# The eBible Corpus: Data and Model Benchmarks for Bible Translation for Low-Resource Languages

Vesa Åkerman[a], David Baines[a], Damien Daspit[a], Ulf Hermjakob[b], Taeho Jang[c], Michael Martin[a]*, Joel Mathew[b,d], Jonathan Robie[e], Marcus Schwarting[f]

[a]*SIL International, 7500 West Camp Wisdom Road, Dallas, TX 75236*
[b]*University of Southern California Information Sciences Institute, 4676 Admiralty Way #1001, Marina Del Rey, CA 90292*
[c]*Payap University Linguistics Department, Super-highway Chiang Mai – Lumpang Road, Amphur Muang, Chiang Mai, 50000 Thailand*
[d]*Bridge Connectivity Solutions, Plot Nos. 11- 13, Amberhai Main Road, Sector 19, Dwarka, New Delhi 110075, India*
[e]*Clear Bible, 2990 Franklin Ave SW #8, Grandville, MI 49418*
[f]*University of Chicago Department of Computer Science, 5801 South Ellis Avenue, Chicago, IL 60637*

*Corresponding author email: michael_martin@sil.org

**Abstract**
Efficiently and accurately translating into a low-resource language remains a challenge, regardless of the strategies employed, whether manual or automated. Many Christian organizations are dedicated to the task of translating the Holy Bible into languages that lack a modern translation. Bible translation (BT) work is currently underway for over 3000 extremely low resource languages. We introduce the eBible corpus: a dataset containing 1009 translations of portions of the Bible with data in 833 different languages across 75 language families. In addition to this dataset, we design and introduce model performance benchmark tasks and report various metrics using the No Language Left Behind (NLLB) neural machine translation (NMT) models. We describe several problems specific to the domain of BT and consider how the established data and model benchmarks might be used for future translation efforts. On one of the tasks, the fine-tuned NLLB model on the Austronesian and Trans-New Guinea language families achieve 35.1 and 31.6 BLEU scores respectively, which spurs future innovations for NMT for low-resource languages in Papua New Guinea.

**Keywords:**Bible translation, natural language processing, low-resource languages, machine translation, NLLB, large language models, multilingual modeling

## 1. Introduction

There has been significant progress recently towards solving multiple problems in the field of Natural Language Processing (NLP). Most of these advances, however, are skewed towards languages of wider communication (LWCs). Though there is ongoing work in low resource languages, the scarcity of training data and benchmarks for meaningful comparison of proposed techniques in such languages slow down the pace of research.

The Holy Bible has been translated to a very large number of languages of the world with continued work to modernize multiple translations. Historically, Bible translations have been foundational to the standardizing and revitalization of language for various communities. Therefore, this data has the potential to be the starting point for NLP research in many extremely low resource languages. It would especially be useful for benchmarking model performance for NLP researchers working in the Biblical domain against modern techniques. Though not all translations are published under a permissive license for reuse, eBible.org has curated more than 1000 translations in various formats that are unencumbered. Yet there are domain specific nuances and issues in the data format (cite USFM), structure (versification) and encoding that need careful handling and have been observed by the authors as an impedance to use the data efficiently for NLP.

In this work we tackle the problem of scarcity of data and a model benchmark (for machine translation) in low resource languages by:

1. Collecting, parsing and cleaning 1009 translations from eBible.org which have been automatically verified to be under a permissive license. These are made available as a verse(footnote on definition of verse)-wise parallel corpus across 833 languages.
2. Designing domain relevant benchmarking tasks that take into consideration the textual and stylistic variations in the content, having multiple related languages to a target language and the realities of the progress of a Bible translation project on the ground.

In our knowledge, this is the first time such a large unencumbered multilingual corpus and carefully designed benchmark has been released to the NLP community. This work draws heavily from our own experience and work with multiple recognized Bible translation teams, organizations and languages.

The following sections are organized such that we first review existing relevant work on developing large Bible corpora and low resource machine translation. We then detail the eBible corpus and its statistics along with the steps we took to parse and clean the data. We share our experimental setup for the benchmark and the models we used. We then share the experiment results and discuss findings to attract researchers to the issues faced in Bible translation (BT). Finally, we propose multiple research directions for future work using this dataset and provide concluding remarks.

## 2. Background

In this section we briefly review previous efforts to aggregate Biblical corpora. We then consider previous NLP-driven strategies for multilingual translation to such low-resource languages, including those specific to Bible translation tasks.

### 2.1. Previously Aggregated Data

The number of languages represented in Biblical corpora has been rising steadily over the last few decades. Resnik et. al. produced a parallel corpus with 13 languages in 1999 [1], and by 2015 the corpus of Christodouloupoulos et. al. contained Bibles spanning 100 languages [2]. In 2020, McCarthy et. al. reported on an effort to scrape and align Scriptures from various sources. With over 1600 languages and 4000 unique translations represented it is most likely the largest Biblical corpus ever compiled, unfortunately this corpus is not publicly available [3]. Other online archives of open-license partial and full Bible translations exist, but have not been made available in a format amenable to statistical or deep learning driven translation tasks. Outside of Bible translation, the general problem of translating text into extremely low resource languages remains a challenge, primarily due to a lack of high-quality open-source data. Datasets such as FLORES-101 [T], SALT [X,Y] and AmericasNLI [Z] previously provided a starting point for model training. With the release of NLLB also came the FLORES-200 dataset [NLLB], which contains 3001 sentences sampled across 204 total languages. FLORES-200 provides a many-to-many multilingual data benchmark which is the largest to date.

### 2.2. Previous Translation Models

We define a translation task as follows. Suppose we are given a passage which is readily available in one or multiple source languages. Suppose we also have a target language for which the passage has not been previously translated. We define the translation task for this passage as the creation of a mapping between the source(s) and the target. In the case of Bible translation, a mapping between passages is carried out by verse, but in the general case this can be performed by sentence. The first non-classical machine translation models relied on statistical machine translation (SMT), and include alignment-based strategies [4], Markovian methods [5], and many other frequentist and Bayesian approaches [6]. Of particular interest is the work of Wu et. al. [Q], which employs an SMT approach on the Bible corpus compiled by McCarthy et. al. [4]. Neural machine translation (NMT) is a natural extension of SMT, and utilizes a neural network architecture to train directly on source and target texts. Basic NMT implementations use an encoder and a decoder structure, and may forgo the recurrence and attention mechanism characteristic of transformers. The OpenNMT package provides a turnkey implementation for fine-tuning NMT models for specific translation tasks [cite: OpenNMT].

Transformers designed for translation tasks can be considered an extension of earlier NMT models through the inclusion of recurrent layers and an attention mechanism. Many transformer architectures have been modified for translation, including fairseq [cite Ott fairseq paper] and BERT [cite Zhu et. al. 2020]. Of particular interest is the work of Leides, which uses a fairseq architecture trained on Bible translations across 50 different languages and available for general-purpose BT tasks [cite GH repo and blog post]. Finally, Meta's NLLB model represents the current state-of-the-art NMT transformer, trained on the FLORES-200 dataset representing over 200 different languages.

## 3. Methods

In this section we describe the content of the eBible corpus and our pipeline for aggregating and preprocessing Bible translations. We also present summary statistics describing the contents of the eBible corpus. We then detail several benchmark translation tasks we performed across eight language families within the corpus. Finally, we describe the model architectures and scoring methods by which we will evaluate model translation performance.

### 3.1. The eBible Corpus

We gathered and aligned 1,009 Scripture translations in 833 languages from eBible.org which are provided under a Creative Commons or similarly permissive licenses. This includes 113 files under Attribution ShareAlike (CC BY-SA), two files under Attribution Non-Commercial (CC BY-NC), 106 files under Attribution No Derivs (CC BY-ND), 699 files under Attribution Non-Commercial No Derivs (CC BY-NC-ND), and 84 files under Public Domain. After downloading these Bibles, the versification scheme (*Original*, *English*, *Russian Orthodox*, *Russian Protestant*, *Septuagint*, or *Vulgate*) for each Bible was inferred based on its content. The text of each verse was extracted and all formatting, cross-references, footnotes, and other markup was removed. The verse text was placed into an extract file with a verse-per-line format with 41,899 lines per file; the placement of each verse in the extract file was normalized to the Original versification scheme, allowing ready comparison of verses across translations. A separate index file (`vref.txt`) records the verse reference for each line of all verse extract files. Verse ranges were preserved by placing the verse range text on the first line of the range, and tagging subsequent lines from the verse range with the `<range>` indicator in the verse extract file. The corpus and code are available on Github [X]. Additional tools used in this process include the SIL-NLP package [Y] and the Wildebeest package [Z]. Figure 1 shows the general format of several extract files, designed to be easily ingested for NLP analysis and machine translation tasks.

**vref.txt**

```
1       GEN 1:1
2       GEN 1:2
3       GEN 1:3
…
```

**eng-engULB.txt**

**1** In the beginning, God created the heavens and the earth.

**2** The earth was without form and empty. Darkness was upon the surface of the deep. The Spirit of God was moving above the surface of the waters.

**3** God said, "Let there be light," and there was light.

…

**fra-frasbl.txt**

**1** Au commencement, Dieu créa les cieux et la terre.

**2** La terre était informe et vide. Les ténèbres étaient à la surface de l'abîme et l'Esprit de Dieu planait au-dessus de la surface des eaux.

**3** Dieu dit : « Que la lumière soit ! » et la lumière fut.

…

**deu-deuelo.txt**

**1** Im Anfang schuf Gott die Himmel und die Erde.

**2** Und die Erde war wüst und leer, und Finsternis war über der Tiefe; und der Geist Gottes schwebte über den Wassern.

**3** Und Gott sprach: Es werde Licht! und es ward Licht.

…

**Figure 1.** Sample verse extract files (vref, English, French, German).

In order to reduce data fragmentation for NLP tasks, we performed character-level cleaning on the eBible corpus using the **Wildebeest** tool, making 3.3 million changes to 220 out of the 1,009 Bible translations, the vast majority of which are not or barely perceptible to the human reader. Changes include bringing complex character sequences into conventional order (e.g. Devanagari primary character, nukta, vowel sign), preferring composed characters per Unicode standard[1]; correcting some look-alike characters, e.g. mapping Latin A (U+0041) to Cyrillic A (U+0410) in Cyrillic-script text, or mapping Latin l (U+006C) to Devanagari danda । (U+0964) where appropriate; character normalization, e.g. reversed c to open o (ↄ → ɔ; U+2184 → U+0254); for one translation, mapping the replacement character  (U+FFFD) to open/close double/single quotes; correcting some comma errors (deleting spaces before a comma, adding a space after a comma, removing duplicate commas); decomposing some ligatures (e.g. fi → fi; U+FB01 → U+0066 U+0069); and more. Slightly over half of these changes were done fully automatically, using the Wildebeest Normalization script `wb_normalize.py`; the remaining changes were made with an eBible-specific script `wb_bible_plus.py` based on a manual review using the Wildebeest Analysis script `wb_analysis.py`. Some issues raised by the Wildebeest Analysis script have not been addressed, such as private-use characters in two Bible translations (dwr-dwrENT, gof-gofENT), and some residual wrong-script characters that can't readily be corrected automatically or semi-automatically.

The eBible corpus exhibits a wide diversity of languages, as shown in Figure 2 and Figure 3. Roughly a quarter of the translations are in languages spoken primarily in Papua New Guinea, which is widely known as the most linguistically diverse country in the world. Many translations are in languages considered ultra-low-resource, and additional texts in some languages may not be readily available. A large portion of the Bible translations are partially complete, as shown in Figure 4. Experts often start with translating the New Testament before proceeding to the Old Testament, which is reflected in the availability of full New Testament translations versus full Old Testament translations. Translations of the deuterocanon are included in the eBible corpus but are excluded from further analysis due to the sparsity of examples.

[1] Wildebeest complex character order normalization generally matches the dominant forms of most original Bible translations and other corpora; it is closest to Unicode's NFC, but unlike NFC follows conventional order. E.g., NFC order of the above pattern is: Devanagari primary character, vowel sign, nukta.

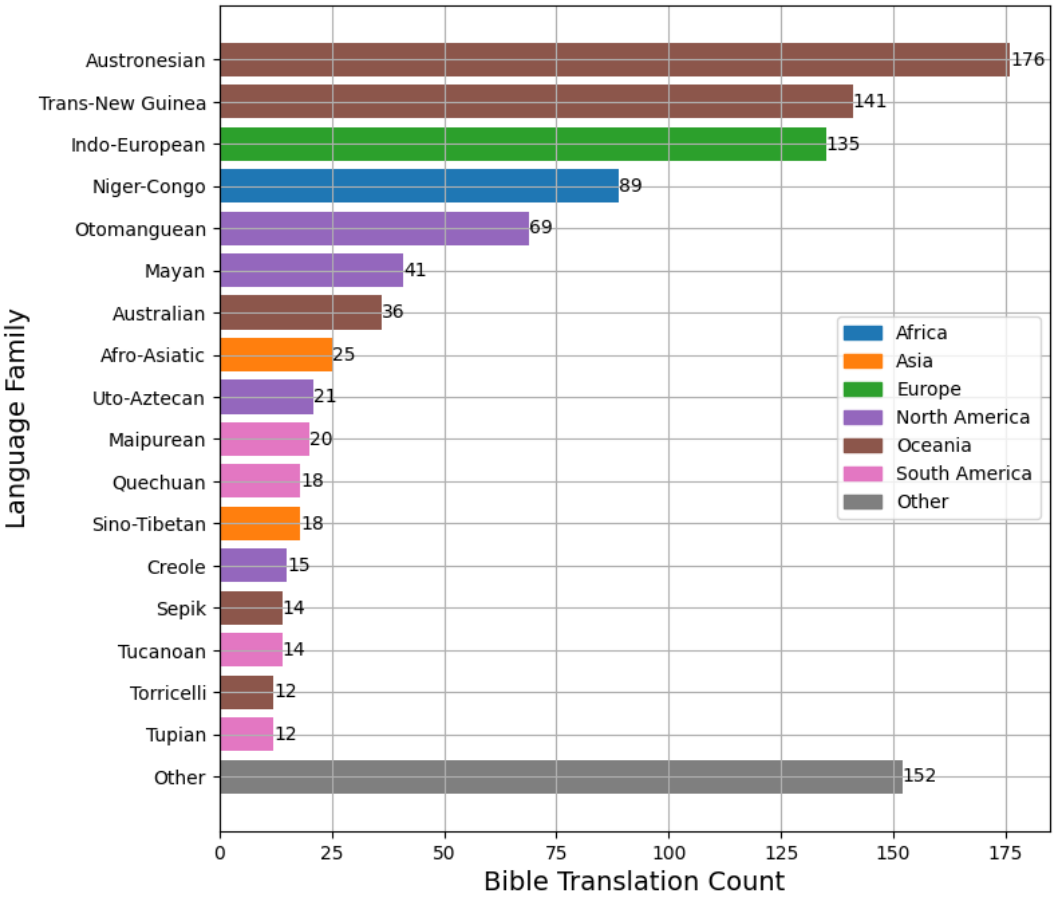


**Figure 2.** Count of Bible translations by language family.

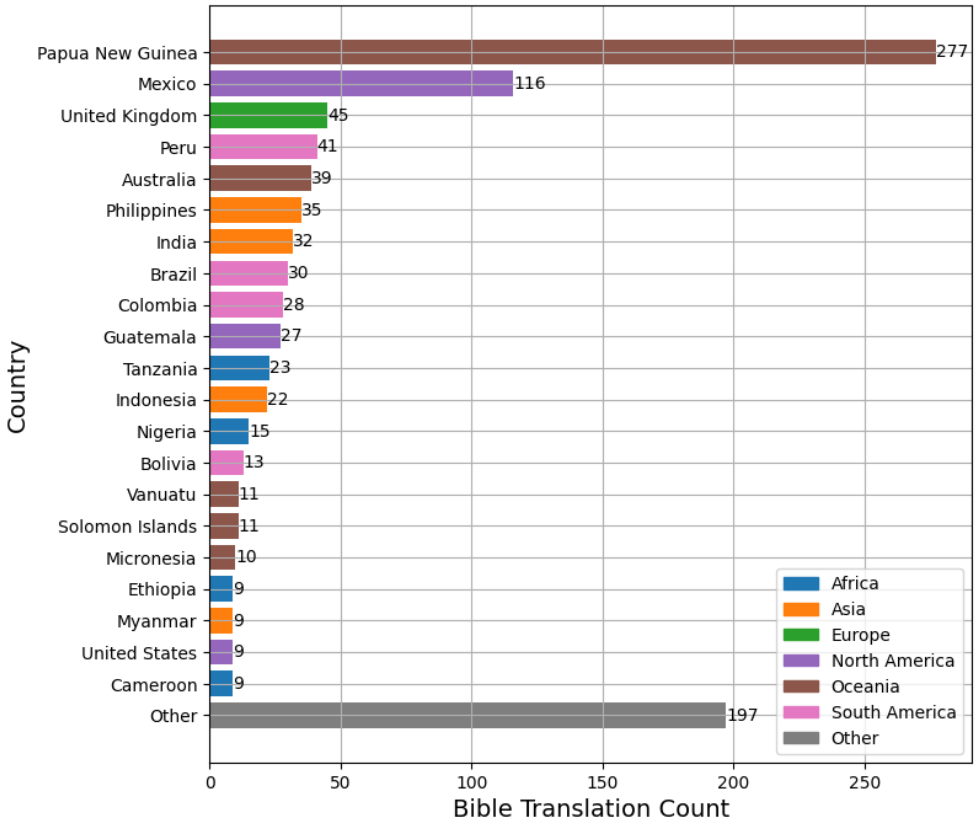


**Figure 3.** Count of Bible translations by country.

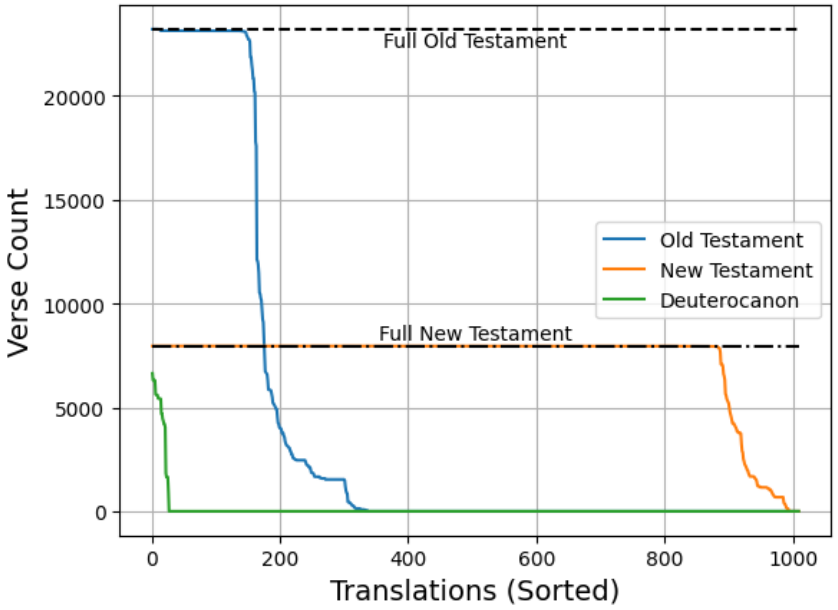


**Figure 4.** Sorted count of available verses per translation, separated by OT, NT, and Deuterocanon (DT).

### 3.2. Benchmark Translation Tasks

In addition to a standard benchmark train/test/validation splitting of our paired-verse corpora, we also benchmark model performance based on more realistic translation approaches. We chose several additional tasks motivated by plausible trajectories taken by those seeking to translate the Bible into a low-resource language. For example, the progression of a translation may begin with the Gospels, then certain epistles, followed by portions of the Pentateuch, and so on. In this fashion, we can train a model on content obtained earlier in the translation progression, then measure performance on content typically translated later. Our translation tasks include:

- *Randomized cross-validation (CV) task.* Translation pairs are delimited by Bible verses available in both source and target corpora. Due to the small size of the translation corpora (relative to those normally employed for NMT), we do not use a standard 80%/10%/10% split. We instead set aside 250 randomly selected verses for testing and validation sets respectively, with the training set being the remaining verses, and running a five-fold CV for scoring. This task will represent an upper bound for possible model translation performance, with other tasks likely to perform worse overall.
- *Gospel Translation task.* Train a model on the Gospel of Mark (MRK), test the model on the Gospel of Matthew (MAT). We selected this task because Mark is often the first book to be translated, with Matthew to follow. Some sections of Mark are also found in Matthew, however Mark is also a shorter book. For three-letter Biblical book codes, readers are directed to Appendix C. Due to the small size of the training data set (roughly 675 verses), the validation set size is reduced to 75 verses for this task.
- *Epistles Translation task.* Train a model on the Gospels (MAT, MRK, LUK, JHN) and Acts of the Apostles (ACT), test the model on the five epistles (1TH, 2TH, 1TI, 2TI, TIT, collectively abbreviated as 5T).
- *New Testament (NT) Completion task.* Train a model on the entire New Testament except Romans (ROM) and Revelation (REV), test the model on the books of Romans (ROM) and Revelation (REV). Due to their translation difficulty, these books are often among the last books of the NT to be translated.
- *Early Old Testament (OT) translation task.* Train a model on the entire NT, test the model on selected books of the OT (GEN, EXO, LEV, NUM, DEU, RUT, PSA, JON, collectively abbreviated as Early OT). These books are often the first from the Old Testament to be translated.
- *Late OT translation task.* Train a model on the entire Bible excluding minor prophets, test the model on OT minor prophets (HOS, JOL, AMO, OBA, MIC, NAH, HAB, ZEP, HAG, ZEC, MAL). These books were chosen as books that are typically among the last to be translated.
- *Related Language task.* Train a model to translate from the source language into the target language and into a related language using the same train/test splits used for the *Gospel Translation, Epistles Translation,* and *NT Completion* tasks. This task is intended to explore the potential for improving translation accuracy through the use of other completed translations.

We focus on eight specific translation pairings with a translation source and target, spanning unique language families, as described in Table 1. These translation pairings are selected to represent countries and language families with a significant number of active Bible translation projects and with reasonable representation in the eBible corpus. Within each selected language family, source and target language translations were selected to simulate the work of a Bible translation team using a reference translation from a national or gateway language as a guide for translating into their target language, with access to a related language translation for further guidance. In creating these source / target / related language translation pairings, the target language translation was selected by identifying a language family branch with multiple translations in the corpus,

preferably from the same country or in close geographic proximity [EL]. Preference was given to branches containing more total languages, some with translations and some without. Within this branch, preference was given to languages with a full Bible translation, or with a New Testament and partial Old Testament content. The next step was identifying a translation from the corpus in a national or gateway language to act as the source language translation, with priority given to more recent translations from these languages. Symmetric HMM word alignment models were trained between each candidate national / gateway language translation and a small group of candidate target language translations; the source / target translation pairing with the highest overall word alignment score was then chosen. Finally, symmetric HMM word alignment models were trained between the target language translation and the available related language translations from the language family branch. Generally, the translation with the highest overall alignment score was selected for the related language. However, in some cases a related language translation with a particularly high alignment score was excluded, based on the assumption that it may have been translated as an adaptation or using some other less generalized translation methodology. Detailed information on each translation pairing is available in Appendix C.

**Table 1.** Translation pairings for machine translation benchmarks.

| **Lang. Family (ISO-639-5)** | **Branch(es)** | **Purpose** | **Language (ISO-639-3)** | **Country** |
|---|---|---|---|---|
| Afro-Asiatic (afa) | Chadic (afa : cdc) | Source | Hausa (hau) | Nigeria |
| | | Target | Dangaléat (daa) | Chad |
| | | Related | Fulfulde, Western Niger (fuh) | Niger |
| Austronesian (map) | Malayo-Polynesian, Central Eastern Malayo-Poly. Eastern Malayo-Polynesian (map : poz : pqe) | Source | Kuanua (ksd) | Papua New Guinea |
| | | Target | Kandas (kqw) | Papua New Guinea |
| | | Related | Ramoaaina (rai) | Papua New Guinea |
| Dravidian (dra) | N/A | Source | Tamil (tam) | India |
| | | Target | Malayalam (mal) | India |
| | | Related | Kannada (kan) | India |
| Indo-European (ine) | Indo-Iranian, Indo-Aryan (ine : iir : inc) | Source | Hindi (hin) | India |
| | | Target | Eastern Panjabi (pan) | India |
| | | Related | Gujarati (guj) | India |
| Niger-Congo (nic) | Atlantic-Congo (nic : alv) | Source | Swahili (swh) | Tanzania |
| | | Target | Kwere (cwe) | Tanzania |
| | | Related | Vidunda (vid) | Tanzania |
| Otomanguean (cai : omq) | Eastern Otomanguean | Source | Spanish (spa) | Spain |
| | | Target | Zapotec, Tabaa (zat) | Mexico |
| | | Related | Tapotec, Cajonos (zad) | Mexico |
| Sino-Tibetan (sit) | N/A | Source | Nepali (npi) | Nepal |
| | | Target | Tamang, Eastern (taj) | Nepal |
| | | Related | Limbu (lif) | Nepal |
| Trans-New Guinea (paa : ngf) | N/A | Source | Tok Pisin (tpi) | Papua New Guinea |
| | | Target | Yopno (yut) | Papua New Guinea |
| | | Related | Iyo (nca) | Papua New Guinea |

Our benchmarking tasks will be approached using four different models. First, we use a SMT technique; namely, the "fast_align" package from Dyer et. al. [4] which uses a fast implementation IBM2 word alignment strategy. Second, we perform training on the OpenNMT TransformerBase architecture from Klein et. al. [7] with a SentencePiece unigram tokenization. Next, we perform fine-tuning on the "No Language Left Behind" (NLLB) model architecture from Meta [8], both the small version consisting of 600 million tunable parameters (NLLB-600M) and the medium distilled version consisting of 1.3 billion tunable parameters (NLLB-1.3B-distilled), which have pre-trained weights available from HuggingFace.

We scored models across translation tasks using three different metrics: BLEU [9], Sentence Piece BLEU (spBLEU) [10], character 3-gram F-score (chrF3) [11]. While BLEU and spBLEU scores are correlated to some extent, spBLEU more readily accounts for language variations in scripting and agglutination. Models were trained with an early stopping criteria of +0.1 BLEU over four checkpoints (1000 steps per checkpoint). Models used a batch size of 16 with four gradient accumulation steps, 4000 warm-up steps, and label smoothing of 0.2. For languages unknown to the NLLB tokenizer, we added a new language code as a special token to the tokenizer. All models were trained and evaluated on an NVIDIA A-100 with 40 GB of

available VRAM. This hardware is sufficient for fine-tuning the small and medium size NLLB architectures. For the random cross-validation task, we train five distinct models on a different train/test/validation verse pair splits. For all other tasks, only a single model of each type is trained with the identified splits.

## 4. Results

We divide our results into four sections according to tasks. We first present results for a random cross-validation task, then results for tasks specific to translating NT books, then results for tasks specific to translating OT books, and finally results for tasks specific to the use of a related language translation. Further results are available for analysis in the ebible-experiment repository [cite].

### 4.1. Cross-Validation Benchmark Task

We first consider the benchmark task for CV by verse. This effectively provides an upper bound on possible model performance on later translation tasks. We run our five-fold CV across various models and present the results in Figure 5. For the three language families shown in Table 2, we see a clear increase in performance from SMT to OpenNMT in nearly all tests, with NLLB-600M out-performing both techniques across all metrics. Likewise, the fine-tuned NLLB-1.3B-distilled architecture out-performs its smaller counterpart in all instances.

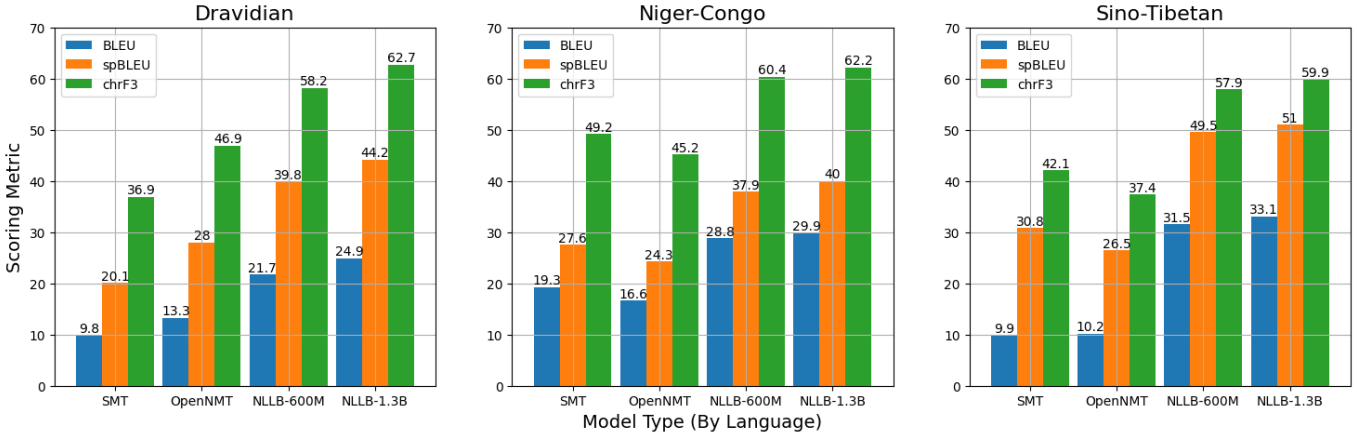


**Figure 5.** Bar chart of median BLEU, spBLEU, and chrF3 scores from SMT, NMT, and fine-tuned NLLB-600M models for the Dravidian, Niger-Congo, and Sino-Tibetan translation pairings. Note that five-fold CV is not performed on the NLLB-1.3B-distilled model due to training overheads.

We also present the test set results across all language families for the *CV* task using NLLB-600M. Figure 6 shows a bar chart of the median BLEU, spBLEU, and chrF3 scores for all eight translation pairings. Interestingly, we find no clear correlation between the scope (NT-only, NT with partial OT, or full Bible) of the translation pairing and our selected scoring metrics. The wide differences between word-level and subword-level metrics (BLEU and spBLEU) for some translation pairings such as Dravidian (+18.1), Sino-Tibetan (+18.0), and Trans-New Guinea (+17.4), compared to other translation pairings such as Austronesian (+4.3), highlight the value of examining a range of translation accuracy metrics when evaluating results on these benchmarks.

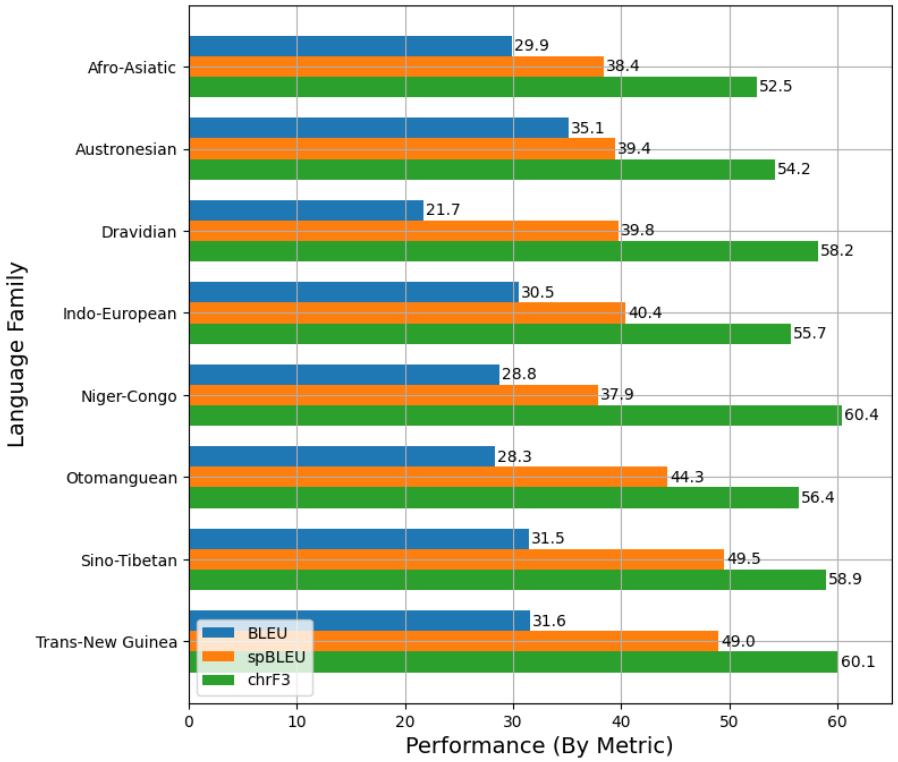


**Figure 6.** Bar chart of median BLEU, spBLEU, and chrF3 scores from a fine-tuned NLLB-600M model across eight language families for the *CV* task.

### 4.2. New Testament Benchmark Tasks

We first consider the *Gospel Translation* task of fine-tuning an NLLB-600M model using MRK as the training set and MAT as the test set. Table 5 shows the BLEU scores for this task across our eight language families, including a comparison to the *CV* task results. We see that in general, there is a drop-off in performance compared to the *CV* task for 6 of the 8 translation pairings. This is attributed to the increased data heterogeneity and larger corpus used for training during the *CV* task. These factors outweigh the benefits of the subject matter overlap between MRK and MAT. However, results from these models for MAT are significantly better than their results for Epistles (see Table 2, *Gospel Translation* portion).

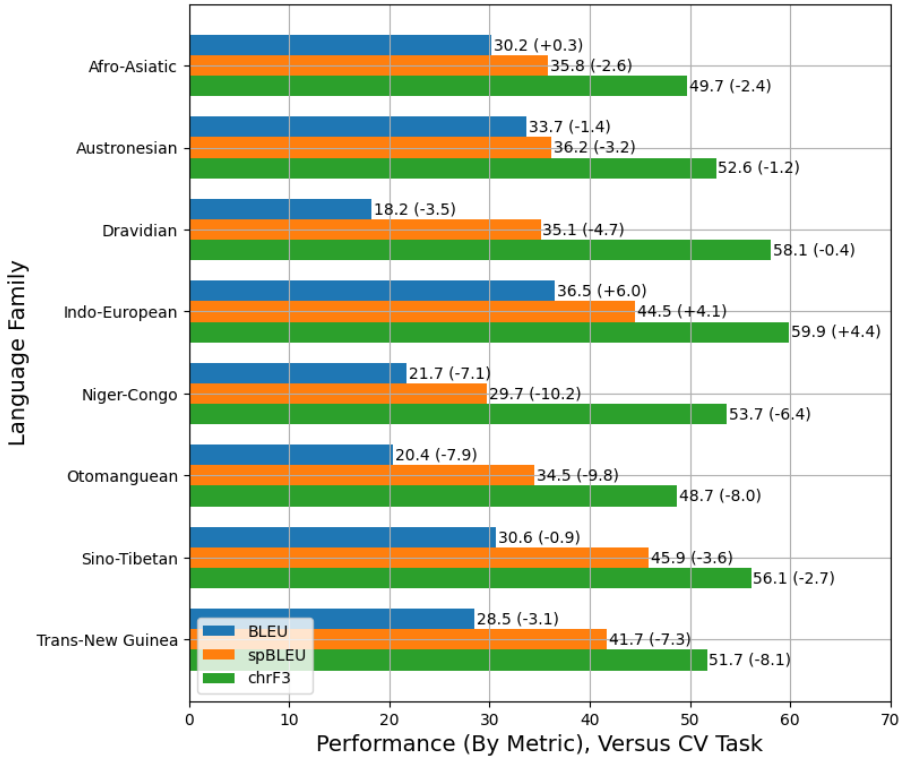


**Figure 7.** Bar chart of median BLEU, spBLEU, and chrF3 scores from a fine-tuned NLLB-600M model across eight language families for the *Gospel Translation* task, as compared to the *CV* task.

Next, we consider the *Epistle Translation* task where NLLB-600M models are either trained with MRK, or trained with the Gospels plus Acts of the Apostles, and then used to translate a selection of the Epistles (1TH, 2TH, 1TI, 2TI, and TIT). These results are summarized in Table 2, which demonstrates that models perform significantly better when translating the selected Epistles when their training sets include these four additional books (MAT, LUK, JHN, ACT).

**Table 2.** BLEU scores for NLLB-600M fine-tuned models: *Gospel Translation* and *Epistle Translation* tasks.

| Translation Pairing | Gospel Translation | | | | | Epistle Translation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1TH | 2TH | 1TI | 2TI | TIT | 1TH | 2TH | 1TI | 2TI | TIT |
| Afro-Asiatic | 6.3 | 5.5 | 4.8 | 6.9 | 2.9 | 14.8 (+6.5) | 17.3 (+11.8) | 11.3 (+6.5) | 14.3 (+7.4) | 9.3 (+6.4) |
| Austronesian | 12.7 | 15.4 | 13.3 | 13.3 | 13.9 | 20.1 (+7.4) | 21.5 (+6.1) | 19.9 (+6.6) | 23.5 (+10.2) | 19.8 (+5.9) |
| Dravidian | 10.1 | 12.0 | 4.2 | 7.4 | 2.0 | 11.4 (+1.3) | 13.4 (+1.4) | 7.0 (+2.8) | 8.1 (+0.7) | 5.2 (+3.2) |
| Indo-European | 24.1 | 22.5 | 17.1 | 22.3 | 15.5 | 28.7 (+4.6) | 26.2 (+3.7) | 21.3 (+4.2) | 26.7 (+4.4) | 21.6 (+6.1) |
| Niger-Congo | 9.4 | 13.4 | 10.7 | 10.8 | 7.3 | 14.9 (+5.5) | 16.7 (+3.3) | 15.1 (+4.4) | 14.5 (+3.7) | 13.6 (+6.3) |
| Otomanguean | 6.5 | 7.4 | 8.4 | 8.4 | 6.9 | 16.2 (+9.7) | 14.0 (+6.6) | 15.5 (+7.1) | 13.1 (+4.7) | 15.6 (+8.7) |
| Sino-Tibetan | 13.2 | 10.6 | 9.4 | 10.3 | 8.7 | 18.8 (+5.6) | 17.2 (+6.6) | 16.2 (+6.8) | 18.2 (+7.9) | 18.3 (+9.6) |
| Trans-New Guinea | 6.9 | 8.4 | 7.2 | 7.6 | 6.5 | 18.7 (+11.8) | 20.5 (+12.1) | 16.9 (+9.7) | 17.6 (+10.0) | 15.0 (+8.5) |

Finally, we consider the *NT Completion* task using a fine-tuned NLLB-600M model. We fine-tune models first on MRK only, then on the Gospels plus ACT, then on the entire New Testament except Romans (ROM) and Revelation (REV). In this fashion we form training sets with corpora of increasing size and literary breadth. For each translation pairing, BLEU scores for both ROM and REV increase across these three tasks, indicating that the increased size and literary breadth of the training set is beneficial for the translation of these challenging NT books.

**Table 3.** BLEU scores for NLLB-600M fine-tuned models: *Gospel Translation*, *Epistle Translation*, *NT Completion* tasks.

| Translation Pairing | Gospel Translation | | Epistle Translation (vs Gospel Translation) | | NT Completion (vs Epistle Translation) | |
|---|---|---|---|---|---|---|
| | ROM | REV | ROM | REV | ROM | REV |
| Afro-Asiatic | 7.0 | 7.6 | 14.1 (+7.1) | 15.1 (+7.5) | 18.9 (+4.8) | 16.0 (+0.9) |
| Austronesian | 11.5 | 15.3 | 18.1 (+6.6) | 26.2 (+10.9) | 23.2 (+5.1) | 28.6 (+2.4) |
| Dravidian | 8.2 | 9.7 | 10.9 (+2.7) | 13.0 (+3.3) | 12.8 (+1.9) | 13.7 (+0.7) |
| Indo-European | 23.2 | 21.7 | 26.5 (+3.3) | 26.5 (+4.8) | 30.0 (+3.5) | 29.2 (2.7) |
| Niger-Congo | 11.4 | 11.4 | 16.9 (+5.5) | 20.3 (+8.9) | 20.1 (+3.2) | 23.5 (+3.2) |
| Otomanguean | 7.6 | 9.8 | 15.2 (+7.6) | 19.0 (+10.2) | 22.1 (+6.9) | 20.6 (+1.6) |
| Sino-Tibetan | 12.0 | 13.5 | 19.5 (+7.5) | 21.2 (+7.7) | 23.3 (+3.8) | 23.3 (+2.1) |
| Trans-New Guinea | 8.8 | 7.4 | 18.5 (+9.7) | 18.3 (+10.9) | 24.9 (+6.4) | 22.1 (+3.8) |

**4.3. Old Testament Benchmark Tasks**

For the *Early OT* translation task, the entire NT is used as the training set; the fine-tuned model is then evaluated on various books that are typically translated early in an Old Testament translation project (GEN-DEU, RUT, PSA, JON). Figure 8 shows the BLEU scores of NLLB-600M models trained across five translation pairs on the NT, with translations performed across books in the Early OT as the test set. Empty cells indicate a lack of an available translation for a particular book in the target language. For the tested Early OT books, the highest BLEU scores within a translation pair are observed for GEN, while the lowest scores are observed for LEV, NUM, or DEU.

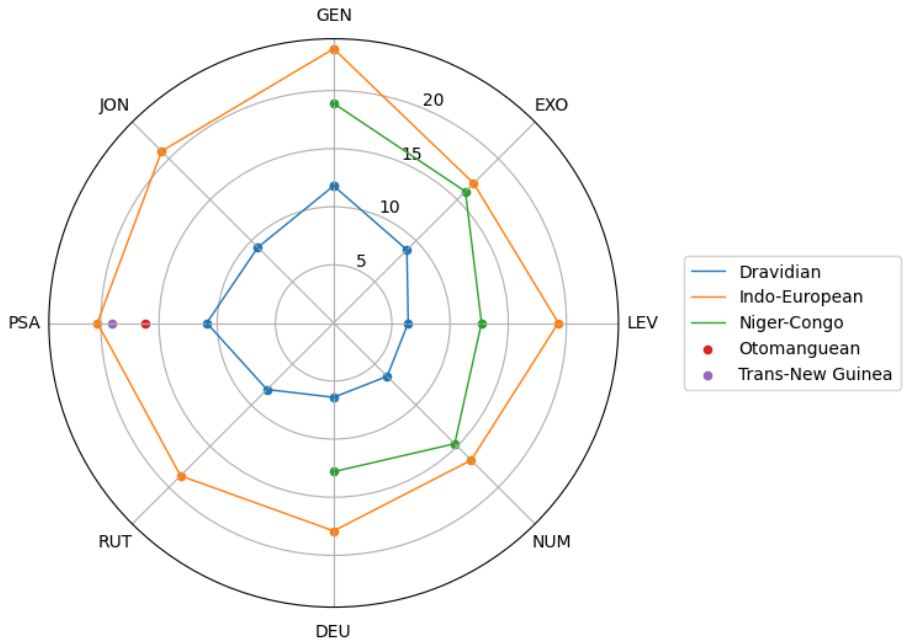


**Figure 8.** Radar plot of NLLB-600M BLEU scores for five language families for the *Early OT* task. Missing points for Niger-Congo, Otomanguean, and Trans-New Guinea indicate books unavailable in the target language translation.

Next, we assess the performance of NLLB-600M on the *Late OT* translation task. The books in this test set (the minor prophets) tend to be among the last books of the Bible to be translated. Figure 9 shows the BLEU scores for OT minor prophet books across Dravidian and Indo-European language families, both in the case of models trained on the entire NT and models trained on the entire Bible (excluding the minor prophets). As before, a larger corpus of training data led to a large improvement in BLEU scores.

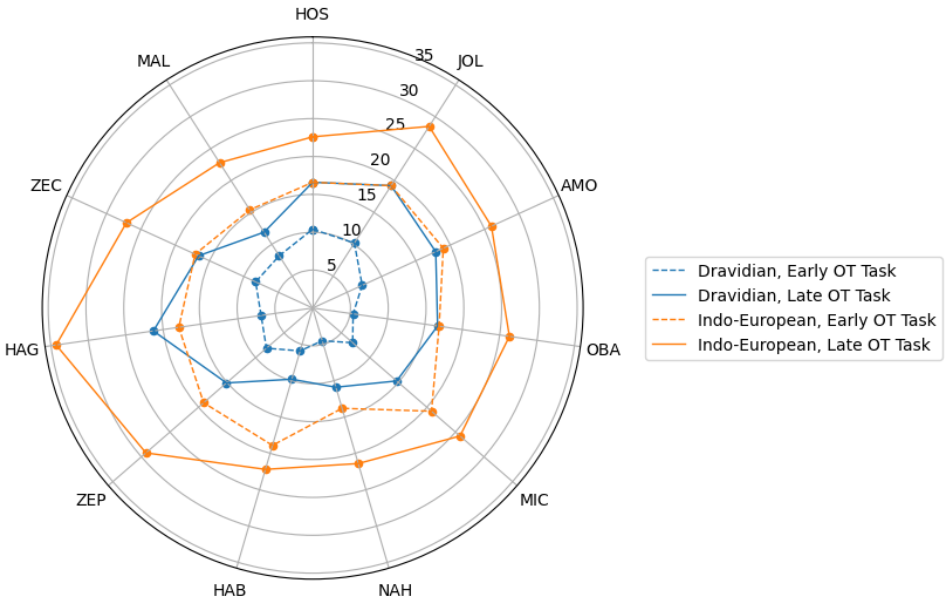


**Figure 9.** Radar plot of NLLB-600M BLEU scores for Dravidian and Indo-European for *Early* and *Late OT* tasks.

### 4.4. Related Language Benchmark Tasks

For the *Related Language* task, an additional translation is selected from the eBible corpus for each of the eight language families. HMM word alignment models are trained between the target language and each related language translation in the eBible corpus from the same branch of the language family; the translation with the best alignment to the target language is selected as the related language translation. Then, the *Gospel Translation, Epistle Translation,* and *NT Completion* tasks are repeated using both the target language and the related language on the target side of the model. For each task, the related language training data included the same verse pairs used for the target language training data; additionally, the related language training data included the verse pairs from the target language test set.

Figure 10 compares the BLEU score deltas for the *Related Language* version of each task compared to the original version of the task. BLEU score deltas across the eight language families are widely divergent, ranging from -2.9 to +3.2 BLEU (*Gospel Translation* task), -4.0 to +10.6 BLEU (*Epistle Translation* task), and -2.7 to +11.5 BLEU (*NT Completion* task). Results for

the Austronesian and Niger-Congo language families are strongest, while results for the Afro-Asiatic, Indo-European, Sino-Tibetan and Trans-New Guinea language families are the weakest. These BLEU score deltas correlate well with the HMM alignment scores between the target language and related language, with the Austronesian (0.68) and Niger-Congo (0.49) translation pairings exhibiting the highest alignment, and the Afro-Asiatic (0.21), Indo-European (0.30), Sino-Tibetan (0.27) and Trans-New Guinea (0.27) exhibiting the lowest alignment among the selected translation pairings.

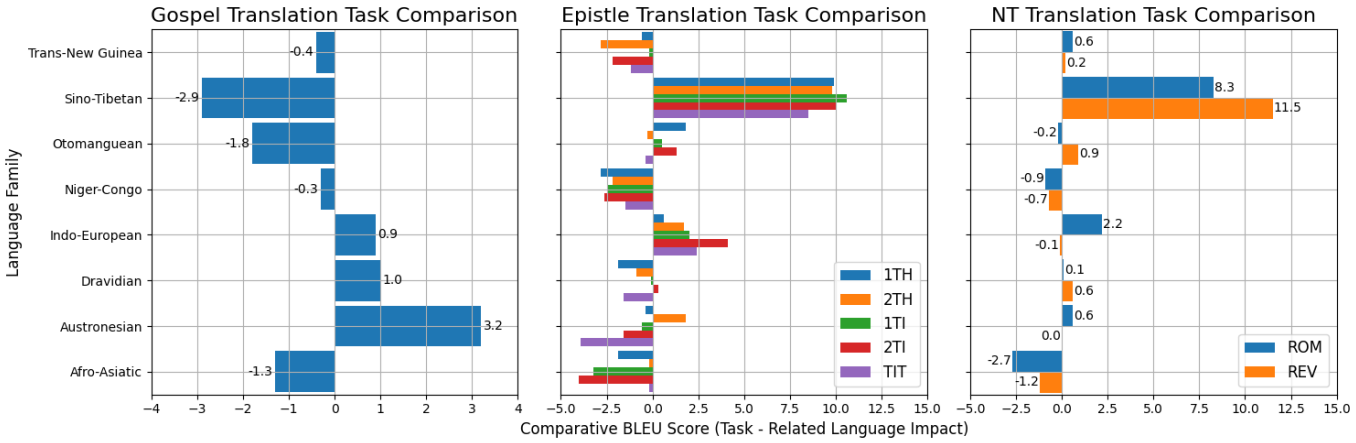


**Figure 10.** Impact of a Related Language Translation on the *Gospel Translation*, *Epistle Translation*, and *NT Completion* tasks.

## 5. Discussion

On the *CV* benchmark task, median accuracy metrics across the eight selected language families varied widely, ranging from 21.7 - 35.1 (BLEU), 37.9 - 49.5 (spBLEU), and 52.5 - 60.4 (chrF3). No individual characteristic of the selected languages and translations correlated closely with the distribution of these results for any of these three accuracy metrics. For instance, the scope of the source / target translation pairings varied from NT-only, to NT with OT portions, to full Bible translations; while the highest BLEU score was seen for an NT-only pairing (Austronesian (35.1)) and the lowest BLEU score was seen for a full Bible pairing (Dravidian (21.7)), other NT-only pairings (Afro-Asiatic (29.9); Sino-Tibetan (31.5)) and full Bible pairings (Indo-European (30.5)) scored comparably. Similar diversity was seen when comparing spBLEU and chrF3 scores to the scope of the translation pairings. HMM word alignment scores for the source / target translation pairings also do not correlate closely with the accuracy metrics on this benchmark; translation pairings with the highest (Niger-Congo (0.44)) and lowest (Otomanguean (0.19)) word alignment scores resulted in comparable BLEU scores (28.8 and 28.3, respectively).

NLLB characterized the resource level of each supported language as either high or low, with low resource languages being trained on less than 1M bitexts. In our *CV* benchmark, the source languages were a mix of high resource (Hindi, Swahili, Spanish), low resource (Hausa, Tamil, Nepali, Tok Pisin), and unsupported (Kuanua). However, there was no clear correlation between the NLLB resource level of the source language and the resulting metrics for the translation pairing. Similarly, while the best chrF3 results were seen for translation pairings with languages using the Latin script (Niger-Congo (60.4); Trans-New Guinea (60.1)), which is well-represented in the NLLB vocabulary, other Latin script translation pairings performed relatively poorly (Afro-Asiatic (52.5); Austronesian (54.2)).

While BLEU is a widely used, language agnostic metric for assessing machine translation accuracy, the fact that it is a word-level metric means that it can be difficult to interpret the metric across languages, particularly when attempting to judge the usefulness of a translation model for a less well known language. Combining a word-level metric (e.g., BLEU) with a subword-level metric (e.g., spBLEU) and a character-level metric (e.g., chrF3) provides a more nuanced view.

Table 4 presents several sample predictions from the *CV* task with median BLEU, spBLEU, and chrF3 scores for their respective model. The predictions are color-coded at the word level to give a general sense of the accuracy of each prediction, and suggest that, although the NLLB models represent a strong improvement over earlier MT approaches, further improvement is needed.

**Table 4.** Sample Predictions With Median Verse Accuracy, *CV* Task Models
(Legend: Correct, Partially Correct, Moved, Incorrect)

| | |
|---|---|
| Afro-Asiatic<br>GAL 4:21 | **Source**: Ku gaya mini, ku da kuke so ku zauna a ƙarƙashin Doka, ba ku san abin da Doka ta ce ba ne?<br>**Target**: Kaawondu ja! Kuŋ kuuk raka goye ɗo serre ta gaanuundi, ku dorgitɗo kaaw ta gaanuundi walla?<br>**Prediction**: Kuŋ kuuk raka goye ɗo serpey ta gaanuundi, kaawoŋdu, ku ibingig gaanuun okiŋji walla?<br>**BLEU**: 29.6; **spBLEU**: 38.1; **chrF3**: 57.4; **WER**: 44.0; **TER**: 50.0 |
| Niger-Congo<br>EXO 5:9 | **Source**: Fanyeni kazi kuwa ngumu zaidi kwa watu hao ili kwamba wakazane na kazi na kuacha kusikiliza uongo."<br>**Target**: Wongezeni wanhu wano sang'hano ndala muladi wagaye na waleke kuhulika mbuli za uvwizi."<br>**Prediction**: Watendeleni wanhu wawo sang'hano ndala muladi wadahe kuiguma muna isang'hano yawo na |

| | |
|---|---|
| | waleke kutegeleza mbuli za uvwizi."<br>**BLEU**: 28.5; **spBLEU**: 38.4; **chrF3**: 70.0; **WER**: 48.3; **TER**: 61.5 |
| Otomanguean PSA 112:4 | **Source**: En las tinieblas resplandece luz a los rectos: Él es clemente, misericordioso y justo.<br>**Target**: Dxusaní' tu xi' lu da chul-la naga zaj zra' benne' li lazre' ca', benne' ca' zaj naquę' zren lazre', ne xrlátaje, ne dxelexache lazrę' bénneache.<br>**Prediction**: Lu da chul-la tu xi' dxusení na benne' ca' zaj naquę' xrlátaje. Dios naquę' zren lazre', ne dxexache lazrę' bénneache cheę', ne naquę' xrlátaje.<br>**BLEU**: 28.5; **spBLEU**: 38.4; **chrF3**: 70.0; **WER**: 48.3; **TER**: 61.5 |

Evaluating median verse accuracy across a range of word, subword and character-level metrics can help to establish a broad intuition about the usefulness of a translation model, but it is also important to evaluate the distribution of these metrics. Generally, the distribution can be quite broad, as shown in Figure 11 for the Afro-Asiatic, Niger-Congo, and Otomanguean *CV* models. When the same or similar verse text occurs in multiple passages (e.g., the parables in the Gospels), or when the verse text follows a repeating pattern ("from the tribe of Joseph, …"; "from the tribe of Dan, …"), accuracy can be relatively higher. Translations for longer, more complex verses tend to be relatively lower. In the context of Bible translation, presenting the model's confidence level to the translator may be as helpful as presenting the suggested verse text. Augmenting the model's predictions with external evaluation metrics may also be helpful for focusing the translator's attention on low-confidence verse drafts. Empirically collected data on challenging verses could also provide a valuable means of focusing the translator's efforts.

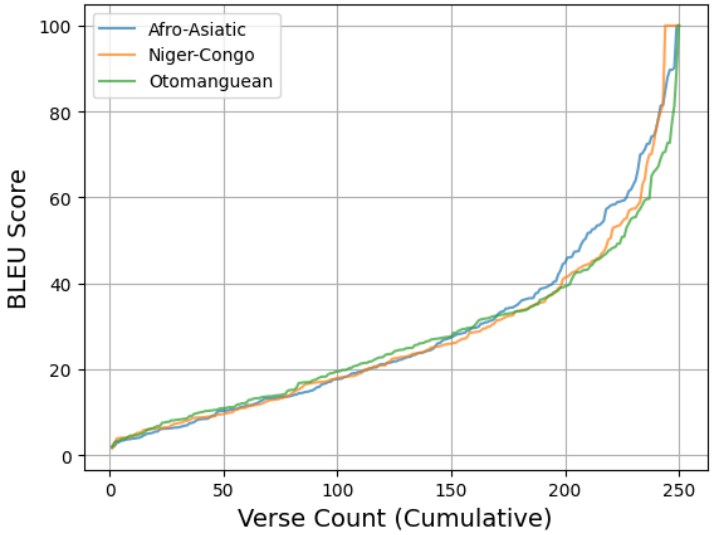


**Figure 11.** Cumulative Per-Verse BLEU Score Distribution for Afro-Asiatic, Niger-Congo, Otomanguean Models on *CV* task.

## 6. Future Work

We envision a number of opportunities to expand and improve on this current work. From a linguistic point of view, our results open up questions about how spBLEU scoring might capture agglutination and scripting more effectively than a BLEU score, as well as broader questions about how the viability of scoring methods might change based on language morphology. Furthermore, it is not usually clear why certain language families seem more amenable to fine-tuning than others using NLLB (as measured by BLEU scores). It is also unclear whether larger NLLB models (such as the next largest architecture with 3.3B tunable parameters) will lead to continued improvements, or yield diminishing returns.

Our work also highlights potential benefits from incorporating multiple Bible translations during model training, although the mixed results indicate that more sophisticated strategies should be investigated, while considering factors such as language and translation characteristics (e.g. translation age, reading level, style) and other metrics (e.g. word alignment scores, subword evenness). Automated methods for selecting the best source and related language translations may improve performance over the heuristics used in this work.

In this work, changes to the NLLB tokenizer were limited to the introduction of special tokens for new language codes. Some of the translations, such as the Yopno (yut-yut) and Zapotec Tabaa (zat-zatNTps) translations, included a number of characters that were not known to the NLLB tokenizer, slightly reducing the accuracy of the fine-tuned model. Other translations, such as the Limbu translation in Limbu script (lif-lifNT) and the Greek SBL (grc-grcsbl) included many unknown characters, and prevented their use in this work. Effective methods for augmenting NLLB and the NLLB tokenizer for new scripts and characters will be necessary to achieve the full benefits of the eBible corpus.

Moreover, given the limitations of available data and compute resources for target communities there is a gap in developing

techniques that overcome the ‘low-resource double bind’ (cite https://aclanthology.org/2021.findings-emnlp.282). Towards this end, implementing and studying efficient fine-tuning methods (cite the newer work in this table) would provide significant support for the more advanced neural models to be useful for translators on the ground.

The wide distribution of per-verse translation accuracy indicates the need for effective human-in-the-loop strategies to support the Bible translator, ensuring that these differences are accurately and intuitively presented to guide their work.

CONTENT TBD. Points include:

- Use of Greek / Hebrew source texts as the source language translation. Modern Greek and Hebrew already part of FLORES-200.

## 7. Conclusion

In this work we present the eBible corpus: an open-source NLP-ready dataset of over a thousand partial and full Bible translations spanning more than 800 languages. With the aim of leveraging machine translation to aid expert Bible translators, we introduce a number of benchmark tasks based on the translation ordering often used by these experts. These tasks include a randomized CV along with NT- and OT-specific objectives. In addition to benchmark tasks, we provide benchmark model results for selected language families, with models ranging from SMT methods to a fine-tuned NLLB architecture.

## 8. Contributions

J.M. cast the vision and along with M.M. organized the work on the eBible corpus within PAB-NLP. V.A wrote the initial scripts to download the corpus from eBible.org while D.B. improved upon these. U.H. provided significant support for pre-processing through manual and automatic data checks and cleanup. M.M, J.M, T.J, D.B and M.S designed the benchmark experiments. These were run and reported by M.M., D.B and D.D. M.S., M.M., U.H. and J.M. wrote and edited the manuscript.

## 9. Acknowledgements

The authors would like to thank their respective organizations for encouraging their participation in this cross-organization collaboration. They would also like to acknowledge the foundational and ongoing work of Michael Johnson in creating and maintaining the eBible.org site. Finally, we thank the ETEN Innovation Lab (EIL) for sponsoring the computing resources used in this project.

## 10. Conflicts of Interest

The authors have no conflicts of interest to declare.

## 11. References

[1] P. Resnik, M. B. Olsen, and M. Diab, “The Bible as a Parallel Corpus: Annotating the ‘Book of 2000 Tongues,’” *Comput. Humanit.*, vol. 33, no. 1/2, pp. 129–153, 1999, doi: 10.1023/A:1001798929185.

[2] C. Christodouloupoulos and M. Steedman, “A massively parallel corpus: the Bible in 100 languages,” *Lang. Resour. Eval.*, vol. 49, no. 2, pp. 375–395, Jun. 2015, doi: 10.1007/s10579-014-9287-y.

[3] A. McCarthy *et al.*, “The Johns Hopkins University Bible Corpus: 1600+ Tongues for Typological Exploration,” *Proc. Twelfth Lang. Resour. Eval. Conf.*, pp. 2884–2892, 2020.

[4] C. Dyer, V. Chaheuneau, and N. Smith, “A simple, fast, and effective reparameterization of IBM model 2.,” *Proc. 2013 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.*, 2013.

[5] Yonggang Deng and W. Byrne, “HMM Word and Phrase Alignment for Statistical Machine Translation,” *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 3, pp. 494–507, Mar. 2008, doi: 10.1109/TASL.2008.916056.

[6] P. Koehn, *Statistical machine translation.* Cambridge University Press, 2009.

[7] G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush, “OpenNMT: Open-Source Toolkit for Neural Machine Translation,” 2017, doi: 10.48550/ARXIV.1701.02810.

[8] NLLB Team *et al.*, “No Language Left Behind: Scaling Human-Centered Machine Translation,” 2022, doi: 10.48550/ARXIV.2207.04672.

[9] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[10] N. Goyal *et al.*, “The Flores-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation,” *Trans. Assoc. Comput. Linguist.*, vol. 10, pp. 522–538, May 2022, doi: 10.1162/tacl_a_00474.

[11] M. Popović, “chrF: character n-gram F-score for automatic MT evaluation,” in *Proceedings of the tenth workshop on statistical machine translation*, 2015, pp. 392–395.

OTHER CITATIONS

Winston Wu, Nidhi Vyas and David Yarowsky, 2018. Creating a Translation Matrix of the Bible’s Names Across 591 Languages. In Proceedings *of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Sami Liedes. 2018. Data and general life geekery. https://samiliedes.wordpress.com/2018/03/07/machine-translating-the-bible-into-new-languages/ Accessed 16/11/2022

Sami Liedes. https://github.com/sliedes/fairseq-py Accessed 16/11/2022

https://christos-c.com/bible/

Speecon Consortium. (2014). Dutch Speecon Database. Speecon Project, distributed via ELRA, Speecon resources, 1.0, ISLRN 613-489-674-355-0.

## Appendix A: Characteristics of Selected Bible Translations

*TODO: Fix the section numbering in these appendices.*
*TODO: Include reference to Ethnologue as source for Script/Typology/Country information.*

**Table A.1.** Bible Translation Characteristics

| Language Family | Purpose | Language | ISO-639-3 | eBible Translation | Scope | Verses | Script | Typology | Country |
|---|---|---|---|---|---|---|---|---|---|
| Afro-Asiatic | Source | Hausa | hau | hau-hausa.txt | Bible | 31,082 | Latn | SVO | Nigeria |
| | Target | Dangaléat | daa | daa-daaNT.txt | NT | 7,957 | Latn | Unk | Chad |
| | Related | Fulfulde, Western Niger | fuh | fuh-fuhbkf.txt | NT | 7,57 | Latn | SVO | Niger |
| Austronesian | Source | Kuanua | ksd | ksd-ksd.txt | Bible | 31,098 | Latn | Unk | PNG* |
| | Target | Kandas | kqw | kqw-kqw.txt | NT | 7,957 | Latn | Unk | PNG |
| | Related | Ramoaaina | rai | rai-rai.txt | NT | 7,957 | Latn | SVO | PNG |
| Dravidian | Source | Tamil | tam | tam-tam2017.txt | Bible | 31,099 | Taml | SOV | India |
| | Target | Malayalam | mal | mal-mal.txt | Bible | 31,089 | Mlym | SOV | India |
| | Related | Kannada | kan | kan-kan2017.txt | Bible | 31,099 | Knda | SOV | India |
| Indo-European | Source | Hindi | hin | hin-hin2017.txt | Bible | 31,099 | Deva | SOV | India |
| | Target | Eastern Panjabi | pan | pan-pan.txt | Bible | 31,099 | Gurm | SOV | India |
| | Related | Gujarati | guj | guj-guj2017.txt | Bible | 31,099 | Gujr | SOV | India |
| Niger-Congo | Source | Swahili | swh | swh-swhonen.txt | Bible | 31,098 | Latn | SVO | Tanzania |
| | Target | Kwere | cwe | cwe-cwe.txt | GEN-DEU, NT | 13,806 | Latn | Unk | Tanzania |
| | Related | Vidunda | vid | vid-vid.txt | GEN-DEU, NT | 13,809 | Latn | Unk | Tanzania |
| Otomanguean | Source | Spanish | spa | spa-sparvg.txt | Bible | 31,097 | Latn | SVO | Spain |
| | Target | Zapotec, Tabaa | zat | zat-zatNTps.txt | PSA, NT | 10,416 | Latn | VSO | Mexico |
| | Related | Tapotec, Cajonos | zad | zad-zadNT.txt | NT | 7,957 | Latn | VSO | Mexico |
| Sino-Tibetan | Source | Nepali | npi | npi-npiulb.txt | Bible | 31,099 | Deva | SOV | Nepal |
| | Target | Tamang, Eastern | taj | taj-taj.txt | NT | 7,957 | Deva | SOV | Nepal |
| | Related | Limbu | lif | lif-lifNT2.txt | NT | 7,957 | Deva | SOV | Nepal |
| Trans-New Guine | Source | Tok Pisin | tpi | tpi-tpiOTNT.txt | Bible | 31,099 | Latn | SOV | PNG |
| | Target | Yopno | yut | yut-yut.txt | PSA, NT | 10,417 | Latn | SOV | PNG |
| | Related | Iyo | nca | nca-nca.txt | NT | 7,957 | Latn | SOV | PNG |

*PNG is an abbreviation for Papua New Guinea.

## Appendix B: NLLB Language Codes

**Table B.1.** NLLB Language Codes

| Language Family | Language | ISO-639-3 | NLLB Language Code |
|---|---|---|---|
| Afro-Asiatic | Hausa | hau | hau_Latn |

| | | | |
|---|---|---|---|
| | Dangaléat | daa | (!) daa_Latn |
| | Fulfulde, Western Niger | fuh | (!) fuh_Latn |
| Austronesian | Kuanua | ksd | (!) ksd_Latn |
| | Kandas | kqw | (!) kqw_Latn |
| | Ramoaaina | rai | (!) rai_Latn |
| Dravidian | Tamil | tam | tam_Taml |
| | Malayalam | mal | mal_Mlym |
| | Kannada | kan | kan_Knda |
| Indo-European | Hindi | hin | hin_Deva |
| | Eastern Panjabi | pan | pan_Gurm |
| | Gujarati | guj | guj_Gujr |
| Niger-Congo | Swahili | swh | swh_Latn |
| | Kwere | cwe | (!) cwe_Latn |
| | Vidunda | vid | (!) vid_Latn |
| Otomanguean | Spanish | spa | spa_Latn |
| | Zapotec, Tabaa | zat | (!) zat_Latn |
| | Tapotec, Cajonos | zad | (!) zad_Latn |
| Sino-Tibetan | Nepali | npi | npi_Deva |
| | Tamang, Eastern | taj | (!) taj_Deva |
| | Limbu | lif | (!) lif_Deva |
| Trans-New Guinea | Tok Pisin | tpi | tpi_Latn |
| | Yopno | yut | (!) yut_Latn |
| | Iyo | nca | (!) nca_Latn |

(!) indicates a language code not supported by NLLB

## Appendix C: Source / Target / Related Language Alignment Scores (HMM)

### Appendix C.1. Afro-Asiatic

For the Afro-Asiatic language family, the *Chadic > Biu-Mandara* branch contains 79 languages (4 translations in the corpus) and the *Chadic > East* branch contains 36 languages (1 translation in the corpus). These five translations only contain the New Testament portion of the Bible, and are languages spoken in either Cameroon or Chad. English and French are national languages for Cameroon. French and Arabic are national languages for Chad. The best alignment results were achieved using a Hausa translation (hau-hausa) with the Dangaléat translation (daa-daaNT). The Western Niger Fulfulde translation (ful-fuhbkf) aligned best with the Dangaléat translation.

**Table C.1.** Source / Target / Related Language Alignment Scores (Afro-Asiatic).

| **Target Language(s)** | | **National/Gateway Language(s)** | | | | **Related Language(s)** | |
|---|---|---|---|---|---|---|---|
| **Language** | **Translation** | **eng-engULB** | **fra-frasbl** | **hau-hausa** | **hau-hauulb** | **ffm-ffm** | **fuh-fhubkf** |
| Hdi | xed-xed | 0.1792 | 0.1691 | 0.1860 | 0.1736 | 0.1915 | 0.1982 |
| Mbuko | mqb-mqbNT | 0.1366 | 0.1291 | 0.1465 | 0.1298 | 0.1560 | 0.1613 |
| Merey | meq-meq | 0.1422 | 0.1368 | 0.1546 | 0.1376 | 0.1598 | 0.1640 |
| Muyang | muy-muy | 0.1513 | 0.1451 | 0.1537 | 0.1403 | 0.1818 | 0.1862 |
| Dangaléat | daa-daaNT | 0.1912 | 0.1760 | **0.1929** | 0.1705 | 0.2044 | **0.214**8 |

### Appendix C.2. Austronesian

For the Austronesian language family, the *Malayo-Polynesian > Central-Eastern Malayo-Polynesian > Eastern Malayo-Polynesian > Oceanic* branch contains 513 languages. Of these 513 languages, the eBible corpus contains 4 translations from the *Western Oceanic* sub-branch. All 4 of these are for languages spoken in Papua New Guinea (PNG). Three of these translations contain the New Testament only, and one contains the New Testament and a portion of the Old Testament.(4 translations in the corpus) and the *Chadic > East* branch contains 36 languages (1 translation in the corpus).

English and Tok Pisin (tpi) are national languages in PNG, while Dobu (dob), Kuanua (ksd), Suau (swp) and Tawala (tbo) are gateway languages from the *Malayo-Polynesian* branch with translations in the corpus.
Among these translations, the Kuanua translation (ksd-ksd) was selected as the source and the Kandas translation (kqw-kqw) was selected as the target due to their strong alignment score. The Label (lbb-lbb) and Ramoaaina (rai-rai) translations both aligned well with the Kandas translation (kqw-kqw); Ramoaaina was selected as the related language translation.

**Table C.2.** Source / Target / Related Language Alignment Scores (Austronesian).

| **Target Language(s)** | | **National/Gateway Language(s)** | | | | | **Related Language(s)** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Language** | **Translation** | **tpi-tpiOTNT** | **dob-dob** | **swp-swp** | **tbo-tbo** | **ksd-ksd** | **kqw-kqw** | **lbb-lbb** | **gfk-gfk** | **rai-rai** | **sgq-sgq** |
| Fanamarket | bjp-bjp | 0.2581 | 0.1761 | 0.2068 | 0.1990 | 0.2701 | 0.3994 | 0.4134 | 0.3025 | 0.3974 | 0.2576 |
| Kandas | kqw-kqw | 0.2213 | 0.1629 | 0.2003 | 0.1897 | **0.3303** | | **0.6847** | 0.2910 | **0.6849** | 0.2516 |
| Label | lbb-lbb | 0.2152 | 0.1657 | 0.2127 | 0.1979 | 0.3229 | | | 0.2789 | 0.6583 | 0.2538 |
| Patpatar | gfk-gfk | 0.1966 | 0.1462 | 01662 | 0.1567 | 0.2707 | | | | 0.2938 | 0.1919 |
| Ramoaaina | rai-rai | 0.2120 | 0.1613 | 0.2069 | 0.1874 | 0.3916 | | | | | 0.2475 |
| Sursurunga | sgq-sgq | 0.1775 | 0.1353 | 0.1440 | 0.1498 | 0.1692 | | | | | |

**Appendix C.3. Dravidian**

For the Dravidian language family (85 total languages), the *Southern > Tamil-Kannada* branch contains 31 languages. Of these 31 languages, the eBible corpus contains 5 translations, 1 from the *Kannada* sub-branch and 4 from the *Tamil-Kodagu* sub-branch. All 5 of these are for languages spoken in India, and each translation is a full Bible translation. Each of these translations is for a national language of India (Kannada, Malayalam, Tamil, and Telugu). There are no translations for low-resource languages from this language family and geography available in the corpus. As a result, the Tamil (tam-tam2017) and Malayalam (mal-mal) translations were chosen for the source and target translation pairing, with the Kannada translation (kan-kan2017) as the related language translation.

**Table C.3.** Source / Target / Related Language Alignment Scores (Dravdian).

| **Target Language(s)** | | **National/Gateway Language(s)** | **Related Language(s)** | | |
|---|---|---|---|---|---|
| **Language** | **Translation** | **hin-hin2017** | **kan-kan2017** | **mal-mal** | **mal-malc** |
| Tamil | tam-tam2017 | 0.2063 | **0.3466** | **0.4295** | 0.3396 |
| Telugu | tel-tel2017 | 0.2097 | 0.3423 | 0.3229 | 0.3068 |

**Appendix C.4. Indo-European**

For the Indo-European language family, the *Indo-Iranian > Indo-Aryan* branch contains 220 languages, with 92 languages in the *Intermediate > Western* sub-branch, 94 languages in the *Outer* sub-branch, and 11 languages in the *Western Hindi* sub-branch. The eBible corpus contains 2 translations from the *Intermediate > Western* sub-branch, 4 from the *Outer* sub-branch and 5 from the *Western Hindi* sub-branch (representing 2 languages). Each of these translations is for a language spoken in Bangladesh, India, Nepal, and/or Pakistan, and each is a full Bible translation. There are no translations for low-resource languages from this language family and geography available in the corpus. As a result, the Hindi (hin-hin2017) and Eastern Panjabi (pan-pan) translations were chosen for the source and target translation pairing, with the Gujarati translation (guj-gju2017) as the related language translation.

**Table C.4.** Source / Target / Related Language Alignment Scores (Indo-European).

| **Target Language(s)** | | **National/Gateway Language(s)** | | **Related Language(s)** | |
|---|---|---|---|---|---|
| **Language** | **Translation** | **hin-hin2017** | **npi-npiulb** | **pan-pan** | **ben-ben2017** |
| Gujarati | guj-guj2017 | 0.3012 | 0.3200 | **0.2973** | 0.3041 |
| E. Panjabi | pan-pan | **0.4240** | 0.2623 | | 0.3665 |
| Assamese | asm-asmfb | 0.2796 | 0.3172 | 0.2616 | 0.3666 |
| Bengali | ben-ben2017 | 0.2904 | 0.3242 | 0.2774 | |
| Marathi | mar-mar | 0.2521 | 0.2989 | 0.2461 | 0.2767 |
| Orya | ory-ory | 0.3036 | 0.3098 | 0.2942 | 0.4031 |

| | | | | | |
|---|---|---|---|---|---|
| Urdu | urd-urd | 0.4622 | 0.2576 | 0.4415 | 0.2765 |
| Urdu | urd-urdgvh | 0.3123 | 0.2346 | 0.2949 | 0.2270 |
| Urdu | urd-urdgvr | 0.3092 | 0.2293 | 0.3012 | 0.2238 |
| Urdu | urd-urdgvu | 0.3041 | 0.2253 | 0.2964 | 0.2205 |

**Appendix C.5. Niger-Congo**

For the Niger-Congo language family, the *Volta-Congo* > *Benue-Congo* > *Bantoid* > *Southern* > *Narrow-Bantu* > *Central* branch contains 354 languages. Of these 354 languages, the eBible corpus contains 19 translations for languages spoken in Tanzania, including 3 Swahili full Bible translations, 2 NT+ translations (Kwere (cwe-cwe) and Vidunda (vid-vid)), and 14 NT-only translations. Among these translations, the Swahili translation (swh-swhonen) was selected as the source and the Kwere translation (cwe-cwe) was selected as the target, and the Vidunda translation (vid-vid) was selected as the related language translation. Preference was given to the Kwere and Vidunda translations due to their partial Old Testament content.

**Table C.5.** Source / Target / Related Language Alignment Scores (Niger-Congo).

| **Target Language(s)** | | **National/Gateway Language(s)** | | **Related Language(s)** | |
|---|---|---|---|---|---|
| **Language** | **Translation** | **swh-swhonen** | **swh-swhulb** | **cwe-cwe** | **vid-vid** |
| Kwere* | cwe-cwe | **0.4382** | 0.3758 | | **0.4912** |
| Isanzu | isn-isn | 0.3245 | 0.5625 | 0.2638 | 0.2424 |
| Kutu | kdc-kdc | 0.4326 | 0.3685 | 0.5347 | 0.4760 |
| Makonda | kde-kde | 0.4509 | 0.3672 | 0.4451 | 0.4315 |
| Kisi | kiz-kiz | 0.4320 | 0.7855 | 0.3446 | 0.3304 |
| Mwera | mwe-mwe | 0.4216 | 0.3560 | 0.4287 | 0.4382 |
| Ndamba | ndj-ndj | 0.4098 | 0.3544 | 0.4609 | 0.4808 |
| Ngulu | ngp-ngp | 0.3891 | 0.3383 | 0.4459 | 0.4112 |
| Ngindo | nnq-nnq | 0.4118 | 0.3517 | 0.4602 | 0.4754 |
| Pogolo | poy-poy | 0.4235 | 0.3608 | 0.4745 | 0.4781 |
| Kara | reg-reg | 0.3732 | 0.5998 | 0.3143 | 0.2850 |
| Luguru | ruf-ruf | 0.4457 | 0.3807 | 0.5570 | 0.4701 |
| Vidunda* | vid-vid | 0.4045 | 0.3462 | **0.4913** | |
| Vwanji | wbi-wbi | 0.3477 | 0.4811 | 0.3169 | 0.2933 |
| Zaramo | zaj-zaj | 0.4652 | 0.3965 | 0.4784 | 0.4331 |
| Zigula | ziw-ziw | 0.4149 | 0.3624 | 0.4748 | 0.4451 |

**Appendix C.6. Otomanguean**

For the Otomanguean language family with 179 total languages, the *Eastern-Otomanguean* > *Popolocan-Zapotecan* > *Zapotecan* branch contains 64 languages. Of these 64 languages, the eBible corpus contains 28 translations for languages spoken in Mexico, including 3 NT+ translations[2], and 25 NT-only translations; there are no full Bible translations from this branch. Among these translations, a Spanish translation (spa-sparvg) translation was selected as the source and the Zapotec Tabaa translation (zat-zatNTps) was selected as the target; preference was given to the Zaopotec Tabaa translation due to its partial Old Testament content. The Zapotec Cajonos translation (zad-zadNT) was selected as the related language translation.

**Table C.6.** Source / Target / Related Language Alignment Scores (Otomanguean).

| **Target Language(s)** | | **National/Gateway Language(s)** | **Related Language(s)** | |
|---|---|---|---|---|
| **Language** | **Translation** | **spa-sparvg** | **zar-zarNT** | **zat-zatNTps** |
| Chatino, Tataltepec | cta-ctaNT | 0.1029 | 0.1229 | 0.1214 |
| Chatino, Western Highland | ctp-ctpNT | 0.0947 | 0.1165 | 0.1157 |
| Chatino, Nopala | cya-cya | 0.1718 | 0.1718 | 0.1667 |

[2] Zapotec Rincón (zar-zarNT), Zapotec Tabaa (zat-zatNTps), and Zapotec Yatee (zty-ztyNTps)

| | | | | |
|---|---|---|---|---|
| Zapotec, Sierra de Juárez | zaa-zaaNT | 0.1662 | 0.1663 | 0.1653 |
| Zapotec, Western Tlacolula Valley | zab-zabNT | 0.2092 | 0.2281 | 0.2456 |
| Zapotec, Ocotlán | zac-zacNT | 0.1733 | 0.1896 | 0.2022 |
| Zapotec, Cajonos | zad-zadNT | 0.1877 | 0.2485 | **0.2757** |
| Zapotec, Isthmus | zai-zaiNT | 0.1835 | 0.2000 | 0.1957 |
| Zapotec, Miahuatlán | zam-zamNT | 0.1161 | 0.1362 | 0.1369 |
| Zapotec, Ozolotepec | zao-zaoNT | 0.1810 | 0.2056 | 0.2119 |
| Zapotec, Rincón | zar-zarNT* | 0.1901 | | |
| Zapotec, Santo Domingo Albarradas | zas-zasNT | 0.1858 | 0.2301 | 02420 |
| Zapotec, Tabaa | zat-zatNTps* | **0.1869** | | |
| Zapotec, Yatzachi | zav-zavNT | 0.1536 | 0.1960 | 0.2049 |
| Zapotec, Mitla | zaw-zawNT | 0.2094 | 0.2202 | 0.2263 |
| Zapotec, Coatecas Altas | zca-zcaNT | 0.1567 | 0.1855 | 0.1825 |
| Zapotec, Choapan | zpc-zpcNT | 0.1439 | 0.2057 | 0.2097 |
| Zapotec, Mixtepec | zpm-zpmNT | 0.1215 | 0.1387 | 0.1370 |
| Zapotec, Amatlán | zpo-zpoNT | 0.1718 | 0.1985 | 0.2027 |
| Zapotec, Zoogocho | zpq-zpqNT | 0.1549 | 0.2162 | 0.2371 |
| Zapotec, Yalálag | zpu-zpuNT | 0.1557 | 0.2148 | 0.2440 |
| Zapotec, Chichicapan | zpv-zpvNT | 0.1642 | 0.1836 | 0.1948 |
| Zapotec, Texmelucan | zpz-zpzNTpp | 0.1438 | 0.1727 | 0.1745 |
| Zapotec, Southern Rincon | zsr-zsrNT | 0.1988 | 0.9000 | 0.3517 |
| Zapotec, Quioquitani-Quierí | ztq-ztqNT | 0.1676 | 0.1944 | 0.2111 |
| Zapotec, Yatee | zty-ztyNTps* | 0.1782 | 0.5623 | 0.4381 |

**Appendix C.7. Sino-Tibetan**

For the Sino-Tibetan language family (458 total languages), the *Tibeto-Burman* > *Western Tibeto-Burman* branch contains 442 languages. Of these 442 languages, the eBible corpus contains 15 translations spread across the *Kuki-Chin* (8), *Ngwi-Burmese* (3), and *Western Tibeto-Burman* (4) sub-branches. Translations from these sub-branches are for languages spoken in China, India, Myanmar, and Nepal, with 8 full Bible, 1 NT+, and 6 NT-only translations. The best alignment results were seen with a Nepali translation (npi-npiulb) as the source and the Eastern Tamang translation (taj-taj) as the target. The Limbu translation in Devanagari script (lif-lifNT2) was selected for the related language translation.

**Table C.7.** Source / Target / Related Language Alignment Scores (Sino-Tibetan).

| **Target Language(s)** | | **National/Gateway Language(s)** | | | | **Related Language(s)** |
|---|---|---|---|---|---|---|
| **Language** | **Translation** | **hin-hin2017** | **npi-npiulb** | **mya-mya** | **mya-myajvb** | **taj-taj** |
| Zaiwa | atb-atbNT | 0.1827 | 0.1640 | 0.1621 | 0.1703 | 0.1545 |
| Chin, Eastern Khumi | cek-cekak | 0.2112 | 0.2074 | 0.1259 | 0.2237 | 0.1655 |
| Chin, Thaiphum | cth-cth | 0.1821 | 0.1806 | 0.1237 | 0.2478 | 0.1463 |
| Chin, Siyin | csy-csy | 0.2078 | 0.1843 | 0.1423 | 0.1854 | 0.1590 |
| Chin, Matu | hlt-hlt | 0.2281 | 0.2170 | 0.1229 | 0.2208 | 0.1620 |
| Chin, Matu | hlt-hltmcsb | 0.2296 | 0.2191 | 0.1237 | 0.2216 | 0.1642 |
| Chin, Matu | hlt-hltthb | 0.1678 | 0.1698 | 0.1195 | 0.1559 | 0.1449 |
| Limbu (Limbu*) | lif-lifNT | 0.2108 | 0.2665 | 0.1142 | 0.1513 | 0.2664 |
| Limbu (Deva) | lif-lifNT2 | 0.2093 | 0.2660 | 0.1131 | 0.1509 | **0.2650** |
| Sunwar | suz-suzBl | 0.2008 | 0.2065 | 0.1107 | 0.1563 | 0.1805 |
| Tamang, Eastern | taj-taj | 0.1959 | **0.2657** | 0.1204 | 0.1344 | |
| Chin, Thado | tcz-tczchongthu | 0.1271 | 0.1303 | 0.0976 | 0.1207 | 0.1487 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Chin, Zyphe | zyp-zypNT | 0.1765 | 0.1824 | 0.1399 | 0.1583 | 0.1662 |

(*) Limbu script is not supported by NLLB

**Appendix C.8. Trans-New Guinea**

For the Trans-New Guinea language family (481 total languages), the *Finisterre-Huon > Finisterre* branch contains 40 languages and the *Madang > Croisilles* branch contains 57 languages. The eBible corpus contains 10 translations from the *Finisterre-Huon > Finisterre* branch; 5 of these are NT-only translations and 5 are NT+ translations. For the *Madang > Croisilles* branch, the eBible corpus contains 8 translations; 6 of these are NT-only translations and 2 are NT+ translations. These translations are for languages spoken in Papua New Guinea, where national languages are English and Tok Pisin (tpi). The best alignment results were observed with a Tok Pisin translation (tpi-tpiOTNT) as the source and the Yopno translation (yut-yut) as the target, while the Iyo translation (nca-nca) was selected for the related language translation.

**Table C.8.** Source / Target / Related Language Alignment Scores (Trans New-Guinea).

| **Target Language(s)** | | **National/Gateway Language(s)** | | **Related Language(s)** | |
|---|---|---|---|---|---|
| **Language** | **Translation** | **tpi-tpi** | **tpi-tpiOTNT** | **yut-yut** | **aey-aey** |
| *Finisterre-Huon > Finisterre Sub-branch* | | | | | |
| Gwahatike | dah-dah | 0.1354 | 0.1354 | 0.1545 | |
| Tuma-Irumu | iou-iou* | 0.1573 | 0.1573 | 0.1902 | |
| Iyo | nca-nca | 0.2034 | 0.2034 | **0.2677** | |
| Numanggang | nop-nop* | 0.1790 | 0.1790 | 0.2261 | |
| Rawa | rwo-rwo-karo | 0.1913 | 0.1913 | 0.1859 | |
| Rawa | rwo-rwo-rawa | 0.1917 | 0.1917 | 0.1854 | |
| Uri | uvh-uvh | 0.1307 | 0.1307 | 0.1296 | |
| Wantoat | wnc-wnc* | 0.1840 | 0.1840 | 0.1831 | |
| Yau | yuw-yuw | 0.1774 | 0.1774 | 0.2161 | |
| Yopno | yut-yut* | **0.2252** | **0.2252** | | |
| *Madang > Croisilles Sub-branch* | | | | | |
| Amele | aey-aey* | **0.2089** | **0.2089** | | |
| Girawa | bbr-bbr | 0.1948 | 0.1948 | | 0.2298 |
| Nobonob | gaw-gaw | 0.1875 | 0.1875 | | **0.2432** |
| Kein | bmh-bmh | 0.1934 | 0.1934 | | 0.2171 |
| Mauwake | mhl-mhl | 0.1659 | 0.1659 | | 0.1809 |
| Bargam | mlp-mlp | 0.1868 | 0.1868 | | 0.1971 |
| Usan | wnu-wnu | 0.1592 | 0.1592 | | 0.1765 |
| Waskia | wsk-wsk* | 0.1866 | 0.1866 | | 0.2353 |

## Appendix D: Biblical Book Identifiers

Table D.1 lists the standard three-character book identifiers utilized as part of the Unified Standard Format Markers (USFM) format.

**Table D.1.** Three-letter identifiers for Biblical books (excluding those not used in this work).

| **Book Name** | **Identifier** |
|---|---|
| Genesis | GEN |
| Exodus | EXO |
| Leviticus | LEV |
| Numbers | NUM |
| Deuteronomy | DEU |

| | |
|---|---|
| Joshua | JOS |
| Judges | JDG |
| Ruth | RUT |
| 1 Samuel | 1SA |
| 2 Samuel | 2SA |
| 1 Kings | 1KI |
| 2 Kings | 2KI |
| 1 Chronicles | 1CH |
| 2 Chronicles | 2CH |
| Ezra | EZR |
| Nehemiah | NEH |
| Esther | EST |
| Job | JOB |
| Psalms | PSA |
| Proverbs | PRO |
| Ecclesiastes | ECC |
| Song of Songs | SNG |
| Isaiah | ISA |
| Jeremiah | JER |
| Lamentations | LAM |
| Ezekiel | EZK |
| Daniel | DAN |
| Hosea | HOS |
| Joel | JOL |
| Amos | AMO |
| Obadiah | OBA |
| Jonah | JON |
| Micah | MIC |
| Nahum | NAH |
| Habakkuk | HAB |
| Zephaniah | ZEP |
| Haggai | HAG |

| | |
|---|---|
| Zechariah | ZEC |
| Malachi | MAL |
| Matthew | MAT |
| Mark | MRK |
| Luke | LUK |
| John | JHN |
| Acts of the Apostles | ACT |
| Romans | ROM |
| 1 Corinthians | 1CO |
| 2 Corinthians | 2CO |
| Galatians | GAL |
| Ephesians | EPH |
| Philippians | PHP |
| Colossians | COL |
| 1 Thessalonians | 1TH |
| 2 Thessalonians | 2TH |
| 1 Timothy | 1TI |
| 2 Timothy | 2TI |
| Titus | TIT |
| Philemon | PHM |
| Hebrews | HEB |
| James | JAS |
| 1 Peter | 1PE |
| 2 Peter | 2PE |
| 1 John | 1JN |
| 2 John | 2JN |
| 3 John | 3JN |
| Jude | JUD |
| Revelation | REV |

## Appendix E: Extended Results

In this section, we include additional results of interest, represented in tabular form. Several of these tables have figure analogs in the main body of the paper.

First, we consider the effect of increasing model size for Meta's NLLB model, which is available in five different configurations, ranging from 600M to 54.5B parameters. To evaluate the potential benefits of working with the larger models, the NLLB-1.3B (distilled) model was fine-tuned on one of the same train/test/validation splits from the previous tests. This

was done for the Dravidian, Niger-Congo, and Sino-Tibetan translation pairings. Translation accuracy metrics with the NLLB-1.3B-distilled models were consistently better for each translation pairing than with the NLLB-600 model, as shown in Table E.1 below.

**Table E.1.** Model size impact, NLLB-600M versus NLLB-1.3B (distilled).

| Translation Pairing | Model | BLEU | spBLEU | chrF3 | WER | TER |
|---|---|---|---|---|---|---|
| Dravidian | NLLB-600M | 21.7 | 40.0 | 59.5 | 46.6 | 69.7 |
| | NLLB-1.3B-distilled | **24.9 (+3.2)** | **44.2 (+4.2)** | **62.7 (+3.2)** | **43.9 (-2.7)** | **65.6 (-4.1)** |
| Niger-Congo | NLLB-600M | 28.3 | 37.9 | 60.5 | 43.6 | 65.2 |
| | NLLB-1.3B-distilled | **29.9 (+1.6)** | **40.0 (+2.1)** | **62.2 (+1.7)** | **41.7 (-1.9)** | **63.1 (-2.1)** |
| Sino-Tibetan | NLLB-600M | 31.5 | 49.5 | 57.9 | **52.0** | 64.0 |
| | NLLB-1.3B-distilled | **33.1 (+1.6)** | **51.0 (+1.5)** | **59.9 (+2.0)** | 52.1 (+0.1) | **62.3 (-1.7)** |

Next, we consider how model complexity improves performance across all five scoring metrics.

**Table E.2.** Machine translation across model types (SMT, OpenNMT, and NLLB-600M).

| Translation Pairing | MT Technology | BLEU | spBLEU | chrF3 | WER | TER |
|---|---|---|---|---|---|---|
| Dravidian | SMT | 9.8 | 20.1 | 36.9 | 98.6 | 84.4 |
| | OpenNMT | 13.3 | 28.0 | 46.9 | 56.3 | 79.0 |
| | NLLB-600M | 21.7 | 39.8 | 58.2 | 47.0 | 68.6 |
| | NLLB-1.3B-distilled* | **24.9** | **44.2** | **62.7** | **43.9** | **65.6** |
| Niger-Congo | SMT | 19.3 | 27.6 | 49.2 | 81.6 | 73.8 |
| | OpenNMT | 16.6 | 24.3 | 45.2 | 55.2 | 77.9 |
| | NLLB-600M | 28.8 | 37.9 | 60.4 | 43.9 | 65.2 |
| | NLLB-1.3B-distilled* | **29.9** | **40.0** | **62.2** | **41.7** | **63.1** |
| Sino-Tibetan | SMT | 9.9 | 30.8 | 42.1 | 60.2 | 84.4 |
| | OpenNMT | 10.2 | 26.5 | 37.4 | 66.9 | 86.8 |
| | NLLB-600M | 31.5 | 49.5 | 57.9 | **52.0** | 64.0 |
| | NLLB-1.3B-distilled* | **33.1** | **51.0** | **59.9** | 52.1 | **62.3** |

*Five-fold CV is not performed on NLLB-1.3B-distilled due to training overheads.

We also reiterate the median performance across scoring metrics, along with counts for the total number of available verses per each target translation.

**Table E.3.** Median translation accuracy scores by translation pairing.

| Translation Pairing | Training Verses | BLEU | spBLEU | chrF3 | WER | TER |
|---|---|---|---|---|---|---|
| Afro-Asiatic | 7,394 | 29.9±1.4 | 38.4±1.5 | 52.5±1.2 | 51.6±1.3 | 67.2±1.6 |
| Austronesian | 7,404 | **35.1±1.1** | 39.4±1.1 | 54.2±0.6 | 49.1±1.6 | 62.0±1.4 |
| Dravidian | 30,589 | 21.7±0.8 | 39.8±0.8 | 58.2±0.6 | 47.0±0.9 | 68.6±0.9 |
| Indo-European | 30,599 | 30.5±1.0 | 40.4±1.1 | 55.7±0.7 | **42.5±1.1** | **55.8±1.4** |
| Niger-Congo | 13,304 | 28.8±1.0 | 37.9±0.7 | **60.4±0.6** | 43.9±0.7 | 65.2±1.1 |
| Otomanguean | 9,901 | 28.3±0.9 | 44.3±0.8 | 56.4±1.2 | 50.9±0.6 | 68.1±0.7 |
| Sino-Tibetan | 7,419 | 31.5±1.6 | **49.5±1.5** | 58.9±0.7 | 50.5±1.3 | 63.9±1.7 |
| Trans-New Guinea | 9,775 | 31.6±0.7 | 49.0±0.5 | 60.1±0.9 | 48.0±0.8 | 61.6±0.4 |

Table E.4 shows BLEU, spBLEU, and chrF3 scores for the *CV* task as well as the *Gospel Translation* task.

**Table E.4.** Scores for NLLB-600M fine-tuned models for the *CV* task and the *Gospel Translation* task.

| | **CV Task** | | | **Gospel Translation Task** | | |
|---|---|---|---|---|---|---|
| **Translation Pairing** | **BLEU** | **spBLEU** | **chrF3** | **BLEU** | **spBLEU** | **chrF3** |
| Afro-Asiatic | 29.9 | **38.4** | **52.1** | **30.2 (+0.3)** | 35.8 (-2.6) | 49.7 (-2.4) |
| Austronesian | **35.1** | **39.4** | **53.8** | 33.7 (-1.4) | 36.2 (-3.2) | 52.6 (-1.2) |
| Dravidian | **21.7** | **39.8** | **58.5** | 18.2 (-3.5) | 35.1 (-4.7) | 58.1 (-0.4) |
| Indo-European | 30.5 | 40.4 | 55.5 | **36.5 (+6.0)** | **44.5 (+4.1)** | **59.9 (+4.4)** |
| Niger-Congo | **28.8** | **37.9** | **60.1** | 21.7 (-7.1) | 29.7 (-10.2) | 53.7 (-6.4) |
| Otomanguean | **28.3** | **44.3** | **56.7** | 20.4 (-7.9) | 34.5 (-9.8) | 48.7 (-8.0) |
| Sino-Tibetan | **31.5** | **49.5** | **58.8** | 30.6 (-0.9) | 45.9 (-3.6) | 56.1 (-2.7) |
| Trans-New Guinea | **31.6** | **49.0** | **59.8** | 28.5 (-3.1) | 41.7 (-7.3) | 51.7 (-8.1) |

Table E.5 shows BLEU scores for various minor-prophet books with both the *Early OT* task (training set includes the full NT and several OT books) and the *Late OT* task (training set includes the full NT and OT books except the minor prophets in the test set, shown below).

**Table E.5.** BLEU scores for NLLB-600M fine-tuned models on *Early OT* and *Late OT* tasks.

| | **Dravidian (BLEU)** | | **Indo-European (BLEU)** | |
|---|---|---|---|---|
| **Minor Prophet** | **Early OT Task** | **Late OT Task** | **Early OT Task** | **Late OT Task** |
| HOS | 10.4 | 16.6 (+6.2) | 16.6 | 22.6 (+6.0) |
| JOL | 10.2 | 19.2 (+9.0) | 19.2 | 28.5 (+9.3) |
| AMO | 7.2 | 17.9 (+10.7) | 19.0 | 26.0 (+7.0) |
| OBA | 5.4 | 16.7 (+11.3) | 16.9 | 26.3 (+9.4) |
| MIC | 7.0 | 14.7 (+7.7) | 20.8 | 25.8 (+5.0) |
| NAH | 4.5 | 10.9 (+6.4) | 13.8 | 21.4 (+7.6) |
| HAB | 5.9 | 9.8 (+3.9) | 18.9 | 22.2 (+3.3) |
| ZEP | 8.1 | 15.1 (+7.0) | 19.1 | 29.2 (+10.1) |
| HAG | 6.9 | 21.3 (+14.4) | 17.9 | 34.3 (+16.4) |
| ZEC | 8.3 | 16.6 (+8.3) | 17.1 | 27.1 (+10.0) |
| MAL | 8.2 | 11.9 (+3.7) | 15.4 | 22.8 (+7.4) |

**Table E.6.** BLEU scores for NLLB-600M fine-tuned models on the *Early OT* task.

| **Translation Pairing** | **GEN** | **EXO** | **LEV** | **NUM** | **DEU** | **RUT** | **PSA** | **JON** |
|---|---|---|---|---|---|---|---|---|
| Dravidian | 11.8 | 8.9 | 6.4 | 6.5 | 6.4 | 8.1 | 10.9 | 9.2 |
| Indo-European | 23.6 | 17.0 | 19.3 | 16.7 | 17.9 | 18.6 | 20.3 | 20.9 |
| Niger-Congo | 18.9 | 16.0 | 12.7 | 14.7 | 12.8 | | | |
| Otomanguean | | | | | | | 16.2 | |
| Trans-New Guinea | | | | | | | 19.0 | |

**Table E.7.** BLEU, spBLEU, and chrF3 scores for NLLB-600M fine-tuned models on *Gospel Translation (without Related Language)* task and the *Gospel Translation (with Related Language)* task.

| **Translation Pairing** | **Gospel Translation (MAT) (Without RL)** | | | **Gospel Translation (MAT) (With RL)** | | |
|---|---|---|---|---|---|---|
| | **BLEU** | **spBLEU** | **chrF3** | **BLEU** | **spBLEU** | **chrF3** |

| | | | | | | |
|---|---|---|---|---|---|---|
| Afro-Asiatic | **30.2** | **35.8** | **49.7** | 28.5 | 33.8 | 48.0 |
| Austronesian | 33.7 | 36.2 | 52.6 | **36.2** | **39.0** | **54.2** |
| Dravidian | 18.2 | 35.1 | 58.1 | **19.4** | **36.5** | **58.7** |
| Indo-European | **36.5** | **44.5** | 59.9 | 36.0 | 44.2 | **60.1** |
| Niger-Congo | **21.7** | **29.7** | **53.7** | 21.1 | 28.7 | 52.2 |
| Otomanguean | **20.4** | **34.5** | **48.7** | 18.1 | 31.2 | 47.9 |
| Sino-Tibetan | **30.6** | **45.9** | **56.1** | 27.3 | 41.6 | 53.9 |
| Trans-New Guinea | **28.5** | **41.7** | **51.7** | 27.8 | 40.5 | 51.3 |

**Table E.8.** BLEU scores for NLLB-600M fine-tuned models on *Epistle Translation (without Related Language)* task and the *Epistle Translation (with Related Language)* task.

| **Translation Pairing** | **Epistle Translation (Without RL)** | | | | | **Epistle Translation (With RL)** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **1TH** | **2TH** | **1TI** | **2TI** | **TIT** | **1TH** | **2TH** | **1TI** | **2TI** | **TIT** |
| Afro-Asiatic | **14.8** | **17.3** | **11.3** | **14.3** | **9.3** | 14.2 | 14.6 | 11.2 | 12.3 | 8.1 |
| Austronesian | 20.1 | 21.5 | 19.9 | 23.5 | 19.8 | **30.0** | **31.3** | **30.5** | **33.4** | **28.3** |
| Dravidian | 11.4 | **13.4** | 7.0 | 8.1 | **5.2** | **13.2** | 13.1 | **7.5** | **9.**4 | 4.8 |
| Indo-European | **28.7** | **26.2** | **21.3** | **26.7** | **21.6** | 26.0 | 24.0 | 18.8 | 24.2 | 20.2 |
| Niger-Congo | 14.9 | 16.7 | 15.1 | 14.5 | 13.6 | **15.5** | **18.4** | **17.1** | **18.6** | **16.0** |
| Otomanguean | **16.2** | **14.0** | **15.5** | 13.1 | **15.6** | 14.2 | 13.1 | 15.4 | **13.4** | 14.0 |
| Sino-Tibetan | **18.8** | 17.2 | **16.2** | **18.2** | **18.3** | 18.3 | **18.9** | 15.7 | 16.6 | 14.4 |
| Trans-New Guinea | **18.7** | **20.5** | **16.9** | **17.6** | **15.0** | 16.8 | 20.4 | 13.7 | 13.6 | 14.8 |

**Table E.9.** BLEU, spBLEU, and chrF3 scores for NLLB-600M fine-tuned models on *NT Completion (without Related Language)* task and the *NT Completion (with Related Language)* task.

| **Translation Pairing** | **NT Completion (Without RL)** | | | | | | **NT Completion (With RL)** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **ROM** | | | **REV** | | | **ROM** | | | **REV** | | |
| | **BLEU** | **spBLEU** | **chrF3** | **BLEU** | **spBLEU** | **chrF3** | **BLEU** | **spBLEU** | **chrF3** | **BLEU** | **spBLEU** | **chrF3** |
| Afro-Asiatic | 18.9 | 28.2 | 43.7 | 16.0 | **25.6** | 43.3 | **19.4** | **29.1** | **44.3** | **16.2** | 25.1 | **43.7** |
| Austronesian | 23.2 | 27.3 | 44.9 | 28.6 | 33.3 | 50.3 | **31.5** | **35.0** | **50.7** | **40.1** | **44.0** | **58.0** |
| Dravidian | **12.8** | 29.3 | 53.1 | 13.7 | 31.9 | 54.6 | 12.6 | **29.5** | 53.1 | **14.6** | **32.6** | **55.7** |
| Indo-European | **30.0** | **38.2** | **56.1** | **29.2** | **38.1** | **54.3** | 29.1 | 37.9 | 55.5 | 28.4 | 37.7 | 54.2 |
| Niger-Congo | 20.1 | 29.9 | 54.6 | **23.5** | **32.3** | 55.3 | **22.3** | **32.1** | **57.1** | 23.4 | 31.1 | **56.2** |
| Otomanguean | 22.1 | **38.8** | 53.1 | 20.6 | 35.5 | 51.1 | 22.1 | 38.6 | **53.5** | **21.2** | **36.0** | **52.3** |
| Sino-Tibetan | 23.3 | 41.4 | 51.6 | 23.3 | **42.5** | 53.2 | **23.9** | **41.5** | **52.9** | **23.4** | 42.3 | **54.0** |
| Trans-New Guinea | **24.9** | **42.3** | **56.2** | **22.1** | **42.1** | **53.5** | 22.2 | 40.0 | 53.2 | 20.9 | 41.2 | 52.7 |