to-many and many-to-many model architectures under multilingual and bidirectional training setups and found that multilingual training consistently outperforms monolingual setups, with their best model achieving chrF++ scores exceeding 50 and BLEU scores near 29 for Spanish-to-Yucatec Maya translation. These scores are higher than those achieved in this thesis across most experimental conditions, with the gap being substantially smaller for the T5S model under low-Bible-proportion conditions, as discussed in Chapter 6 and Chapter 7. Notably, Rangel and Kobayashi (2024) employ T5S and M2M100 as their primary model architectures. The present thesis replicates the seven experimental conditions using T5S in addition to TowerInstruct-7B-v.02, enabling a more direct architectural comparison with their work.

The present thesis uses a different dataset than Rangel and Kobayashi (2024) for two reasons. First, the non-Bible data used here was obtained directly from Yucatec Maya community members – César Can and Silvia Fernandez Sabido – reflecting a commitment to community partnership and data sovereignty that is central to the ethical framing of this thesis. The Bible data used here was obtained from the New World Translation, recommended by César Can and Silvia Fernandez Sabido as one of the most widely used version in the community. Second, the CPLM corpus was not identified as a data source during the initial design of this study. As a result, direct comparison with Rangel and Kobayashi (2024) is complicated not only by differences in dataset composition and size, but also by differences in translation direction and model architecture. Nonetheless, their results provide a useful point of reference for situating the performance levels achieved here within the broader landscape of low-resource Yucatec Maya NMT.

3.4 Corpus Comparison and Semantic Similarity Methods

The experimental conditions described in this thesis vary the proportion of Bible-derived parallel data used during fine-tuning, raising the question of how exposure to Bible text may affect the model's outputs to drift toward Bible-like language even when translating non-Bible input. This phenomenon, referred to here as *semantic drift*, encompasses both stylistic overfitting to Bible phrasing and subtler shifts in semantic content toward religious themes. Two complementary methods are used to detect it: n-gram distribution analysis, which captures surface-level stylistic contamination, and semantic similarity analysis, which captures deeper content-level drift.

The n-gram distribution analysis employs the log-likelihood ratio (LLR) statistic to identify n-grams significantly overrepresented in the Bible corpus relative to the non-Bible corpus. This approach follows Dunning (1993), who demonstrated that likelihood ratio tests are more appropriate than chi-square or z-score methods for statistical text analysis, particularly when the items of interest are rare, as is the case with natural language. Dunning showed that normality-based tests dramatically overestimate the significance of rare events, whereas likelihood ratio tests yield reliable results even with small counts, and demonstrated the method's effectiveness for identifying domain-characteristic bigrams through corpus comparison. Bible-distinctive n-grams identified through this method serve as markers of stylistic contamination: outputs containing significantly more of these n-grams than their reference translations are flagged as having drifted toward Bible-like phrasing.

For the semantic similarity component, sentence embeddings are computed using BETO (Cañete et al., 2020), a BERT model pretrained exclusively on Spanish data that has been shown to outperform multilingual BERT on a range of Spanish NLP tasks. Cosine similarity is used to measure the proximity of each output and its reference to the Bible corpus. Following Reimers and Gurevych (2019), who caution that raw BERT embeddings can be unreliable for absolute semantic similarity judgments, contamination is assessed through relative comparison, which means an output is flagged only when its Bible similarity score exceeds that of its corresponding reference by a predefined margin, controlling for baseline similarity patterns in the embeddings. This relative framing is important because even non-

Bible text may share some vocabulary or phrasing with the Bible; the goal is to identify outputs that have shifted *toward* the Bible domain more than their references, not simply outputs that resemble the Bible in absolute terms.

3.5 Summary

The literature reviewed in this chapter covers three themes that together situate the present thesis. First, recent ethical scholarship has documented the risks of using Bible-derived data in NLP for Indigenous languages, framing such data not as neutral training material but as a carrier of colonial history and potential cultural harm (Domingues et al., 2024; Hutchinson, 2024; Mager et al., 2023; Pinhanez et al., 2023). These works collectively call for more critically self-aware, community-centered data practices, but none provides a systematic computational framework for detecting semantic drift toward Bible-like language and content as a function of training data composition. Second, the low-resource MT methodology literature establishes fine-tuning pretrained models on small parallel corpora as a viable approach for languages like Yucatec Maya, while also noting that the religious or lexicographic nature of most available parallel corpora introduces domain mismatch challenges that must be carefully managed (Haddow et al., 2022). Third, prior NMT work directly targeting Yucatec Maya (Rangel and Kobayashi, 2024) demonstrates that usable translation quality is achievable despite severe data scarcity, providing a useful performance reference point, though it does not address the ethical questions surrounding Bible data use that motivate this thesis. Together, these threads motivate the core contribution of this thesis: a controlled experimental study that systematically varies the proportion of Bible-derived training data for Yucatec Maya-to-Spanish machine translation across two model architectures – TowerInstruct-7B-v.02 and T5S – pairing translation quality evaluation with computational detection of Bible-related contamination in model outputs, and grounding both in a survey of the perspectives held by the Yucatec Maya community.

Chapter 4

SURVEY

I conducted a survey of adult Yucatec Maya speakers to collect their opinions about machine translation generally, the types of texts used as training data for machine translation systems, and data sovereignty.[1]

4.1 Survey Methodology

The survey was conducted via Google Forms. All questions were optional to ensure that participants could skip any questions that they found confusing or sensitive. All responses were also anonymous; the Google Form settings were such that email addresses of participants were not collected.

4.1.1 Participant Recruitment

Participants were recruited via email with the help of Yucatec Maya language community leaders Silvia Fernandez Sabido from CentroGeo Yucatán and César Can from the Indigenous language collective T'aantsil.

4.1.2 Language Privacy Concerns

The survey was conducted in Spanish rather than Yucatec Maya, as members of the Yucatec Maya language community deem it important to keep their language data out of the hands of big tech corporations like Google. Additionally, collecting the survey responses in Spanish made the analysis process easier given that I speak no Yucatec Maya but comprehend Spanish at an intermediate level. For Spanish responses beyond my comprehension, I used

[1]The University of Washington Human Subjects Division granted this survey Institutional Review Board exempt status due to its minimal risk to participants.

Google Translate, which would have been neither possible at a high level of accuracy nor ethical had the responses been collected in Yucatec Maya.

4.1.3 Participant Roles and Language Background

This question was intended to contextualize participants' responses in light of their experience with the Yucatec Maya language and possible professional roles that could influence their expertise and opinions about machine translation, data sources, and data ownership.

Question as Stated in Survey:

Seleccione todas las opciones que lo describan y/o utilice "Otro" para describir su rol:

- ☐ El maya yucateco es mi lengua materna
- ☐ El maya yucateco es mi segunda (o tercera, etc.) lengua
- ☐ Soy miembro de la comunidad lingüística maya yucateca
- ☐ Soy lingüista
- ☐ Soy activista de la comunidad maya yucateca
- ☐ Soy especialista en informática
- ☐ Soy docente de maya yucateco
- ☐ Soy intérprete/traductor(a) de maya yucateco
- ☐ Otro:

English Translation:

Select all options that describe you and/or use "Other" to describe your role:

- ☐ Yucatec Maya is my first language

☐ Yucatec Maya is my second (or third, etc.) language

☐ I am a member of the Yucatec Maya language community

☐ I am a linguist

☐ I am a Yucatec Maya language community activist

☐ I am a computer scientist

☐ I am a teacher of the Yucatec Maya language

☐ I am an interpreter / translator of Yucatec Maya

☐ Other:

4.1.4 Language Use Frequency across Contexts

These questions aim to identify how often Yucatec Maya is spoken by participants across different areas of life, in order to contextualize patterns of bilingualism in the community. It is important to note that language use frequency and desire for machine translation do not map onto each other in a straightforward way. Machine translation for Yucatec Maya could serve a range of purposes that are not reducible to compensating for individual speaker fluency: it could help Spanish-dominant speakers or language learners access Yucatec Maya content, support Yucatec Maya speakers in navigating Spanish-language institutions such as healthcare, legal, or government services, assist in language documentation and revitalization efforts, or provide researchers and educators outside the community with access to Yucatec Maya texts. Accordingly, high frequency of Yucatec Maya use in a given context does not necessarily indicate that MT would be unwanted there, and low frequency does not necessarily indicate that MT would be welcome.

Questions as Stated in Survey:

Pregunta		**Escala**						
¿Cuánto utilizas el maya yucateco...		0	1	2	3	4	5	
en casa o en familia?	nunca	○	○	○	○	○	○	todo el tiempo
con amigos?	nunca	○	○	○	○	○	○	todo el tiempo
en la escuela o en el trabajo?	nunca	○	○	○	○	○	○	todo el tiempo
en el médico?	nunca	○	○	○	○	○	○	todo el tiempo
con las autoridades locales?	nunca	○	○	○	○	○	○	todo el tiempo
en la iglesia?	nunca	○	○	○	○	○	○	todo el tiempo
en ceremonias?	nunca	○	○	○	○	○	○	todo el tiempo
en redes sociales?	nunca	○	○	○	○	○	○	todo el tiempo

English Translation:

Question		**Scale**						
How often do you use Yucatec Maya...		0	1	2	3	4	5	
at home or with family?	never	○	○	○	○	○	○	always
with friends?	never	○	○	○	○	○	○	always
in school or at work?	never	○	○	○	○	○	○	always
at the doctor?	never	○	○	○	○	○	○	always
with local authorities?	never	○	○	○	○	○	○	always
at church?	never	○	○	○	○	○	○	always
in ceremonies?	never	○	○	○	○	○	○	always
on social media?	never	○	○	○	○	○	○	always

4.1.5 Perceived Applications, Benefits, and Risks of Machine Translation

These questions were intended to collect information about Yucatec Maya speakers' opinions about machine translation systems for their language. Participants' responses help to answer my research question about the specific ethical considerations that ought to be taken into account when developing a machine translation system for an Indigenous language like Yucatec Maya.

Question as Stated in Survey:

La traducción automática se refiere al uso de tecnología informática para traducir automáticamente de un idioma a otro. La traducción automática podría ser usada para traducir desde o hacia el maya yucateco, hacia o desde varios otros idiomas (español, portugués, etc.). ¿Apoya el desarrollo de tecnología de traducción automática hacia o desde el maya yucateco?

Se opone muy fuertemente ○ ○ ○ ○ ○ ○ Apoya muy fuertemente

English Translation:

Machine translation refers to the use of computer technology to automatically translate from one language to another. Machine translation could be used to translate to or from Yucatec Maya, or to or from various other languages (Spanish, Portuguese, etc.). Do you support the development of machine translation technology to or from Yucatec Maya?

Strongly oppose ○ ○ ○ ○ ○ ○ Strongly support

Question as Stated in Survey:

Utilice este espacio para explicar por qué respondió así en la pregunta anterior.

English Translation:

Use this space to explain your answer in the previous question.

Question as Stated in Survey:

¿Qué usos considera que podrían darse a una tecnología de traducción automática para la lengua maya?

English Translation:

How do you think machine translation technology for the Mayan language would be used?

Question as Stated in Survey:

¿Qué preocupaciones tiene usted sobre la tecnología de traducción automática?

English Translation:

What concerns do you have about machine translation technology?

Question as Stated in Survey:

¿Cuáles cree que son los beneficios y/o posibles afectaciones que podría tener la tecnología de traducción automática para las comunidades mayas?

English Translation:

What do you think are the benefits and/or potential impacts that machine translation technology could have on Mayan communities?

Question as Stated in Survey:

¿Cuáles considera que son los desafíos tecnológicos que enfrenta la comunidad lingüística maya yucateca?

English Translation:

What do you consider to be the technological challenges facing the Yucatecan Mayan linguistic community?

4.1.6 Data Ownership

These questions gather participants' opinions about data sovereignty, which also informs my research question about the ethical considerations of building a Yucatec Maya machine translation system. Machine translation systems require sufficient amounts of training data, so it is important that the methods of sourcing, handling, and storing the data are considered fair and just by the language community.

Question as Stated in Survey:

¿Qué opina sobre los casos en que los datos recolectados en comunidades mayas, han pasado a ser propiedad de las instituciones que participan en los proyectos académicos o sociales?

English Translation:

What is your opinion on cases where data collected in Mayan communities[2] has become the property of the institutions participating in academic or social projects?

Question as Stated in Survey:

¿En su opinión, quién o quiénes deberían tener la propiedad de los datos recolectados en comunidades mayas, para controlar y decidir sobre los posibles usos?

English Translation:

In your opinion, who should own the data collected in Mayan communities, in order to control and decide on its possible uses?

4.1.7 Training Data Sources

Participants' answers to these questions give information about which types of documents – including the Bible – the Yucatec Maya language community views as appropriate or inap-

[2] "Mayan communities" is used here rather than "Yucatec Maya" because the institutional appropriation of community-collected data is not specific to Yucatec Maya speakers but affects speakers of Mayan languages broadly. The question asks respondents to draw on awareness of regional patterns, not only their direct experience as Yucatec Maya speakers.

propriate for use as training data in machine translation systems. I was advised not to ask about participants' opinions of the Bible more directly than as worded in the question below due to the cultural sensitivity of the topic. Note that I also listed "colonial texts" separately from "religious texts (Bibles)", since it is possible that not all participants would classify the Bible as a "colonial text", and because there are non-biblical documents that could also be considered "colonial texts".

Question as Stated in Survey:

¿Le preocupa que se usen algunos de estos textos para entrenar los traductores automáticos?

- ☐ Textos coloniales
- ☐ Textos traducidos del español al maya (no pensados en maya)
- ☐ Textos religiosos (biblias)
- ☐ Textos literarios (poesías, cuentos, etc.)
- ☐ Narrativas cotidianas de la comunidad (conversaciones reales)
- ☐ Otro:

English Translation:

Are you concerned that some of these texts might be used to train machine translators?

- ☐ Colonial texts
- ☐ Texts translated from Spanish to Mayan (not written in Mayan)
- ☐ Religious texts (Bibles)
- ☐ Literary texts (poems, short stories, etc.)

☐ Everyday community narratives (real conversations)

☐ Other:

Question as Stated in Survey:

Utilice este espacio para explicar por qué respondió así en la pregunta anterior.

English Translation:

Use this space to explain why you answered that way in the previous question.

4.2 Survey Results

There were 84 responses collected. Every question was optional, so the number of respondents for a given question did not always equal 84.

4.2.1 Participant Roles and Language Background

Figure 4.1 shows the number of respondents who identified themselves with each of the roles or language backgrounds listed in the legend on the right-hand side of the bar graph. The order of the roles or backgrounds from top to bottom in the legend matches the order of the bars from left to right.

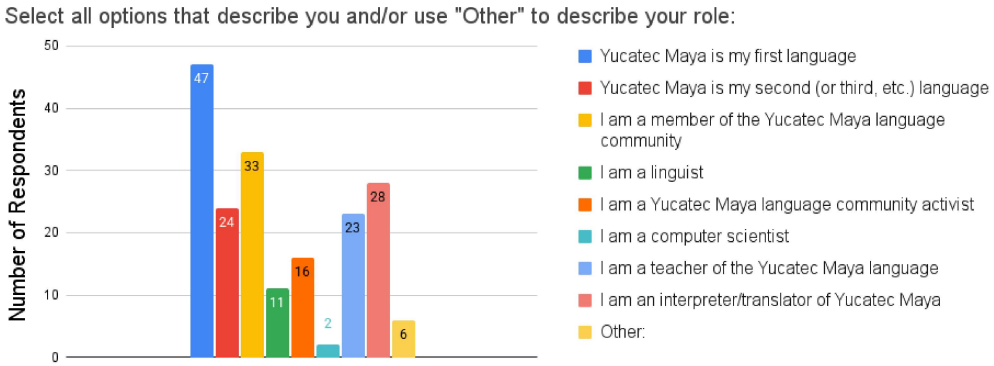

Figure 4.1: Distribution of Participant Roles and Language Background (n=79).

It should be noted that 47 respondents indicated that Yucatec Maya was their first language, while only 33 respondents considered themselves part of the Yucatec Maya language community. I intended for all L1 speakers to be considered part of the language community, along with any L2+ speakers who consider themselves to be culturally and/or socially part of the language community. The wording of the survey question left "language community" underspecified. It is possible that some participants may not have realized that this question allowed for multiple responses. Eleven participants checked only the box to indicate that Yucatec Maya was their mother tongue. Other native speakers evidently realized that this question allowed for multiple responses, as fourteen participants indicated that Yucatec Maya was their mother tongue and indicated that another choice applied to them, but did not indicate that they were part of the language community. Although Yucatec Maya is their L1, they may not feel connected to the "language community" according to their interpretation of what the "language community" is, or they may not use Yucatec Maya as their primary language in their day-to-day life anymore. Even more perplexing is the fact that three participants answered that Yucatec Maya was both their mother tongue and their second or third language. It is possible that the meaning of "lengua materna" or "segunda (o tercera) lengua" was unclear to these individuals or that these respondents made a mistake. Three respondents also indicated that they were part of the Yucatec Maya language community, but did not identify themselves as any of the other options listed.

There were six participants who indicated that their relationship to the Yucatec Maya language was something "other" than the given options, including one blank response:

1. "Soy estudiante en un contexto intercultural donde el maya yucateco es la principal lengua materna hablada." meaning "I am a student in an intercultural context where Yucatec Maya is the main native language spoken."

2. "No hablo ni entiendo la lengua maya." meaning "I do not speak or understand the Mayan language."

3. "Soy antropólogo." meaning "I am an anthropologist."

4. “Estudiante” meaning “student”

5. “Soy cineasta documentalista audiovisual, documento la memoria histórica de los pueblos mayas de Yucatán, mayormente en lengua maya desde su forma de vida, música, costumbres, tradiciones, etc. Promuevo el uso de la lengua maya con mi familia y en espacios en donde me desarrollo de manera más independiente y por intereses de la propia comunidad, diferente a las que realiza las instituciones gubernamentales” meaning “I am an audiovisual documentary filmmaker, I document the historical memory of the Mayan peoples of Yucatán, mostly in the Mayan language from their way of life, music, customs, traditions, etc. I promote the use of the Mayan language with my family and in spaces where I develop more independently and by the interests of the community itself, different from those carried out by government institutions”.

These responses suggest that the predefined relationship categories did not fully capture the diversity of ways people relate to Yucatec Maya. Notably, the responses by the two students and the anthropologist represent people who are actively trying to engage with the language from outside the core speaker community, pointing to a potential use case for machine translation as a tool for language learning and access. The filmmaker’s response frames language use not merely in terms of fluency or identity, but in terms of active cultural stewardship and community-driven revitalization.

4.2.2 Language Use Frequency across Contexts

Figure 4.2 shows the number of respondents who answered 0 to 5 in each of the contexts listed below the x-axis. 0 indicates that the respondent never uses Yucatec Maya in that context, while 5 indicates that the respondent always uses Yucatec Maya in that context.

The context in which the greatest number of respondents reported that they never use Yucatec Maya is at the doctor. The context in which the greatest number of respondents reported that they use Yucatec Maya all of the time is at home or with family. There is also a significant number of respondents who answered that they never use Yucatec Maya with local authorities, at church, or in ceremonies.

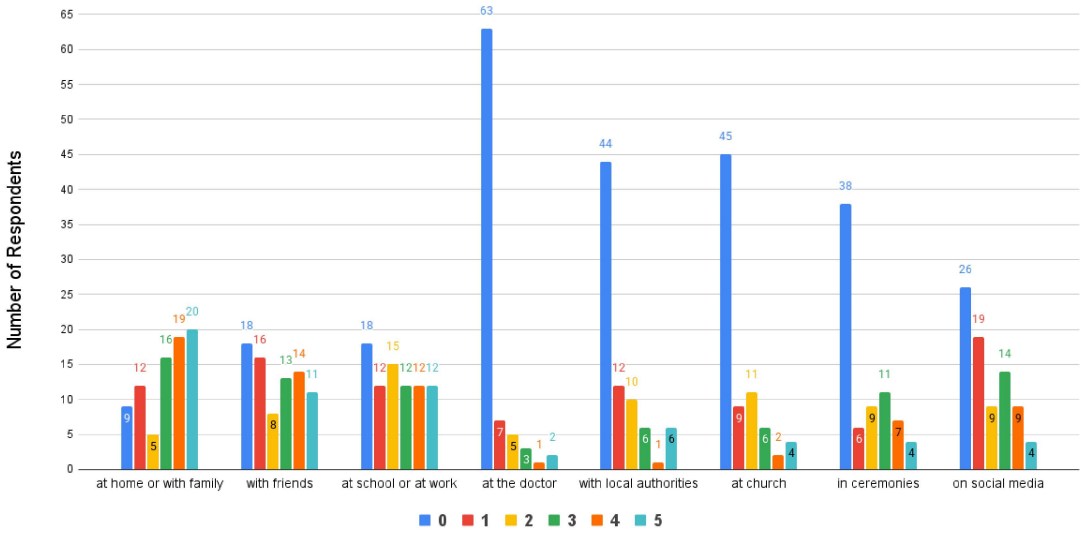

Figure 4.2: Yucatec Maya Usage across Contexts

These results suggest a pattern in which Yucatec Maya use is most frequent in intimate and domestic contexts – at home, with family, and with friends – and most suppressed in institutional ones, particularly at the doctor. This domestic/institutional divide has direct implications for the potential role of machine translation in this community. The contexts in which Yucatec Maya is used least are precisely those in which translation support could be most consequential: navigating healthcare, legal systems, and other Spanish-dominant institutions where language barriers carry real stakes and where both the accuracy of translations and transparency about their reliability are paramount.

4.2.3 Perceived Applications, Benefits, and Risks of Machine Translation

I qualitatively analyzed open-ended survey responses regarding participants' reasons for supporting the development of Yucatec Maya machine translation, their reasons for opposing it, and their perceptions of potential uses, benefits, and risks. I read all responses systematically and identified recurring topics and concerns, noting the frequency with which each appeared across the responses. This process followed the initial phases of thematic analysis

as described by Braun and Clarke (2006), which involved familiarization with the data by reading through it multiple times and identification of recurring patterns.

I also report quantitative results from Likert-scale items measuring participants' degree of support for Yucatec Maya machine translation development. Figure 4.3 shows the number of participants who answered 0 to 5 indicating their degree of support for the development of Yucatec Maya machine translation, with 0 indicating strong opposition, and 5 indicating strong support.

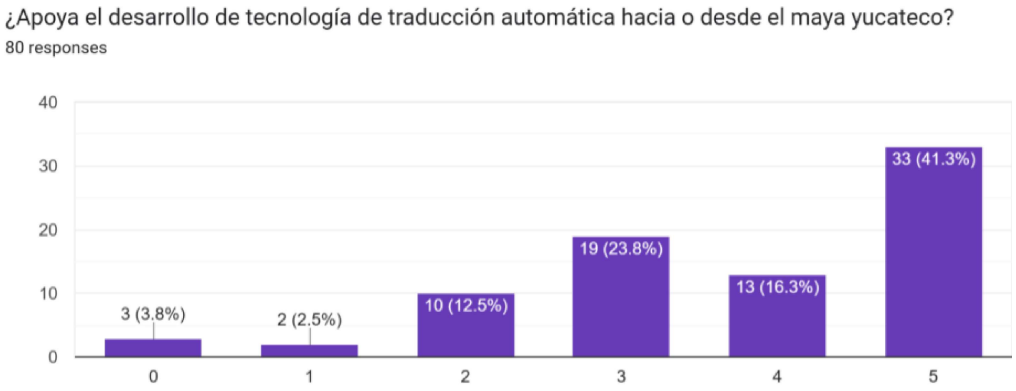

Figure 4.3: Degree of Support for Yucatec Maya Machine Translation

The majority of respondents indicated some degree of support, with the most common response being 5. To contextualize how language background or professional role may influence one's support for or opposition to the development of Yucatec Maya machine translation, I analyzed what percentage of L1 speakers, language community members, and linguists answered at the extremes of the 0 to 5 range. Approximately 47% of L1 speakers and/or Yucatec Maya language community members expressed strong support for machine translation in Yucatec Maya, indicated by a 5. Only 5% expressed strong opposition, indicated by a 0. Approximately 36% of linguists expressed strong support, while only 9% expressed strong opposition. The following subsections describe the reasons participants gave for their support or opposition.

Reasons for Supporting Yucatec Maya Machine Translation

The most common reason for support was the belief that machine translation could help revitalize and preserve the Yucatec Maya language by promoting its usage and making the language more accessible to a broader audience, which was mentioned by 13 respondents. One respondent captured this sentiment: "Considero que la lengua maya debe ser impulsada en todos los aspectos, incluyendo el tecnológico, ya que esto contribuye a valorarla, preservarla y transmitirla a las futuras generaciones. Incorporar el maya yucateco en herramientas digitales no solo fortalece su uso cotidiano, sino que también le da visibilidad, respeto y utilidad práctica en la vida moderna. La tecnología puede ser una gran aliada para revitalizar lenguas originarias y asegurar que sigan vivas, activas y accesibles para todos" [I believe the Mayan language should be promoted in all aspects, including technology, as this contributes to valuing, preserving, and transmitting it to future generations. Incorporating Yucatec Maya into digital tools not only strengthens its everyday use but also gives it visibility, respect, and practical utility in modern life. Technology can be a great ally in revitalizing indigenous languages and ensuring that they remain alive, active, and accessible to all.] Twelve respondents answered that they supported the development of Yucatec Maya machine translation systems because of its potential to aid in language learning, and eleven participants described how they thought machine translation could help individuals who speak different languages avoid misunderstandings when communicating with each other. The lack of Yucatec Maya speakers in healthcare, government, schools, and other public institutions was another common reason for support due to the belief that machine translation could help make these public services more accessible to Yucatec Maya speakers. Six respondents also discussed the ways that machine translation could increase the representation and visibility of the Yucatec Maya community by translating the writings written by Yucatec Maya speakers into other languages. Certain other respondents would likely strongly disagree with this sentiment, as many emphasized how there are many cultural and philosophical concepts in the Yucatec Maya community that simply cannot be translated to other languages. Participants gave no specific examples of cultural or philosophical concepts that cannot be translated, possibly because they were writing their responses in

Spanish rather than Yucatec Maya. Relatedly, two respondents expressed support for how machine translation would increase young Yucatec Maya speakers' ability to engage online, as illustrated by one respondent who wrote: "Se necesitan nuevos métodos para poder hacer que el maya se siga hablando, que se siga usando no solo en casa, si no en todos los espacios necesarios, las redes sociales podrían ser puntos clave para que este idioma siga o incluso aumente de personas que lo hablan" [New methods are needed to ensure that Mayan continues to be spoken, to continue being used not only at home but in all necessary spaces. Social media could be key to ensuring that this language continues to be spoken, or even increases its number of speakers.]

Reasons for Opposing Yucatec Maya Machine Translation

The most common reason for opposition was the concern that machine-generated translations to or from Yucatec Maya would be inaccurate, low-quality, or too literal, since many Yucatec Maya words do not have a direct translation and must be interpreted instead. Another common reason, mentioned by 3 participants, was the replacement of translator jobs and the de-prioritization of the human element of interpretation. Several respondents felt that the use of machine translation is a superficial solution for complex problems, such as the problem of how to preserve and revitalize the Yucatec Maya language, which was the most common reason for support as mentioned above. One respondent argued that the most effective way to preserve and revitalize a language is to motivate the use of the language in one's own locality or home, and that the widespread use of machine translation would demotivate individuals from learning the language since they could rely on the convenience of the machines instead. Other respondents expressed concern that increased implementation of machine translation systems may result in new inequality gaps, the loss of dialogue between young people and older Yucatec Maya speakers, and the propagation of misinformation and bias.

One respondent synthesized several of these concerns while also noting that the appropriateness of machine translation cannot be evaluated in the abstract, but depends on the specific sociolinguistic situation of the speakers involved: "Por la situación de racismo y

discriminación, una traducción automatizada podría invisibilizar a personas hablantes y sus funciones como intérpretes o traductores. Adicionalmente, pese a que es normal que un traductor automático, pase por fases iniciales de correcciones, cabe la posibilidad que las traducciones producidas en estas etapas, sean tomadas por correctas sin revisión o consideraciones previas, abonando a la desinformación y también a que se atienda a necesidades de las personas hablantes de lengua maya de manera superficial Interpretaciones orales por otra parte; para lenguas como el tsotsil en un contexto citadino como de Mérida Yucatán, en el que no cuenta con hablantes bilingües que puedan hacer interpetación, sería de mucha ayudad para resolver situaciones emergente y de emergencia. En conclusión, es difícil que una respuesta pueda ser generalizable porque hay que tomar en cuenta la situación sociolingüística de los hablantes, la vitalidad de la lengua, y el contexto (social, económico, político, etc.) inmediato. Cada intención de realizarlo, merece un análisis profundo y concienzudo de sus implicaciones y posibles afectaciones, además de los posibles beneficios" [Due to the situation of racism and discrimination, an automated translation could render speakers and their roles as interpreters or translators invisible. Additionally, although it is normal for an automated translator to go through initial correction phases, it is possible that the translations produced in these stages are taken as correct without prior review or consideration, contributing to misinformation and also causing the needs of Mayan speakers to be superficially addressed. Oral interpretations, on the other hand, for languages like Tsotsil in an urban context like Mérida, Yucatán, where there are no bilingual speakers capable of interpreting, would be very helpful in resolving emergent and emergency situations. In conclusion, it is difficult for a response to be generalizable because the sociolinguistic situation of the speakers, the vitality of the language, and the immediate context (social, economic, political, etc.) must be taken into account. Every attempt to carry out this type of interpretation deserves a thorough analysis and aware of its implications and possible effects, as well as the possible benefits.] This sentiment reflects a broader pattern among opposing respondents, who tended to emphasize that the value and risks of machine translation for Yucatec Maya cannot be assessed without careful consideration of the community's specific social, political, and linguistic context.

Perceived Uses of Yucatec Maya Machine Translation

The most commonly mentioned use for both Spanish-to-Yucatec Maya and Yucatec Maya-to-Spanish machine translation was academic and educational use, in the form of language learning in schools or linguistic research via interviews and surveys. Relatedly, one participant explained that machine translation could increase Yucatec Maya speakers' access to knowledge and resources that are only available in dominant languages. The next most commonly mentioned theme was using machine translation to translate individual words or act as a spellchecker in Yucatec Maya.

Another common theme was facilitating interactions between people in a variety of public service settings. One respondent described the potential impact across several of these settings: "En los servicios médicos, facilitaría la comunicación entre personal de salud y pacientes mayahablantes, mejorando el diagnóstico, el seguimiento y la confianza en la atención médica, especialmente en comunidades rurales. En el ámbito legal, permitiría que las personas que solo hablan maya comprendan mejor sus derechos, procesos judiciales y documentos legales, garantizando un acceso más justo a la justicia. En la educación, sería una herramienta útil tanto para estudiantes mayahablantes como para docentes, ayudando en la comprensión de materiales escolares y promoviendo el bilingüismo desde una perspectiva de respeto cultural.En resumen, esta tecnología puede ser un puente para la inclusión, el respeto y la equidad lingüística, ayudando a preservar la lengua maya y darle un lugar digno en la vida pública y digital" [In medical services, it would facilitate communication between healthcare personnel and Mayan-speaking patients, improving diagnosis, follow-up, and trust in medical care, especially in rural communities. In the legal field, it would allow people who only speak Mayan to better understand their rights, judicial processes, and legal documents, ensuring fairer access to justice. In education, it would be a useful tool for both Mayan-speaking students and teachers, aiding in the comprehension of school materials and promoting bilingualism from a perspective of cultural respect. In short, this technology can be a bridge to inclusion, respect, and linguistic equity, helping to preserve the Mayan language and give it a worthy place in public and digital life.] Other respondents agreed and similarly stated that machine translation could be used to translate legal documents,

justice system proceedings, public safety announcements, reports by law enforcement, communication between social workers and clients, and diplomatic communication with other countries. Other notable responses included using machine translation to help translate recipes, subtitles of movies, and song lyrics.

Concerns about Yucatec Maya Machine Translation

The most common theme among the reported concerns was the possibility of inaccurate translations. Specifically, six respondents were concerned that machine translation systems would not have the ability to consider regional variation of the language – whether Spanish or Yucatec Maya. For example, regarding Spanish, one respondent wrote: "Que no considere la variación de la lengua en la comunidad. Por ejemplo si se estuviera construyendo un traductor de español pero solo se enfoque en el castellano o use palabras muy técnicas que la poblacion en general no entendería." [That doesn't take into account language variation in the community. For example, if a Spanish translator were being created but it only focused on Castilian Spanish or used highly technical terms that the general population wouldn't understand.] Relatedly, regarding Yucatec Maya, another respondent wrote: "Que pueda perderse el sentido original de la traducción, ya que la lengua maya varía de acuerdo a la región." [The original meaning of the translation may be lost, since the Mayan language varies according to the region.] Similarly, another respondent was concerned about MT's general ability to handle colloquial usages of words and phrases.

Two respondents were concerned that misplaced or absent apostrophes or accents could change the meaning of a word or phrase. One respondent similarly expressed concern that the outputted orthography of Yucatec Maya would be incorrect. One respondent pointed out that many words in Spanish do not have an equivalent idea in the Mayan language, and that Yucatec Maya must be interpreted rather than translated. Relatedly, four respondents were concerned that machine translation systems may not have the ability to consider context with enough complexity and would output translations that were too literal.

The next most common theme among the reported concerns was the possibility of misuse. Specifically, respondents were concerned that machine translation could be used for

linguistic extraction, for biocultural knowledge extraction, for organized crime, for plagiarism, to discriminate against minority language communities, as a replacement for in-person language learning, as a medium for bias and discrimination, and as a replacement for human translators. One participant was also concerned that machine translation could aid corporations in taking ownership of the language for commercial use, partially by perpetuating the impoverishment of the language through the loss of the subtle differences in meaning between similar words.

Four respondents were also concerned about the training process for machine translation systems. Specifically, two participants described how data might be used to train the systems without permission of the people who produced the data. One participant expressed concern that native speakers might not be involved in the development of the machine translation systems, and another was concerned that the training data would be insufficient to produce a high-quality and versatile translation system.

4.2.4 Training Data Sources

Figure 4.4 shows the number of respondents who expressed concern about each type of training data source. Thirty-two participants were concerned that texts not written in Yucatec Maya but instead translated from Spanish into Yucatec Maya would be used to train a Yucatec Maya-to-Spanish translation system. Reasons for this concern included that this type of training text might be too literal or simplistic and would not represent the reality of the Yucatec Maya language, since it was not originally written in the language, which is highly complex in its morphology and interpretive meaning. Another respondent emphasized how a key element of speaking Yucatec Maya is thinking with Mayan philosophy and culture in mind.

Twenty-eight participants were concerned about colonial texts being used as training data. A key reason for this was the role of colonial texts in Indigenous peoples' history of oppression. Accordingly, one respondent wrote "Si el uso de estas herramientas será para continuar reproduciendo discursos colonialistas bajo un mismo sistema de pensamiento homogéneo, perdería el alcance real o esencia de la lengua y se volvería otro instrumento

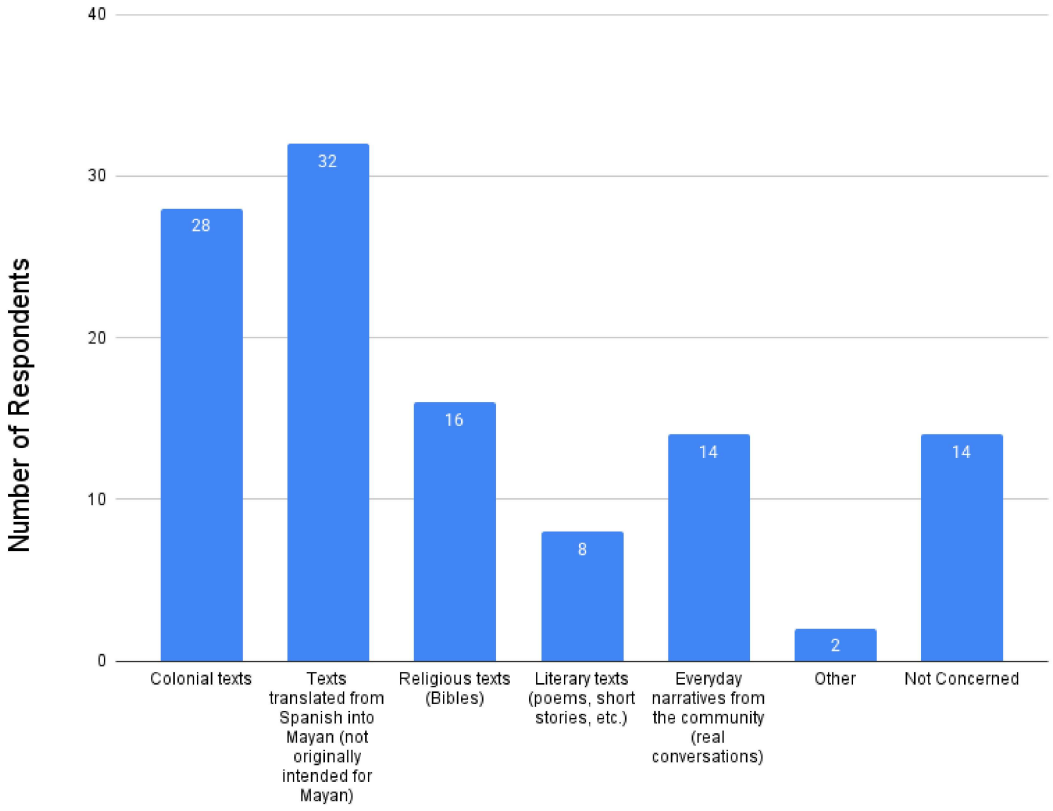

Figure 4.4: Training Data Sources of Concern

Question: Are you concerned that some of these texts might be used to train machine translators?

de opresión." [If these tools are used to continue reproducing colonialist discourses under the same homogeneous system of thought, the true scope or essence of language would be lost and would become another instrument of oppression.] Similarly, other respondents emphasized how colonial texts tell stories in a distorted manner from the perspective of the conquistadors and have the potential to manipulate people.

Sixteen respondents expressed concern over religious texts, such as the Bible, being used as machine translation training data. Their reasons for this included how it could cause discord in the community, how religious texts do not represent everyday language use in terms of vocabulary and syntactic structures, how certain religious texts are connected to a history of oppression and cultural erasure for Indigenous communities, how the Mayan community has its own culture and religion, and how religious content also has the potential to manipulate people.

In regard to everyday community narratives and real conversational data being used as training data in machine translation systems, fourteen respondents expressed concern. Two respondents emphasized the need to ensure that privacy is maintained and consent is clearly obtained when using conversational data from community members. Another two respondents explained that different regions have different cultural contexts and different pronunciations of Yucatec Maya, which means the conversational data from one region may not be appropriate to train a translation system for another region.

Another fourteen respondents expressed no concern about using these types of training data to develop Yucatec Maya-to-Spanish machine translation systems, explaining their view that more training data from any source could lead to better translations. One respondent did contextualize this by saying there is no reason for concern so long as both the originators of the data and the users of the system have clear insight into how the data was used.

Eight participants expressed concern about literary texts being used as training data. The main reason for this was that certain genres of literary texts may not be representative of everyday language use.

Two respondents listed other text categories that would be concerning if used as training data: Payalchi'ob texts (healing ritual prayers and agrarian ritual prayers) and vulgar comedy.

4.2.5 Data Ownership

Regarding the case where academic institutions collect linguistic data from Mayan communities and take ownership of the data, forty-one survey participants expressed disapproval because they view the data as a product of the Mayan community, which therefore belongs to the Mayan community. Four participants specifically described this scenario as epistemic extractivism, a pattern whereby dominant institutions appropriate Indigenous knowledge while disregarding the rights, recognition, and authority of the communities from which that knowledge originates. One participant further emphasized that epistemic extractivism is a way of perpetuating the colonial violence that has long affected Indigenous communities.

Eighteen participants indicated that they would approve of academic institutions owning data collected from Mayan communities if and only if certain conditions were met, such as rendering authorship credits and benefits to the community members involved, returning the results of the research and the data files to the community in an accessible manner, and the non-profit nature of the institution's project.

Twelve participants indicated that they believe it is a beneficial thing for academic institutions to own the linguistic data they collect from the Mayan community because of the ways it could aid in language preservation, language revitalization, and perpetuating Mayan knowledge and culture.

When asked the related question of who should own and control the usage of the data collected in Mayan communities, twenty-six participants answered that the Mayan community collectively ought to own and control the uses of the data. Other participants answered similarly but more specifically, with one stating that only highly qualified people from the Mayan community should own and control the data, and two stating that only the individuals from whom the data was collected should own and control it.

One respondent replied that an association created for the language community ought to control the data, and another responded that a legal institution should own the data. Only one participant, notably an L2+ speaker, answered that the government should own the data. This participant did not specify whether that meant the local or national government. Four participants described that the ownership rights should be shared between

the collectors of the data and the community members from whom the data was collected. Five stated that only the researchers who collected the data should own it, as long as they provide compensation to the participating community members.

4.2.6 Technological Challenges Faced by the Yucatec Maya Community

The most commonly reported theme among technological challenges faced by the Mayan community was a lack of access. This included a lack of access to devices, limited internet access in certain regions, and the lack of digitalization of the language. Respondents reported a lack of adapted keyboards for Yucatec Maya, a lack of user interfaces in Yucatec Maya in operating systems, and a lack of effective machine translation systems. Other respondents clarified that one of the reasons these tools do not yet exist is that the Mayan language varies by region and has non-standardized orthography. Additionally, many Yucatec Maya speakers are illiterate – many are bilingual and literate only in Spanish, and some are monolingual and cannot read or write – which poses a barrier to engaging with many technological devices that rely on written language. Another subset of respondents explained that most Yucatec Maya speakers are of older generations and do not have knowledge of how to use technology, and do not have an incentive to learn.

Another commonly reported theme was the lack of respect for the Yucatec Maya community in broader society as a both a cause and effect of the lack of access that Yucatec Maya speakers have to technology. For example, monolingual Yucatec Maya speakers are less likely to get jobs than speakers of more dominant languages are, so many young people are focusing on learning dominant languages, leading to the endangerment of the language and fewer technologically savvy young people motivated to create technology for the Yucatec Maya language community. Relatedly, another respondent reported that laws protecting Indigenous languages are not enforced.

4.2.7 Summary

In general, survey respondents support the development of machine translation for the Yucatec Maya language because of its potential to revitalize and preserve the language,

largely through language learning. The likelihood of inaccurate translations due to Yucatec Maya's interpretive nature was the most common reason for opposing the development of Yucatec Maya machine translation among those who oppose. Of the possible types of training data, the highest number of participants expressed concern about the use of texts translated from Spanish into Yucatec Maya as training data for Yucatec Maya-to-Spanish translation systems due to the lack of direction translation for many Yucatec Maya words into Spanish.

Regarding broader technological challenges, survey respondents generally reported limited access to technology in the Yucatec Maya language. Regarding data ownership, survey respondents generally agreed that the Mayan language community should own, control, and benefit from the use of any data collected from community members.

Chapter 5

EXPERIMENTAL METHODOLOGY

This chapter describes the methodology used to conduct the fine-tuning experiments central to this thesis. I begin by describing the models selected for fine-tuning: TowerInstruct-7B-v0.2 and T5S. I then describe the Bible and non-Bible parallel corpora used as training data and the preprocessing steps applied to both, followed by the experimental design, including the seven training conditions manipulating the ratio of Bible to non-Bible data. I describe the fine-tuning procedure used for each model and the evaluation metrics used to assess both translation quality and the degree of semantic drift toward Bible-related content in model outputs.

5.1 Model Selection

TowerInstruct-7B-v.02 The first model I fine-tuned was TowerInstruct-7B-v.02 (Alves et al., 2024), a state-of-the-art open language model specifically designed for translation-related tasks. TowerInstruct was developed by fine-tuning TowerBase—a model created by continuing the pretraining of LLaMA-2 on multilingual monolingual and parallel data across 10 languages (English, German, French, Dutch, Italian, Spanish, Portuguese, Korean, Russian, and Chinese) using translation-specific instructions from the TowerBlocks dataset. This 7B parameter version remains competitive with other open systems and achieves GPT-3.5-turbo translation quality for some language pairs, outperforming ALMA-R 7B, the only system of comparable size. I chose this model because of its state-of-the-art performance, its smaller size relative to other LLMs, its translation-specific training, its exposure to Spanish, and its lack of exposure to Yucatec Maya.

T5S To assess whether the patterns observed with TowerInstruct-7B-v.02 were model-specific or generalizable across architectures, I replicated the seven experimental conditions

using T5S (Raffel et al., 2020), the encoder-decoder model used by Rangel and Kobayashi (2024) in their prior work on Yucatec Maya-to-Spanish machine translation. Specifically, I fine-tuned the `vgaraujov/t5-base-spanish` checkpoint, a T5-base model that has been further pretrained on Spanish text, making it better suited to Spanish-output translation tasks than the original multilingual T5. T5S is substantially smaller than TowerInstruct-7B-v.02, with approximately 250 million parameters, and uses an encoder-decoder sequence-to-sequence architecture rather than the decoder-only architecture of TowerInstruct. Unlike TowerInstruct, T5S is not instruction-tuned and does not use a prompt template; instead, input sentences are passed directly to the encoder with a task prefix. I chose T5S for this replication because its use in prior work on this language pair provides a point of architectural comparison, and because its smaller size allowed fine-tuning under the same hardware constraints without LoRA adaptation.

5.2 Data

5.2.1 Bible Corpus

The Bible data used in the experiments described below is comprised of a parallel corpus of the Bible in Yucatec Maya and Spanish, including complete books from both the Old and New Testaments. The specific version of the Yucatec Maya Bible used is the New World Translation[1] due to its widespread use in the community, as recommended by Silvia Fernandez Sabido and César Can who are linguists in the Yucatec Maya language community. This translation includes the 66 books of the Protestant canon and does not include the Deuterocanonical books. The Bible in Spanish has numerous available translations, but for the sake of consistency and translation quality parallel with the Yucatec Maya Bible, the New World Translation was also used.[2] The parallel Bible corpus is aligned by verse, resulting in 31,100 pairs. A Python script was used to webscrape these sites and create a .txt file with one Bible verse per line.

---

[1] `https://www.jw.org/yua/publicacionoob/biblia/nwt/libroob/`

[2] `https://www.jw.org/es/biblioteca/biblia/biblia-estudio/libros/`

5.2.2 Non-Bible Corpus

The non-Bible data comes from two sources, the T'aantsil corpus[3] provided by César Can and the UNAM corpus provided by Silvia Fernandez Sabido of the Yucatec Maya language community. The T'aantstil corpus contains 15,600 Yucatec Maya and Spanish sentence pairs from everyday conversations in Yucatec Maya within the community, transcribed into text and translated to Spanish. The UNAM corpus was extracted from the corpus management system described by Sierra Martínez et al. (2017) in 2021-2022 and contains 2,972 sentence pairs from narratives, stories, administrative and instructional texts in Yucatec Maya and Spanish. The community's data privacy protocols, built to maintain data sovereignty, prohibit open posting of these corpora. Scholars interested in reproducing the results in this paper would need to form a relationship with the community in order to gain access.

5.2.3 Preprocessing and Splits

For each experimental condition, I used a fixed 80–10–10 split on the non-Bible data, reserving 2,000 sentence pairs each for validation and test sets. These sets were held constant across all conditions for both models. The validation set was excluded and not used during training. The test set was reserved for final evaluation after training was complete. Bible data was excluded from both the validation and test sets. After splitting, all data was preprocessed uniformly. Preprocessing steps included NFC Unicode normalization, lowercasing, and normalization of whitespace to a single space between tokens. No additional segmentation was applied, as each line corresponds to a single verse or sentence pair.

Each training set for experimental conditions A-E consisted of 14,634 sentence pairs, with varying Bible-to–non-Bible ratios depending on the experimental condition. In experimental condition F, all available training data was used from both the non-Bible training dataset and the Bible dataset, resulting in 45,712 total pairs. In experimental condition G, all available Bible data was used, resulting in 31,078 total pairs. These conditions explore the question of how the quantity of training data impacts translation quality and the frequency of semantic drift towards Bible-related content.

---

[3] `https://taantsil.com.mx/info`

For TowerInstruct-7B-v.02, tokenization was performed using the model's fast LLaMA-style byte-level BPE tokenizer, implemented via the Hugging Face tokenizers library. For T5S, tokenization was performed using the SentencePiece tokenizer associated with the `vgaraujov/t5-base-spanish` checkpoint, also accessed via the Hugging Face tokenizers library.

5.3 Experimental Design

5.3.1 Training Conditions

Condition	Bible Pairs	Non-Bible Pairs	Total Pairs
A (0%)	0	14,634	14,634
B (25%)	3,658	10,976	14,634
C (50%)	7,317	7,317	14,634
D (75%)	10,976	3,658	14,634
E (100%)	14,634	0	14,634

Table 5.1: Training data composition for each experimental condition. All conditions manipulating the proportion of Bible vs non-Bible training data use 14,634 sentence pairs and share the same validation and test sets.

Condition	Bible Pairs	Non-Bible Pairs	Total Pairs
F	31,078	14,634	45,712
G	31,078	0	31,078

Table 5.2: Condition F uses all available training data from both the Bible dataset and the non-Bible dataset. Condition G uses all available Bible data.

For each training condition, I randomly sampled sentence pairs at the individual pair level to construct training sets of the target Bible-to-non-Bible ratios described above. Sampling was performed independently for the Bible and non-Bible corpora using fixed random

seed 73 to ensure reproducibility. No stratification was applied. After sampling, sentence pairs were shuffled within each training set.

5.3.2 Fine-Tuning Procedure

TowerInstruct-7B-v.02

I performed the fine-tuning procedure for each experiment using the Hugging Face Transformers library together with the PEFT framework for parameter-efficient adaptation via LoRA (Hu et al., 2022). I selected LoRA instead of full-model fine-tuning in order to limit overfitting and catastrophic forgetting when adapting a large instruction-tuned model to low-resource parallel data, while also making training feasible under GPU memory constraints.

Following Hu et al. (2022), I applied low-rank adapters with rank 8 and scaling factor 16 to the query and value projection matrices of the attention layers. These components have been shown to provide high leverage for task adaptation while preserving the pretrained model's linguistic representations. I used a modest LoRA dropout of 0.1 to further reduce overfitting on small training sets.

I fine-tuned all TowerInstruct-based experimental models for a fixed budget of 5,000 training steps with an effective batch size of 128, achieved via gradient accumulation. I chose a fixed training schedule without early stopping to ensure comparability across experimental conditions, allowing differences in model behavior to be attributed to data composition rather than training duration. I enabled mixed-precision training and gradient checkpointing to reduce memory usage during fine-tuning. I saved intermediate checkpoints every 500 training steps, with storage limited to the two most recent checkpoints, in order to allow training to be resumed in the event of job interruption on the cluster. I conducted the training on the University of Washington's Hyak compute cluster.

T5S

I performed the fine-tuning of the T5S model using the Hugging Face Transformers library with the `Seq2SeqTrainer` class. Unlike TowerInstruct-7B-v.02, I fine-tuned T5S without

LoRA adaptation, as its smaller parameter count made full-parameter fine-tuning feasible under the available GPU memory constraints. Rather than using a prompt template, I prefixed each source sentence with the task instruction `translate maya to spanish:` following the standard T5 input format.

Unlike the fixed-step training schedule I used for TowerInstruct-7B-v.02, I trained T5S for up to 200 epochs with early stopping applied at a patience of 2 epochs, halting training when validation loss failed to improve for two consecutive epochs. The best-performing checkpoint, defined as the one with the lowest validation loss, was saved and used for evaluation. I used a learning rate of 2e-5 with a weight decay of 0.01 and a per-device batch size of 32. I enabled mixed-precision training via fp16. These hyperparameters were selected to replicate the training procedure of Rangel and Kobayashi (2024), who used the same base model for Spanish-to-Yucatec Maya translation. As with TowerInstruct-7B-v.02, I conducted the training procedure on the University of Washington's Hyak compute cluster.

5.3.3 Evaluation Metrics

Translation Quality Evaluation

To enable direct comparison with the results reported by Domingues et al. (2024), translation quality was evaluated using BLEU and chrF, despite well-known limitations of BLEU as a metric. Evaluation was extended by including COMET, a neural-based metric that uses multilingual sentence embeddings to assess semantic adequacy (Rei et al., 2020), and METEOR, which accounts for synonymy, stemming, and word order variation (Banerjee and Lavie, 2005). BLEU and chrF were computed using the sacreBLEU implementation (Post, 2018) via Hugging Face's *evaluate* library, while METEOR was computed using the corresponding *evaluate* module. COMET scores were computed using the *unbabel-comet* toolkit with the *wmt22-comet-da* model checkpoint. All metrics were computed once over the full test set for each experimental condition to ensure reproducibility and comparability.

Bible-Related Content Detection

N-Gram Distribution Analysis To identify stylistic overfitting to Bible phrasing, I analyzed n-gram distributions in the Bible and non-Bible training corpora. I extracted all unigrams, bigrams, and trigrams and computed their relative frequencies across both corpora. Using log-likelihood ratio (LLR) statistics, I identified Bible-distinctive n-grams – sequences that were significantly overrepresented in the Bible corpus (e.g., *en el nombre de*, *he aquí*, *por siempre jamás*). These were compiled into a list of Bible-associated n-grams. Then, Silvia Fernandez Sabido and César Can, who speak both Yucatec Maya and Spanish, reviewed the list and identified any n-grams that were likely to occur in everyday Spanish conversations so that I could remove them from the list. With their feedback, I removed 189 unigrams, 86 bigrams, 234 trigrams, and 214 4-grams, leaving 300 unigrams, 413 bigrams, 265 trigrams, and 285 4-grams. For each test output and its reference translation, I counted how many Bible-associated n-grams occurred. Outputs that contained a significantly higher number than their reference (e.g., z-score > 2) were flagged as stylistically contaminated.

Semantic Similarity Analysis To detect subtle semantic drift toward religious content, I compared each test output's similarity to Bible content relative to its gold-standard reference. Prior to embedding, all texts – including the YM model outputs, gold-standard Spanish references, and Bible corpus – had been lowercased and Unicode-normalized (NFC) as part of the shared preprocessing pipeline. Sentence embeddings were computed using the BETO Spanish BERT model with whole-word masking (cased), specifically `dccuchile/bert-base-spanish-wwm-cased`[4] accessed via the Hugging Face Transformers library. Although all input texts were lowercase at the point of embedding, the cased variant was selected over its uncased counterpart because cased BETO has demonstrated stronger performance on Spanish semantic similarity and downstream NLP tasks (Cañete et al., 2020); crucially, since lowercasing was applied uniformly across all compared texts, the absence of case variation does not introduce any asymmetry into the similarity scores. BETO does not provide an explicit version number; therefore, the experiments use the latest available

[4] `https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased`

model revision at the time of evaluation. For each test output and its reference translation, I computed their cosine similarity against a set of 2,000 sentence embeddings randomly sampled from the Spanish Bible corpus. I retained the maximum similarity to any Bible sentence as the similarity score for each. An output was flagged as Bible-contaminated if its cosine similarity to the Bible set exceeded that of its reference by a margin greater than 0.2 (on a scale of 0 to 1). This comparison ensures that only outputs whose content has shifted toward religious domains more than their reference counterparts are identified as contaminated.

5.4 Summary

This chapter described the methodology underlying the fine-tuning experiments reported in this thesis. The first model selected for fine-tuning was TowerInstruct-7B-v0.2, a 7-billion parameter open language model designed for translation tasks, chosen for its state-of-the-art performance, translation-specific training, exposure to Spanish, and lack of prior exposure to Yucatec Maya. The seven experimental conditions were subsequently replicated using T5S, an encoder-decoder sequence-to-sequence model fine-tuned without LoRA adaptation, to assess whether the observed patterns were model-specific or generalizable across architectures. The same training data, experimental conditions, evaluation metrics, and test set were used for both models. Training data consisted of two parallel corpora: a Bible corpus of 31,078 Yucatec Maya–Spanish verse-aligned pairs and a non-Bible corpus of 14,634 sentence pairs drawn from everyday conversation and narrative texts. All data was preprocessed uniformly via NFC Unicode normalization, lowercasing, and whitespace normalization. Seven experimental conditions were constructed by varying the ratio of Bible to non-Bible training data, ranging from 0% to 100% Bible, with two additional conditions using all available data from one or both corpora. Fine-tuning was performed using LoRA applied to the attention layers of TowerInstruct, with a fixed training budget of 5,000 steps to ensure comparability across conditions. Translation quality was assessed using BLEU, chrF, COMET, and METEOR. Bible-related content contamination in model outputs was assessed through two complementary methods: an n-gram distribution analysis using log-likelihood ratio statistics to detect stylistic overfitting to Bible phrasing, and a semantic similarity analysis using

BETO sentence embeddings to detect subtler semantic drift toward religious content.

Chapter 6

EXPERIMENTAL RESULTS

This chapter presents the results of the fine-tuning experiments described in the previous chapter. I report translation quality scores across all seven experimental conditions, followed by the results of the semantic drift analysis. Validation checks are included to contextualize the metric scores. Overall, for the TowerInstruct-7B-v.02 model, translation quality degrades as the proportion of Bible training data increases. To assess whether this pattern was model-specific or generalizable across architectures, I replicated the seven experimental conditions using T5S, an encoder-decoder model used in prior work on this language pair. These experiments also showed that increasing proportions of Bible data generally harm translation quality, but the relationship is not strictly monotonic. Semantic drift toward Bible-related content remains negligible across all conditions for both models.

6.1 Translation Quality

Table 6.1 summarizes translation quality across experimental conditions A-E with varying Bible-to–non-Bible training data ratios, evaluated using BLEU, chrF, METEOR, and COMET.

For the TowerInstruct-7B-v.02 model, translation quality decreases monotonically as the proportion of Bible training data increases. Condition A (0% Bible training data) achieves the highest scores across all four metrics, while Condition E (100% Bible training data) performs substantially worse than all other conditions. Table 6.1 also summarizes translation quality across experimental conditions F and G with varying amounts of training data. Experiment F included all available training data and contained approximately 68% Bible data out of 45,712 total training pairs. Experiment G was fine-tuned on all available 31,078 pairs of Bible data. Experiments F and G show that increased quantities of training data do not result in additional translation quality gains.

For the T5S model, translation quality generally decreases as the proportion of Bible training data increases. Condition A (0% Bible training data) achieves the highest METEOR score. Condition B (25% Bible training data) achieves the highest BLEU score, which is marginally better than Condition A's BLEU score. Condition A and Condition B achieve the same scores for chrF and COMET. Condition E (100% Bible training data) performs substantially worse than the other 4 conditions with the fixed amount of training data across all metrics. Condition F for the T5S model represents a near-total failure: every output in the 2,000-sentence test set consisted solely of the word *la*, resulting in a BLEU score of 0.0 and a chrF score of 2.554. Condition G from the T5S model show similar results to those from the TowerInstruct-7B-v.02 model: increased quantities of training data do not result in additional translation quality gains.

Base Model	Experiment	BLEU	chrF	METEOR	COMET
TowerInstruct-7B-v.02	A (0%)	3.49	22.56	0.206	0.543
TowerInstruct-7B-v.02	B (25%)	2.97	20.18	0.174	0.510
TowerInstruct-7B-v.02	C (50%)	2.42	19.19	0.165	0.511
TowerInstruct-7B-v.02	D (75%)	2.14	18.71	0.157	0.489
TowerInstruct-7B-v.02	E (100%)	0.56	14.40	0.109	0.418
TowerInstruct-7B-v.02	F (All Combined)	2.05	18.35	0.163	0.478
TowerInstruct-7B-v.02	G (All Bible)	0.405	13.89	0.101	0.400
TowerInstruct-7B-v.02	baseline	0.47	16.61	0.106	0.466
T5S	A (0%)	13.00	35.39	0.394	0.653
T5S	B (25%)	13.84	35.39	0.390	0.653
T5S	C (50%)	12.48	33.12	0.366	0.639
T5S	D (75%)	9.189	28.97	0.331	0.609
T5S	E (100%)	1.233	16.33	0.145	0.456
T5S	F (All Combined)	0.0	2.554	0.00985	0.433
T5S	G (All Bible)	1.783	18.82	0.184	0.479
T5S	baseline	0.878	13.37	0.102	0.409

Table 6.1: Translation quality across ordered experimental conditions with varying Bible-to–non-Bible training data ratios (A–E). Condition F uses all available training data from both the Bible dataset and the non-Bible dataset and is therefore approximately 68% Bible data. Condition G uses all available Bible data and is therefore 100% Bible data. Higher scores indicate better performance. *baseline* refers to the respective unfine-tuned model performing the Yucatec Maya-to-Spanish translation task with the same testing set that was used for experimental conditions A-G. The T5S and the TowerInstruct model baselines perform similarly, further highlighting the significance of the performance gains achieved by the T5S model after fine-tuning.

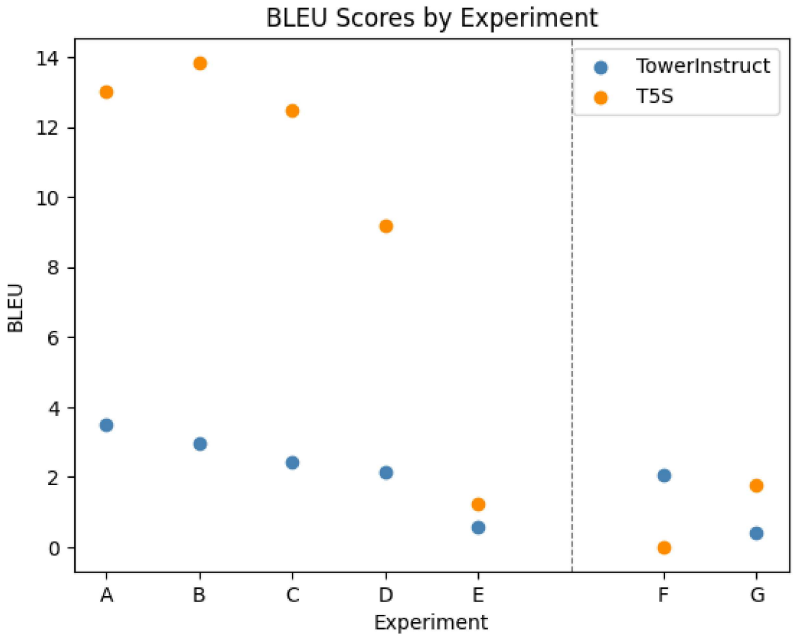

Figure 6.1: BLEU Scores Across Experimental Conditions for TowerInstruct and T5S

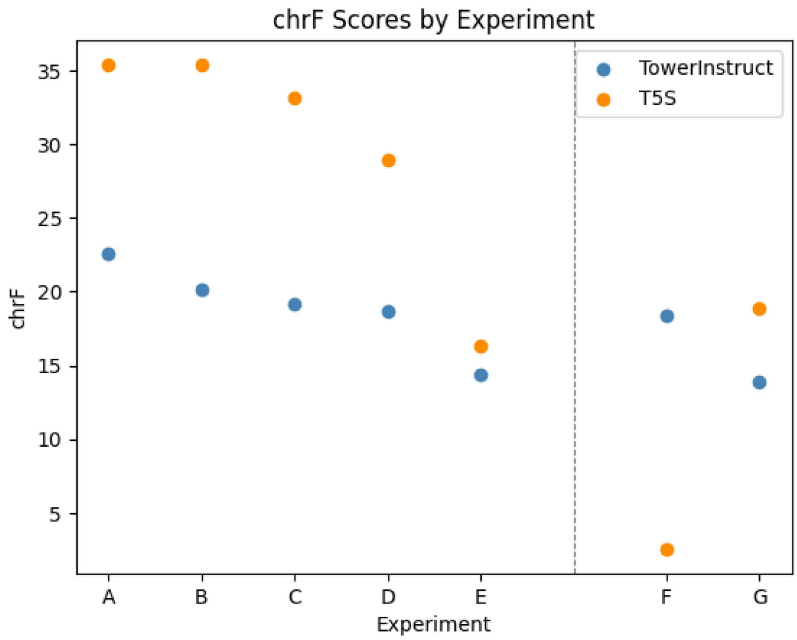

Figure 6.2: chrF Scores Across Experimental Conditions for TowerInstruct and T5S

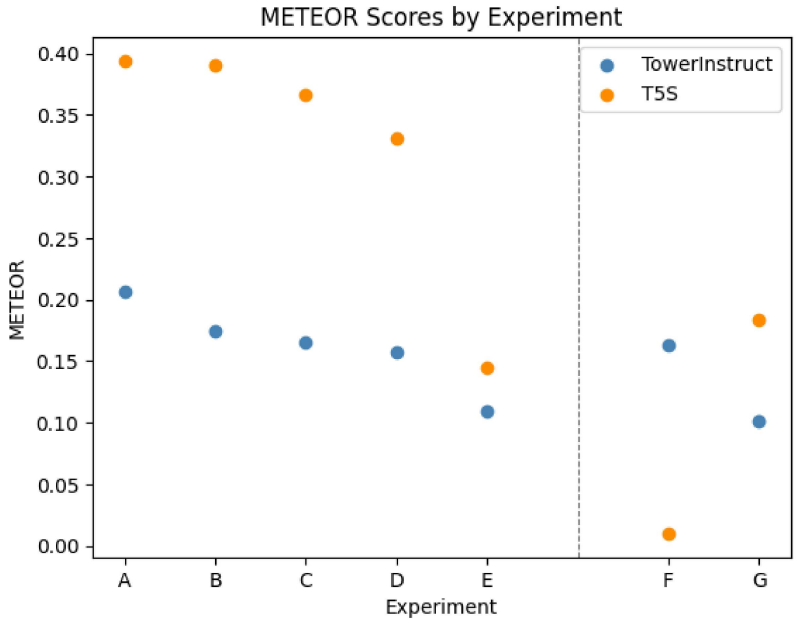

Figure 6.3: METEOR Scores Across Experimental Conditions for TowerInstruct and T5S

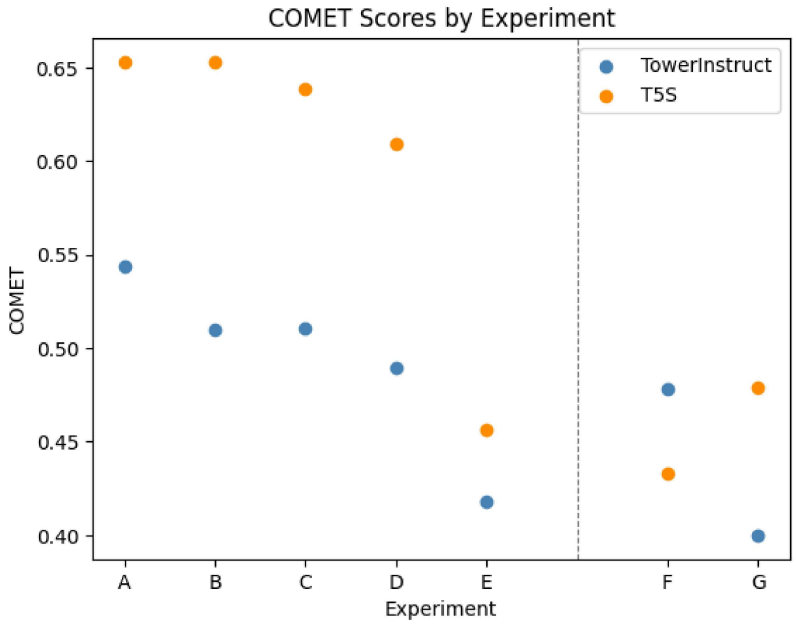

Figure 6.4: COMET Scores Across Experimental Conditions for TowerInstruct and T5S

6.1.1 Validation Checks

To obtain a better understanding of the meaning and quality of these scores, I performed several validation checks against the 2000 sentences of the gold standard Spanish translation from the test set. To establish the maximum values for each of the evaluation metrics, I compared the gold standard Spanish translation against itself. To establish minimum values for each of the evaluation metrics when there is a completely different script used in the machine-produced translation vs the gold standard Spanish translation, I compared a set of Japanese sentences against the gold standard Spanish translation. To establish minimum values for each of the metrics when the script and language are the same, but the content is misaligned, I compared a random reordering of sentences from the gold standard Spanish translation against the original ordering of the gold standard. To establish scores in the case that the model outputs the input without translating it at all, I compared the Yucatec Maya source text from the testing set to the target Spanish translation. As Table 6.2 indicates, the range of possible BLEU scores was 0.00 - 100.00, the range of possible chrF scores was 0.05 - 100.00, the range of possible METEOR scores was 0.000 - 0.989, and the range of possible COMET scores was 0.391 - 0.951.

Another variable that could have affected the performance of the model was the output language. The baseline TowerInstruct model was trained to produce an output language of English. I compared the performance of the baseline model on a French-to-English translation task and a French-to-Spanish translation task. The baseline model's training data included French source text. As shown in Table 6.2, the baseline model performs worse across all metrics on the French-to-Spanish task, indicating that, regardless of the fine-tuning process on the experimental conditions with different proportions and quantities of Bible training data, translation quality is substantially affected by whether the target language aligns with the model's original training.

Model	Experiment	BLEU	chrF	METEOR	COMET
	gold2gold	100.00	100.00	0.989	0.951
	jpn2gold	0.00	0.05	0.000	0.391
	rand2gold	0.18	11.69	0.054	0.411
	src2targ	1.47	14.94	0.115	0.410
TowerInstruct-7B-v.02	fr–en	31.03	59.42	0.590	0.791
TowerInstruct-7B-v.02	fr–es	25.45	57.86	0.522	0.779
TowerInstruct-7B-v.02	A (0%)	3.49	22.56	0.206	0.543
TowerInstruct-7B-v.02	B (25%)	2.97	20.18	0.174	0.510
TowerInstruct-7B-v.02	C (50%)	2.42	19.19	0.165	0.511
TowerInstruct-7B-v.02	D (75%)	2.14	18.71	0.157	0.489
TowerInstruct-7B-v.02	E (100%)	0.56	14.40	0.109	0.418
TowerInstruct-7B-v.02	F (All Combined)	2.05	18.35	0.163	0.478
TowerInstruct-7B-v.02	G (All Bible)	0.405	13.89	0.101	0.400
TowerInstruct-7B-v.02	baseline	0.47	16.61	0.106	0.466
T5S	A (0%)	13.00	35.39	0.394	0.653
T5S	B (25%)	13.84	35.39	0.390	0.653
T5S	C (50%)	12.48	33.12	0.366	0.639
T5S	D (75%)	9.189	28.97	0.331	0.609
T5S	E (100%)	1.233	16.33	0.145	0.456
T5S	F (All Combined)	0.0	2.554	0.00985	0.433
T5S	G (All Bible)	1.783	18.82	0.184	0.479
T5S	baseline	0.878	13.37	0.102	0.409

Table 6.2: *gold2gold* refers to the gold standard Spanish translation being evaluated against itself. *jpn2gold* refers to a set of Japanese sentences evaluated against the gold standard Spanish translation. *rand2gold* refers to a random reordering of sentences from the gold standard Spanish translation being compared to the original ordering of sentences from the gold standard Spanish translation. *src2targ* refers to the testing set in Yucatec Maya being compared against the gold standard Spanish translation. *fr–en* and *fr–es* refer to the French pivot experiments.

To illustrate that low metric scores do not uniformly indicate unintelligible output, Table 6.3 presents selected examples of translations produced under each experimental condition. I manually selected examples from each condition by inspecting model outputs and choosing representative cases where the output was intelligible and bore a plausible semantic relationship to the source.

Table 6.3: Selected examples of intelligible model outputs across experimental conditions.

Condition	**Model**	**Output**	**Reference**
A	TowerInstruct-7B-v.02	*hacen el camino para que puedan pasar,*	*y andaba pidiendo dónde quedarse,*
A	T5S	*pasa a buscarlo donde está,*	*y andaba pidiendo dónde quedarse,*
B	TowerInstruct-7B-v.02	*no me gustó el cambio de vestido y la ropa de juan así que me quedé.*	*no es mi suerte quedarme con josé así y lo dejé.*
B	T5S	*no era mi suerte quedarme con el josé y lo deshilé.*	*no es mi suerte quedarme con josé así y lo dejé.*
C	TowerInstruct-7B-v.02	*sólo el conocimiento y la sabiduría de las cosas,*	*es muy feo si no sabes nada le digo,*
C	T5S	*es muy complicado o no sabes nada le digo,*	*es muy feo si no sabes nada le digo,*
D	TowerInstruct-7B-v.02	*te quemaste, y si no es por el fuego, si no es por el calor, si no es por el sol,*	*en xcail, calentura o lo que sea, debes ir a comprar la medicina,*

Condition	Model	Output	Reference
D	T5S	*en la milpa, si no te gusta algo, vas a ir a dar de comer,*	*en xcail, calentura o lo que sea, debes ir a comprar la medicina,*
E	TowerInstruct-7B-v.02	*la gente se olvidó de la ley de sus antepasados, y sus hijos no se acuerdan de ella.*	*yo entonces puse a la mano todo cuanto consideré que era necesario para el alumbramiento.*
E	T5S	*después de eso, me levanté y les dije a mis hermanos lo que iba a suceder.*	*yo entonces puse a la mano todo cuanto consideré que era necesario para el alumbramiento.*
F	TowerInstruct-7B-v.02	*yo lo aprendí.*	*y pues huyeron de ahí.*
F	T5S	*la*	*y pues huyeron de ahí.*
G	TowerInstruct-7B-v.02	*“ahora bien”, dijo el señor. “todo está listo”.*	*ha si, aqui en estás zonas es así,*
G	T5S	*¡amén! ¡qué rico es el pan!.*	*ha si, aqui en estás zonas es así,*

6.1.2 Summary of Translation Quality Results

Across both models, increasing the proportion of Bible-sourced training data is associated with decreased translation quality. For the TowerInstruct-7B-v.02 model, this relationship

is strictly monotonic across Conditions A–E: every increase in Bible data proportion corresponds to a decrease across all four evaluation metrics. For the T5S model, the relationship is broadly similar but not strictly monotonic: Condition B achieves a marginally higher BLEU score than Condition A, and Condition A and Condition B are tied on chrF and COMET, suggesting that small amounts of Bible data do not substantially harm T5S performance. In both models, Condition E (100% Bible data) represents a clear performance floor among the fixed-dataset-size conditions. The T5S model substantially outperforms TowerInstruct-7B-v.02 across all comparable conditions, with BLEU scores roughly three to four times higher under equivalent training data compositions. Both models show that increasing the total quantity of training data does not compensate for poor data composition: Conditions F and G perform worse than or comparably to Condition A despite containing more training pairs. Condition F represents a particularly severe failure for the T5S model, producing only the word *la* for every test sentence. The selected output examples in Table 6.3 confirm that intelligible translations are produced under the better-performing conditions despite the low absolute metric scores. The metric scores should be interpreted in light of the validation checks in Table 6.2.

6.2 Semantic Drift

This section presents the results of two complementary analyses designed to detect whether fine-tuning on Bible data caused model outputs to drift toward biblical Spanish register. The first analysis uses BETO embeddings to measure semantic similarity between model outputs and a Spanish Bible corpus; the second examines the presence of Bible-associated n-grams in model outputs. Both analyses are applied to both the TowerInstruct model and the T5S model across all seven experimental conditions, ranging from 0% to 100% Bible-sourced training data.

6.2.1 Semantic Similarity Measured with BETO

To evaluate whether model outputs exhibited semantic drift toward Bible-related content, I measured the relative semantic similarity of each model output to a Spanish Bible corpus, compared against its corresponding gold-standard reference translation. For each output,

the maximum cosine similarity to any of 2,000 sampled Bible sentence embeddings was computed and compared to the reference's similarity to the same set. Outputs were flagged as Bible-influenced if this relative difference exceeded a threshold of 0.2 cosine similarity on a scale of 0 to 1. The results displayed in Figure 6.5 and Figure 6.6 show that the number of semantically Bible-related outputs remains relatively stable and extremely small across experimental conditions. Even though Condition E on the T5S model has the highest percentage of flagged outputs of all conditions across all models, it is still extremely small: 0.15%, representing only 3 flagged outputs out of 2000 outputs.

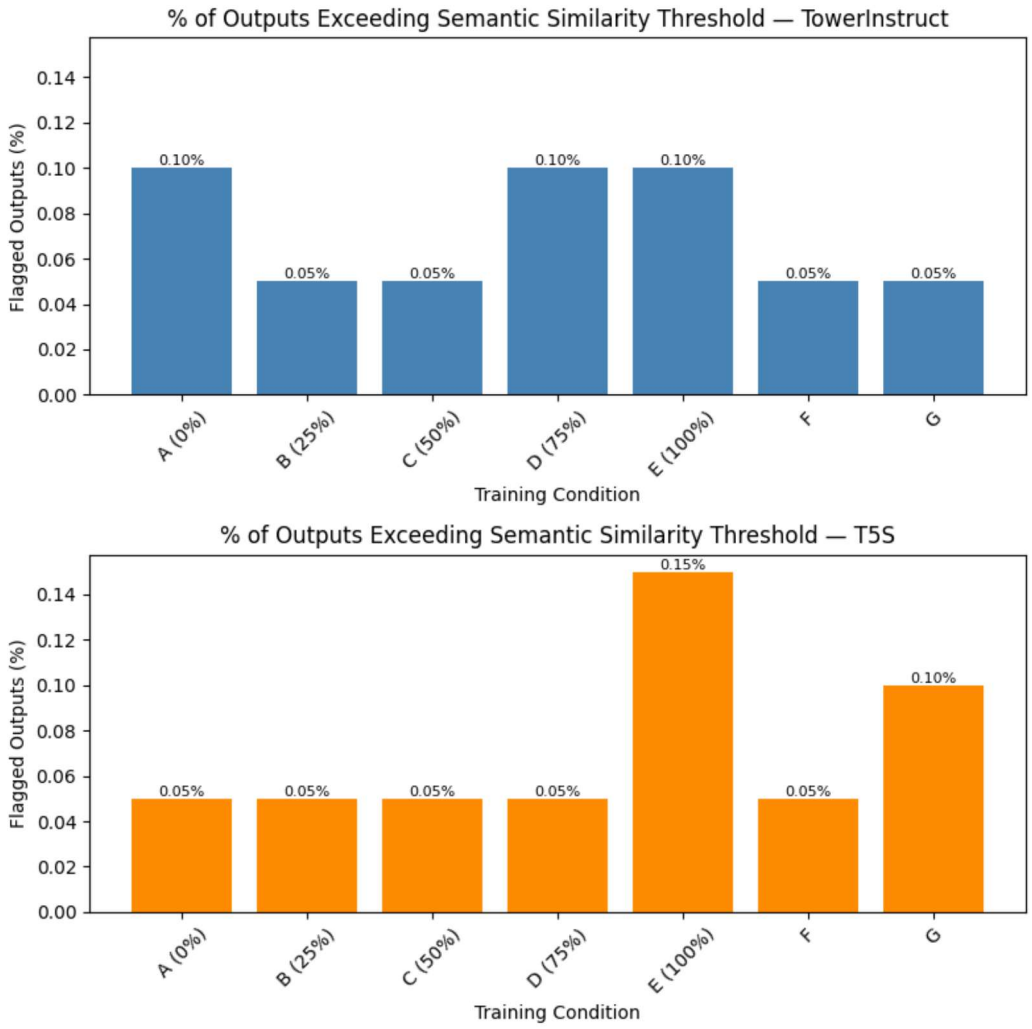

Figure 6.5: Proportion of test outputs flagged as semantically Bible-influenced by training condition. An output was flagged if its semantic similarity to a Spanish Bible sentence exceeded that of its gold-standard reference by more than 0.2 measured by cosine similarity on a scale of 0 to 1. Flagged outputs are rare across all conditions, indicating that strong semantic drift toward Bible content is uncommon even when substantial amounts of Bible data are included in training.

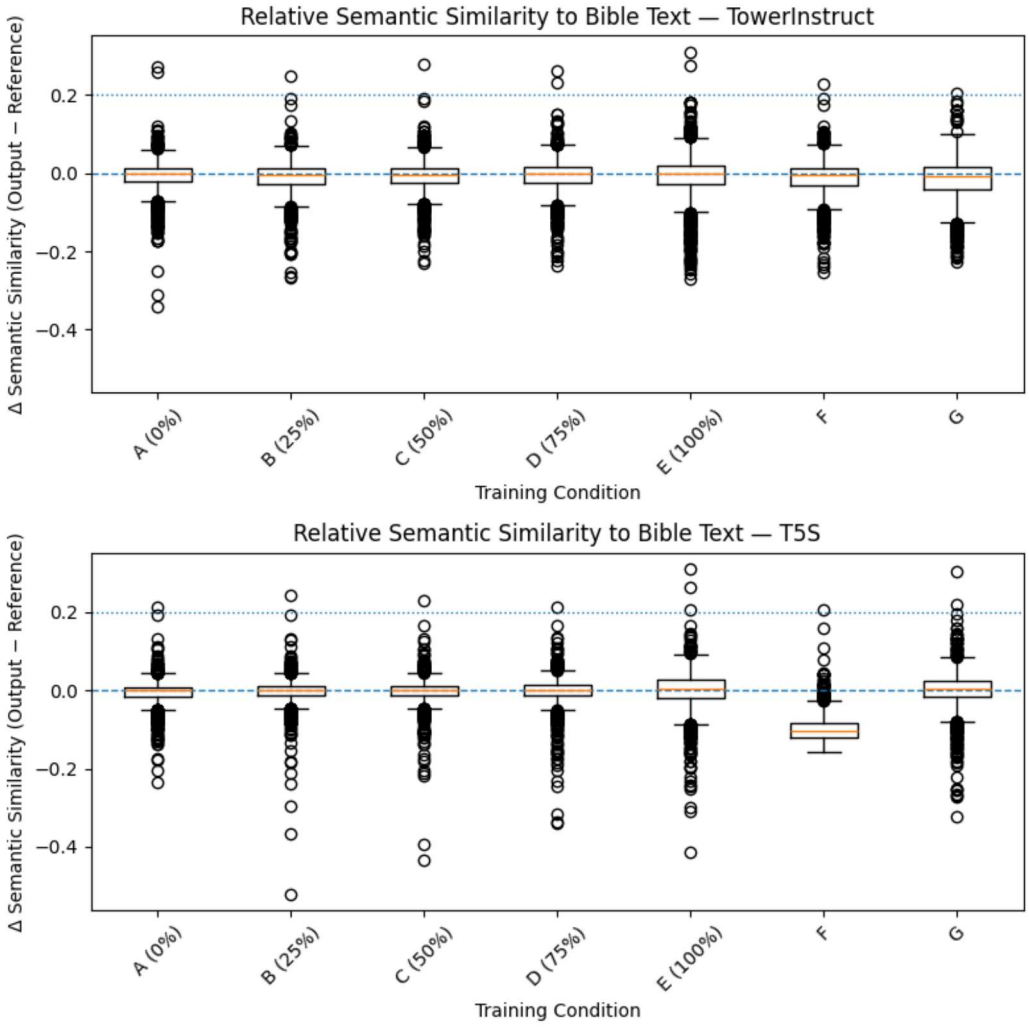

Figure 6.6: Distribution of per-output semantic similarity deltas (model output minus gold-standard reference) across experimental conditions. Each box shows the interquartile range, with the median marked by a horizontal line and whiskers extending to 1.5 times the interquartile range. Outliers are plotted individually. The distributions are tightly clustered around zero across all conditions, confirming that outputs are not systematically more similar to Bible content than their gold-standard references regardless of the proportion of Bible data used in training.

Flagged outputs for both models are shown in Table 6.4 below. All but one flagged outputs come from one or two of the same lines of the source file. The output written in red is the only output that could be considered genuine semantic drift towards Bible-related content.

Condition	**Model**	**Source**	**Output**	**Reference**
A	TowerInstruct	*jáaj*	*no,*	*que*
A	TowerInstruct	*u tsolts'fibil xook*	*te lo enseñan en la escuela.*	*índice*
A	T5S	*u tsolts'fibil xook*	*la lengua maya*	*índice*
B	TowerInstruct	*u tsolts'fibil xook*	*u tsolts'fibil xook*	*índice*
B	T5S	*u tsolts'fibil xook*	*el estudio de la lengua maya*	*índice*
C	TowerInstruct	*jáaj*	*se llama así.*	*que*
C	T5S	*u tsolts'fibil xook*	*la lectura de los libros*	*índice*
D	TowerInstruct	*jáaj*	*sí.*	*que*
D	TowerInstruct	*u tsolts'fibil xook*	*tienes mucha paciencia.*	*índice*
D	T5S	*u tsolts'fibil xook*	*la enseñanza de la lectura*	*índice*
E	TowerInstruct	*jáaj*	*Escribe únicamente la traducción [...] Spanish:*	*que*
E	TowerInstruct	*u tsolts'fibil xook*	*Traduce el siguiente texto [...] repeticiones.*	*índice*
E	T5S	*jáaj*	*já, ja, ja.*	*que*
E	T5S	*Ma'alo'*	*ma'alo.*	*bien*
E	T5S	*u tsolts'fibil xook*	*esto es lo que dice el señor soberano jehová:*	*índice*
F	TowerInstruct	*u tsolts'fibil xook*	*es muy difícil aprender a leer*	*índice*

Condition	Model	Source	Output	Reference
F	T5S	*jáaj*	*la*	*que*
G	TowerInstruct	*u tsolts'fibil xook*	*Traduce el siguiente texto del maya yucateco al español. [...] Maya yucateco: u tsolts'fibil xook "u tsolts'fibil xook yéetel, [...]"*	*índice*
G	T5S	*jáaj*	*¡mira! ¡mira! ¡mira!*	*que*
G	T5S	*u tsolts'fibil xook*	*la vara de la vara de la vara de la vara de la vara.*	*índice*

Table 6.4: Outputs Flagged as Semantically Similar to Bible Data

6.2.2 Bible N-Gram Contamination Analysis

As a complementary check on semantic drift, the evaluation pipeline also measured the presence of Bible-associated n-grams in model outputs using a set of precomputed high-LLR n-gram lists (300 unigrams, 413 bigrams, 265 trigrams, and 285 4-grams) reviewed by community members familiar with Yucatec Maya religious language. For each output sentence, the pipeline counted matches against these lists and computed a z-score relative to a 2,000-sentence Bible reference sample, flagging any output whose count of n-grams from this list exceeded two standard deviations above the reference mean. Because no condition produced any n-gram matches, the standard deviation of the delta distribution was zero across all conditions, precluding z-score differentiation; the flagged count of zero is therefore a direct count rather than a threshold-based result.

6.2.3 Summary of Semantic Drift Findings

Taken together, the BETO semantic similarity analysis and the Bible n-gram contamination analysis provide converging evidence that fine-tuning on Bible-parallel data does not induce measurable stylistic or lexical drift toward biblical Spanish. Across all seven conditions, the number of semantically flagged outputs remained extremely small and stable, never exceeding three per condition out of 2,000 test sentences. The n-gram analysis strengthens this conclusion further: no output sentence in any condition produced lexical patterns statistically associated with biblical register. Inspection of the flagged outputs suggests that the small number of BETO-flagged cases are largely attributable to two recurring source phrases – *jáaj* and *u tsolts'fibil xook* – rather than to systematic Bible-influenced generation. The reference translation of *jáaj* is *que*, and the reference translation of *u tsolts'fibil xook* is *índice*, as shown in Table 6.4. Neither of these reference translations are religious in meaning. The evidence consistently supports the conclusion that the proportion of Bible data in fine-tuning has no detectable effect on semantic drift in model outputs.

6.3 Summary

In general, as the proportion of Bible-sourced training data in the fine-tuning process increases, the output translation quality decreases. Increased quantities of training data do not result in additional translation quality gains. There is no observable change in semantic drift towards Bible-related content across experimental conditions.

Chapter 7

DISCUSSION

This chapter interprets the results presented in the previous chapter, organizing the discussion around the two primary research questions motivating this study: Whether increasing the proportion of Bible training data affects translation quality, and whether Bible-heavy training induces semantic drift toward religious content in model outputs. These findings are examined across both the TowerInstruct-7B-v.02 model and the T5S model, which was fine-tuned on the same seven experimental conditions to assess whether the pattern of Bible-data-induced quality degradation was model-specific or generalizable across architectures. The findings are clear across both models: Bible training data degrades translation quality, and Bible training data does not induce meaningful semantic drift toward religious content in model outputs. Translation quality generally decreases as the proportion of Bible training data increases, with Bible-exclusive training conditions degrading performance below even the unfine-tuned TowerInstruct baseline model. Evidence of semantic drift toward Bible-influenced content, by contrast, remained limited and rare across all conditions for both models; manual inspection of flagged outputs revealed that only one constituted a genuine case of religious content contamination, while the others originated from the same two source phrases.

7.1 Translation Quality

The following subsections discuss the translation quality results in detail. I first examine the effects of Bible data proportion on model performance across experimental conditions, then interpret the evaluation metrics, and finally contextualize the scores via validation checks.

7.1.1 Effects of Bible Data on Translation Quality

These results from the TowerInstruct-7B-v.02 model suggest that increasing reliance on Bible-aligned parallel data introduces domain-specific biases that do not transfer well to tasks in other domains. Although academic linguists and tech industry workers have commonly used Bible translations in the development of language models to address data scarcity in low-resource language settings, the findings here indicate that more data is not necessarily better when that data is drawn from a highly narrow and stylistically constrained domain. Data scarcity alone does not fully explain poor translation performance.

The T5S model broadly replicates this finding. Translation quality generally decreases as the proportion of Bible training data increases, though the relationship is not strictly monotonic: Condition B achieves a marginally higher BLEU score than Condition A, and the two conditions are tied on chrF and COMET. This suggests that small amounts of Bible data do not substantially destabilize the T5S model, in contrast to TowerInstruct-7B-v.02 where any increase in Bible proportion corresponds to a measurable decline across all metrics. Despite this difference, both models agree on the core finding: Condition E (100% Bible data) performs substantially worse than all other fixed-dataset-size conditions, and increasing data quantity does not compensate for poor data composition.

A particularly notable result is the performance of Experiment E with the TowerInstruct-7B-v.02 model, which underperforms even the unfine-tuned baseline model. This condition fine-tuned the model exclusively on Bible data. The resulting degradation of translation quality suggests that such fine-tuning may overwrite or distort previously learned multilingual representations rather than strengthen them.

Qualitative observations of the outputs from the TowerInstruct-7B-v.02 model under Experiment E further support this interpretation. Several outputs include gibberish strings or repeated tokens. For example, quotation marks were absent from the non-Bible training data but among the most frequent unigrams in the Bible corpus; correspondingly, some outputs consist entirely of repeated quotation marks. Other outputs include repeated words or syllables, such as *jejejejejejejejeje* or *más allá, más allá, más allá.* These patterns are consistent with overfitting to highly frequent surface forms in the training data.

The T5S model under Condition E does not exhibit the same prompt-repetition failures as TowerInstruct-7B-v.02, but does produce degraded and sometimes nonsensical outputs, including *já, ja, ja* as a translation of *jáaj* and *esto es lo que dice el señor soberano jehová:* as a translation of *u tsolts'fibil xook*. The latter is the only flagged output across either model that constitutes a plausible case of Bible-register content.

The results from Conditions F and G also support this interpretation. Condition F, which included all available training data from both corpora, did not outperform Condition A despite containing more than three times as much training data. This suggests that the addition of Bible data actively degrades model performance rather than simply failing to help. Condition G, trained exclusively on all available Bible data, performs comparably to Condition E, the worst-performing fixed-dataset-size condition, indicating that increasing the total quantity of training data does not yield translation quality gains when that data is drawn from or dominated by the Bible corpus.

Condition F represents the most dramatic failure across both models, but in different ways. For TowerInstruct-7B-v.02, Condition F underperforms Condition A despite containing more than three times as much training data. For T5S, the failure is categorical: every output in the 2,000-sentence test set consists solely of the word *la*, producing a BLEU score of 0.0. This suggests that the combination of a large, heterogeneous training set dominated by Bible data is particularly destabilizing for the smaller encoder-decoder T5S architecture, which may lack the capacity to recover coherent translation behavior from such a noisy training signal. Condition G for T5S performs better than Condition F despite containing only Bible data, which further supports the interpretation that data heterogeneity combined with domain mismatch—rather than Bible data alone—is responsible for the catastrophic failure in Condition F.

These findings have practical implications for low-resource language MT development. When curating training data for endangered or low-resource languages like Yucatec Maya, the Bible corpus is often one of the largest and most readily available parallel resources. However, the results presented here suggest that uncritical reliance on Bible data can actively harm translation quality for non-religious domains. Practitioners working with low-resource languages should carefully consider the domain match between available parallel

data and their intended use case, even when data scarcity creates pressure to use all available resources.

7.1.2 Contextualizing Low Evaluation Scores

The low absolute scores observed in this study are best understood in the context of broader trends in low-resource and Indigenous language machine translation, as well as key differences in experimental design between this study and prior work. Although recent work reports substantially higher BLEU and chrF scores for Spanish-to-Yucatec Maya translation—reaching BLEU values near 29 and chrF++ above 50 (Rangel and Kobayashi, 2024)—the results obtained in this study are markedly lower, with BLEU ranging from 0.47 to 3.49 and chrF from 14.40 to 22.56 with the TowerInstruct-7B-v.02 model, and with BLEU ranging from 0.0 to 13.84 and chrF from 2.554 to 35.39 with the T5S model. The substantially higher scores reported by Rangel and Kobayashi (2024) are likely attributable to several key differences in experimental design. First, while their study evaluates translation from Spanish into Yucatec Maya, the present study evaluates translation from Yucatec Maya into Spanish; these directions are not directly comparable, and performance differences across directions are well-documented in the MT literature. Second, Rangel and Kobayashi (2024) train and evaluate on data drawn from the same source, the CPLM corpus, whereas the present study deliberately evaluates on naturalistic conversational text from the T'aantsil and UNAM corpora that is domain-distinct from the Bible training data used in several experimental conditions, with the exception of experiment A where both the entirety of the training data and the testing data come from the T'aantsil and UNAM corpora. This domain mismatch between training and test data is expected to suppress metric scores independently of model quality. Third, while the present study shares the T5S architecture with Rangel and Kobayashi (2024), it also employs TowerInstruct-7B-v.02, a decoder-only instruction-tuned LLM that differs substantially from the encoder-decoder models used in their work. The T5S experiments reported in the present study allow for a more direct architectural comparison with Rangel and Kobayashi (2024), as both use the same base model family. Nevertheless, the differences in translation direction and evaluation domain

remain, and the T5S scores obtained here remain substantially lower than those reported by Rangel and Kobayashi (2024), consistent with the domain mismatch hypothesis.

Recent large-scale evaluations of machine translation for Indigenous languages of the Americas further contextualize the low absolute scores observed in this study. Results from the AmericasNLP 2021 and 2024 shared tasks report average BLEU scores typically below 30 and chrF++ values in the range of 15–30 across a wide range of Indigenous language pairs, even when leveraging large pretrained multilingual models and extensive data augmentation (Mager et al., 2021; Ebrahimi et al., 2024). Despite substantial advances in modeling and training strategies, these studies consistently demonstrate that machine translation involving Indigenous languages remains highly challenging regardless of direction, particularly under realistic low-resource conditions. In this context, the BLEU, chrF, METEOR, and COMET values reported in the present work fall well within the broader distribution observed across recent shared-task evaluations, especially when accounting for the domain mismatch introduced by training on Bible data while evaluating on naturalistic conversational text.

BLEU and METEOR both rely heavily on surface-level overlap between the model output and the reference translation. BLEU measures n-gram precision, rewarding outputs that share exact word sequences with the reference, while METEOR additionally accounts for stemming, synonymy, and word order variation, making it somewhat more flexible. In this study, BLEU scores range from 0.405 to 3.49 for TowerInstruct-7B-v.02 and from 0.0 to 13.84 for T5S across experimental conditions, and METEOR scores range from 0.101 to 0.206 for TowerInstruct-7B-v.02 and from 0.00985 to 0.394 for T5S. These low values reflect not only the difficulty of the Yucatec Maya-to-Spanish translation task, but also the sensitivity of surface-based metrics to the kinds of lexical and structural variation that naturally arise when translating from a low-resource language with limited training data. Even semantically adequate translations may receive low BLEU and METEOR scores if they differ from the reference in word choice or phrasing, which is particularly likely when the model has been fine-tuned on domain-mismatched data such as the Bible corpus.

chrF operates at the character level rather than the word level, measuring character n-gram overlap between the output and the reference. This makes it more tolerant of

morphological variation and partial matches. chrF scores in this study range from 13.89 to 22.56 for TowerInstruct-7B-v.02 and from 2.554 to 35.39 for T5S, broadly consistent with the range reported for other Indigenous language-Spanish pairs in recent shared task evaluations (Mager et al., 2021; Ebrahimi et al., 2024). The relative stability of chrF compared to BLEU across experimental conditions suggests that while exact word-level matches deteriorate as Bible data increases, some character-level similarity to the reference is preserved even in lower-quality outputs.

COMET differs from the other three metrics in that it does not rely on surface overlap at all. Instead, it uses a pretrained multilingual neural model to assess semantic adequacy and fluency, designed to compare the model output to both the reference translation and the source sentence. However, COMET's underlying model – `wmt22-comet-da` – was trained on human judgment data from WMT shared tasks, which does not include Yucatec Maya. As a result, the source-side representations produced for YM input are likely unreliable, meaning that in practice, COMET functions here as a reference-based measure rather than a full source-aware one. This is a limitation of applying COMET to low-resource Indigenous language pairs for which the underlying model has no pretraining coverage. With this caveat in mind, COMET scores in this study range from 0.400 to 0.543 for TowerInstruct-7B-v.02 and from 0.409 to 0.653 for T5S, with the validation checks establishing a practical minimum of approximately 0.391 for completely mismatched output and a maximum of 0.951 for perfect output. The fact that COMET scores remain above the minimum even for the worst-performing conditions suggests that while some outputs are degenerate – including repeated tokens, gibberish strings, and untranslated source text – these failures represent a subset of outputs, while another subset of outputs retain sufficient surface and semantic similarity to the reference to keep the aggregate score above the floor. The monotonic decrease in COMET scores as Bible data increases is consistent with a broad deterioration in translation quality across conditions, though given the source-side limitation noted above, this trend is best interpreted as reflecting reference-based similarity rather than semantic adequacy in the fullest sense.

7.1.3 Establishing Score Bounds via Validation Checks

The validation check experiments confirm that the evaluation metrics respond appropriately to extreme cases of perfect alignment, script mismatch, sentence misalignment, and untranslated input. The gold2gold comparison, in which the reference translation is evaluated against itself, establishes the upper bound for each metric: BLEU of 100.00, chrF of 100.00, METEOR of 0.989, and COMET of 0.951. The jpn2gold comparison, in which a set of Japanese sentences is evaluated against the Spanish reference, establishes near-zero lower bounds reflecting complete script and language mismatch: BLEU of 0.00, chrF of 0.05, METEOR of 0.000, and COMET of 0.391. The rand2gold comparison, in which a random reordering of the reference sentences is evaluated against the original ordering, isolates the effect of content misalignment while preserving language and style, yielding BLEU of 0.18, chrF of 11.69, METEOR of 0.054, and COMET of 0.411. Finally, the src2targ comparison, in which the Yucatec Maya source text is evaluated against the Spanish reference, identifies the score range expected when the model outputs the input without translating it, yielding BLEU of 1.47, chrF of 14.94, METEOR of 0.115, and COMET of 0.410. The relatively high COMET lower bound of 0.391 – compared to BLEU of 0.00 and METEOR of 0.000 for the jpn2gold comparison – reflects a general property of the `wmt22-comet-da` model: its neural embedding space does not produce near-zero similarity scores even for completely mismatched inputs. The score 0.391 represents a practical floor for the COMET metric rather than a meaningful measure of similarity. Scores below this threshold are therefore not interpretable as indicating degrees of mismatch, and the experimental scores should be understood as occupying the range between this practical floor and the gold2gold upper bound of 0.951.

7.1.4 Summary

For the TowerInstruct-7B-v.02 model, translation quality decreased monotonically as the proportion of Bible training data increased, with conditions trained exclusively on Bible data falling below even the unfine-tuned baseline. The T5S model replicates this finding broadly, though without the strict monotonicity observed in TowerInstruct-7B-v.02. The

T5S model also substantially outperforms TowerInstruct-7B-v.02 under comparable conditions, suggesting that encoder-decoder architectures pretrained explicitly for translation may be better suited to this task than decoder-only instruction-tuned LLMs. Increasing the total quantity of training data did not improve performance when that data was drawn from or dominated by the Bible corpus, suggesting that domain mismatch is a more important factor than data quantity in this setting. The low absolute metric scores are consistent with results reported for other Indigenous language–Spanish pairs in recent shared task evaluations, and are further explained by the domain mismatch between Bible training data and conversational test data in all experimental conditions except condition A. Interpretation of COMET scores is subject to an additional caveat: because the underlying model has no pretraining coverage of Yucatec Maya, its source-side signal is unreliable, and COMET should be treated as a reference-based measure in this setting. Taken together, the evaluation metrics point to a consistent pattern of degradation as Bible data increases, operating at both the surface and reference-similarity levels.

7.2 Semantic Drift

Perhaps the most interesting finding of this study is the near-complete absence of detectable semantic drift toward Bible-related content across all experimental conditions. Despite training on datasets ranging from 0% to 100% Bible data, the semantic similarity detection method flagged no more than three outputs per condition out of 2,000 test sentences, and in most conditions flagged only one or two. Manual inspection of flagged outputs revealed that all but one constitute methodological false positives rather than genuine cases of religious content contamination. The flagged outputs fall into three categories, which apply across both models. The first category consists of very short segments – including *no* (TowerInstruct, Condition A) and *sí* (TowerInstruct, Condition D) – which are flagged as a consequence of a known limitation of embedding-based similarity methods. When a segment contains minimal lexical content, its sentence embedding is poorly anchored semantically, making it susceptible to false similarity with arbitrary reference sentences. These cases should be interpreted as false positives arising from the detection methodology rather than evidence of Bible influence on model outputs.

The second category consists of ordinary conversational Spanish sentences, including *te lo enseñan en la escuela* (TowerInstruct, Condition A), *se llama así* (TowerInstruct, Condition C), *tienes mucha paciencia* (TowerInstruct, Condition D), *es muy difícil aprender a leer* (TowerInstruct, Condition F), *la lengua maya* (T5S, Condition A), *el estudio de la lengua maya* (T5S, Condition B), *la lectura de los libros* (T5S, Condition C), *la enseñanza de la lectura* (T5S, Condition D), and *ma'alo* (T5S, Condition E). None of these outputs contain religious content, archaic phrasing, or stylistic features associated with Biblical Spanish. Their flagging likely reflects noise in the similarity computation rather than meaningful semantic overlap with the Bible corpus.

The third category, which is the most analytically interesting, consists of output failures in Conditions B, E, and G. In Condition B, the TowerInstruct-7B-v.02 flagged output *u tsolts'fibil xook* is untranslated Yucatec Maya source text, suggesting a failure of the translation process rather than semantic drift; the embedding model's behavior when applied to out-of-vocabulary Maya text is unpredictable, and the flagging of this output should be attributed to the detection method's inability to handle non-Spanish text rather than to any Bible-related content in the output. Conditions E and G present the most severe failures across both models. In TowerInstruct Condition E, the flagged outputs consist of the prompt template alone, with no translation attempt. In TowerInstruct Condition G, the flagged output consists of the prompt template followed by a repetitive loop of the Maya source phrase *u tsolts'fibil xook yéetel*, consistent with a degenerate generation pattern in which the model fails to terminate. The T5S model under Conditions E and G produces analogous failures: *já, ja, ja* reflects phonetic echoing rather than translation, and *¡mira! ¡mira! ¡mira!* and *la vara de la vara de la vara de la vara de la vara* exhibit the same repetitive degenerate pattern seen in TowerInstruct. Across both models, these failures indicate catastrophic forgetting of translation behavior rather than semantic drift, and are most severe under the conditions with the highest proportions of Bible training data, suggesting that fine-tuning on large quantities of Bible data undermines the models' ability to produce coherent output. The single exception is *esto es lo que dice el señor soberano jehová:* produced by T5S in Condition E, which contains recognizable Biblical Spanish register and is best interpreted as a case of genuine semantic contamination: a model fine-tuned almost

entirely on Bible text defaulting to biblical phrasing when it cannot produce a coherent translation.

The Bible n-gram contamination analysis provides independent corroboration of these findings. Across all seven experimental conditions, no output sentence produced lexical patterns statistically associated with biblical Spanish register, as measured against a set of high-LLR Bible n-grams covering unigrams, bigrams, trigrams, and 4-grams. This negative result is particularly notable in Conditions E and G, where instruction-following failures and degenerate outputs confirm that Bible-heavy training causes significant model degradation. Even in these extreme cases, the degraded outputs do not contain Bible-associated vocabulary or phrasing identified by the n-grams lists; the models fail in ways that reflect overfitting to surface forms and loss of instruction-following behavior, not in ways that reflect absorption of biblical n-grams. This distinction matters: it suggests that the influence of Bible training data on model outputs is better characterized as a training stability problem than as a content contamination problem.

These findings suggest that the semantic drift hypothesis, while theoretically well-motivated, is not supported by the empirical evidence from this study. With the singular exception of one T5S output under Condition E, neither model appears to generate Bible-influenced content in response to Bible-heavy training; instead, Bible-heavy training degrades translation quality and, at the extreme, causes instruction-following failures.

7.3 Limitations and Future Work

The evaluation metrics used here, while standard in the MT literature, are imperfect proxies for translation quality in low-resource settings. BLEU in particular is known to correlate poorly with human judgments at low score ranges, and all experimental conditions fall well below the ranges typically observed for high-resource pairs. Human evaluation by fluent Yucatec Maya and Spanish speakers would be a valuable complement to the metric-based findings reported here.

The semantic drift detection methodology also has limitations. The false positive rate in this study is non-trivial, and the threshold of 0.2 cosine similarity difference was set a priori without calibration to this specific language pair or domain. Adaptive thresholding

or alternative embedding models with stronger multilingual coverage could address this in future work.

The findings are also specific to Yucatec Maya-to-Spanish translation and should not be generalized to other low-resource language pairs, other domains of Bible text, or other base models. The result that Bible data harms non-Bible translation quality is a property of this experimental setting, and replication across other language pairs is needed before drawing broader conclusions.

Finally, the LoRA fine-tuning approach used for TowerInstruct-7B-v.02 was appropriate given GPU memory constraints, but it introduces variables relative to full-parameter fine-tuning. The T5S experiments used full-parameter fine-tuning, which removes this concern for that model, but whether the TowerInstruct findings would hold under full-parameter fine-tuning remains an open question worth revisiting with greater computational resources.

7.4 Summary

This chapter has interpreted the experimental results in relation to the two primary research questions. On the question of translation quality, the evidence from both models is consistent: increasing the proportion of Bible-sourced training data degrades translation performance, and this degradation cannot be offset by increasing data quantity alone. The T5S model outperforms TowerInstruct-7B-v.02 substantially under comparable conditions, suggesting that encoder-decoder architectures pretrained for translation may be better suited to this task than decoder-only instruction-tuned LLMs. However, T5S is not immune to the effects of Bible data: Condition F produces a catastrophic failure unique to T5S, and the general downward trend across Conditions A-E holds for both models. Domain mismatch between training and evaluation data is a more consequential factor than data volume in this setting, with practical implications for how low-resource MT practitioners approach corpus curation. On the question of semantic drift, the evidence argues against the hypothesis: neither the BETO semantic similarity analysis nor the Bible n-gram contamination analysis found meaningful evidence of Bible-influenced content in model outputs across any experimental condition for either model, with a singular exception in Condition E for the T5S model. The influence of Bible training data on model behavior is better under-

stood as a training stability problem than a content contamination problem. These findings are subject to the methodological limitations discussed above, and replication across other language pairs and model architectures remains an important direction for future work.

Chapter 8

CONCLUSION

8.1 Summary of Findings

This thesis investigated the trade-offs associated with using Bible-derived parallel data to fine-tune machine translation models for the Yucatec Maya-to-Spanish translation task. Across seven experimental conditions varying the proportion and quantity of Bible training data, a consistent relationship emerged across both the TowerInstruct-7B-v.02 model and the T5S model: as Bible data increased, translation quality generally decreased across all four evaluation metrics – BLEU, chrF, METEOR, and COMET. For TowerInstruct-7B-v.02, this relationship was strictly monotonic across the fixed-dataset-size conditions. For T5S, the relationship was broadly similar but not strictly monotonic, with Condition B performing marginally better than Condition A on BLEU. Both models trained exclusively on Bible data, whether at the fixed training budget (Condition E) or with the full available corpus (Condition G), performed at or below the unfine-tuned baseline. The T5S model additionally exhibited a catastrophic failure under Condition F, in which every output in the test set consisted solely of the word *la*, suggesting that large heterogeneous training sets dominated by Bible data may be particularly destabilizing for smaller encoder-decoder architectures. Across both models, the addition of Bible data to all available non-Bible data (Condition F) also failed to improve upon Condition A, suggesting that Bible data actively degrades rather than merely fails to contribute to translation quality for conversational, everyday text.

Contrary to the hypothesis motivating the semantic drift detection framework, no meaningful evidence of Bible-influenced content emerged in model outputs across any experimental condition. Two complementary analyses – BETO semantic similarity and Bible n-gram contamination – both returned largely negative results: with a single exception, flagged outputs were attributable to embedding instability on short segments, non-Spanish output,

and generation failures rather than genuine semantic drift toward religious content. The sole exception was one T5S output under Condition E, which contained recognizable biblical Spanish register and is interpreted as a case of genuine semantic contamination under extreme Bible-data fine-tuning conditions.

These computational findings were contextualized by a community survey of 84 Yucatec Maya speakers, which revealed broad support for machine translation technology alongside specific concerns about training data provenance, data sovereignty, and the risk of perpetuating colonial dynamics through the uncritical use of religious and colonial texts. Survey respondents broadly agreed that the Mayan community should own and govern the linguistic data produced by its members, and a majority expressed concern about the use of texts translated from Spanish into Yucatec Maya – the inverse of the translation direction studied here – as training data.

8.2 Broader Significance

These findings carry practical and ethical implications for low-resource MT development more broadly. The common practice of supplementing scarce parallel data with Bible corpora is not merely an ethical concern in the abstract; this thesis demonstrates empirically that it can measurably harm the translation quality that communities actually need. For practitioners working with endangered languages, these results argue for prioritizing domain-matched community-generated data over large-scale religious corpora, even when the latter are more readily available.

The survey findings extend the thesis's contributions beyond computational measurement by situating the technical results within the expressed priorities of the community most directly affected. Respondents articulated specific visions for what machine translation systems should do and whom they should serve, citing language revitalization, accessibility in healthcare and legal settings, and increased digital presence for younger speakers as motivating goals. At the same time, they identified clear limits: concern about epistemic extractivism, the risks of perpetuating colonialist frameworks through training data choices, and the fundamental principle that the Mayan community should govern its own linguistic data. The survey thus reframes the ethical dimension of this thesis: the question is not

only whether Bible data harms translation quality in a linguistic sense, but also whether the practices through which MT systems are built can be made accountable to the communities whose languages and cultural histories are at stake. Answering that question requires researchers to treat community consultation not as an afterthought but as a constitutive part of the research process itself.

BIBLIOGRAPHY

Alves, D. M., Pombal, J., Guerreiro, N. M., Martins, P. H., Alves, J., Farajian, A., Peters, B., Rei, R., Fernandes, P., Agrawal, S., Colombo, P., de Souza, J. G. C., and Martins, A. F. T. (2024). Tower: An open multilingual large language model for translation-related tasks. arXiv preprint arXiv:2402.17733.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Goldstein, J., Lavie, A., Lin, C.-Y., and Voss, C., editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data. In *Proceedings of the Workshop on Practical ML for Developing Countries (PML4DC) at ICLR 2020*.

Christodouloupoulos, C. and Steedman, M. (2015). A massively parallel corpus: The Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Domingues, P. H., Pinhanez, C. S., Cavalin, P., and Nogima, J. (2024). Quantifying the ethical dilemma of using culturally toxic training data in AI tools for Indigenous languages. In Melero, M., Sakti, S., and Soria, C., editors, *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 283–293, Torino, Italia. ELRA and ICCL.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Ebrahimi, A., de Gibert, O., Vazquez, R., Coto-Solano, R., Denisov, P., Pugh, R., Mager, M., Oncevay, A., Chiruzzo, L., von der Wense, K., and Rijhwani, S. (2024). Findings of the AmericasNLP 2024 shared task on machine translation into Indigenous languages. In Mager, M., Ebrahimi, A., Rijhwani, S., Oncevay, A., Chiruzzo, L., Pugh, R., and von der Wense, K., editors, *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 236–246, Mexico City, Mexico. Association for Computational Linguistics.

Errington, J. (2001). Colonial linguistics. *Annual Review of Anthropology*, 30:19–39.

Fadaee, M., Bisazza, A., and Monz, C. (2017). Data augmentation for low-resource neural machine translation. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Gordon, M. A., Duh, K., and Kaplan, J. (2021). Data and parameter scaling laws for neural machine translation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5915–5922, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haddow, B., Bawden, R., Miceli Barone, A. V., Helcl, J., and Birch, A. (2022). Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Hanks, W. F. (2012). Birth of a language: The formation and spread of colonial Yucatec Maya. *Journal of Anthropological Research*, 68(4):481–504.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hutchinson, B. (2024). Modeling the sacred: Considerations when using religious texts in natural language processing. In Duh, K., Gomez, H., and Bethard, S., editors, *Findings of*

the Association for Computational Linguistics: NAACL 2024, pages 1029–1043, Mexico City, Mexico. Association for Computational Linguistics.

Instituto Nacional de Estadística y Geografía (2020). Censo de población y vivienda 2020: Tabulados del cuestionario básico. Instituto Nacional de Estadística y Geografía. Accessed March 2026.

Mager, M., Mager, E., Kann, K., and Vu, N. T. (2023). Ethical considerations for machine translation of Indigenous languages: Giving a voice to the speakers. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada. Association for Computational Linguistics.

Mager, M., Oncevay, A., Ebrahimi, A., Ortega, J., Rios, A., Fan, A., Gutierrez-Vasques, X., Chiruzzo, L., Giménez-Lugo, G., Ramos, R., Meza Ruiz, I. V., Coto-Solano, R., Palmer, A., Mager-Hois, E., Chaudhary, V., Neubig, G., Vu, N. T., and Kann, K. (2021). Findings of the AmericasNLP 2021 shared task on open machine translation for Indigenous languages of the Americas. In Mager, M., Oncevay, A., Rios, A., Ruiz, I. V. M., Palmer, A., Neubig, G., and Kann, K., editors, *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. (2019). Facebook FAIR's WMT19 news translation task submission. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Turchi, M., and Verspoor, K., editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 314–319, Florence, Italy. Association for Computational Linguistics.

Pinhanez, C. S., Cavalin, P., Vasconcelos, M., and Nogima, J. (2023). Balancing social impact, opportunities, and ethical constraints of using AI in the documentation and

vitalization of Indigenous languages. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Post, M. (2018). A call for clarity in reporting BLEU scores. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1):5485–5551.

Rangel, J. and Kobayashi, N. (2024). Advancing NMT for Indigenous languages: A case study on Yucatec Mayan and Chol. In Mager, M., Ebrahimi, A., Rijhwani, S., Oncevay, A., Chiruzzo, L., Pugh, R., and von der Wense, K., editors, *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 138–142, Mexico City, Mexico. Association for Computational Linguistics.

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A neural framework for MT evaluation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Resnik, P., Diab, M., and Olsen, M. B. (1999). The Bible as a parallel corpus: Annotating the 'book of 2000 tongues'. *Computers and the Humanities*, 33(1–2):129–153.

Sierra Martínez, G., Montaño, C., Bel-Enguix, G., Córdova, D., and Mota Montoya, M. (2020). CPLM, a parallel corpus for Mexican languages: Development and interface. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2947–2952, Marseille, France. European Language Resources Association.

Sierra Martínez, G., Solórzano Soto, J., and Curiel Díaz, A. (2017). GECO, un Gestor de Corpus colaborativo basado en web. *Linguamática*, 9(2):57–72.

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

Wonderly, W. L. and Nida, E. A. (1963). Linguistics and Christian missions. *Anthropological Linguistics*, 5(1):104–144.

Yamasaki, E. (2019). *Yucatec Maya Language on the Move: A Cross-disciplinary Approach to Indigenous Language Maintenance in an Age of Globalization*. PhD thesis, Rheinische Friedrich-Wilhelms-Universität Bonn.