BUILDING MACHINE TRANSLATION SYSTEMS FOR THE NEXT THOUSAND LANGUAGES

Ankur Bapna* Isaac Caswell* Julia Kreutzer Orhan Firat Daan van Esch
Aditya Siddhant Mengmeng Niu Pallavi Baljekar Xavier Garcia
Wolfgang Macherey Theresa Breiner Vera Axelrod Jason Riesa Yuan Cao
Mia Xu Chen Klaus Macherey Maxim Krikun Pidong Wang Alexander Gutkin
Apurva Shah Yanping Huang Zhifeng Chen Yonghui Wu Macduff Hughes

ABSTRACT

In this paper we share findings from our effort to build practical machine translation (MT) systems capable of translating across over one thousand languages. We describe results in three research domains: (i) Building clean, web-mined datasets for 1500+ languages by leveraging semi-supervised pre-training for language identification and developing data-driven filtering techniques; (ii) Developing practical MT models for under-served languages by leveraging massively multilingual models trained with supervised parallel data for over 100 high-resource languages and monolingual datasets for an additional 1000+ languages; and (iii) Studying the limitations of evaluation metrics for these languages and conducting qualitative analysis of the outputs from our MT models, highlighting several frequent error modes of these types of models. Using this approach, we add 24 new languages to Google Translate, the product’s largest increase in language coverage to-date. We hope that our work provides useful insights to practitioners working towards building MT systems for currently understudied languages, and highlights research directions that can complement the weaknesses of massively multilingual models in data-sparse settings.

*Equal contributions. Correspondence to {ankurbpn,icaswell}@google.com. All authors affiliated with Google Research.
1 AN OVERVIEW

The past decade has seen tremendous improvements in the quality of academic and commercial machine translation (MT) systems. These improvements have largely been driven by advances in machine learning and the availability of large-scale web-mined datasets (Resnik & Smith, 2003; Uszkoreit et al., 2010; Esplà-Gomis, 2009; Esplà et al., 2019; Bañón et al., 2020; Schwenk et al., 2021). The advent of deep learning and sequence-to-sequence models (Sutskever et al., 2014; Bahdanau et al., 2015; Luong et al., 2015; Vaswani et al., 2017), large parallel (and monolingual) datasets mined from the web, data augmentation approaches like back-translation (Sennrich et al., 2016; Edunov et al., 2018) and self-training (He et al., 2019) and massively multilingual modeling (Firat et al., 2016a; Johnson et al., 2017; Aharoni et al., 2019a; Arivazhagan et al., 2019; Tang et al., 2021; Fan et al., 2021) have enabled high quality machine translation systems that can support over 100 languages.

However, despite tremendous progress in low-resource MT, the number of languages for which widely-available, general-domain MT systems have been built has been limited to around 100, which is a small fraction of the over 7000+ languages that are spoken in the world today. Apart from the limited number of languages, the distribution of languages supported by current MT systems is highly skewed in favour of European languages. Despite high speaker populations, languages spoken in Africa, South and South-East Asia and indigenous languages of the Americas are relatively under-served. For example, Google Translate supports Frisian, Maltese, Icelandic, and Corsican, each with fewer than 1M L1 speakers, but not (up until this work) Bhojpuri (~51M speakers), Oromo (~24M speakers), Quechua (~9M speakers), or Tigrinya (~9M speakers) (van Esch et al., 2022). We will refer to these languages as long-tail languages, since the data scarcity requires the application of machine learning techniques that can generalize beyond the languages for which ample training data is available.

Web-crawled datasets for 1000 languages: The progress towards building machine translation systems in these languages has largely been limited by the lack of digitized and accessible datasets and NLP tools like language identification (LangID) models; such resources are ubiquitous for higher resource languages. The first stage of this paper describes our approach to building monolingual web text corpora in over 1500 languages, with a particular focus on dealing with common noise, data quality and scale challenges encountered when building a dataset from the web (Section 2).

Towards this goal, we first scale LangID models to 1500+ languages (2.1.1), both using traditional n-gram models and semi-supervised approaches. We next describe several practical approaches and filtering techniques that enable using these LangID models to identify long-tail language data on the web. To minimize recall loss during mining, we cluster languages by error rate (2.1.3). To reduce noise from LangID mis-predictions, we leverage document-level LangID consistency to filter our data (2.1.4), followed by percent-threshold wordlist filtering (2.1.5), TFiF filtering (2.1.7), and hand-designed filters to address specific noise issues with certain languages (2.1.8). We describe the resulting dataset in Section 2.2 and perform an audit of 72 language corpora (2.2.2), finding that the data is between 70% and 100% in-language, with a median score of 80%.

Machine Translation for long-tail languages: With monolingual data mined from the web, the next challenge is to build high quality, general-domain MT models from limited amounts of monolingual training data. We follow a practical approach, utilizing all the parallel data that is available for higher resource languages to boost the quality of long-tail languages where only monolingual data is available[1]. We will refer to this setting as zero-resource since no direct supervision is available for our long-tail languages. We want to emphasize that this term is used only in the technical sense, meaning that the model itself is not able to see parallel text; in reality, there is a richness of resources for these languages, including tens of millions of native speakers, centuries (or in some cases millennia) of scholarship, and even large segments of text inaccessible to digital methods. See Bird (2020) for some more reflections on this term.

We leverage several techniques that have been developed for MT over the last few years in order to boost the quality of zero-resource translation for long-tail languages. These include self-supervised learning from monolingual data, massively multilingual supervised training, large-scale

---

[1] Preliminary experiments indicated no quality improvements on our evaluation sets when incorporating widely available limited-domain parallel corpora in our models.
back-translation and self-training, and high capacity models. We utilize these tools to build MT models capable of translating across over 1000 languages, utilizing our existing parallel corpus spanning around 100 languages and the 1000-language monolingual dataset built from the web (Section 3).

We first highlight the importance of model capacity in highly multilingual models by comparing the performance of 1.5B and 6B parameter Transformers on zero-resource translation (3.2). We then scale up the number of self-supervised languages to 1000, demonstrating that the performance of most long-tail languages improves as more monolingual data from similar languages becomes available (3.3). While our 1000-language models demonstrate reasonable performance, in order to understand the strengths and limitations of the approach we incorporate large-scale data augmentation. For practical purposes we fine-tune the resulting model on a subset of 30 languages with large amounts of synthetic data via self-training and back-translation (3.4). We further describe practical approaches to filter synthetic data to increase the robustness of these fine-tuned models to hallucinations and wrong-language translations (3.5). We also describe our efforts to distill these models into smaller, more inference-friendly architectures using sequence level distillation, and highlight the performance gaps between the teacher and student models (3.6).

**Evaluating MT for 1000 languages:** Existing MT systems heavily rely on n-gram overlap based lexical metrics (BLEU) (Papineni et al., 2002) or the new and emerging model-based metrics like YtSi (Lo, 2019), BLEURT (Sellam et al., 2020) or COMET (Rei et al., 2020) to evaluate translation quality. These metrics are usually computed on static evaluation sets with fixed references obtained from professional translators or crowd-workers.

To evaluate our MT models, we first build an evaluation set for 38 selected languages from the long-tail (4.1), by translating English sentences into these languages. We highlight the limitations of BLEU in the long-tail setting and make the case for evaluating these languages with CHR (4.2). We also propose an approximate, round-trip translation based, reference-free metric to understand the quality of our models on languages where reference sets were unavailable, and report the quality of our models as measured on this metric (4.3). We conduct and report the findings from human evaluations of our models (on a subset of 28 languages), confirming that it is possible to build functioning MT systems by following the recipe described in this paper (4.4).

To understand the weaknesses of our massively-multilingual zero-resource models, we perform qualitative error analysis on several languages. We find that our models often confuse distributionally similar words and concepts, e.g. “tiger” becomes a “miniature crocodile” (4.5) and their ability to translate tokens deteriorates on less frequent tokens for lower-resource settings (4.6). We also find that these models often fail to adequately translate short, or single word inputs (4.7). A study of our distilled models also reveals that all models are more likely to magnify biases or noise present in the training data (4.8).

**Other miscellaneous findings and experiments:** We perform a few additional experiments on these models, demonstrating that they often perform better when translating directly between similar languages — rather than using English as a pivot (5.1) and that they can be utilized for zero-shot transliteration between different scripts (5.2). We describe a practical technique of appending terminal punctuation to any input, called the “period trick”, that improves the quality of translation (5.3). Furthermore, we demonstrate that these models are robust to nonstandard Unicode glyph usage for some but not all languages (5.4), and we explore several non-Unicode fonts (5.5).

Throughout the course of this work, we rely extensively on native speakers of these languages to guide, evaluate and often contribute technically to the development of these systems. Section 6 highlights the important role they play — a role that researchers not familiar with the language and community could not have contributed.

As a result of the work outlined in this paper, we add support for 24 new languages to Google Translate. Adding a language to a user-facing product requires considerable effort and individual attention to evaluate each language thoroughly and consult members of affected communities; as a result we limited this effort to 30 languages, of which 24 met our launch bar. This set of languages was chosen to cover languages with large speaker populations in regions that are under-represented in technology, like the Americas, Africa, and South Asia.
2 BUILDING A 1000-LANGUAGE WEB TEXT DATASET

It is difficult to uncover large, clean, and highly multilingual corpora from the web. As explored in Caswell et al. (2020), the task of using LangID models for web-mining is exceedingly challenging for low-resource languages, with noise arising from domain mismatch between LangID training data and web text, idiosyncrasies of n-gram based LangID models, and the massive class imbalance between high- and low-resource languages. This is further borne out by Kreutzer et al. (2022), who demonstrate that a variety of public multilingual corpora, created using techniques that work for high-resource languages, are unusably noisy for many low-resource languages.

This problem is compounded by the size of the internet. For lower-resource languages, it is important to get as much data as possible to have any usable signal for a model to learn from. However, the sheer quantity of data online makes it infeasible to apply computationally intensive methods, like Transformer-based models; and even storing unfiltered data on disk can be challenging.

The following subsections explain in detail the approach we took to crawl a monolingual text dataset for 1500+ languages. Our approach focused on recovering high-precision data (high percentage clean, in-language text), so a large portion of the steps are various filtration approaches. The work in this section is an extension of and improvement over the methods presented in Caswell et al. (2020).

In summary, our approach is as follows:

2.1.2 Omit languages from the LangID model with poor quality training data and poor LangID performance; train both a 1,629-language CLD3 (n-gram) LangID model and Semi-supervised LangID (SSLID) model
2.1.3 Cluster languages by their error rate in the CLD3 model
2.1.4 Perform a first-pass webcrawlex with the CLD3 model
2.1.5 Filter sentences with document-consistency filtering
2.1.6 Filter all corpora with percent-threshold wordlist filtering
2.1.7 Filter all corpora with Semi-Supervised LangID (SSLID) filtering
2.1.8 Detect outliers languages with Relative Recall Rate and filter them with Term-Frequency-Inverse-Internet-Frequency (TF-IIF) filtering
2.1.9 Detect outliers with Token-Frequency Anomalousness score and hand-design filters for them
2.1.10 De-duplicate all corpora at a sentence level

2.1 STEPS NECESSARY TO CREATE THE DATASET

2.1.1 LANGID MODELING

As described in Caswell et al. (2020), the task of reliably detecting what language a given portion of text is in, known as Language Identification or LangID, is surprisingly challenging for low-resource languages on the web.

There are two types of LangID models used in this work: CLD3 models (Bakalov et al., 2016) and Semi-Supervised LangID (SSLID) models (Caswell et al., 2020). The CLD3 models are a multi-layer perceptron on top of a bag of character ngrams, and are very efficient, but suffer from various error pathologies. The semi-supervised models use the MASS objective (Song et al., 2019) on noisy web text in addition to the LangID prediction task, and use the Transformer Big (Vaswani et al., 2017) architecture. They are generally more accurate and also more robust to web noise, but are much slower to run inference with.

Overall, three LangID models were used for this work: 1) a 1,745-language CLD3 model to measure the quality of the training data (Section 2.1.2); 2) a 1,629-language CLD3 model to cluster languages and label sentences when performing a deep crawl of the web (Sections 2.1.3 and 2.1.4); and 3) a 1,629-language semi-supervised LangID model for filtering (Section 2.1.6).

The training and evaluation data for our LangID models is identical to that described in Caswell et al. (2020); namely, we trained on an aggregation of proprietary and publicly available text corpora, with an average of 800K tokens per language. Some of the data came from sources with language tags.
like Wikipedia or Corpus Crawler (Brawer, 2017), while another subset was created using a text elicitation task where we prompted native speakers to write sentences in their language (van Esch et al., 2019).

2.1.2 Parsing down the set of languages

An important first step is determining what languages to crawl the web for. In an ideal world we would crawl all possible languages, but some of them may have poor-quality training data, or be very close to each other and be hard to distinguish.

We trained a CLD3 LangID model on 1,745 language varieties, and investigated languages that seemed to be having large quality losses. Our metrics for considering a language as possibly presenting difficulties were 1) LangID precision was under 33%; 2) language had over 50% confusion (False Negative Rate, FNR, or False Discovery Rate, FDR) with respect to another language; 3) a preliminary web-crawl with this model indicated severe overtriggering with a high-resource language; 4) LangID training data had under 2000 examples. Following this analysis we removed 116 languages from the model, primarily based on poor-quality training data. Among these most were regional dialects in Europe.

We additionally did not try to crawl data for the highest-resource 83 languages, in part because of storage space limitations, and in part as we already had monolingual data for these languages from a previous web crawl.

2.1.3 False Negative Rate clustering

For the first-pass, we wanted to avoid lost recall between related dialects, of which many existed in the 1,629 languages. For instance, our training data contains many different varieties of Hindustani, e.g. Haryanvi, Garhwali, Magahi, and so on.

To mitigate this, we clustered languages based on their False Negative Rate (FNR), as measured on our LangID evaluation sets. Of the many error rates one could cluster on, we chose to cluster on FNR to minimize the loss in recall for related languages. We clustered using the Sklearn (Pedregosa et al., 2011) implementation of Hierarchical Agglomerative Clustering, using distance_threshold=None and affinity="precomputed", and linkage=average. These parameters were chosen largely because they seemed to produce the nicest distribution in cluster sizes. Since some of the clusters were still excessively large, we re-split the largest ones by hand such that no cluster had more than 20 languages. We split these clusters by hand since re-splitting using clustering algorithms still resulted in very uneven sizes, probably because these clusters have hubs with poor or noisy data.

To observe the effect of clustering predictions, we compare the size of a dataset for a language if document-consistency filtering (Section 2.1.4) were applied on the language-code level (which would result in lower recall) or on the cluster level. The median ratio between these sizes was 1.006, meaning that for most languages, clustering didn’t significantly improve recall. However, for some languages it improved recall significantly, with 57 languages showing a dataset size increase of greater than 20x. The languages with large wins here were by-and-large to be expected, including Hindustani and Arabic varieties, and a variety of cases like Oromo (om) and Eastern Oromo (hae).

The higher-resource languages omitted in this crawl (Section 2.1.2) were all put in their own clusters.

2.1.4 Document consistency filtering and first-pass LangID

To start with, we performed LangID prediction on every sentence in every document with the 1,629-language CLD3 n-gram LangID model. Having obtained these predictions, we applied document consistency filtering (Caswell et al., 2020). This is one of the simplest and by far most effective filtering steps. We simply discarded any sentence whose sentence-level LangID cluster prediction did not match the document-level LangID cluster prediction. We defined the document-level LangID cluster prediction as the most-often predicted cluster among all sentences — e.g. if a document had 20 sentences in cluster A, 19 sentences in cluster B, and 18 in cluster C, we gave it a document-level ID of cluster A.
Figure 1: Histogram of document consistency scores, using a 1,745-language CLD3 LangID model, on web text. The large majority of sentence-level LangID predictions had a score of under 10%, indicating that they were likely noisy predictions on pages in other languages or non-linguistic content. Of the 0.55% with document consistency score over 90%, over half (0.3%) had a score of 100%.

To investigate the effectiveness of document consistency filtering, we looked at the document consistency score. For a given sentence in a document, the document consistency score is simply the percent of sentences in that document sharing the same LangID prediction as that sentence. (Therefore, if the document consistency score is over 50%, it will never be filtered out with document consistency filtering.) Figure 1 shows the deciles for document consistency score across all languages, for a sample web crawl with the initial, 1,745-language CLD3 LangID model. The distribution is heavily weighted towards the lower values, with a whole 83% of sentences having a score under 10%. The large mass of sentences with a low document consistency score indicates that there are many single, random sentences in larger documents, including non-linguistic content, whose language is mis-predicted. This suggests a more refined approach to document-consistency filtering. Using the method outlined above, a page with content in multiple languages could only yield data for one language; if instead any sentence were preserved with a document consistency score under a particular threshold (say, 0.3), documents with multilingual content would be handled much better, while preserving most of the benefits of filtering. This is left for future work.

2.1.5 HIGH-RECALL WORDLIST-FILTERING

Following Caswell et al. (2020), we apply percent-threshold wordlist filtering to all languages. A sentence was discarded if it had < 20% in-language words for any of the languages in the cluster, where the wordlists were the most frequent 800 words from the LangID training data.

2.1.6 FILTERING AND DECLUSTERING WITH SEMI-SUPERVISED LANGID

Once the first-pass, clustered filtering was completed, we classified each sentence with the more computationally expensive Semi-Supervised LangID model (SSLID), as described in Caswell et al. (2020), resulting in per-language corpora. If the SSLID model predicted any language outside of the cluster, we filtered that sentence. The CLD3 predictions were ignored.

2.1.7 TF-IIF FILTERING

Even after these three rounds of filtering — CLD3, percent-threshold, and SSLID — many languages still had extremely noisy data. The worst case was Tok Pisin (tpi), whose dataset consisted of 1.3B sentences, of which over 99.7% were in Standard English (mostly containing the word “long”,
which is also a common function word in Tok Pisin). Therefore, as in (Caswell et al. 2020), we applied percent-threshold filtering with TF-IIF (Term-Frequency Inverse-Internet-Frequency) lists meaning that we retained any sentence which contained at least 20% of its tokens in our TF-IIF wordlists. However, we optimize the approach used in (Caswell et al. 2020), and furthermore develop a heuristic metric to determine whether to apply the filtering on a per-language basis.

In contrast to (Caswell et al. 2020), we omitted the IDF term entirely, since this term is influenced by the set of languages one considers, and becomes less helpful as the number of languages scales. We also adjusted the values of the parameters \( \kappa \) and \( \tau \) (described below). To understand these changes, we can revisit the formulation for TF-IIF. For a token \( t \) in a language \( l \), with a frequency function \( f(\text{term}, \text{corpus}) \) and language-specific corpora \( D_l \):

\[
\text{TFIIF}_{t,l} = \frac{f(t, D_l)}{\max(f(t, \text{internet}), \alpha)}
\]

Note the clipping parameter \( \alpha \) in the IIF term, which is introduced to account for OOV tokens (which are common) and noise near the tail of the distribution. For this work we set \( \alpha = f(w_\kappa) \) — in other words, we only consider the top \( \kappa \) terms in the empirical frequency distribution, and give all less common terms the same weight as the \( \kappa \)th most common token. A higher value of \( \kappa \) means that more words are covered by the frequency distribution — which seems good — but it also means that OOV words will have a higher weight, since the IIF term is an inverse. In the worst case this could mean that the resulting TF-IIF wordlist would contain too many rare words. An especially low value of \( \alpha \), e.g. \( \alpha < 1.0 \), would essentially guarantee this, pushing up only OOV words to the top of the list.

There is another parameter hidden in this formulation, namely \( \tau \). This parameter is simply how many words to include in the wordlist. For instance, \( \tau = 1000 \) means that the wordlist used for filtering includes the top 1000 words by TF-IIF score.

In order to investigate the effect of these parameters, we define a heuristic measure of distractibility as a proxy for precision, which measures how much a given language’s data is likely to be polluted by a common high-resource language. For a language \( l \) and a set of distractor languages \( D \), we define the distractibility \( \delta_l \):

\[
\delta_l = \max_{d \in D}(\text{FDR}(d, l))
\]

Where, for our purposes, we have chosen \( D = \{\text{en, de, es, hi, id, ar, ru}\} \), and the False Discovery Rate (FDR) of a distractor language \( d \) with regard to a language \( l \) is defined in the standard way, as (on a balanced eval set)

\[
\text{FDR}(d, l) = \frac{\#\text{examples in language } d \text{ mis-predicted as } l}{\#\text{true examples in language } l}
\]

Despite the concern that it would elevate rare and OOV words, we found that increasing the value of \( \kappa \) steadily decreased the distractibility with little loss in recall on the target languages. Increasing the size of the wordlists, \( \tau \), recouped the loss in recall, but increased distractibility to more dangerous levels. These results can be seen in Table 1.

Therefore, for our TF-IIF lists, we use values of \( \kappa = 80000 \) and \( \tau = 1000 \). For comparison, (Caswell et al. 2020) use \( \kappa = 10000 \), and the value of \( \tau \) was set on a per-language basis from the recall on the dev set. Furthermore, (Caswell et al. 2020) use an IIF list with only 980k unique tokens (7M webpages), whereas this work uses 41M unique tokens (250M web pages). The public GitHub repo has been updated to reflect these improvements.

Deciding when to apply TF-IIF filtering: Many languages do not need extra filtering, and it is important not to overfilter and decrease recall. Therefore we needed an outlier detection metric to determine whether filtering needed to be applied. We wanted to filter those corpora where the loss in recall was small, but the reduction in dataset size was large (indicating that many out-of-language sentences were removed).

---

2Open-sourced at https://github.com/google-research-datasets/TF-IDF-IIF-top100-wordlists/
We looked at the percent of our LangID eval sets that remained after filtering \(r_{\text{TFIIF}}(\text{gold})\), aka the filtering recall, and compared it to the percent of our web-crawled corpora remaining after filtering \(r_{\text{TFIIF}}(\text{crawl})\). We used these to define the heuristic Relative Recall Rate (RRR) as follows:

\[
\text{RRR}_{\text{TFIIF}} = \frac{r_{\text{TFIIF}}(\text{gold})^\rho}{r_{\text{TFIIF}}(\text{crawl})}
\]

This quantity measures, approximately, how many true positives are kept by this method for every false positive filtered out. The exponent \(\rho\) is the trade-off between recall and percent filtered. A value of \(\rho > 1\) means that we weight a loss in recall (undesired outcome) more than an equivalently sized reduction in data size (desired outcome). One way of interpreting this is how much we trust the recall on our eval set. If \(\rho = 1\), we trust it perfectly. However, in practice, since we fear that the recall on our eval set overestimates the true recall on natural text on the web, we set \(\rho > 1\). For our experiments we used a value of \(\rho = 2\). This flagged 895 out of 1503 languages. However since we were only concerned about datasets that were more severely polluted, we only filtered datasets where the filtering would remove 20% or more of the data \(r_{\text{TFIIF}}(\text{crawl}) \leq 80\). We also decided not to filter any language where the recall \(r_{\text{TFIIF}}(\text{gold})\) was less than 80%, in case we lost too much usable data. The result of this was that we applied Tf-IIF filtering to 210 out of 1503 language corpora.

### 2.1.8 Token distribution anomaly detection and Negative Token Filters

Even after the filtering steps described in the previous sections, a variety of languages still had data-quality issues. One issue that was not easily filtered out by approaches like Tf-IIF was templated content, which is technically in the right language, but is not useful for training data. To a lesser extent we also saw issues like the “unlucky n-gram” [Caswell et al., 2020] effect. Examples of the type of content we found were:

1. **Scottish Gaelic (gd)** found 570M in-language sentences even after Tf-IIF filtering. It turned out that this was mostly from one site, and the most common token was “Luchdaich a-nois” (“download”).
2. **Darija (ar-MA)** came up with a dataset of over a billion sentences, but 94.9% contained some reference to “casinos”, “gambling”, etc.
3. **Kurukh (kru-Mlym)** was mainly A N T S P E A K. (And it later turned out that the training data was spurious.)
4. **The Arabic-script Indo-Aryan languages of Kalami (gwc)**, Hindko (hnd), and Torwali (trw) had picked up masses of templates from unit conversion websites (“X lbs is Y ounces”, “convert Euro to US American Dollar”, etc.)
5. **Many Latin-script Indic languages** (hi-Latn, ml-Latn, etc.) had large amounts of content that were just download links or titles for videos, songs, and so on.
6. **Cree (cr-Latn)** was almost 100% “Lorem ipsum” sentences

While some of these are actually incorrect content, much of it is technically in-language, and therefore can’t be filtered out by Tf-IIF filtering or straightforward application of LangID. Therefore, we

| \(\kappa\) | \(\tau\) | mean recall | median recall | mean \(\delta_l\) | median \(\delta_l\) | \(\delta_l > 10\) |
|--------|---------|-------------|--------------|----------------|----------------|----------------|
| 5000   | 1000    | 94.3        | 98.9         | 8.6            | 1.2            | 20             |
| 10000  | 1000    | 94.2        | 98.9         | 7.3            | 0.7            | 16             |
| 20000  | 1000    | 94.1        | 98.7         | 5.7            | 0.4            | 14             |
| 80000  | 1000    | 93.4        | 98.5         | 2.9            | 0.1            | 8              |
| 80000  | 2000    | 94.9        | 99.3         | 6.8            | 0.4            | 21             |

**Table 1:** Investigating the effect of varying the IIF-thresholding parameter \(\kappa\) on the recall and distractibility \(\delta_l\) of Tf-IIF filtering, over 1598 languages. As we increase \(\kappa\), the distractibility steadily trends down, with only small losses in recall. Increasing \(\tau\) recoups lost recall but increases distractibility.
decided to develop an approach to detect anomalous data and investigate datasets that looked like they had issues.

To detect these extreme domain shifts, we hypothesized that the token distribution would be severely skewed. Therefore, we compared the distribution of the tokens in the LangID train data (the reference distribution) to the token distribution in the crawled data (the empirical distribution). To compare these distributions we looked at several scores for the top N=40 tokens in the empirical distribution:

- **2n-overlap**: This is simply the percentage of the top N tokens that appear in the top 2N tokens of the reference distribution; this metric is very simple and highly interpretable.

- **Euclidean**: This is the Euclidean distance between the frequencies of the top N tokens and their corresponding frequencies from the reference distribution.

Different scores, like Jenson-Shannon divergence and Pearson R, were initially considered but found to be ineffective.

We then combined these two scores together with the harmonic mean, yielding the *Harmonic Token Anomalousness Score*. This is a very approximate measure, but still gives a useful score for outlier detection. Based on some qualitative analysis, we determined that a score of < 0.70 is a sign of a questionable dataset quality. Interestingly enough, however, datasets with scores that were too high also had issues: a score of > 0.97 often indicated that the web crawl had merely recovered the training data, which in these cases was often religious material.

After computing this score for all datasets, we manually observed samples of the data for all languages with Harmonic Token Anomalousness Score < 0.7 and more than 20,000 sentences. There were 179 languages flagged as suspicious in this way. It was relatively straightforward to make filters for 62 of these, for instance excluding sentences containing “casino” in Arabic dialects. For some of the others, we made notes that they were the wrong language. For many others, there was no clear or obvious solution, so we left them as-is. These filters removed on median 21% of the data for these 62 languages.

We acknowledge that this measure of quality is very approximate, so it should be used judiciously.

### 2.1.9 Deduplication

The last step was simply to remove duplicate sentences in all datasources. This reduced the median dataset size by a factor of 1.8x, and the average dataset by a factor of 1.4x.

### 2.2 Description of Resultant Dataset

#### 2.2.1 Monolingual Data

The result of the process described in section 2.1 was a dataset with corpora for 1503 low-resource languages, ranging in size from one sentence (Mape) to 83 million sentences (Sabah Malay). For our experiments we chose to experiment only on those 1057 languages where we recovered more than 25,000 monolingual sentences (before deduplication). We combined this with our existing monolingual datasources for the 83 high-resource languages. Table 2 shows statistics for these three datasets — the full low-resource dataset (“LRL-full”), the portion of the full low-resource dataset used for model training (“LRL-train”), and the full training dataset (“all-train”), which combines in the 83 higher-resource languages.

#### 2.2.2 Monolingual Data Quality Audit

We conducted an audit of our data as in [Kreutzer et al. 2022](#) with a variety of volunteers, comprising native speakers and non-speakers willing to do detective work. From the sample of 72 languages we audited, the median data score was 80%. The “score” is a simple heuristic to estimate the percent usable data given the different error codes, and is defined as $1.0 * cc + 0.5 * cb + 0.3 * ca + 0.2 * wd$, where cc is the percent of the sample labeled as “Correct”, cb is the percent labeled as “Correct, Low quality”, ca is the percent labeled as “Correct, ambiguous dialect”, and wd is the percent labeled as “Correct, wrong dialect” (See Appendix Section B for details).
The largest error category tended to be Wrong Language, with an average value of 16%, followed by Low Quality/Boilerplate, with an average of 10%. Languages with the poorest quality tended to be close dialects of major languages, especially varieties of Arabic (Sa’idi, Moroccan, Mesopotamian, Latinized, Algerian: ace, ar-MA, acm, ar-Latn, arg), but also varieties of French (Saint Lucian Creole: acf) and English (Nigerian Pidgin: pcm). There were also some close varieties that had higher quality, including varieties of French (Seychellois Creole: crs), Hindi (Bhojpuri: bho) and English (Krio: kri). Of these three, kri and crs use quite different orthographies than their high-resource relatives. In addition to close varieties, some African languages (Ndebele, Anaang, Kwanyama, Wolof: nd, anw, kj, wo) also had very poor quality. Less common languages with extremely high quality, e.g. Northeastern Dinka (dip), Zarma (dje), and Dombe (dov), each with a score of 100, may also be viewed with suspicion, as this may mean that a very narrow domain has been recovered (usually religious text). Details on the per-language performance and the set of error codes we used can be seen in Appendix Section B.

### 2.2.3 Parallel Data

In addition to the monolingual data described above, we also utilize the web-crawled parallel corpora available to us. This corpus is a slightly extended version of the corpus described in Arivazhagan et al. (2019), containing approximately 25 billion sentence pairs spanning 112 languages, to and from English.
3 Building Machine Translation Models for Long-Tail Languages

Next, we utilize the datasets described in Section 2.2 to build our MT models. As we see from the monolingual data statistics in Table 2, while there are over 1M sentences per language present in our corpus for the highest-resource 205 languages, the median number of sentences in our full training corpus is only around 43K sentences per language. This is a relatively small amount of data, given that traditional MT systems often require a few hundred thousand to a few million parallel sentences to reach good quality, while we are limited to around 43K monolingual sentences per language.

Given the highly data sparse setting, it is clear that the approaches that work for high resource languages cannot apply directly in our setting. In order to build high quality models for long-tail languages, we build on the approach developed previously in Siddhant et al. (2022) to enable zero-resource translation, by leveraging (i) Self-supervised training on in-language monolingual data, (ii) Massively multilingual supervised translation on out-of-language data, (iii) Large-scale data augmentation via back-translation and self-training and (iv) Model scaling. In this section we describe our recipe for training MT models for long-tail languages, starting with massive massively multilingual models and concluding with smaller, inference-friendly models trained on a selected subset of languages.

We start by highlighting the role of capacity in our highly multilingual setting by comparing the performance of 1.5B and 6B parameter models in Section 3.2. These experiments were conducted on an earlier version of our dataset, spanning 112 supervised and 94 zero-resource languages. In Section 3.3 we scale up our models to cover over 1000 languages and highlight the effect of increasing multilinguality in the long-tail setting. To further increase performance, we incorporate large-scale back-translation and self-training. In Section 3.4 we evaluate these approaches on a subset of 30 languages for practical considerations. In Section 3.5 we further highlight the role of filtering synthetic data and its effect on model quality for close dialects. Finally we describe our distillation approach and compare the resulting student models against the fine-tuned teachers, in Section 3.6.

Of the languages we evaluate on in this work, only Sorani Kurdish (ckb) uses any in-language supervised data; all other languages are evaluated under a zero-resource setting. While small amounts of parallel resources weren’t difficult to obtain for many of these languages, preliminary experiments did not result in any quality improvements with including limited amounts of (often religious-domain) parallel text.

3.1 Experimental Setup

We utilize the 1000-language monolingual corpus available to us, together with large amounts of massively multilingual parallel data in 112 languages (including English), to build zero-resource MT models for 1000 languages following the approach described in Siddhant et al. (2022). To elaborate, we train Transformer-based MT models on the translation task on 112 languages, simultaneously with the MASS masked denoising task (Song et al., 2019) to first build zero-resource models capable of translating from xx → en with reasonable quality. These models follow a two-stage training procedure, where the first stage consists of training on just the MASS and supervised translation tasks, and the second stage also incorporates data generated using online translation from the model’s latest checkpoints (in the xx → en direction). This data is used as back-translated and self-training (aka forward-translated) data simultaneously.

All our models are based on the standard Transformer architecture, with 32 layer encoders and decoders. We use two variants of the Transformer; a smaller model, with approximately 1.5B parameters, and a larger one with 2x larger model dimension and 2x wider hidden dimension, with approximately 6B parameters. Our models utilize the vocabulary from a 64K-token SentencePiece model (Kudo & Richardson, 2018) trained on monolingual data from the entire set of languages covered by the model. Data is upsampled with a temperature of $T = 5$ (Arivazhagan et al., 2019) before creation of the SPM. An analysis of the vocabulary suggests that our vocabulary is close to character-level for all but the highest resource languages. We use GPipe pipeline parallelism (Huang et al., 2019) and the Tensorflow-Lingvo (Shen et al., 2019) framework for all our MT experiments.

Different from Siddhant et al. (2022), in addition to the <2xx> token that was prepended to the source sequence to signify the target language for both translation and MASS tasks, we add a <2task> token (<2translation> for the translation task, and <2mass> for the MASS task) that specifies
the task to be performed by the model. We find this to be critical for zero-resource performance, especially when model sizes are scaled up. In the absence of this task token, our models learnt to ‘infer’ the task from the source languages instead of relying on the \(<2xx>\) token, resulting in copying sentences when provided zero-resource language sentences with the \(<2en>\) token.

All our models are evaluated with CHRF on standard evaluation sets built by translating 1200 English sentences into the target language, for 38 languages. Details of our evaluation metrics and datasets are provided in Section 5.

### 3.2 Experiments comparing effect of model capacity

We first compare the effect of model capacity on zero-resource performance. In this experiment we train our 1.5B and 6B Transformer variants on supervised data from 112 languages, and from English, and monolingual data from 206 languages (including the above 112 languages). These models are first trained with MASS and translation, and as a second stage, additionally trained on online-translated monolingual sentences as back-translation and self-training data. The results of these experiments are listed in Table 3. We demonstrate results for 38 languages, of which we didn’t include any monolingual data for 6 languages (which weren’t present in our initial scrape). We find that increasing the capacity of the model from 1.5B to 6B has a significant effect on the quality of translation, improving by an average of 2.7 CHRF on the xx→en translation direction. On the other hand, the en→xx direction regresses in quality with an average loss of -1.0 CHRF. This degradation was associated with the same issue that required us to append extra \(<2task>\) tokens to the source input (models learning to infer the task from the input language, resulting in copying instead of translation).

Looking at the overall results, we notice that the performance of these models reaches above 40 CHRF in the xx→en direction for a majority of the languages. Performance is especially high for languages which either have similar languages in our parallel corpus (including South-Asian languages and Pidgins like Bhojpuri (bho), Dogri (doi), and Nigerian Pidgin (pcm)), while CHRF scores are relatively low for lower-resource and those with no related languages in the model (including Native American languages like Quechua (qu), ‘K’iche’ (quc), Aymara (ay), and Kalaallisut (kl)). We observe that our model learns to translate Goan Konkani (gom) into English with surprisingly high quality despite the lack of gom monolingual data seen by the model, perhaps owing to its similarity with Marathi (mr), which is present in our supervised data.

### 3.3 Continual learning, extending models from 200 to 1000 languages

We next compare our 6B model trained on monolingual data from 200 languages against one trained on our entire monolingual set, spanning 1000 languages. To train the 1000-language translation model we utilize the continual learning technique described in Garcia et al. (2021a). To elaborate, we replace the vocabulary used by our 200-language MT model with one trained on monolingual data from all 1000 languages. To be able to reuse the model learnt on 200 languages to initialize the newer, 1000 language model, we align the SPM tokens shared across the two vocabularies and assign them the same IDs. The tokens in the new vocabulary that do not map to any token in the original vocabulary are assigned random IDs not assigned to the aligned tokens. This allows to continue training our 1000-language model from the model trained on 200-language subset.

We compare the performance of our 200-language model against the 1000 language MT model in Table 4. Unsurprisingly, languages that were not covered by our 200-language model but are now covered by our new monolingual data, including Goan Konkani (gom), Isoko (iso), Kalaallisut (kl) and Tsonga (ts), see large quality improvements. However, we also observe quality improvements almost for all languages in our evaluation set, including both, the xx→en and en→xx translation directions with average improvements of +2.5 and +5.3 CHRF points respectively. This is counter-intuitive since we are increasing the number of self-supervised languages, which should presumably increase the amount of interference and worsen capacity contention within this model, similar to what was observed in Siddhart et al. (2022). What is different from earlier studies is the extent of multilinguality in the model; any new added language in our model is likely to be very similar to one of the existing languages, resulting in stronger transfer between those languages. This transfer effect is strengthened by the low quantity of monolingual data for long-tail languages; in the massively multilingual and resource constrained setting the cross-lingual transfer effect dominates the
Table 3: Comparison of our 206-language 1.5B and 6B models trained on supervised data from 112 languages to and from English and monolingual data from 206 languages. Languages marked with * weren’t included in the set of 206 languages, so monolingual data was not used for these languages in this experiment.

interference observed in massively multitask models. This is also supported by the fact that languages that benefit the least from increased multilinguality are Native American languages like Quechua (qu), K’iche’ (quc), Aymara (ay), Yucatec Maya (yua), and Kalaallisut (kl), which tend to have few or no similar languages in the training data.
| Language 1  | Language 2  | Language 3  | Language 4  | Language 5  | Language 6  |
|------------|------------|------------|------------|------------|------------|
| Mono Size  | en-mfa     | en-ff      | en-mni     | en-kri     | en-mad     |
|            | 7k         | 86k        | 106k       | 129k       | 138k       |
| Direction  | ← →        | ← →        | ← →        | ← →        | ← →        |
| 200L       | 62.1       | 17.9       | 34.3       | 25.0       | 54.8       | 52.0       | 25.9       |
| 1000L      | 65.4       | 27.5       | 41.2       | 32.3       | 56.4       | 38.2       | 56.5       | 34.9       | 50.5       | 30.2       |
| Mono Size  | en-do      | en-bm      | en-ba      | en-qu      | en-ady     |
|            | 179k       | 187k       | 188k       | 250k       | 296k       |
| Direction  | ← →        | ← →        | ← →        | ← →        | ← →        |
| 200L       | 60.4       | 24.4       | 33.4       | 28.5       | 35.5       | 29.4       | 27.5       | 22.2       | 56.0       | 21.4       |
| 1000L      | 63.1       | 25.5       | 36.3       | 34.3       | 35.4       | 33.1       | 33.1       | 26.7       | 22.9       | 54.8       | 28.2       |
| Mono Size  | en-av      | en-gom*    | en-iso*    | en-yua     | en-min     |
|            | 301k       | 311k       | 409k       | 419k       | 533k       |
| Direction  | ← →        | ← →        | ← →        | ← →        | ← →        |
| 200L       | 50.5       | 27.3       | 53.3       | 0.9        | 19.3       | 16.8       | 39.3       | 31.3       | 62.6       | 54.1       |
| 1000L      | 48.1       | 28.1       | 55.5       | 39.1       | 35.4       | 33.1       | 33.1       | 40.7       | 31.6       | 62.4       | 56.1       |
| Mono Size  | en-bho     | en-kl*     | en-qu      | en-gn      | en-bbc     |
|            | 734k       | 741k       | 842k       | 861k       | 923k       |
| Direction  | ← →        | ← →        | ← →        | ← →        | ← →        |
| 200L       | 56.9       | 38.8       | 20.0       | 13.0       | 32.7       | 30.2       | 35.6       | 28.8       | 43.9       | 35.6       |
| 1000L      | 58.3       | 40.6       | 29.7       | 23.1       | 32.5       | 33.1       | 38.9       | 32.2       | 44.0       | 35.4       |
| Mono Size  | en-skr     | en-ak      | en-ts*     | en-mai     | en-ce      |
|            | 974k       | 1.29m      | 1.3m       | 1.35m      | 1.36m      |
| Direction  | ← →        | ← →        | ← →        | ← →        | ← →        |
| 200L       | 46.1       | 28.5       | 34.6       | 32.8       | 28.3       | 17.5       | 59.9       | 32.5       | 48.1       | 24.4       |
| 1000L      | 48.4       | 31.3       | 36.3       | 34.3       | 43.0       | 45.5       | 61.6       | 37.6       | 44.9       | 23.7       |
| Mono Size  | en-pcm     | en-nso     | en-il*     | en-cv      | en-ti      |
|            | 1.59m      | 1.87m      | 2.6m       | 2.9m       | 3.9m       |
| Direction  | ← →        | ← →        | ← →        | ← →        | ← →        |
| 200L       | 52.5       | 52.4       | 49.1       | 46.9       | 39.9       | 47.8       | 44.4       | 27.5       | 40.9       | 20.0       |
| 1000L      | 51.2       | 53.5       | 51.3       | 41.6       | 43.4       | 52.4       | 46.3       | 32.1       | 44.2       | 21.1       |
| Mono Size  | en-om      | en-sa      | en-dv      | en-lus     | en-as      |
|            | 5.6m       | 6.2m       | 7.9m       | 8.3m       | 9.3m       |
| Direction  | ← →        | ← →        | ← →        | ← →        | ← →        |
| 200L       | 37.0       | 36.7       | 45.8       | 24.7       | 35.8       | 39.2       | 33.0       | 19.1       | 49.6       | 36.2       |
| 1000L      | 38.1       | 39.1       | 46.3       | 28.4       | 45.2       | 43.7       | 34.5       | 39.3       | 58.6       | 36.7       |
| Mono Size  | en-mzn     | en-bew     | en-ckb*    | en-mad     |
|            | 11.6m      | 33.3m      | 25.1m      |
| Direction  | ← →        | ← →        | ← →        |
| 200L       | 53.3       | 40.2       | 51.6       | 44.0       | 51.7       | 41.6       |
| 1000L      | 55.2       | 41.4       | 51.8       | 46.0       | 54.4       | 41.6       |

Table 4: Comparison of our 206-language 6B and 1000-language 6B models trained on monolingual data from 206 languages and 1000 languages respectively, while sharing supervised data from 112 languages. Languages marked with * weren’t included in the set of 206 languages, so monolingual data was not used for these languages in the baseline model, and the 1000-language model consequently sees larger performance improvements.

While our evaluation sets are limited to 38 languages, we provide RTTLANGIDCHRF as a highly approximate measure of the quality of the model on all 1000 languages in Appendix Table 24. More details of the metric and the performance of the model on all languages are described in Section 4.3.
Table 5: Comparison of our 1000-language 6B models against the 30-language version fine-tuned from this model. Scores are shown for the languages that were fine-tuned on.

### 3.4 Effects of Large Scale Data Augmentation

In order to understand the limits of quality with our zero-resource approach, we select a subset of 30-languages for large-scale data augmentation. We continue training our 1000-language model on this subset of 30-languages with the MASS, translation and online back-translation objectives. To leverage the full power of data augmentation, we translate all the available monolingual data (up to 10 million sentences) for these languages into English, and sample around 10 million English web sentences which are translated into each of these languages. The model is then continued training with MASS, translation and offline back-translation and self-training with this synthetic data. This process is repeated twice since quality improvements with subsequent stages were observed to be incremental. This model is then compared against our vanilla 1000-language model that wasn’t trained with large-scale augmented data.

The results of this comparison are depicted in Table 5. We notice that fine-tuning with augmented data shows substantial quality improvements across the board, with especially large improvements for en→xx translation for several languages. Excepting our Native American subset, most languages reach greater than 40 CHRF on xx→en translation and greater than 35 CHRF in the reverse direction.

### 3.5 Effect of Filtering Synthetic Data

One challenge with using the model’s own predictions for improving model quality is the presence of positive feedback loops which can magnify any problems with the model outputs, as elaborated in Section 4.8. To reduce these effects we compare our fine-tuned model against a version that was fine-tuned with a round-trip translation and LangID filtered version of our synthetic data for training.
The results of this comparison are depicted in Table 6. For most languages this additional filtering has no major impact on model performance, often performing within 0.5 CHRF of the non-filtered model. However, for certain languages like Nigerian Pidgin (pcm) and Kalaallisut (kl) we observe large improvements in quality. Closer inspection reveals that our pcm data suffered from mixing with African-American Vernacular English (AAVE), which gets magnified as the model trained on its own outputs. Similarly, our kl monolingual set was polluted with a small fraction of Danish (da) data, which gets magnified with self-training. Filtering with round-trip-translation and LangID reduces the instances of data pollution improving the correctness of the model outputs. Section 4.8 explores this in greater depth.

### 3.6 Distillation

We next describe our approach for distilling our best 6B parameter Transformer model fine-tuned on 30 languages into smaller, more efficient, architectures. This process also yields further quality gains.

#### 3.6.1 Data Generation for Distillation

We follow the sequence-distillation approach [Kim & Rush, 2016] to distill our teacher model into smaller students. To this end, we generate large amounts of synthetic forward- and backward-translated data with the teacher model, which is then used to train a smaller student model. We start with the 30-language fine-tuned model and translate roughly 15-25 million English sentences into each of the 30 languages. We translate our entire corpus in these 30 languages (up to a maximum of 20 million sentences) into English. This synthetic data is then filtered using the same round-trip and LangID filtering approach described in Section 3.5. We also applied a few manual regex-based
filters for specific languages where we observed particular data pollution and noise issues, as further elaborated in Section [3.8]

3.6.2 Distillation Approach and Hyper-parameters

We looked at two candidate student architectures with increasing encoder depth; we refer to them as shallow encoder (330M parameters) and deep encoder (850M parameters). All our models are sequence-to-sequence models with attention (Bahdanau et al., 2015), using Transformer encoders and LSTM (Hochreiter & Schmidhuber, 1997) decoders, as described in Chen et al. (2018). All our student models are multilingual, with separate models for xx→en and en→xx translation.

Effect of student model capacity: Although there are some quality improvements from optimizing hyperparameters, we found that most distilled models performed similarly, having little sensitivity to the hyperparameters we experimented with. Nonetheless, one important take-away is that trends that appeared to hold for the shallow encoder model, for instance the impact of increased amounts of back-translated data, were often erased when experimenting with the deeper model.

In all cases, the shallow encoder model was noticeably worse. For 30-language models, the deep encoder saw gains of about +0.3 and +1.0 CHRF in the en→xx and xx→en directions respectively. With respect to multilinguality, we found that increasing the multilinguality of the students from 6 languages to 30 yielded small quality losses of 0.1 − 0.3 median CHRF.

Effect of amounts of synthetic en→xx data used for distillation: Another important hyperparameter was the amount of English-original synthetic data used for distillation (the non-English datasets were small enough that we could just translate the entire dataset). In the en→xx direction, where English-original data is forward-translated data, we varied the number of forward-translated sentences from 1M to 8M, but found no significant differences in model performance. In the xx→en direction, where English-original data is back-translated data, we saw consistent but small gains across all languages, with CHRF rising by about +0.6 when increasing from 1M to 2M synthetic sentences. Increasing past 2M back-translated sentences saw minimal gains. However, these experiments were carried out on the lower-capacity shallow encoder models (with 14 languages each), so more gains from higher quantities of back-translated data may be seen on a higher-capacity model.

Although the teacher model may be trained with back-translation (BT), sequence-level distillation is typically only conducted with forward-translated data (FT). However, in this case there are some interesting implications that arise from a) the very small data sizes, and b) the asymmetrical teacher model quality, where xx→en quality tends to be better than en→xx quality. For en→xx, FT (aka English-original) data is more abundant, but also lower quality; for xx→en, the FT data is higher quality but rather scarce. Therefore, we experiment with using different proportions of FT and BT data for distillation.

For en→xx, we saw very small differences in performance between different ratios, and preferred using 80% FT and 20% BT data.

For an initial experiment on a 7-language shallow encoder model in the xx→en direction, we saw noticeable losses with under 50% FT data, losing about 1.7 CHRF when going to 33% FT, and losing a further 3.7 CHRF when going to 20% FT. Values above 50% FT were not significantly different. However, when observing the CHRF curve over time, we saw that the models with more BT data were learning more slowly and probably underfitting. Replicating this experiment with the deep encoder model, the performance on all languages increases, and the aggregate differences between different ratios of synthetic data are minimal. However, there is a slight trend on a per-language basis, with higher-resource languages benefiting from more back-translated data, with a Kendall tau of 0.28 between the number of monolingual training sentences and the difference between the 70% FT and 20% model. Since the differences are very slight, we favor models with a smaller percentage of FT data (20%-50%), motivated by the intuition that the increased amount of natural target side English compared to the small number of natural source-side sentences may have benefits we can’t measure on our eval sets and metrics.

3.6.3 Comparison against Teacher Model

A comparison of our best deep encoder student models against the teacher on all 30 languages can be seen in Table [7]. The student outperforms the teacher on the en→xx direction, with an average gain of
Table 7: Performance of teacher model versus student model in CHRF.

+1.1 CHRF, and minor gains of +0.2 average CHRF on xx→en. These gains are probably in part due to the filtering applied to the RTT data. However, for the most part the differences are likely an artifact of the fact that the eval sets are English-Original, meaning that the English sentences are natural sentences, and the non-English sentences were produced via translation. Since distilled models tend to produce more translationese than their teachers, reference-based metrics like CHRF will tend to overestimate their performance in the source-original direction (en→xx), and underestimate it in the target-original direction (xx→en). This phenomenon is investigated in depth in Freitag et al. (2019).

For the 30 languages in the distilled model, we looked for correlations between the amount of monolingual data, the number of speakers, the percentage of data removed by deduplication, the harmonic data quality score (Section 2.1.8), and the CHRF score and human-rated scores in en→xx and xx→en directions. The correlations were all quite low, with Kendall’s Tau under 0.3. The only exception to this is that there was a larger correlation (τ = 0.59) between CHRF in the xx→en direction and a heuristic measure of closeness to supervised languages (e.g. Bhojpuri/bho, being very close to Hindi, gets a score of 0.9; Aymara/ay, with some loanwords from Spanish but otherwise entirely unrelated, gets a score of 0.1).
4 Evaluation

Traditional metrics like BLEU, which have enough problems with higher-resource languages [Freitag et al., 2019, 2020], have even more problems with the languages studied in the present work. Reference translations are hard to come by, and tail languages are often less standardized with respect to dialect, orthography, and even sometimes Unicode encoding. Furthermore, the frequent presence of close varieties complicates evaluation: automatic metrics like CHRF can give very high scores to outputs which are entirely in the wrong variety. Finally, as [Marchisio et al., 2021] find, outputs of unsupervised NMT are less monotonic and more natural than outputs from supervised NMT, which, like the findings on paraphrased references from [Freitag et al., 2020], produce BLEU scores that are much lower — although sometimes better correlated with human judgements of quality. Model-based metrics like YI[Lo, 2019], BLEURT [Sellam et al., 2020] or COMET [Rei et al., 2020] cannot be used for these languages due to a lack of human ratings and pretrained models in these languages.

The following sections analyze performance along a variety of axes. First we describe the evaluation sets we collected (Section 4.1). We analyze the models starting with more quantitative methods, including the performance of CHRF versus BLEU (4.2) and human evaluations (4.4). We explore RTL-LANGIDCHRF, a reference-free metric supervised for very low-resource languages, which shows reasonable correlation with CHRF. We next perform qualitative analysis of our model outputs, and highlight several patterns of errors including confusing between distributionally similar words and concepts like “tiger” and “miniature crocodile” (4.5), errors on single word inputs (4.7), and investigation of magnification of error modes in distilled models (4.8).

4.1 Evaluation Sets

In order to measure the quality of translation for development and experiments, we collected reference translations for 38 languages. For ease of comparison we collected a multi-way parallel data set, with the same English side for all languages. The xx→en eval sets were made by reversing this dataset. Rather than opting for larger evaluation sets for a small number of languages, we decided to collect relatively small evaluation sets containing 1200 sentences in 38 linguistically and geographically diverse languages. 50% of the English sentences were drawn from a corpus of simpler and more colloquial language (average length: 12.0 tokens; $\sigma = 4.45$), and the remaining 50% from more technical web content (average length: 20.1 tokens; $\sigma = 11.6$). The resulting dataset was shuffled and then split into a 600-sentence development set and a 600-sentence test set. The tables in this paper report the score across the combined set except where otherwise mentioned. The average absolute difference between the CHRF scores on both sets was about 0.5.

4.2 Evaluation Metrics

For this work we deem token-level metrics like BLEU unsuitable, and rely almost entirely on the character level CHRF [Popovic, 2015], reported on a scale from 0 to 100. This section explains this decision, and gives examples of where the two metrics differ on the languages we studied.

Many of the languages studied in this work have complex morphologies, including the agglutinating Bantu languages and the polysynthetic Native American languages. This means that they can inflect to form very long tokens. An extreme example can be seen in Table[2], which shows how full sentences in English translate to only a few tokens in Kalaallisut (k.j). For such highly inflecting languages, it is to be expected that a character-based metric, like CHRF, would correlate better with quality than a token-based metric like BLEU. This expectation is borne out by our observations. These observations align with Mirzakhalov et al. [2021b], who observed similar trends for the agglutinating Turkic languages.

It is difficult to make any sort of direct comparison between BLEU and CHRF. Not only do they measure different quantities and have different score distributions (e.g. CHRF = 0 is exceedingly unlikely), but they also are influenced by different artifacts. The BLEU score will be affected by infection, and a slightly wrong infection will nullify all affected n-grams. On the other hand, CHRF is affected by the writing system – for languages with abugidas or abjads, for instance, it will...
we have defined a very simplistic conversion metric based on performance of the less-inflecting (Quechua/To compute this score, we simply round-trip-translate a corpus of English sentences through some language according to our SSLID model. If fewer than 10% of the intermediate translations receive

| lp | BLEU | CHRF | ratio | lp | BLEU | CHRF | ratio |
|----|------|------|-------|----|------|------|-------|
| en→om | 3.8 | 40.1 | 4.0 | en→ts | 12.8 | 46.7 | 1.6 |
| en→lg | 3.7 | 39.8 | 4.0 | en→gom | 10.8 | 42.3 | 1.6 |
| en→dv | 5.4 | 45.5 | 3.5 | en→see | 11.9 | 40.5 | 1.3 |
| en→sa | 2.8 | 33.1 | 3.5 | en→ff | 9.9 | 37.2 | 1.3 |
| en→squ | 3.6 | 36.2 | 3.4 | en→bm | 9.5 | 36.4 | 1.3 |
| en→ln | 3.5 | 34.9 | 3.2 | en→lus | 13.2 | 41.5 | 1.2 |
| en→kl | 5.3 | 40.6 | 2.9 | en→kri | 10.6 | 36.2 | 1.1 |
| en→ay | 4.6 | 34.3 | 2.4 | en→ilo | 24.2 | 54.9 | 1.1 |
| en→ckb | 9.4 | 44.2 | 1.9 | en→ak | 10.4 | 34.6 | 1.0 |
| en→gn | 5.2 | 32.1 | 1.7 | en→nsa | 20.3 | 47.6 | 1.0 |
| en→mai | 8.6 | 39.8 | 1.7 | en→bho | 16.7 | 42.7 | 1.0 |
| en→as | 9.6 | 40.6 | 1.6 | en→doi | 15.3 | 40.7 | 1.0 |
| en→mni-Mtei | 12.9 | 47.1 | 1.6 | en→ti | 3.7 | 22.1 | 0.4 |

Table 8: CHRF versus BLEU scores on en→xx language pairs in the distilled model. Although it is almost impossible to compare these scores directly, as explained for a variety of reasons in the text, it is still clear that they give a very different picture of the performance on the languages. In order to make an approximate comparison between the two, we have included the ratio of the SCALEDCHRF score to the BLEU score. Language pairs in this table are sorted from the highest ratio (where BLEU underestimates performance) on the top to the lowest ratio on the bottom. As expected, polysynthetic and agglutinative languages are misjudged my BLEU.

However, since this metric is trivially fooled by error modes like copying or translating into the wrong language, we omit any round-trip translation where the intermediate translation is not in the correct language according to our SSLID model. If fewer than 10% of the intermediate translations receive

Table 9: Example translations from English to Kalaallisut from our evaluation set. A token-based metric like BLEU is unsuitable for such a highly inflecting language, as exact match is very unlikely. Character-based metrics, like CHRF, are more appropriate.

| source | translation |
|--------|-------------|
| I can’t sing so well so I don’t want to be a singer. Without my car, I wouldn’t be able to get to work. I would like to raise a dog instead of a cat I have a cat who is lazy. | Erinangippallaanginnama erinarsortartunngorusunngilanga. Bileqangikkuma sulialtortananaanvianngilanga. Qimmuteqaruasunnerussangaluarpunga qitsuuteqarnissannit. Qitsuuteqarpunga eqigasutuuoq. |

behave differently than for languages with alphabets, as there will be fewer characters to match. (On languages with writing systems like Chinese or Japanese, it will function very similarly to BLEU.) Nonetheless, in order to make some attempt to demonstrate how and where the two metrics differ, we have defined a very simplistic conversion metric based on performance of the less-inflecting low-resourced languages we studied: \( \text{SCALEDCHRF} = 0.75 \times \text{CHRF} - 0.15 \). The purpose of this metric is solely to have some way of flagging which languages lead to very different performance under the two metrics. Table 8 shows the language pairs in our distilled models, along with their BLEU score, their CHRF score (unscaled) and the ratio between SCALEDCHRF and BLEU. The results correspond with intuitions: The languages with the highest ratio tend to be polysynthetic (Quechua/qu, Aymara/ay, Kalaallisut/kl), agglutinative (Luganda/lg, Lingala/l), or otherwise highly fusional (Sanskrit/sa ). Oromo (om) is worth a special mention, as its orthography seems to have higher character usage per morpheme (because of the many doubled letters), which may inflate CHRF. It is not clear why Dhivehi (dv) has such a high ratio. The performance of both metrics on Tigrinya (ti) is also something of a mystery, given that the model translations were rated very highly by humans, and native speakers validated that the reference translations were also high quality.

4.3 RTT LANGID CHRF

Because of the infeasibility of collecting reference translations for 1000+ languages, some sort of reference-free evaluation is increasingly important. Round-trip translations have been utilized as a metric for MT Quality Evaluation several times over the last few decades (Huang, 1990; Aiken & Park, 2010; Moon et al., 2020). We experimented with a simple modification we call RTTLANGIDCHRF. To compute this score, we simply round-trip-translate a corpus of English sentences through some language, and compute the CHRF score of these translations with respect to the original sentences. However, since this metric is trivially fooled by error modes like copying or translating into the wrong language, we omit any round-trip translation where the intermediate translation is not in the correct language according to our SSLID model. If fewer than 10% of the intermediate translations receive

22
the correct LangID score, we consider the score invalid, as this may also be a result of errors in the LangID model.

We computed correlation between CHRFF and RTTLANGIDCHRFF over 30 languages, and found it to correlate moderately well both in the en→xx direction (ρ = 0.46, τ = 0.34) and the xx→en direction (ρ = 0.30, τ = 0.28). When we recalculated only on the scores from the distilled models, the correlations were much better in the xx→en direction (ρ = 0.80, τ = 0.69), but similar in the en→xx direction (ρ = 0.96, τ = 0.22).

In addition to this, we computed this score over all languages in the 1000-language model, of which 630 passed the LangID > 10% threshold. Figure 2 shows these scores as a function of log data size. There is a relatively clear trend (ρ = 0.80, τ = 0.63), and the large majority of languages with over 100,000 monolingual sentences have relatively high scores. In general, the languages above the trend line are close dialects to high-resource languages, most notably variants of English written in different scripts. Languages below the trend line tend to be unrelated to high-resource languages or have poor-quality data according to our data audit.

Of the 630 languages with valid RTTLANGIDCHRFF scores, 268 have a RTTLANGIDCHRFF score of over 30.0, which we tentatively deem of “hopeful quality”, and 147 have RTTLANGIDCHRFF ≥ 50, which is the minimum score from any of our supervised languages. Interestingly enough there is actually fairly low correlation between RTTLANGIDCHRFF and the percent of intermediate translations that were assigned the correct LangID score (ρ = 0.18, τ = 0.10). One possible explanation is that frequent error modes are outputs in the wrong language and copying, phenomena that we observed in Section 4.8.

We include RTTLANGIDCHRFF for all 630 languages in Appendix Table 24, as an approximate measure of translation quality of the model. However, since we have done only cursory analysis of how effective this score is at measuring quality, we advise readers to view it with appropriate skepticism. While a low RTTLANGIDCHRFF probably means low translation quality, a high value may well mean something other than high quality.

The version of this score described above we call the loose version of RTTLANGIDCHRFF, since it does not penalize the model for intermediate sentences in the wrong language. To get the strict RTTLANGIDCHRFF, we multiply the loose RTTLANGIDCHRFF by the percent of intermediate translations assigned the correct LangID score, thereby penalizing wrong-language translations. This version does not correlate well with CHRFF on the 30 languages where we have evaluation sets, and in fact has negative correlation in the xx→en direction (ρ = −0.28, τ = −0.20), and only a weak correlation in the en→xx direction (ρ = 0.12, τ = 0.09), likely due to noise from the LangID model. It also has weaker correlation with the size of the monolingual dataset over all 630 applicable languages (ρ = 0.43, τ = 0.55), and is noticeably messed up by close dialects, e.g. assigning Bosnian (bs) a low score because intermediate translations were frequently LangID’d as Croatian (hr), a mistake that should not be penalized as the two languages are frequently indistinguishable. Despite these failings, it has the attractive property that it penalizes the common error mode of wrong-language outputs; and furthermore, the supervised and zero-shot languages appear more clearly separated when plotting them versus monolingual data size. The graph is therefore included in Appendix B.1.

4.4 HUMAN EVALUATIONS

Any decision on translation quality ultimately cannot be made with automatic metrics like CHRFF alone. In order to understand the quality of our distilled models, we asked human raters to rate the quality of the translations from our test set on a scale from 0 (nonsense or wrong language) to 6 (perfect). Full results may be seen in Appendix Table 20.

Although we made an attempt to calibrate raters and explain each point in the scale very clearly, each rater will naturally have a different understanding of “a good translation”. For this reason, it is very difficult to interpret these results in any sort of holistic way. However, with some diving into the results together with native speakers, a few things stood out.

The biggest takeaway is that automatic metrics overestimate performance on related dialects. Nigerian Pidgin (pcm), a dialect of English, had very high BLEU and CHRFF scores, of around 35 and 60 respectively. However, humans rated the translations very harshly, with a full 20% judged as
Figure 2: Plot of RTTLangIDCHRF scores (loose) for languages as a function of log monolingual data size. With over 100,000 sentences, almost any language does reasonably well. Outliers are labeled with their language code. The largest outliers are English in Cyrillic script (en-Cyr1), which has excellent RTTLangIDCHRF score but very little monolingual data, and Tibetan (bo), which has plenty monolingual text but very poor performance. In general, the languages above the trend line are close to high-resource languages (where the metric may also be fooled), and the languages below the trend line are linguistically distant from other languages in the model or have poor-quality data. Languages added to Google Translate as part of this effort (all unsupervised except Sorani Kurdish (ckb)) are marked with stars.

“Nonsense/Wrong Language”, with trusted native speakers confirming that the translations were unusable. Krio (kri – close to English), Maithili (mai – close to Hindi), and Bhojpuri (bho – close to Hindi) were in a similar boat, though trusted native speakers agreed that the translation quality, though borderline, was usable. What’s happening here that the model translates into (a corrupted version of) the wrong dialect, but it is close enough on a character n-gram level that the CHRF is still high. In our cases, this is the result of a data pollution problem. Since these languages are so close to other much more common languages on the web – in this case, English and Hindi – the training data is much more likely to be mixed with either corrupted versions of the higher-resource language, or other varieties. As a result, many model outputs that were supposed to be in Dogri (doi) were actually in misspelled or ungrammatical Hindi (hi), outputs supposed to be in Nigerian Pidgin (pcm) were sometimes in other English dialects like AAVE, and so on.

4.5 Mistakes on Distributionally Similar Words

We observe that our zero-resource models exhibit some characteristic error modes. The most common one relates to translating nouns that occur in distributionally similar contexts in the training data. This occurs even for relatively common nouns like “tiger” – which is often translated as another kind of animal, showing that the model learned the distributional context in which this noun occurs, but was unable to acquire the exact mappings from one language to another with enough detail within this category. This may be related to the relatively small amounts of training data that were used, alongside the unsupervised nature of training. Common nouns suffering from these mistakes include animal names, colors, and times of day. This was also an issue with adjectives, but we observed few such errors with verbs. Sometimes, words were translated into sentences that might be considered culturally analogous concepts – for example, translating “cheese and butter” into “curd and yogurt” when translating from Sanskrit (sa). Surprisingly, the model hypotheses were often strings that would probably yield a high perplexity under most language models. Table 10 provides a variety of
Combinatorial Method

"red", mediocre at translating "yellow", and poor at translating "orange". This observation is consistent with the cross-lingual hierarchy of color terms described in Berlin & Kay (1969); Saunders & Brakel (2002), which find that terms for color generally arise in specific orders across the world’s cultures, e.g. into Tsonga as *ririmi ra xitsonga* or into Oromo as *Afaan Oromoo*

Table 10: Examples of correct translations (blue) and mistranslations (orange), illustrating the model’s tendency to make mistakes on distributionally similar nouns.

| sl | reference | translation |
|----|-----------|-------------|
| ak | I believe a *lion* is stronger than a *tiger*. | I think the *hyena* is hotter than the *elephant*. |
| dv | I believe a *lion* is stronger than a *tiger*. | I believe a *lion* would be *stronger* than a miniature *crocodile*. |
| mni | I believe a *snake* is stronger than a *crocodile*. | I believe seizures are more severe than epilepsy. |
| dos | I believe a *lion* is stronger than a *tiger*. | I think a *rabbit* is stronger than a *squirrel*. |
| ff | The first three colors are *red*, *orange*, and *yellow*. | The first three colors are *red*, *yellow*, and *blue*. |
| mui | The first three colors are *red*, *orange*, and *yellow*. | The first three colors are *red*, fire red, and *yellow*. |
| qu | The first three colors are *red*, *orange*, and *yellow*. | The first three colors are *red*, yellow and *safron*, and *the first three colors are red, purple and pink. |
| sa | The first three colors are *red*, *orange*, and *yellow*. | The first three colors are *red*, yellow and safron, and the first three colors are *red, purple and pink. |
| ts | The first three colors are *red*, *orange*, and *yellow*. | The first three colors are *red, orange*, and *yellow*. |
| ysa | In this I use *capsicum*, tomatoes, onions, garlic, *green chilies*, olives, etc. and do not use *cheese* and *butter* in it. | Here I deal with greater marjoram, blood fruit, plantain, lassí, green marjoram, jujube, etc., but I do not deal with curd and yogurt. |
| ak | I would want to be a *dog* for a day. | I want to be a *crocodile* just one day. |
| ak | I would ask my *cat* some questions about what he’s always trying to tell him by meowing. | I will ask my *friend* some questions about why she is crying. |
| ak | My *dog* keeps me moving and enjoying life. He is man’s best friend. | My pet *cat* teaches me how to live a healthy lifestyle and enjoy being with people. |
| ak | I went to a *carnival* yesterday that was located in the middle of nowhere under a huge *red* and white circular striped tent. | I went to a *place of pleasure* in a desolate place, unknown to me, where there was a parrot that had a golden-yellow coat. |
| ak | John was working in the *lighthouse* and went for a walk on the beach one night. | John was working in the *synagogue* and he was sound asleep one afternoon. |
| ak | This is why *hair* turns grey with age. | This is the reason why ticks change into *womans* after a certain period of time. |
| sa | My bad habit is that I *eat* too much. | My bad behaviour is that I eat poison. |
| sa | *Dogs* are very intelligent animals, they understand very much about humans. | Cockroaches are extremely sharp minded animals, they know about humans correctly. |
| sa | Susy loved her *space book* and would ask her parents to read it to her every night. | Susie loved her *newspaper*, and she asked her parents to read it to her in the morning. |

examples of mistakes from these models. These examples are from the 6B parameter 1000-language Transformer model, described in Section 3.3.

A good example of what it means for words to have different meanings but to be distributionally similar given the usage of the language is the translation of the string “English Language”. Around 25% of the languages translated the string “English Language” into the name of their own language, e.g. into Tsonga as *ririmi ra xitsonga* or into Oromo as *Afaan Oromoo*.

Another interesting phenomenon evident from Table 10 is that models seem to be good at translating “red”, mediocre at translating “yellow”, and poor at translating “orange”. This observation is consistent with cross-lingual hierarchy of color terms described in Berlin & Kay (1969); Saunders & Brakel (2002), which find that terms for color generally arise in specific orders across the world’s cultures, with words for “red” occurring in stage II of color development, “yellow” in stage III/IV, and “orange” in stage VII.

An interesting parallel can be found between this problem and an issue of “unsupervised” translation in the real world – the case of the words for *four* and *six* in Etruscan, a language that went extinct about 2,000 years ago (Freeman 1999). There does not currently exist any surviving parallel text with Etruscan, excepting a single 37-word tablet, but there do exist some 13,000 monolingual inscriptions. Etruscan does not seem obviously related to any living languages, so meanings usually cannot be discovered by their similarity to cognates. As a result, modern scholars have to use a sort of “unsupervised translation” called the Combinatorial Method, a first-principles approach to discovering word sense and grammar. And in the case of Etruscan, there are some words whose meaning cannot be teased apart with existing context. As an example, the words for all numerals from one to ten have been established with some certainty – with the exception of *four* and *six*. For these two there is no clear contextual clue to separate them, and there is no academic consensus of which of *huθ* and *śa* mean *four* and *six* (Artioli et al. 2011).
4.6 PERFORMANCE ON TOKENS BY TOKEN FREQUENCY

In order to quantitatively measure the distributional token errors from Section 4.5, we decided to look at the accuracy of our model at translating particular tokens as a function of their frequency in an open-domain English web corpus.

First, we collected a list of the 8000 most common tokens in a large, web-crawled, monolingual English corpus. Then, we separated these into exponentially-spaced bins, based on the exponential distribution of token frequencies in natural text (Zipf, 1935). Therefore, each bin corresponds to a set of English tokens in a certain frequency band. Then, for each of these bins, we made an evaluation set for this bin composed of all sentences in our standard xx→en evaluation set containing these target tokens at least once in their references.

To score a model on a bin-specific evaluation set, we looked at a simple hit-rate metric. For a given reference sentence containing \( k \) tokens in the set of target sentence, the model got one point for each token it produced in its output that was among these \( k \) tokens (a “hit”), for a maximum of \( k \) points. The hit-rate score for that eval set is then simply the number of hits the model got divided by the total number of possible hits. Formally, for a given set of tokens \( B \) (in our case a frequency bin), a list of reference translations \( r_i \in R \) and a corresponding list of model hypotheses \( h_i \in H \), the hit-rate \( \Xi(H, R, B) \) is defined as follows:

\[
\Xi(H, R, B) = \frac{\sum_i \sum_{t \in r_i} \mathbb{1}\{t \in h_i \wedge t \in B\}}{\sum_i \sum_{t \in r_i} \mathbb{1}\{t \in r_i \wedge t \in B\}}
\]

Table 11 shows the result of this analysis, sorted by the number of monolingual sentences seen by the model. Hit-rate is reported on five bins, starting on the most frequent tokens (tokens #0 - #125) and ending with the most infrequent tokens (tokens #8000 - #12800). For each bin the number of sentences in this eval set is included in the column labeled \( \text{sents} \). The most frequent bin includes 1191 of the 1197 sentences in the full eval set; the least frequent bin includes only 958 of them. For each bin the total number of possible points (i.e. hits) is also reported in the column labeled \( \text{pts} \). Whereas the first bin has a total of 9682 possible points – averaging eight per sentence, and constituting mostly function words – the least frequent bin had only 2420 possible points, so slightly above two per sentence. It is also worth mentioning that the least frequent bin had 6,000 possible tokens in it, so only about 30% of them actually occurred in our eval set.

For many languages, higher CHRF means higher hit-rate across the board, and for others, like K’iche’ (quc) low CHRF corresponds with generally lower hit-rate. Things get more interesting for languages that have higher hit-rates for the first several bins of higher frequency tokens, but lower hit-rates for less common tokens. These are the ones that have higher CHRF scores, corresponding to the ability to translate the top 500 tokens or so very well, but that also make frequent mistakes on less common tokens. Perhaps the best example of this is Fulfule (ff), which had the relatively high CHRF of 40.7, but sees a large drop-off in hit-rate from 41% on bin 2 (token #125 - #500) to 26% on bin 4 (token #2000 - #8000). This is also the language where, anecdotally, we observed several of these sorts of mistakes. Mizo (lus) and Bambara (bm) exhibit similar patterns.

| sl  | tl  | source      | translation |
|-----|-----|-------------|-------------|
| en  | fl  | devices     | kabir‘e (dispositifs) |
| en  | lus | fragile     | hring hring (fragile) |
| en  | pcm | removes     | dey remove |
| en  | pcm | blame       | blame wetin? |
| en  | lus | four        | pali a awm a |
| en  | lus | freedom     | zalemma a awm |
| en  | quc | solved      | xsol rij ri |
| en  | quc | dam         | ri k‘o pa ri cho |
| en  | kri | juvenile    | pilik we no rich 18 ia yet |
| en  | kri | notorious   | pipul den we get badnem |

Table 12: Our models frequently displayed exceeding verbosity with single-word inputs. In some cases they would add extra definitions for words in parentheses or commas afterwards (top half of table). In other instances they added function words from that language after the translation (boldface), or gave whole definitions, as with Krio.
Table 11: Token hit-rate (Ξ) for different sets of tokens, binned by frequency. The lower and upper bound of the token frequency rank per bin are given in the lb and ub columns; thus, the top row is the hit-rate on the 125 most frequent tokens, and the bottom row is on the 6,000 least frequent tokens. Columns sents and pts give the number of sentences in each bin-specific eval set and the number of tokens in that bin occurring in those references, respectively. The CHRF score and number of monolingual training sentences is also given. The interesting results are on the languages, like Fulfulde (ff), that have a high hit-rate on more frequent tokens and a lower hit-rate on rarer tokens. All results are only in the xx→en direction.

4.7 Errors on short inputs

Another category of errors we encountered was with single word inputs to the model. The translations tended to be longer, and the model would frequently give alternate translations or append frequent tokens (Table 12). Outputs also frequently had duplicates among the other outputs, suggesting hallucination, or copied inputs. A breakdown of types of errors can be found in Table 13.

This was an issue we observed mainly for lower-resource languages. For higher-resource languages, like Ilocano, the model tended to provide a succinct and correct single-word translation. Ensuring that the MASS training data covers the lower end of the length distribution would likely remedy these issues. However, this is also an inherently difficult problem since we do not provide the model any source language information. For the xx→en direction, the translation task needs to solve LangID and translation simultaneously, which could often be ambiguous and quite challenging for short queries. This can potentially be addressed by providing the model source language information along with the input, and is left to future work.

Table 13: Statistics for single-token translation on the top ten thousand most common tokens from the monolingual datasets, compared for 48 language pairs, for the fine-tuned teacher model. Metrics shown are 1) percent of outputs that copy the source; 2) percent of input that is multi-word; 3) percent of outputs that are identical to other outputs, suggesting incorrect translation; 4) percent of outputs with low character diversity, suggesting repeats; and 5) percent of outputs with none of those features. The entries with full-stops (.) after them use the “period trick” (Section 5.3). Single word translation is a particularly tricky task as the model needs to solve translation and LangID simultaneously, which can be undefined for short queries.


| Language | % Mono | % Synth (Clean) | % Synth (Noisy) |
|----------|--------|----------------|-----------------|
| pcm      | 8%     | 20%            | -               |
| kl       | 14%    | 18%            | 30%             |
| sa       | 13%    | 16%            | 37%             |
| ja       | 0.0004%| -              | 1%              |

Table 14: Percent of each data set removed after applying specialized filters. mono refers to the monolingual data from the corpus mined in Section 2.2. synth(clean) refers to synthetic data generated by forward translating a clean English monolingual corpus and synth (noisy) refers to synthetic data generated from a noisy web-scraped English corpus. We see that the synthetic text generated by the teacher model exhibited these problems to a greater degree than the monolingual web-sourced data, and that these problems intensified on noisier data.

| LP | this method | M2M-100 | LP | this method | M2M-100 |
|----|-------------|---------|----|-------------|---------|
| en→as | 29.1 | 1.22 | as→en | 34.5 | 3.76 |
| en→ckb | 28.5 | 0.23 | ckb→en | 37.5 | 7.65 |
| en→ff | 2.5 | 0.68 | ff→en | 11.2 | 2.4 |
| en→lg | 17.3 | 0.61 | lg→en | 29.3 | 4.45 |
| en→ln | 24.7 | 1.03 | ln→en | 30.2 | 4.57 |
| en→nso | 32.5 | 1.54 | nso→en | 45.0 | 6.76 |
| en→om | 17.2 | 0.4 | om→en | 30.8 | 3.33 |

Table 15: Flores dev-test: comparing SP BLEU between this method and Goyal et al. [2021].

4.8 Magnification of Error Modes in Student Models

After training distilled models, we noticed a variety of unexpected error modes when analyzing the output translations:

- The translations to Nigerian Pidgin (pcm) frequently instead translated to (often offensive) US slang. For instance, the English sentence “She said to herself” was translated to the unacceptable string “da b***** say ta da b*****self.”
- Many of the translations to Kalaallisut (kl) were actually Danish (da)
- Many translations to Sanskrit (sa) were actually Hindi (hi)

We developed filters to remove this content from the forward translated data and distilled the models again. We observed that these problems were more prevalent in the synthetic data used for distillation (generated by the teacher model) than in the monolingual data that had originally been used to train these models, and that the issues were more severe for synthetic text produced by translating a noisier source corpus. The changes in noise level are illustrated in Table 14. We hypothesize that this error magnification could either be an artifact of a positive feedback loop arising from training the model on its own prediction (self-training), or due to a difference in domains between the training and distillation datasets.

This table also includes one entry about distillation from Japanese to English. In this case, we found that one particular Amharic string was often hallucinated. This string occurred occasionally in the target side of the original training data, and then occurred much more often in the distilled data.

4.9 Comparison on Flores Benchmark

Since these models are not trained on the same data as public benchmarks, a comparison on public benchmarks is not necessarily very meaningful. Nonetheless, in Table 15 we provide a comparison between the SP BLEU results from our method (on distilled models) versus the Flores-101 benchmark scores (Goyal et al., 2021) reported for the massively multilingual M2M-124 (Fan et al., 2021) for overlapping languages. Given the higher language coverage in our monolingual dataset, our models yield higher SP BLEU for all language pairs.
5 ADDITIONAL EXPERIMENTS AND NOTES

5.1 NON-ENGLISH-CENTRIC BRIDGING

There is nothing inherently English-centric about the approach to zero-resource translation put forth in this paper. Nonetheless, the model has only seen translated text between English and other languages, so it would be a reasonable hypothesis that it is just inherently better at translating to and from English, even in the zero-shot scenario.

We test this hypothesis by evaluating the model on bridged translations. We first translate the English source sentences to other languages using bilingual supervised models on these language pairs. Then we use the model proposed in this paper to translate these translations directly into the desired target language. For each desired target language we pick 1) the closest mid- or high-resource language (HRL) that we expect our bilingual models to do well on; and 2) if applicable, a lower-resource language (LRL) that may be closer to the desired target language, but have lower quality supervised models. Please note that the definition of close is somewhat approximate. For instance, for Native American languages we choose colonial languages as the “close” languages, because they may share a certain amount of vocabulary, even if the grammar is entirely divergent. Furthermore, for some of these languages the “close” languages are in fact not very close at all, as with the Sino-Tibetan languages of North-East India (Bodo (brx), Meiteilon (mni-Mtei), Mizo (lus)), which are only somewhat related to the “closer LRL” of Myanmar/Burmese (my).

Table 16: Results for bridged translation from English (CHR F) on 1200 sentences/language. Bridging seems to improve quality only when the intermediate language is both higher-resource and close to the target language.

Table 16 shows the results of this investigation, along with which languages were chosen as “close high-resource languages” and “closer low-resource languages”.

We find that in a substantial number of cases, bridged translation scores better on automatic metrics. This is true especially for the languages of India, with bridged translation drawing even or improving even for the non-Indo-European languages. Overall, the largest improvements are seen on close dialects, for instance Maithili (mai: +6.2 CHR F), Bhojpuri (bho: +1.4 CHR F), Kashmiri (ks: +7.4 CHR F), and Cantonese (yue: +1.5 CHR F). Aymara (ay) and Dogri (doi) also saw noticeable gains. However, most Native American languages, Bantu languages, and languages without close relatives saw large losses by bridging.

4For this experiment, we considered Swahili (sw) to be a high resource language. This is not precisely accurate, but it was the closest mid-resource language available.

Table 16
Nonetheless, in 19 out of 28 cases, we find that direct translation from English produces the highest CTHRF score. This may be due in part to the model being inherently better at translating from English, and in part from errors compounding from the two-step process. Evidence for the second hypothesis may be seen in the fact that only two languages (Kashmiri (ks) and Tigrinya (ti)) see any gain from bridging through a lower-resource language. Overall, it seems that bridging only improves quality when the intermediate model is already relatively high quality, and additionally when the intermediate language is close to the target language.

The fact that bridging appears to work relatively well is a good sign. In practice, translation to and from English is not the major use case for many low resource languages, and direct models to local languages (e.g. Hindi for India, Spanish for Latin America, etc.) would likely provide more utility to local communities.

5.2 ZERO-SHOT TRANSLITERATION

Many of the world’s languages are written in multiple scripts, whether because of historical and national changes, informal online usage, non-standardized writing systems, or different ethnic or religious populations. For this reason, especially with under-resourced languages where this is more common, it is especially important to be able to support these languages in their many different writing systems.

For our data collection efforts, we crawled data in a variety of different scripts for several languages, e.g. both Malayalam in Malayalam script (ml) and Malayalam in Latin script (ml-Latn). We treated script-variants of the same language as separate languages, with their distinct tags. To transliterate, we simply asked the model to provide “translations” from text in one script to the same language in another script.

We applied this approach to transliterate from Latinized variants of Indian languages to their native scripts, and found that it worked very well out of the box, appearing to be more robust than existing transliteration libraries to informal or nonstandard spellings. One example is the common abbreviation kr in Latinized Hindi (hi-Latn), as in kya kr rhe ho which our model correctly transliterated as “kar” (in Devanagari script), whereas the rule-based model incorrectly rendered as “ke”. Similarly the model was able to do some spelling correction, e.g. on the the misspelled Konkani swpnatat, which the model correctly transliterated to “swapnaat”. However, the model also had a tendency to change small parts of the input sentence, as well as occasionally hallucinating extra content. This also made it difficult to compare this model to rule-based approaches in a rigorous way, because the lack of guaranteed monotonic alignment made word error rate inapplicable.

One interesting example highlighting issues this approach has is the Konkani sentence Xet-camot ani knenedeam gaucho vaur vo dondo. This was transliterated mostly correctly to Devanagari, but the dialect was changed: the original text is Goan Roman Catholic Konkani, but the transliteration was in Goan or Maharashtrian Konkani, changing “camot” to “camat” and “vo dondo” to “vaa dando”. Future work is necessary to coax this technique into a form that does not take any liberties with the input. One promising direction is to separate the <2xx> tokens (Johnson et al., 2016) into a language subtag and a script subtag. For instance, <2ms> becomes <2ms> <2Latn>, and <2ms-Arab> becomes <2ms> <2Arab>. This enables 0-shot transliteration between any scripts supported by the model, simply by using the desired <2Script> tag at inference time. In order to prevent the model editorializing, it would also be advisable to consider adding synthetic transcribed parallel text with a <2transliteration> tag, to teach the model that this task only involves substituting letters/sounds.

5.3 THE “PERIOD TRICK”

Even after having been distilled on a mix of clean and noisy data, we observed that some of these languages still had lower performance on inputs that lacked terminal punctuations. To study this, we compared performance on the evaluation sets with and without terminal punctuation. Table 17 illustrates the results of this experiment on the distilled models, though we noticed similar trends in the teacher models. The gain is small but consistent, and in the xx→en direction, 100% of the language pairs benefit. We noticed that sentences without terminal punctuation sometimes triggered common error modes, e.g. decoding into Danish (da) instead of Kalaallisut (kl), or misspelled
Table 17: Comparing CHRF on versions of the evaluation sets with and without terminal punctuation (TP) on 26-language distilled models. The Win/Loss ratio is also reported, meaning the fraction of language pairs that saw an increase in CHRF by applying the terminal punctuation.

| direction | no TP | TP | ∆   | W/L |
|-----------|------|----|-----|-----|
| en→xx     | 39.32| **39.56** | +0.24 | 0.77 |
| xx→en     | 48.54| **49.23** | +0.69 | 1.0  |

Hindi (hi) instead of Dogri (doi). We hypothesize that the presence of terminal punctuations might provide a “domain” signal to the model and trigger different translation qualities.

5.4 Robustness to Non-standard Glyph Usage

There are many different ways to write certain letters, especially those where a Unicode standard was introduced after a population was already active online, or before keyboards using this standard were widely available. Common cases include the many Unicode points for the “open o” (œ) and “open e” (œ) used in many African languages; the Palochka (resembling the letter I), used in many Caucasian languages; the apostrophe or ‘Okina, used around the world but especially in many American languages; and many other examples that can be found in the Unicode Confusables list (Davis & Suignard 2021). Table 18 gives an example of this phenomenon in the wild, showing the variation of Unicode points used for the Chechen and Chuvash languages in our webcrawled data. We refer the reader to Prasad et al. (2018) for many more examples of this phenomenon.

We conducted a simple experiment to determine how robust our model was to these different usages. Looking at translations into English, we decoded our evaluation sets (with the non-finetuned, 1000-language teacher model) using each of a variety of different ways of representing each letter. We then compared the CHRF scores on the output. Results can be seen in Table 19. We found that in most cases, this made very little difference in CHRF, even when using rare glyphs like the capital Greek Iota in place of the Palochka in Caucasian languages, indicating that our models were quite robust to these variations. However, we did notice a relatively large change in CHRF for West African languages when using nonstandard glyphs, including the “chatspeak” ASCII characters used when texting or writing very informally – for instance, writing “aho³f3” for “aho Œf¢”.

| Unicode character name | Percent of Data |
|-----------------------|-----------------|
| LATIN CAPITAL LETTER I | 33.6%           |
| DIGIT ONE             | 30.4%           |
| CYRILLIC LETTER PALOCHKA | 3.8%         |
| CYRILLIC CAPITAL LETTER BYELORUSSIAN-UKRAINIAN I | 4.2% |
| LATIN SMALL LETTER L  | 2.8%            |
| GREEK CAPITAL LETTER IOTA | 0.4%       |
| LATIN SMALL LETTER A WITH BREVE | 52.4% |
| LATIN SMALL LETTER E WITH BREVE | 55.0% |
| LATIN SMALL LETTER C WITH CEDILLA | 43.4% |
| CYRILLIC SMALL LETTER IE WITH BREVE | 10.2% |
| CYRILLIC SMALL LETTER A WITH BREVE | 9.4% |
| CYRILLIC SMALL LETTER ES WITH DESCENDER | 8.8% |

Table 18: Examples of the different Unicode points used to encode the Palochka character in the Chechen language (above) and letters with diacritics in the Chuvash language (below), along with their prevalence on web-crawled data. In both cases, the “correct” Unicode point (bolded) is much less common.

5.5 Non-Unicode Fonts

Before easy access to keyboards using the correct Unicode points, or before the Unicode standard itself, it was often not clear how to represent alphabets or characters that were not in the ASCII range. We already saw one consequence of this in Section 5.4. A more difficult consequence is the existence of non-Unicode fonts. The way this often works is that one types in ASCII characters, but downloads a special font that provides glyphs for these the ASCII code points that give the desired visual rendering – for instance, one types “l72is4wo”, and it renders as the Ewe word lâdìs³wọ.
Table 19: CHRF score where the source uses different versions of common Unicode points. The top line of each bloc represents the “correct” codepoints, whereas the lower lines are other ways of representing the same letter. In many cases there is very little difference in performance, but the African languages are affected by nonstandard œ and è. When the apostrophe is removed entirely, the performance also drops noticeably.

Other non-Unicode fonts even assign different values to code points beyond the ASCII block: one well-known example is the Zawgyi encoding for Myanmar (Burmese) (Liao 2017). In the course of this work, we discovered that a wide variety of languages still use non-Unicode fonts. We ran into such fonts for Ewe (ee), Kashmiri (ks), Meiteilon (mni-Mtei), Mooré (mos), Navajo (nv), and Tamazight (ber-Latn), usually in the case of deliveries from professional translators. It is likely that there exists a large, hidden portion of data for some languages using these or other fonts. Our attempted reconstruction of some of these fonts can be seen in Appendix Section D.
6 IMPORTANCE OF NATIVE SPEAKERS

Since the advent of statistical and then neural machine translation, large data sets and improved modeling techniques have driven progress in translation quality. It has sometimes been difficult for the expertise of native speakers and linguists to continue helping to improve models in this environment. However, the participation of speakers of languages, and of members of affected communities, is nonetheless still vital. In the very-low-resource domain, there are more errors in models and data – and consequently, more opportunities for native speakers to help improve the quality.

We stress that where possible, it is important to try to build relationships with native speakers and members of these communities, rather than simply interacting with them as crowd-workers at a distance. For this work, the authors reached out to members of as many communities as we could, having conversations with over 100 members of these communities, many of whom were active in this project (See Acknowledgements).

Here is an incomplete list of ways in which speakers of these languages helped us extend and improve machine translation to their languages:

1. **Understanding Data:** As in Kreutzer et al. (2022), we conducted an extensive review of the quality of our datasets (see Appendix Table 22). In addition to simply giving us an idea of which languages had higher or lower quality data, this also gave us valuable insights about other uses and aspects of the corpora that were useful beyond this project – for instance, which corpora had more colloquial (or religious) text, and which dialects were mixed with other languages.

2. **Understanding errors in reference translations:** A variety of languages had extremely low automatic metrics (e.g. BLEU under 1.0), despite having large and clean corpora. For two such cases, native speakers helped us identify quality and fluency issues with our reference translations – and that the model outputs often looked better than the references.

3. **Specialized Filters:** Our corpora for a few languages were polluted with related high-resource languages that had somehow passed all the previous rounds of filtering. For two cases, natives speakers helped design custom filters to remove the unwanted content, and for a third, they helped remove sensitive content.

4. **Transliteration and political sensitivity around script:** For one language, we were initially unaware that we were using the wrong script. We were using a script that was associated with colonial times, and had since been replaced in the entire region, and was a matter of political sensitivity. Native speakers both pointed out this issue and helped us transliterate the text to the appropriate script. (See Appendix Section C)

5. **Understanding crowd-worker annotations:** when we sent translations to crowd workers to rate their quality, several languages showed unusual rating patterns, or patterns that did not line up with our expectations. We were fortunate to have native speakers who helped us understand the ratings differentiate cases where raters were mis-calibrated against real quality issues. Freitag et al. (2021) explores some of these phenomena more.

6. **Clarifying utility for Community:** Even if one can build a translation model for a language, should one? For some groups, community desires may differ from what many in the machine translation community might expect (Long, 2007; Coffey, 2021; Hiraishi, 2021). And if the translation model is of imperfect quality, is that still helpful for the community – or is it perhaps offensive? These are questions that can only be answered by members of the community. In our interviews, we generally found that native speakers of the languages we spoke to were very enthusiastic for even lower-quality translation offerings. This said, no one person can represent the entire community, and there is much to learn in how to handle situations where opinions and desires differ within a community.

7. **Commenting on Dialects:** Many “languages” have a wide variety of dialects, sometimes hardly mutually intelligible. Native speakers helped us understand when our models were producing a particular dialect, or mixing and matching them.

8. **The correct name to use for a language:** Whereas a language like French has a fairly unambiguous name, many languages have multiple names, some of which may be offensive or exclusive. Frequently there exists a colonial name (e.g. “Oriya” or “Lushai”), which may
be more widely known, but may also be disliked by the members of the community, as well as a native name (e.g. “Odia” and “Mizo”) that is lesser known but preferred. Similarly, there may be names which may feel exclusive to some sub populations – for instance using the name “Manipuri” for the language of the Meitei ethnic group. Although the most common way to refer to the Meitei language is indeed “Manipuri”, this is exclusive to the many other ethnic groups in the state of Manipur with their own languages.
7 Conclusions and Open Problems

7.1 Main Findings

Starting with an initial seed dataset of monolingual sentences spanning over 1500 languages, we demonstrate that it is possible to build relatively clean web-mined monolingual text datasets for over 1000 languages. We highlight the importance of incorporating expressive semi-supervised LangID models, document-level consistency signals, and several word-based and custom filtering techniques to identify and filter web text in long-tail languages (Section 2.1). Using this approach we are able to build a multilingual unlabeled text dataset containing over 1 million sentences for more than 200 languages and over 100 thousand sentences in more than 400 languages (Section 2.2).

Training on this dataset and a parallel corpus spanning 112 languages, we build massively multilingual models capable of translating across 1000 languages (Section 3). We highlight the importance of model capacity when training highly multilingual translation models and the positive effect on zero-resource quality when increasing the number of languages in the model. We also describe the significant quality improvements achievable by incorporating large-scale back-translation and self-training, and share our findings towards developing practical, inference-friendly models for long-tail languages.

We evaluate our models on evaluation sets collected for 38 languages, and highlight the need for choosing the right automatic metric (CHRF) when evaluating long-tail languages (Section 4.2). Apart from automatic metrics on evaluation sets, we additionally release approximate reference-free quality scores of our 1000-language MT model to provide an indicator of web-trained multilingual model quality on hundreds of previously under-studied languages (Section 4.3). We perform human evaluations on a subset of the languages in distilled models, and highlight that it is possible to build high quality, practical MT models for long-tail languages utilizing the approach described in this work (Section 4.4).

Through qualitative and quantitative analysis of the model outputs, we reveal a few characteristic error modes of our models; including confusing distributionally similar and infrequent tokens, and also producing verbose and inaccurate translations for short or single word queries arising from the extreme data sparsity of the zero-resource setting (Sections 4.5 and 4.7). We furthermore highlight several other observations from our studies, including non-English-centric direct translation, zero-shot transliteration, the effect of terminal punctuation on translation quality, and the robustness of the model to the non-standard glyph usages that are common for many languages (Section 5).

Finally, we highlight several indispensable contributions of native speakers who helped us evaluate, understand, filter and improve our datasets and models; and helped us understand the overall context of how these models should fit in with their communities (Section 6).

7.2 Related Work

There is a considerable wealth of literature on building highly multilingual text corpora, LangID models, and MT models. Our work differs largely in the scale, quality, and number of languages covered, together with the integration of many moving parts in the entire data-to-translation-model pipeline.

Access to multilingual datasets for NLP research has vastly improved over the past years. Since 2006, the Web as Corpus workshops have focused on the challenges around identifying relevant pages, extracting clean text, content de-duplication, and many other relevant topics (Barbaresi et al., 2020; Jakubícek et al., 2020). A variety of web-derived collections for hundreds of languages is available for anyone to download, such as the Corpora Collection at Leipzig University (Goldhahn et al., 2012), the Corpus of Global Language Use (Dunn, 2020), ParaCrawl (Esplà et al., 2019; Bañón et al., 2020), WikiMatrix (Schwenk et al., 2019), CCNET (Wenzek et al., 2020) and CCAligned (El-Kishky et al., 2020), OSCAR (Ortiz Suárez et al., 2019; Ortiz Suárez et al., 2020; Abadij et al., 2022), and several others; all of which have between 100 and 300 languages. The largest language coverage is probably An Crúbadán, which does not leverage LangID, and found (small amounts of) web data in about 2,000 languages (Scannell, 2007). These corpora have in turn enabled a variety of highly multilingual models, like mT5 (Xue et al., 2020), M2M-100 (Fan et al., 2020), and M4 (Arivazhagan et al., 2019; Siddhant et al., 2022).
Curating such datasets relies on the websites giving clues about the language of their contents (e.g., a language identifier in the URL) and on automatic language classification (LangID). It is commonly known that these automatically crawled and filtered datasets tend to have overall lower quality than hand-curated collections (Koehn et al., 2020), but their quality is rarely measured directly, and is rather judged through the improvements they bring to downstream applications (Schwenk et al., 2019). Therefore, many of these multilingual web corpora suffer from serious quality issues, especially for low-resource languages. A recent audit conducted by Kreutzer et al. (2022) on five public, multilingual datasets found pervasive issues. Many corpora claiming to be in one particular language in fact contained zero percent in-language content — and sometimes zero percent linguistic content entirely. Of the many issues contributing to this phenomenon, a fundamental one is the poor efficacy of LangID on low-resource languages.

Several works have investigated LangID at the level of multilinguality studied in this work. One relevant LangID implementation is Dunn (2020), achieving an F1 above 0.95 for 464 languages, and offering a thorough evaluation on different data sources and domains. The only LangID systems with higher coverage that we are aware of are those developed by Brown (2012; 2013; 2014), with the most recent version covering as many as 1,366 language varieties, with accuracy above 99%. Finally, Caswell et al. (2020) trains LangID models on 1,629 languages, and demonstrates that although they appear to have very high scores on held-out evaluation sets, in practice, when applied to web text, they produce datasets of almost unusable noisiness. Various error pathologies are detailed, and a few novel filtering techniques are proposed to counteract them, including TF-IIF filtering and semi-supervised LangID (SSLID). The present work can be viewed as an extension of Caswell et al. (2020), and a fusion of it with translation technology.

Our MT modeling approach builds on several previous works on massively multilingual, zero-resource and self-supervised MT; differing primarily in terms of the scale of multilinguality, model capacity and the extreme data sparsity in our experimental setting. We call these approaches that combine different aspects of scale as M4: massively multilingual, massive machine translation.

Multilingual Neural Machine Translation models were first introduced during the last decade (Firat et al., 2016a; Johnson et al., 2017), but the initial versions of these models were limited to a few languages (10–12). Over the last few years, there has been an explosion of work focusing on massively multilingual models that could translate between around 100 languages (Neubig & Hu, 2018; Aharoni et al., 2019b; Arivazhagan et al., 2019; Zhang et al., 2020; Tang et al., 2021; Fan et al., 2021). However, most work on massively multilingual MT has focused on the purely supervised setting. There are a few works that have ventured beyond the limitations of large-scale multilingual corpora, and trained MT models spanning over 1000 languages (Mueller et al., 2020), usually limited to narrow-domain (often religious) corpora.

Another stream of research on unsupervised MT developed modeling approaches to train MT models using monolingual datasets only (Lample et al., 2017; Artetxe et al., 2017; Song et al., 2019). With the advent of multilingual pre-training, with models like multilingual BERT (Devlin et al., 2019), XLM (Lample & Conneau, 2019), mBART (Liu et al., 2020) and others, the focus shifted towards fine-tuning pre-trained models with paired data in a sub-set of the pre-training languages to enable zero-resource translation in the remaining languages. These approaches are often complemented with large-scale back-translation (Sennrich et al., 2016; Edunov et al., 2018) to continue improving the model beyond its initial zero-shot performance.

Our work builds on Siddhant et al. (2020); Garcia et al. (2021b); Siddhant et al. (2022) that combines together multilingual supervised MT, zero-resource MT (Firat et al., 2016b) and self-supervised learning within a single model. We extend the work in Siddhant et al. (2022) by scaling to larger models, a more multilingual dataset, utilizing self-training and a novel filtering technique based on round-trip translation consistency and LangID predictions.

Apart from efforts focused on building highly multilingual web-mined corpora and MT models, another line of NLP research has focused on building datasets and NLP technologies for specific languages, not necessarily from web content. Many of these are grassroots, bottom-up efforts from the affected communities, organized through research collectives like Masakhane (V et al., 2020), Turkish Interlingua (Mirzakhalov et al., 2021a; 2021b), and GhanaNLP (Azunre et al., 2021a); and conferences

https://ai.googleblog.com/2019/10/exploring-massively-multilingual.html
and workshops like AfricaNLP, AmericasNLP (Mager et al., 2021) and ArabNLP. These efforts, in addition to providing datasets, frequently provide models and baselines, or even public interfaces, like the Khaya Translator Web App by GhanaNLP for West African languages, and the lesan.ai translation website for Ethiopian languages.

Participation is especially strong from the African continent, including corpora and models for pan East-African languages (Babirye et al., 2022), languages from the Horn of Africa (Hadgu et al., 2022), Ethiopian languages (Abate et al., 2018; Gezmu et al., 2021), Ugandan languages (Akera et al., 2022), Emakhuwa (Ali et al., 2021), South-African languages (Eiselen & Puttkammer, 2014), Setswana and Sepedi (Marivate et al., 2020), Yoruba (Adelani et al., 2021b; Oshiwanmo (Nekoto et al., 2022), Igbo (Ezeani et al., 2020), Zulu (Mabuya et al., 2021), Twi (Azunre et al., 2021b), Obe (Hacheme, 2021), Bambara (Tapo et al., 2021), and Fon (Emezue & Dossou, 2020). Outside of Africa, corpora have been created for languages of the Americas, including for four indigenous languages of Peru in Bustamante et al. (2020), the numerous results on the largely South- and Central American languages from the first AmericasNLP conference (Mager et al., 2021), and the Inuktitut language of Canada (Joanis et al., 2020). Datasets for lower-resourced languages of India have also sprung up, including the 13-language PMIndia (Haddow & Kirefu, 2020), and datasets focused on languages of the Northeast like Mizo (Thihlum et al., 2020), Khasi (Laskar et al., 2021) and Assamese (Laskar et al., 2020). Finally, a variety of such datasets and models are available for public use on HuggingFace or Zenodo. We believe language-specific efforts to be orthogonal and complementary to massively multilingual approaches for corpora building and modeling, as we elaborate further in Section 7.3.

7.3 Future Work

Barring a dramatic increase in the amount of web text available for long-tail languages, the types of errors produced by our zero-resource models (Section 4.5) are likely to persist. We highlight a few potential directions for future research that could help address the data sparsity that underlies the quality limitations of these models.

Utilizing dictionaries to ground distributionally similar words: One approach to address errors with distributionally similar words could involve helping ground the model’s translations by utilizing bilingual dictionaries or similar resources. Dictionaries are relatively widely available and have already yielded promising results for individual language pairs (Xia et al., 2019; Duan et al., 2020; Karamanolakis et al., 2020; Reid et al., 2021). For languages where dictionaries do not exist or coverage is low, dictionaries are much cheaper to build as compared to a dataset of parallel sentences. Efforts to develop high-coverage dictionaries and modeling approaches to incorporate them in massively multilingual MT models could nicely complement a corpus of monolingual web text and massively multilingual MT.

Complementing massively multilingual modeling with language-specific efforts: The quality of web-mined datasets is unlikely to match that of language-specific, hand-curated datasets; and building hand-curated datasets might be the only way forward to build text datasets for languages with limited presence on the web. However, we believe the two approaches to be orthogonal and complementary. Leveraging highly multilingual web-mined datasets and models significantly reduces the amount of data and research efforts needed to build practical NLP technologies for these languages (Wang et al., 2020; Emezue & Dossou, 2021; Adelani et al., 2022; Alabi et al., 2022; Nekoto et al., 2022), and research efforts could be more efficient by building resources and modeling approaches that complement the weaknesses of web-based massively multilingual models. Furthermore, community-based contributions could yield other useful language-specific tools, like specialized data filters as in Section 4.8, tools to normalize orthography or script (like those described in Appendix C) or pre- and post-processors to correct certain mistakes, improve diacritization, etc.
Leveraging multimodal datasets and models: A large proportion of the 7000+ languages spoken in the world have no written forms or standardized orthographic conventions. For a large majority there is limited amounts of text data available on the web. Being able to train models that can learn from and represent speech and text jointly (Zheng et al., 2021; Bapna et al., 2022; Chen et al., 2022; Bai et al., 2022; Tang et al., 2022) is essential to alleviate data sparsity and building more robust language technologies for such languages.

In future work we plan to investigate the above-mentioned and related threads of research, hope-fully making progress towards building and supporting language and speech technologies for more languages.

ACKNOWLEDGMENTS

We would like to extend our deepest gratitude to the following native speakers and members of the affected communities, who helped us in a wide variety of ways:

Yasser Salah Eddine Bouchareb (Algerian Arabic); Mfoniso Ukwak (Aanaang); Bhaskar Borthakur, Janu Nelakanti, Kishor Barman, Rasika Saikia, Suraj Bharech (Assamese); Annette David, William Merza (Assyrian Neo Aramaic), Ruben Hilarie Quispe (Aymara); Devina Suyanto, Puspa Si Ptinguin (Balinese); AllahseriX Auguste Tapo, Bakary Diarrassouba, Maimouna Siby, Moussa Doumbouya (Bambara); Mohammad Jahangir (Baluchi); Subhajit Naskar (Bengali); Animesh Pathak, Ankur Banpa, Anup Mohan, Chaitanya Joshi, Chandan Dubey, Kapil Kumar, Manish Katiyar, Mayank Srivastava, Neerakika, Saumya Pathak, Tanya Sinha, Vikas Singh (Bhojpuri); Bowen Liang, Ellie Chio, Eric Dong, Frank Tang, Jeff Pitman, John Wong, Kenneth Chang, Manish Goregaokar, Mingfei Lau, Ryan Li, Yiwen Luo (Cantonese); Monang Setyawan (Caribbean Javanese); Craig Cornelius (Cherokee); Anton Prokopiev (Chuvash); Nash Rafeeq (Dhivehi); Rajat Dogra, Sid Dogra (Dogri); Mohamed Kamagate (Dyula); Chris Assigbe, Dan Ameme, Emeafa Doe, Irene Nyavor, Thierry Gnanih, Yvonne Dumor (Ewe); Abdoulaye Burry, Adama Diallo, Fauzia van der Leeuw, Ibrahim Barry (Fulfulde); Isabel Papadimitriou (Greek); Alex Rudnick (Guarani); Mohammad Khdeir (Gulf Arabic); Paul Remollata (Hiligaynon); Ankur Banpa (Hindi); Mfoniso Ukwak (Ibibio); Nze Lawson (Igbo); D.J. Abuy, Miami Cabansay (Ilocano); Archana Koul, Shashwat Razdan, Sujeet Akula (Kashmiri); Jatin Kulkarni, Salil Rajadhyaksha, Sanjeet Hegde Desai, Sharayu Shenoy, Shashank Shanbhag, Shashi Shenoy (Konkani); Ryan Michael, Terrence Taylor (Krio); Bokan Jaff, Medya Ghazizadeht, Roshna Omer Abdulrahman, Saman Vaisipour, Sarchia Khursheed (Kurdish (Surani)); Suphian Tweel (Libyan Arabic); Espoir Murhabazi, Doudou Kisabaka, Marie-Celeste Bagniakanha (Lingala); Colleen Mallahan, John Quinn (Luganda); Cynthia Mboli (Luyia); Abhishek Kumar, Neeraj Mishra, Priyanjanan Jha, Saket Kumar, Snehal Bhilare (Maithili); Lisa Wang (Mandarin Chinese); Cibu Johny (Malayalam); Viresh Ratnakar (Marathi); Abhi Sanoujam, Gautam Thockchom, Pritam Pembi, Sam Chaomai, Shangkar Mayanglambam, Thangjam Hindustani Devi (Meiteilon Manipuri); Hala Ajil (Mesopotamian Arabic); Hamdanil Rasyid (Minangkabau); Elizabeth John, Remi Ratle, S Lallienkawli Gangte, Vaiphei Thatsing, Vanlalzami Vanlalzami (Mizo); Ahmed Kakchakh, Hanaa ElAzizi (Moroccan Arabic); George Ouais (MSA); Ujjwal Rajhandari (Newari); Ebufa Ufere, Gabriel Fynecontry, Onome Ofoman, Titi Akinsanmi (Nigerian Pigdin); Marwa Khost Jarkas (North Levantine Arabic); Abduselam Shaltu, Ace Patterson, Adel Kassem, Mo Ali, Yonas Hambissa (Oromo); Carlos Molina Vital, Elvia Andia Grageda, Helvia Taina, Marisol Necochea (Quechua); Rowena Marin (Romani); AbdelKarim Mardini (Saidi Arabic); Ishank Saxena, Manasa Harish, Manish Godara, Mayank Agrawal, Nitin Kashyap, Ranjani Padmanabhan, Ruchi Lohani, Shilpa Jindal, Shreevatsa Rajagopalan, Vaibhav Agarwal, Vinod Krishnan (Sanskrit); Nabil Shahid (Saraiki); Ayanda Mnyakeni (Sesotho, Sepedi); Landis Baker (Seychellois Creole); Taps Matangira (Shona); Ashraf Elsharif (Sudanese Arabic); Sakhile Dlamini (Swati); Hakim Sidaahmed (Tamazight); Melvin Johnson (Tamil); Sneha Kudugunta (Telugu); Alexander Tekle, Bserat Ghebre-micael, Nami Russom, Naud Ghebre (Tigrinya); Abigail Annkah, Diana Akron, Maame Ofori, Monica Opoku-Geren, Seth Duodu-baah, Yvonne Dumor (Twi); Ousmane Loum (Wolof), and Daniel Virtheim (Yiddish).

The authors also thank Raisonond Doctor and Cibu Johny for their invaluable help with the Meiteilon (Manipuri) transliteration project, and Colin Cherry, Markus Freitag and Johan Schalkwyk for their invaluable comments that helped improve the paper.
REFERENCES

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv e-prints*, art. arXiv:2201.06642, January 2022.

Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwosen Mulugeta, Yaregal Assabie, Hafe Abera, Binyam Ephrem, Tewodros Abebe, Wondimagegnhue Tsegaye, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. Parallel corpora for bi-lingual English-Ethiopian languages statistical machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3102–3111, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1262.

Amika Achom and Anupam Basu. Design and evaluation of Unicode compliance Meitei/Meetei Mayek keyboard layout. In *Proceedings of 2015 International Symposium on Advanced Computing and Communication (ISACC)*, pp. 90–97, Silchar, India, 2015. IEEE. doi: 10.1109/ISACC.2015.7377322.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Awokoya, and Cristina España-Bonet. MENYO-20k: A multi-domain English-Yoruba corpus for machine translation and domain adaptation. *CoRR, arXiv:2103.08647v1*, 03 2021a. doi: 10.48550/ARXIV.2103.08647.

David Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pp. 61–75, Virtual, August 2021b. Association for Machine Translation in the Americas. URL https://aclanthology.org/2021.mtsummit-research.6.

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajudeen Gwasabbe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Chinenye Emezue, Colin Leong, Michael Beukman, Shamsudddeen Hassan Muhammad, Guyo Dub Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ayoade Ajibade, Tunde Oluwaseyi Ajayi, Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millcent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Mendjokam Koagne, Edwin Munkoh-Bubeng, Valencia Wagner, Idriss Abdulmumin, Ayodele Awokoya, Happy Buzababa, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. A few thousand translations go a long way! leveraging pre-trained models for african news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2022. doi: 10.48550/ARXIV.2205.02022. URL https://arxiv.org/abs/2205.02022.

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3874–3884, Minneapolis, Minnesota, June 2019a. Association for Computational Linguistics. doi: 10.18653/v1/N19-1388. URL https://aclanthology.org/N19-1388.

Roee Aharoni, Melvin Johnson, and Orhan Firat. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3874–3884, 2019b.

Milam Aiken and Mina Park. The efficacy of round-trip translation for mt evaluation. *Translation Journal, 14*(1):1–10, 2010.

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggaya, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Enigere Bainomugisha, Ernest Mwebaze, and John Quinn. Machine translation for african languages: Community creation of datasets and models in uganda. In *3rd Workshop on African Natural Language Processing*, 2022. URL https://openreview.net/forum?id=BK-z5qzEU-9.
Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Multilingual language model adaptive fine-tuning: A study on African languages, 2022. URL: https://arxiv.org/abs/2204.06487

Felermino D. M. A. Ali, Andrew Caines, and Jaimito L. A. Malavi. Towards a parallel corpus of Portuguese and the Bantu language Emakhuwa of Mozambique. CoRR, abs/2104.05753, 2021. URL: https://arxiv.org/abs/2104.05753

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. Massively multilingual neural machine translation in the wild: Findings and challenges. arXiv preprint arXiv:1907.05019, 2019.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. arXiv preprint arXiv:1710.11041, 2017.

Gilberto Artioli, V Nociti, and Ivana Angelini. Gambling with Etruscan dice: A tale of numbers and letters. Archaeometry 53(3), pp. 1031–1043, 2011. doi: 10.1111/j.1475-4754.2011.00596.x.

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acuquah. NLP for Ghanaian languages. CoRR, abs/2103.15475, 2021a. URL: https://arxiv.org/abs/2103.15475

Paul Azunre, Salomey Osei, Salomey Addo, Lawrence Asamoah Adu-Gyamfi, Stephen Moore, Bernard Adabankah, Bernard Opoku, Clara Asare-Nyarko, Samuel Nyarko, Cynthia Amoaba, Esther Dansoa Appiah, Felix Akwerh, Richard Nii Lante Lawson, Joel Budu, Emmanuel Debrah, Nana Boateng, Wisdom Ofori, Edwin Buabeng-Munkoh, Franklin Adjei, Isaac Kojo Essel Ampomah, Joseph Otoo, Reindorf Borkor, Standylove Birago Mensah, Lucien Mensah, Mark Amoako Marcel, Anokye Acheampong Amponsah, and James Ben Hayfron-Acuquah. English-Twi parallel corpus for machine translation. CoRR, abs/2103.15625, 2021b. URL: https://arxiv.org/abs/2103.15625

Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, Ronald Ogwang, Jeremy Tusubira Francis, Jonathan Mukiibi, Medadi Ssentanda, Lilian D Wanzare, and Davis David. Building text and speech datasets for low resourced languages: A case of languages in East Africa. In 3rd Workshop on African Natural Language Processing, 2022. URL: https://openreview.net/forum?id=S0-U99z4U-q

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, 2015.

He Bai, Renjie Zheng, Junkun Chen, Xintong Li, Mingbo Ma, and Liang Huang. A3t: Alignment-aware acoustic and text pretraining for speech synthesis and editing. ArXiv, abs/2203.09690, 2022.

Anton Bakalov, Alex Salcianu, Andy Goldberg, Chris Alberti, Daniel Andor, David Weiss, Emily Pitler, Greg Coppola, Jason Riesa, Kuzman Ganchev, Michael Ringgaard, Nan Hua, Ryan McDonald, Slav Petrov, Stefan Istrate, and Terry Koo. Compact Language Detector v3 (CLD3), October 2016. URL: https://github.com/google/cld3

Marta Bañoñ, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrias, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4555–4567, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.417. URL: https://aclanthology.org/2020.acl-main.417
Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. mslam: Massively multilingual joint pre-training for speech and text. CoRR, abs/2202.01374, 2022. URL https://arxiv.org/abs/2202.01374.

Adrien Barbaresi, Felix Bildhauer, Roland Schäfer, and Egon Stemle (eds.). Proceedings of the 12th Web as Corpus Workshop, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-68-9. URL https://aclanthology.org/2020.wac-1.0.

Brent Berlin and Paul Kay. Basic color terms: their universality and evolution. Univ California Press, Berkeley, CA, 1969.

Steven Bird. Decolonising speech and language technology. In Proceedings of the 28th International Conference on Computational Linguistics, pp. 3504–3519, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.313. URL https://aclanthology.org/2020.coling-main.313.

Sascha Brawer. Corpus Crawler. https://github.com/google/corpuscrawler, September 2017. URL https://github.com/google/corpuscrawler.

Ralf Brown. Non-linear mapping for improved identification of 1300+ languages. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 627–632, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1069. URL https://aclanthology.org/D14-1069.

Ralf D Brown. Finding and identifying text in 900+ languages. Digital Investigation, 9:S34–S43, 2012.

Ralf D Brown. Selecting and weighting n-grams to identify 1100 languages. In International Conference on Text, Speech and Dialogue, pp. 475–483. Springer, 2013.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquey. No data to crawl? monolingual corpus creation from PDF files of truly low-resource languages in Peru. In Proceedings of the 12th Language Resources and Evaluation Conference, pp. 2914–2923, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.356.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. arXiv preprint arXiv:2010.14571, 2020. URL https://arxiv.org/abs/2010.14571.

Shobhana Lakshmi Chelliah. A Grammar of Meithei, volume 17 of Mouton Grammar Library [MGL]. Mouton de Gruyter, Berlin, Germany, 1997. doi: 10.1515/9783110801118.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Niki Parmar, Mike Schuster, Zhifeng Chen, et al. The best of both worlds: Combining recent advances in neural machine translation. arXiv preprint arXiv:1804.09849, 2018.

Zhehuai Chen, Yu Zhang, Andrew Rosenberg, Bhuvana Ramabhadran, Pedro J. Moreno, Ankur Bapna, and Heiga Zen. Maestro: Matched speech text representations through modality matching. ArXiv, abs/2204.03409, 2022.

Donavyn Coffey. Māori are trying to save their language from Big Tech. https://www.wired.co.uk/article/maori-language-tech, 2021. Accessed: 2022-04-24.

Mark Davis and Michel Suignard. Unicode security mechanisms. Technical Report (Technical Standard #39), Unicode Consortium, August 2021. URL http://www.unicode.org/reports/tr39/#confusables, Version 14.0, Revision 24.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT, 2019.
Xiangyu Duan, Baijun Ji, Hao Jia, Min Tan, Min Zhang, Boxing Chen, Weihua Luo, and Yue Zhang. Bilingual dictionary based neural machine translation without using parallel sentences. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1570–1579, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.143. URL https://aclanthology.org/2020.acl-main.143

Jonathan Dunn. Mapping languages: the Corpus of Global Language Use. Language Resources and Evaluation, April 2020. doi: 10.1007/s10579-020-09489-2. URL https://doi.org/10.1007/s10579-020-09489-2

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 489–500, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1045. URL https://aclanthology.org/D18-1045

Roald Eiselen and Martin Puttkammer. Developing text resources for ten South African languages. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), pp. 3698–3703, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2014/pdf/1151_Paper.pdf

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. CCAligned: A massive collection of cross-lingual web-document pairs. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), pp. 5960–5969, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.480. URL https://www.aclweb.org/anthology/2020.emnlp-main.480

Chris Chinenye Emezue and Bonaventure F. P. Dossou. MMTAfrica: Multilingual machine translation for African languages. In Proceedings of the Sixth Conference on Machine Translation, pp. 398–411, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.48

Chris Chinenye Emezue and Femi Pancrace Bonaventure Dossou. FFR v1.1: Fon-French neural machine translation. In Proceedings of the The Fourth Widening Natural Language Processing Workshop, pp. 83–87, Seattle, USA, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.winlp-1.21. URL https://aclanthology.org/2020.winlp-1.21

Miquel Esplà, Mikel Forcada, Gema Ramírez-Sánchez, and Hieu Hoang. ParaCrawl: Web-scale parallel corpora for the languages of the EU. In Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks, pp. 118–119, Dublin, Ireland, August 2019. European Association for Machine Translation. URL https://aclanthology.org/W19-6721

Miquel Esplà-Gomis. Bitextor: a free/open-source software to harvest translation memories from multilingual websites. In Beyond Translation Memories: New Tools for Translators Workshop, Ottawa, Canada, August 26-30 2009. URL https://aclanthology.org/2009.mtsummit-btm.6

Michael Everson. Proposal for encoding the Meitei Mayek script in the BMP of the UCS. ISO/IEC JTC1/SC2/WG2 N3206R2, Unicode Consortium, August 2007. URL http://unicode.org/wg2/docs/n3206.pdf

Ignatius Ezeani, Paul Rayson, Ikechukwu E. Onyenwe, Chinedu Uchechukwu, and Mark Hepple. Igbo-english machine translation: An evaluation benchmark. CoRR, abs/2004.00648, 2020. URL https://arxiv.org/abs/2004.00648

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. arXiv preprint arXiv:2010.11125, 2020. URL https://arxiv.org/abs/2010.11125
Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12), pp. 759–765, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/327_Paper.pdf

Kyle Gorman. Pynini: A python library for weighted finite-state grammar compilation. In Proceedings of the SIGFSM Workshop on Statistical NLP and Weighted Automata, pp. 75–80, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2409. URL https://aclanthology.org/W16-2409.

Kyle Gorman and Richard Sproat. Finite-State Text Processing, volume 14 of Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2021.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation. CoRR, abs/2106.03193, 2021. URL https://arxiv.org/abs/2106.03193.

Gilles Hacheme. English2gbe: A multilingual machine translation model for {Fon/Ewe} gbe. arXiv preprint arXiv:2112.11482, 2021.

Barry Haddow and Faheem Kirefu. Pmindia--a collection of parallel corpora of languages of India. arXiv preprint arXiv:2001.09907, 2020.

Asmelash Teka Hadgu, Gebrekirstos G, Gebremeskel, and Abel Aregawi. HornMT. https://github.com/asmeleashadgu/HornMT 2022.

Junxian He, Jiatao Gu, Jiajun Shen, and Marc’Aurelio Ranzato. Revisiting self-training for neural sequence generation. arXiv preprint arXiv:1909.13788, 2019.

Ku’uwéhi Hiraishi. Teaching computers ’ölelo Hawai’i prompts debate on data sovereignty. https://www.hawaiipublicradio.org/local-news/2021-09-15/teaching-computers-olelo-hawai-i-prompts-debate-on-data-sovereignty2021. Accessed: 2022-04-24.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

Xiuming Huang. A machine translation system for the target language inexpert. In COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics, 1990.

Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. Advances in neural information processing systems, 32, 2019.

ISO. ISO 15919: Transliteration of Devanagari and related Indic scripts into Latin characters. https://www.iso.org/standard/28333.html 2001. International Organization for Standardization, Geneva, Switzerland.

ISO. ISO 15924: Codes for the representation of names of scripts. https://www.iso.org/obp/ui/#iso:std:iso:15924:ed-1:v1:en 2004. International Organization for Standardization, Geneva, Switzerland.

Miloš Jakubíček, Vojtěch Kovář, Pavel Rychlík, and Vit Suchomel. Current challenges in web corpus building. In Proceedings of the 12th Web as Corpus Workshop, pp. 1–4, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-68-9. URL https://aclanthology.org/2020.wac-1.1
Eric Joanis, Rebecca Knowles, Roland Kuhn, Samuel Larkin, Patrick Littell, Chi-kiu Lo, Darlene Stewart, and Jeffrey Micher. The Nunavut Hansard Inuktitut–English parallel corpus 3.0 with preliminary machine translation results. In Proceedings of the 12th Language Resources and Evaluation Conference, pp. 2562–2572, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL [https://aclanthology.org/2020.lrec-1.312](https://aclanthology.org/2020.lrec-1.312).

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda B. Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google’s multilingual neural machine translation system: Enabling zero-shot translation. CoRR, abs/1611.04558, 2016. URL [http://arxiv.org/abs/1611.04558](http://arxiv.org/abs/1611.04558).

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google’s multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5:339–351, 2017. doi: 10.1162/tacl_a_00065. URL [https://aclanthology.org/Q17-1024](https://aclanthology.org/Q17-1024).

Cibu Johny, Lawrence Wolf-Sonkin, Alexander Gutkin, and Brian Roark. Finite-state script normalization and processing utilities: The Nisaba Brahmic library. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pp. 14–23, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.3. URL [https://aclanthology.org/2021.eacl-demos.3](https://aclanthology.org/2021.eacl-demos.3).

Giannis Karamanolakis, Daniel Hsu, and Luis Gravano. Cross-lingual text classification with minimal resources by transferring a sparse teacher. In Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3604–3622, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.323. URL [https://aclanthology.org/2020.findings-emnlp.323](https://aclanthology.org/2020.findings-emnlp.323).

K. Kabi Khanganba and Girish Nath Jha. Challenges in Indian language transliteration: a case of Devanagari, Bangla and Manipuri. In Proceedings of the 2nd Workshop on Indian Language Data: Resources and Evaluation (WILDRE2), pp. 77–83, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL [http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-WILDRE%20Proceedings.pdf](http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-WILDRE%20Proceedings.pdf).

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. arXiv preprint arXiv:1606.07947, 2016.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. Findings of the WMT 2020 shared task on parallel corpus filtering and alignment. In Proceedings of the Fifth Conference on Machine Translation, pp. 726–742, Online, November 2020. Association for Computational Linguistics. URL [https://aclanthology.org/2020.wmt-1.78](https://aclanthology.org/2020.wmt-1.78).

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawana, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddin Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhakov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhi Dlamini, Nisansa de Silva, Sakine Çabuk Balli, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Ogheneogheh Ahia, Ogheghene Agbawa, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a glance: An audit of web-crawled multilingual datasets. Transactions of the Association for Computational Linguistics, 10:50–72, 2022. doi: 10.1162/tacl_a_00447. URL [https://doi.org/10.1162%2Ftacl_a_00447](https://doi.org/10.1162%2Ftacl_a_00447).

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Proceedings of the 2018 Conference on
Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*, 2019.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji, Partha Pakray, and Sivaji Bandyopadhyay. EnAsCorp1.0: English-Assamese corpus. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pp. 62–68, Suzhou, China, December 2020. Association for Computational Linguistics. URL [https://aclanthology.org/2020.loresmt-1.9](https://aclanthology.org/2020.loresmt-1.9).

Sahinur Rahman Laskar, Abdullah Faiz Ur Rahman Khilji Darsh Kaushik, Partha Pakray, and Sivaji Bandyopadhyay. EnKhCorp1.0: An English–Khasi corpus. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pp. 89–95, Virtual, August 2021. Association for Machine Translation in the Americas. URL [https://aclanthology.org/2021.mtsummit-loresmt.9](https://aclanthology.org/2021.mtsummit-loresmt.9).

Han-Teng Liao. Encoding for access: how Zawgyi success impedes full participation in digital Myanmar. *ACM SIGCAS Computers and Society*, 46(4):18–24, 2017.

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

Chi-kiu Lo. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pp. 507–513, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-5358. URL [https://aclanthology.org/W19-5358](https://aclanthology.org/W19-5358).

Gideon Long. Chilean Mapuches in language row with Microsoft. [https://www.reuters.com/article/us-chile-mapuches-microsoft-idUSN2238412220061123](https://www.reuters.com/article/us-chile-mapuches-microsoft-idUSN2238412220061123). January 2007. Reuters Technology News, Accessed: 2022-04-24.

Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1166. URL [https://aclanthology.org/D15-1166](https://aclanthology.org/D15-1166).

Rooweither Mabuya, Jade Abbott, and Vukosi Marivate. Umsuka english - isizulu parallel corpus, June 2021. URL [https://doi.org/10.5281/zenodo.5035171](https://doi.org/10.5281/zenodo.5035171). Thank you to Facebook Research for funding the creation of this dataset.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishray Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pp. 202–217, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.americasnlp-1.23. URL [https://aclanthology.org/2021.americasnlp-1.23](https://aclanthology.org/2021.americasnlp-1.23).

Kelly Marchisio, Markus Freitag, and David Grangier. On systematic style differences between unsupervised and supervised MT and an application for high-resource machine translation. *CoRR*, abs/2106.15818, 2021. URL [https://arxiv.org/abs/2106.15818](https://arxiv.org/abs/2106.15818).
Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. In Proceedings of the first workshop on Resources for African Indigenous Languages, pp. 15–20, Marseille, France, May 2020. European Language Resources Association (ELRA). ISBN 979-10-95546-60-3. URL https://aclanthology.org/2020.rail-1.3

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr, et al. A large-scale study of machine translation in the Turkic languages. arXiv preprint arXiv:2109.04593, 2021a.

Jamshidbek Mirzakhalov, Anoop Babu, Aigiz Kunafin, Ahsan Wahab, Behzod Moydinboyev, Sardana Ivanova, Mokhiyakhon Uzokova, Shaxnoza Pulatova, Duygu Ataman, Julia Kreutzer, et al. Evaluating multiway multilingual NMT in the Turkic languages. arXiv preprint arXiv:2109.06262, 2021b.

Mehryar Mohri. Weighted automata algorithms. In Manfred Droste, Werner Kuich, and Heiko Vogler (eds.), Handbook of Weighted Automata, Monographs in Theoretical Computer Science, pp. 213–254. Springer, 2009. doi: https://doi.org/10.1007/978-3-642-01492-5_6.

Gourashyam Moirangthem and Kishorjit Nongmeikapam. A back-transliteration based Manipuri Meetei Mayek keyboard IME. In Proceedings of 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), pp. 1–6, Kuala Lumpur, Malaysia, 2021. IEEE. doi: 10.1109/GUCON50781.2021.9573837.

Jihyung Moon, Hyunchang Cho, and Eunjeong L Park. Revisiting round-trip translation for quality estimation. arXiv preprint arXiv:2004.13937, 2020.

Aaron Mueller, Garrett Nicolai, Arya D. McCarthy, Dylan Lewis, Winston Wu, and David Yarowsky. An analysis of massively multilingual neural machine translation for low-resource languages. In Proceedings of the 12th Language Resources and Evaluation Conference, pp. 3710–3718, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.458

Wilhelmina Nekoto, Julia Kreutzer, Jenalea Rajab, Millicent Ochieng, and Jade Abbott. Participatory translations of oshiwambo: Towards sustainable culture preservation with language technology. In 3rd Workshop on African Natural Language Processing, 2022. URL https://openreview.net/forum?id=BFbg59zVUZc

Graham Neubig and Junjie Hu. Rapid adaptation of neural machine translation to new languages. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 875–880, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1103. URL https://aclanthology.org/D18-1103

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In Piotr Bański, Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder, Simon Clementide, Marc Kupietz, Harald Lüngen, and Caroline Iliadi (eds.), Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pp. 9 – 16, Mannheim, 2019. Leibniz-Institut für Deutsche Sprache. doi: 10.14618/ids-pub-90215. URL http://nbn-resolving.de/urn:nbn:de:bsz:mb39-90215.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. A monolingual approach to contextualized word embeddings for mid-resource languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 1703–1714, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.156. URL https://aclanthology.org/2020.acl-main.156

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318, Philadelphia, Pennsylvania, USA,
F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Maja Popović. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395. Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL [https://aclanthology.org/W15-3049](https://aclanthology.org/W15-3049).

Manasa Prasad, Theresa Breiner, and Daan van Esch. Mining training data for language modeling across the world’s languages. In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2018)*, 2018. URL [https://www.isca-speech.org/archive_v0/SLTU_2018/pdfs/Manasa.pdf](https://www.isca-speech.org/archive_v0/SLTU_2018/pdfs/Manasa.pdf).

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2685–2702, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.213. URL [https://aclanthology.org/2020.emnlp-main.213](https://aclanthology.org/2020.emnlp-main.213).

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 African languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1306–1320, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.99. URL [https://aclanthology.org/2021.emnlp-main.99](https://aclanthology.org/2021.emnlp-main.99).

Philip Resnik and Noah A. Smith. The web as a parallel corpus. *Computational Linguistics*, 29(3): 349–380, 2003. doi: 10.1162/089120103322711578. URL [https://aclanthology.org/J03-3002](https://aclanthology.org/J03-3002).

Barbara Saunders and Jaap van Brakel. The trajectory of color. *Perspectives on Science*, pp. 302–355, 2002.

K. P. Scannell. The Crúbadán Project: Corpus building for under-resourced languages. In *3rd Web as Corpus Workshop, 2007, Louvain-la-Neuve, Belgium*, 2007.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. WiKImatrix: Mining 135m parallel sentences in 1620 language pairs from Wikipedia. *arXiv preprint arXiv:1907.05791*, 2019. URL [https://arxiv.org/abs/1907.05791](https://arxiv.org/abs/1907.05791).

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 6490–6500, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.507. URL [https://aclanthology.org/2021.acl-long.507](https://aclanthology.org/2021.acl-long.507).

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892. Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL [https://aclanthology.org/2020.acl-main.704](https://aclanthology.org/2020.acl-main.704).

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1009. URL [https://aclanthology.org/P16-1009](https://aclanthology.org/P16-1009).
Jonathan Shen, Patrick Nguyen, Yonghui Wu, Zhifeng Chen, Mia X Chen, Ye Jia, Anjuli Kannan, Tara Sainath, Yuan Cao, Chung-Cheng Chiu, et al. Lingvo: a modular and scalable framework for sequence-to-sequence modeling. *arXiv preprint arXiv:1902.08295*, 2019.

Aditya Siddhant, Ankur Bapna, Yuan Cao, Orhan Firat, Mia Chen, Sneha Kudugunta, Naveen Arivazhagan, and Yonghui Wu. Leveraging monolingual data with self-supervision for multilingual neural machine translation. *arXiv preprint arXiv:2005.04816*, 2020.

Aditya Siddhant, Ankur Bapna, Orhan Firat, Yuan Cao, Mia Xu Chen, Isaac Caswell, and Xavier Garcia. Towards the next 1000 languages in multilingual machine translation: Exploring the synergy between supervised and self-supervised learning. *arXiv preprint arXiv:2201.03110*, 2022.

Harimohon Thounaojam Singh. The evolution and recent development of the Meitei Mayek script. In *North East Indian Linguistics*, volume 3, pp. 24–32. Cambridge University Press India, New Delhi, India, 2011.

Leihaorambam Sarbajit Singh, Kabita Thaoroijam, and Pradip Kumar Das. Written Manipuri (Meiteron) – phoneme to grapheme correspondence. *Language in India*, 7(6), June 2007. URL [http://www.languageinindia.com/june2007/meitelphoneme.pdf](http://www.languageinindia.com/june2007/meitelphoneme.pdf).

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: masked sequence to sequence pre-training for language generation. *CoRR*, abs/1905.02450, 2019. URL [http://arxiv.org/abs/1905.02450](http://arxiv.org/abs/1905.02450).

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

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, and Juan Miguel Pino. Unified speech-text pre-training for speech translation and recognition. *ArXiv*, abs/2204.05409, 2022.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrv Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3450–3466, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.304. URL [https://aclanthology.org/2021.findings-acl.304](https://aclanthology.org/2021.findings-acl.304).

Allahsera Auguste Tapo, Michael Leventhal, Sarah Lugger, Christopher M. Homan, and Marcos Zampieri. Domain-specific MT for low-resource languages: The case of bambara-french. *CoRR*, abs/2104.00041, 2021. URL [https://arxiv.org/abs/2104.00041](https://arxiv.org/abs/2104.00041).

Zaitinkhuma Thihlum, Vanlalmuansangi Khenglawt, and Somen Deb Nath. Machine translation of english language to mizo language. In *2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*, pp. 92–97, 2020. doi: 10.1109/CCEM50674.2020.00028.

Jakob Uszkoreit, Jay Ponte, Ashok Popat, and Moshe Dubiner. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 1101–1109, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL [https://aclanthology.org/C10-1124](https://aclanthology.org/C10-1124).

Daan van Esch, Elnaz Sarbar, Tamar Lucassen, Jeremy O’Brien, Theresa Breiner, Manasa Prasad, Evan Elizabeth Crew, Chieu Nguyen, and Francoise Beaufays. Writing across the world’s languages: Deep internationalization for gboard, the google keyboard. Technical report, Google, 2019. URL [https://arxiv.org/abs/1912.01218](https://arxiv.org/abs/1912.01218).

Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara E. Rivera. Writing system and speaker metadata for 2,800+ language varieties. In *Proceedings of LREC*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NIPS)*, 30, 2017.
Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. Extending multilingual BERT to low-resource languages. In Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2649–2656, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.240. URL https://aclanthology.org/2020.findings-emnlp.240.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In Proceedings of the 12th Language Resources and Evaluation Conference, pp. 4003–4012, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.494.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. Generalized data augmentation for low-resource translation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5786–5796, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1579. URL https://aclanthology.org/P19-1579.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934, 2020.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. arXiv preprint arXiv:2004.11867, 2020.

Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation. In ICML, 2021.

George Zipf. The Psychology of Language. Houghton-Mifflin, 1935.
A Complete Human Evaluation and CHRF Results for Distilled Models

Table 20 gives the performance of the distilled models, as measured both by humans on a seven point scale (nonsense to perfect), and in CHRF. Please keep in mind that human evaluation numbers are quite hard to compare across languages.

B Complete Audit Results

We rated samples of 100 sentences for 72 of the languages present in the dataset we crawled. The error metrics used are described in Table 21. The breakdown can be seen in Table 22.

| Language Name (BCP-47) | score | cc | cb | ca | wl | nl | wd |
|------------------------|-------|----|----|----|----|----|----|
| Northeastern Dinka (dip) | 100  | 100 | 0  | 0  | 0  | 0  | 0  |
| Zarma (dje)             | 100  | 100 | 0  | 0  | 0  | 0  | 0  |
| Dombe (dov)             | 100  | 100 | 0  | 0  | 0  | 0  | 0  |
| Dyula (dyu)             | 100  | 100 | 0  | 0  | 0  | 0  | 0  |
| Wayuu (guc)             | 100  | 100 | 0  | 0  | 0  | 0  | 0  |
| Kalenjin (khn)          | 100  | 100 | 0  | 0  | 0  | 0  | 0  |
| Wolaytta (wal)          | 100  | 100 | 0  | 0  | 0  | 0  | 0  |
| Assyrian Neo-Aramaic (aii) | 100 | 100 | 0  | 0  | 0  | 0  | 0  |
| Igbo (ig)               | 99   | 98  | 2  | 0  | 0  | 0  | 0  |
| Balinese (ban)          | 98   | 96  | 4  | 0  | 0  | 0  | 0  |
| Latinized Hindi (hi-Latn) | 98 | 98  | 0  | 0  | 2  | 0  | 0  |
| Twi (ak)                | 97   | 94  | 6  | 0  | 0  | 0  | 0  |
| Luyia (fuy)             | 97   | 97  | 0  | 0  | 0  | 0  | 3  |
| Ewe (ee)                | 96   | 91  | 9  | 0  | 0  | 0  | 0  |
| Seselwa Creole French (crs) | 94 | 92  | 4  | 1  | 3  | 0  | 0  |
| Bhojpuri (bho)          | 94   | 94  | 0  | 0  | 6  | 0  | 0  |
| Ilocano (ilo)           | 94   | 88  | 12 | 0  | 0  | 0  | 0  |
| Caribbean Javanese (jvn) | 94  | 87  | 13 | 0  | 0  | 0  | 0  |
| Meiteilon (Manipuri) (mni) | 93 | 87  | 12 | 0  | 1  | 0  | 0  |
| Luba-Katanga (fu)       | 92   | 83  | 17 | 0  | 0  | 0  | 0  |
| Krio (kri)              | 92   | 89  | 5  | 0  | 6  | 0  | 0  |
| Latinized Tamil (ta-Latn) | 91 | 81  | 19 | 0  | 0  | 0  | 0  |
| Lingala (ln)            | 90   | 80  | 19 | 0  | 1  | 0  | 0  |
| Maharasthra Konkani (knn) | 89 | 85  | 8  | 1  | 6  | 0  | 0  |
| Northern Sami (se)      | 89   | 86  | 5  | 0  | 9  | 0  | 0  |
| Cherokee (chr)          | 88   | 78  | 20 | 0  | 1  | 1  | 0  |
| Latinized Malayalam (ml-Latn) | 87 | 76  | 21 | 1  | 2  | 0  | 0  |
| Latinized Bengali (bn-Latn) | 86 | 72  | 28 | 0  | 0  | 0  | 0  |
| Maithili (mai)          | 86   | 83  | 5  | 1  | 11 | 0  | 0  |
| Hiligaynon (hil)        | 84   | 80  | 8  | 0  | 9  | 3  | 0  |
| Tok Pisin (tpi)         | 84   | 83  | 2  | 0  | 15 | 0  | 0  |
| Latinized Chinese (zh-Latn) | 84 | 80  | 7  | 0  | 13 | 0  | 0  |
| Gulf Arabic (afb)       | 83   | 74  | 17 | 1  | 8  | 0  | 0  |
| Minangkabau (min)       | 81   | 64  | 34 | 0  | 2  | 0  | 0  |
| Chuvash (cv)            | 80   | 78  | 4  | 0  | 2  | 2  | 14 |
| Tamazight (ber-Latn)    | 79   | 69  | 20 | 0  | 10 | 1  | 0  |
| Latinized Telugu (te-Latn) | 79 | 74  | 10 | 0  | 16 | 0  | 0  |
| Libyan Arabic (ayl)     | 78   | 68  | 14 | 10 | 7  | 1  | 0  |
| Newari (new)            | 78   | 76  | 3  | 0  | 21 | 0  | 0  |
| Pangasinan (pag)        | 75   | 67  | 17 | 0  | 0  | 17 | 0  |
| Waray (war)             | 75   | 67  | 17 | 0  | 17 | 0  | 0  |
| Goan Konkani (gom)      | 73   | 67  | 5  | 13 | 13 | 2  | 0  |
| Sanskrit (sa)           | 73   | 66  | 14 | 0  | 15 | 5  | 0  |
| North Levantine Arabic (apc) | 73 | 62  | 21 | 0  | 16 | 1  | 0  |
| Sudanese Arabic (apd-SD)| 72   | 60  | 4  | 34 | 2  | 0  | 0  |
Table 22: Results of an audit of the datasets we collected, conducted on sampled of 100 sentences each by a mix of native speakers and non-native speakers. The values are the percent of the audited sample that head each label. The “score” metric combines these numbers for an approximate notion of the percent of the data that is usable, and is described in Section 2.2.2.

B.1 Strict RTT LANGID CHRIF

In the main section of this paper (4.3), we reported the loose version of RTTLANGIDCHRIF, which seems to correlate better with CHRIF and data size. One issue with this score is that it doesn’t penalize the model for producing content in the wrong language, whereas the strict version does.

Figure 3 shows strict RTTLANGIDCHRIF as a function of log data size. Although the correlation is worse than with the loose version, there is a clear trend, and there appears to be some sort of upper bound in quality as a function of data size.

C Transliteration for Meiteilon

The large majority of the text we found online for Meiteilon (Manipuri) was in the Bengali script. Finding almost no Meiteilon in its native script, Meetei Mayek (mni-Mtei), we initially erroneously assumed that this script was archaic or only used in rare contexts. However, conversations with Meiteilon speakers quickly disillusioned us of this notion – not only is it used, but it is now on its way to being the primary script used in the state. Most likely, the reason our mni-Mtei corpus was so small was because the available text online was largely in non-Unicode fonts, and therefore inaccessible to our LangID model (which had not been trained to detect such non-Unicode data).

Therefore, we needed to convert our Bengali-script corpus to Meetei Mayek. However, this is not a straightforward task for such a non-injective script as Bengali.

Meiteilon is a tonal Tibeto-Burman language that is one of the scheduled languages of India and a lingua franca of the Manipur state (Chelliah, 1997; Singh, 2011). The Meetei Mayek script is an indigenous script that was used to record Meiteilon until the 18th century when it was mostly superseded by the Bengali script. Despite recent Meetei Mayek gradual revitalization efforts by the Indian government, Meetei Mayek literacy is still quite low (Singh et al., 2007) and the available online
Figure 3: Plot of $R_{TLGIDCHRF}$ scores (strict) for languages as a function of log monolingual data size. This score has worse correlation with metrics like CHRF than the loose version of $R_{TLGIDCHRF}$, but shows an interesting trend when plotted versus data size like this. Compared to the loose version, languages much below the trend line on the right-hand side are often close to high-resource languages (E.g. Betawi/bew, Sabah Malay/msi, Godwari/gdx, Darija/ar-MA), indicating that their apparently large monolingual datasets are actually a result of over-triggering with a higher-resource language like Indonesian, and there are many wrong-language intermediate translations.

Figure 4: Bengali (Beng) to Meetei Mayek (Mtei) transliteration components. The script codes are denoted according to ISO 15924 [ISO 2004].

Meiteilon data is mostly in Bengali script [Achom & Basu 2015, Moirangthem & Nongmeikapam, 2021].

Meitei Mayek belongs to the Tibetan family of Brahmic scripts and is well suited for Meiteilon phonology representing a near bijective mapping between graphemes and phonemes of a language [Singh et al. 2007]. Unlike the major Brahmic scripts, this script uses a special class of explicit silent final consonants ($lonsum iyek$) in closed syllable codas, but these consonants are represented as full letters rather than combining signs. In modern Meetei Mayek orthography, the falling tone is often unmarked or sometimes marked with full stop punctuation, whereas in the traditional literature a special $lum iyek$ sign was used [Everson 2007].

Unlike Meetei Mayek, the orthographic conventions for Meiteilon in Bengali script are ambiguous due to its larger letter inventory, where more than one Bengali letter or clusters of letters may map to a single Meiteilon sound [Singh et al. 2007, Khanganba & Jha 2014]. This implies that any Bengali to Meitei Mayek transliteration mechanism needs to implement a many-to-one relation.
Our transliterator uses the open-source Nisaba library of finite-state script normalization and processing utilities (Johny et al., 2021). The script operations in Nisaba are efficiently and succinctly represented as weighted finite-state transducers (WFSTs) using Pynini finite-state grammars (Gorman, 2016; Gorman & Sproat, 2021). The main components of transliteration workflow are shown in Figure 4. The four component WFSTs are compiled into the final transliteration WFST \( T = N \circ R' \circ P \circ R^{-1} \), where \( \circ \) denotes FST composition operation (Mohri, 2009). The first component transducer \( N \) implements visual normalization of the Bengali script input that consists of visually invariant normalization transformations including NFC (Johny et al., 2021). This is followed by the Meiteilon-specific Bengali to Latin script many-to-one mapping \( R' \) that produces Latin script output in ISO 15919 format (ISO, 2001) augmented with some placeholder markers required for the next processing stage.

The third stage implements post-processing transformations \( P \) that are required to resolve ambiguities represented by the placeholder markers (introduced by \( R' \)) based on the orthographic context. One example of such transformation is the resolution of the Bengali virama sign, whose original purpose in the Bengali script is to mark silent consonants pronounced without an inherent vowel in consonant clusters. Its Meetei Mayek counterpart, the apun iyek mark, functions differently — it only applies to a non-silent subset of consonants (i.e., all consonants excluding the set of lonsum iyek mentioned above). Hence, given the virama placeholder in the input, two finite-state context-dependent rewrites are required for the resolution: convert the preceding consonant to lonsum iyek representation and remove the placeholder if the preceding consonant sound is covered by the lonsum iyek set, otherwise simply convert the virama placeholder to apun iyek for all other cases.

The final transducer \( R^{-1} \) implements reverse romanization transliterating unambiguous Latin script input in ISO 15919 format into corresponding representation in Meetei Mayek.

### D Tables of non-Unicode fonts

Table 23 shows the mappings of ASCII characters to Unicode points for a few fonts we ran into. We discovered these mappings by copy-pasting text from an environment where it rendered (e.g. a PDF) to one where the font wasn’t installed (e.g. a text editor), and finding the Unicode character that looked like the way it rendered.

---

12 [https://github.com/google-research/nisaba](https://github.com/google-research/nisaba)

13 This Nisaba operation is denoted \( R' \) to distinguish it from the more general reversible romanization \( R \) provided by Nisaba for Bengali and Assamese.
Table 20: Performance of the distilled student model in CHiRF and human evaluation, on a scale from 0 (nonsense/wrong language) to 6 (perfect). The value in the “avg” column reports the weighted average score across all sentences. The value under each of the numbers from 0 to 6 is the percent of sentences that were given that rating. The entropy is included to flag suspicious rating patterns: a low entropy may mean that most raters are assigning the same score to all sentences.
| Code | Weight | Description |
|------|--------|-------------|
| CC   | 100    | Natural in-language sentence. It’s ok if it has a few small issues, like spelling errors or a few words from another language, or if it’s a sentence fragment of reasonable length (about 5 words or more) |
| CB   | 50     | In-language, but low-quality. This could be ungrammatical text, boilerplate, or very short fragments. |
| CA   | 30     | Correct but ambiguous whether it’s in the correct language. This code is only applicable for dialects that are closely related to a major language. For instance, many short sentences in Gulf Arabic may also be valid in MSA, and many written Cantonese sentences might also be valid in Mandarin. |
| WD   | 20     | This sentence is in a related but different dialect to the language it’s supposed to be in. This code is only applicable for dialects that are closely related to a similar dialect. For instance, it’s supposed to be in Sa’idi Arabic but it’s in Egyptian Arabic. |
| WL   | 0      | Wrong Language, but still linguistic content |
| NL   | 0      | Not a language – any sort of non-linguistic content. Proper nouns like “Ibuprofin”, “Calvin Klein”, or “Washington DC” also count as NL. |

**Table 21:** Instructions descriptions of the error codes we used to rate samples of our datasets, along with the weight each one is given in the combined quality score.

| Tamazight (ber-Latn) | Ewe (ee) | Moore (mos) |
|----------------------|----------|-------------|
| ASCII codepoint(s)   | ASCII codepoint(s) | ASCII codepoint(s) |
| â U+025B              | 0 U+025B U+0303 | à U+0269    |
| ç U+010D              | 1 U+025B | â U+00E3    |
| é U+1E93              | 2 U+0256 | è U+025B    |
| ë U+1E25              | 3 U+028B | ë U+1EBD    |
| ì U+1E6D              | 4 U+0254 | ì U+0129    |
| ò U+01E7              | 5 U+0192 | ï U+0128    |
| ò U+1E5B              | 6 U+0263 | ò U+00F5    |
| ù U+1E63              | 7 U+00E3 | ù U+028B    |
| ã U+1E0D              | 8 U+1EBD | ã U+0169    |
| À U+0190              | - U+0254 U+0303 | À U+0196  |
| Ç U+010C              | [ U+0169 | À U+00C3    |
| É U+1E92              | ) U+0292 | É U+0190    |
| È U+1E24              | @ U+0189 | È U+1EBC    |
| Í U+1E6C              | & U+00C3 | Í U+00D5    |
| Ô U+01E6              | % U+0191 | Ô U+01B2    |
| Õ U+1E5A              | ‘ U+014B | Õ U+0168    |
| Õ U+1E62              | ” U+0194 | Õ U+01B2    |
| V U+1E6C              | = U+0129 | V U+0186    |
| $ U+0263              | - U+014A | $ U+0186    |

**Table 23:** Three (possibly incomplete) fonts we reconstructed for three African languages that use an extended Latin character set. Table shows the mapping from the ASCII character to the “correct” Unicode codepoint.
### E List of Languages with Regions, Approximate Speaker Counts and Data Sizes

Information about the languages and datasets we found on the web, including the name, number of speakers, continent, script (writing system), cluster (see Section 2.1.3), langID F1 score for the SSLID model, and RTTLANGIDCHRF (loose) score. The RTTLANGIDCHRF score provides an approximate measure of quality of the model, but it should not be trusted as a reliable measure of translation quality (Section 4.3). Number of speakers refers to the estimated number of L1 speakers, following the estimates from van Esch et al. (2022).

| BCP-47 | Mono | Language Name | Speak. | Cont. | Script | Clust. | F1 | RTTLANGIDCHRF |
|--------|------|---------------|--------|-------|--------|--------|----|----------------|
| en     | 7388M| English       | 550M   | Europe| Latn   | en     | 98.2| NA             |
| es     | 1751M| Spanish       | 490M   | Europe| Latn   | es     | 98.8| 77.8           |
| de     | 1693M| German        | 83M    | Europe| Latn   | de     | 98.4| 78.2           |
| id     | 950M | Indonesian    | 200M   | Asia  | Latn   | id     | 97.3| 76.9           |
| hu     | 724M | Hungarian     | 13M    | Europe| Latin  | hu     | 98.8| 68.1           |
| pl     | 687M | Polish        | 38M    | Europe| Latin  | pl     | 98.0| 72.8           |
| zh     | 676M | Chinese       | 1000M  | Asia  | Hans   | zh     | 99.1| 64.0           |
| ko     | 672M | Korean        | 52M    | Asia  | Kore   | ko     | 99.9| 64.0           |
| ja     | 652M | Japanese      | 130M   | Asia  | Jpan   | ja     | 99.6| 65.2           |
| ru     | 570M | Russian       | 150M   | Europe| Cyril  | ru     | 97.3| 74.0           |
| tr     | 565M | Turkish       | 82M    | Europe| Latin  | tr     | 99.0| 70.9           |
| th     | 531M | Thai          | 20M    | Asia  | Thai   | th     | 98.9| 67.9           |
| it     | 491M | Italian       | 61M    | Europe| Latin  | it     | 98.9| 77.3           |
| pt     | 485M | Portuguese    | 220M   | Europe| Latin  | pt     | 98.0| 80.0           |
| vi     | 477M | Vietnamese    | 96M    | Asia  | Latin  | vi     | 99.7| 73.8           |
| fr     | 450M | French        | 75M    | Europe| Latin  | fr     | 97.6| 77.7           |
| ms     | 394M | Malay         | 80M    | Asia  | Latin  | ms     | 92.3| 77.5           |
| nl     | 320M | Dutch         | 21M    | Europe| Latin  | nl     | 99.2| 79.8           |
| uk     | 320M | Ukrainian     | 32M    | Europe| Cyril  | uk     | 99.2| 72.1           |
| sv     | 311M | Swedish       | 8M     | Europe| Latin  | sv     | 98.7| 79.0           |
| ro     | 301M | Romanian      | 19M    | Europe| Latin  | ro     | 96.8| 75.1           |
| cs     | 300M | Czech         | 10M    | Europe| Latin  | cs     | 98.9| 73.0           |
| fa     | 275M | Persian       | 40M    | Asia  | Arabic | fa     | 95.0| 70.3           |
| lv     | 260M | Standard Latvian | 2M   | Europe| Latin | lv     | 98.5| 75.5           |
| ar     | 258M | Arabic        | 180M   | Africa| Arab   | ar     | 97.9| 71.8           |
| no     | 231M | Norwegian     | 5M     | Europe| Latin  | no     | 87.1| 79.9           |
| ta     | 220M | Tamil         | 76M    | Asia  | Tamil  | ta     | 100 | 67.9           |
| el     | 173M | Modern Greek  | 13M    | Europe| Greek  | el     | 99.4| 75.8           |
| fi     | 171M | Finnish       | 5M     | Europe| Latin  | fi     | 97.0| 72.3           |
| hi     | 165M | Hindi         | 720M   | Asia  | Deva   | hi     | 99.7| 75.5           |
| mr     | 156M | Marathi       | 83M    | Asia  | Deva   | mr     | 99.5| 69.8           |
| sk     | 156M | Slovak        | 4M     | Europe| Latin  | sk     | 99.5| 72.6           |
| hy     | 154M | Eastern Armenian | 6M | Europe| Armn  | hy     | 100 | 70.5           |
| kk     | 153M | Kazakh        | 11M    | Europe| Cyril  | kk     | 99.6| 60.6           |
| da     | 152M | Danish        | 5M     | Europe| Latin  | da     | 97.0| 79.3           |
| mk     | 150M | Macedonian    | 1M     | Europe| Cyril  | mk     | 98.8| 77.6           |
| bg     | 149M | Bulgarian     | 7M     | Europe| Cyril  | bg     | 98.3| 74.4           |
| sr     | 141M | Serbian Standard | 6M  | Europe| Cyril | sr     | 98.4| -              |
| ml     | 139M | Malayalam     | 35M    | Asia  | Mlym   | ml     | 100 | 62.5           |
| az     | 135M | Azerbaijani   | 18M    | Asia  | Latin  | az     | 98.3| 70.9           |
| is     | 134M | Icelandic     | 300K   | Europe| Latin  | is     | 97.5| 78.6           |
| te     | 127M | Telugu        | 82M    | Asia  | Telu   | te     | 100 | 72.8           |
| ne     | 116M | Nepali        | 16M    | Asia  | Deva   | ne     | 99.3| 65.6           |
| et     | 112M | Estonian      | 1M     | Europe| Latin  | et     | 99.4| 73.5           |
| he     | 108M | Modern Hebrew | 4M     | Europe| Hebr   | he     | 100 | -              |
| mn     | 100M | Mongolian     | 5M     | Asia  | Cyril | mn     | 97.3| 61.5           |
| ur     | 99M  | Urdu          | 120M   | Asia  | Arab   | ur     | 98.4| 75.5           |
| hr     | 97M  | Croatian Standard | 4M | Europe| Latin | hr     | 79.4| 75.3           |
| fil    | 95M  | Filipino      | 22M    | Asia  | Latin  | fil    | 97.3| 80.6           |
| lt     | 86M  | Lithuanian    | 2M     | Europe| Latin  | lt     | 99.9| 72.8           |
| msi    | 84M  | Brunei-Sabah  | 3M     | Asia  | Latin  | msi    | 94.3| 69.6           |
| be     | 83M  | Belarusian    | 3M     | Europe| Cyril | be     | 99.2| 73.9           |
| BCP-47 | Mono  | Language Name      | Speak. | Cont.  | Script | Clust. | F1    | RTT |
|--------|-------|--------------------|--------|--------|--------|--------|-------|-----|
| kn     | 80M   | Kannada            | 43M    | Asia   | Knda   | kn     | 100   | 71.3 |
| my     | 78M   | Burmese            | 33M    | Asia   | Mymr   | my     | 99.7  | 53.3 |
| bn     | 73M   | Bengali            | 260M   | Asia   | Beng   | bn     | 98.0  | 70.4 |
| af     | 71M   | Afrikaans          | 17M    | Africa | Latn   | af     | 97.7  | 82.1 |
| eu     | 70M   | Basque             | 500K   | Europe | Latn   | eu     | 99.9  | 72.9 |
| gu     | 69M   | Gujarati           | 56M    | Asia   | Gujr   | gu     | 98.0  | 75.2 |
| mi     | 67M   | Maori              | 200K   | Oceania| Latn   | mi     | 99.6  | 67.2 |
| gdx    | 67M   | Godwari            | 3M     | Asia   | Deva   | gdx    | 70.5  | 74.2 |
| gl     | 64M   | Galician           | 2M     | Europe | Latn   | gl     | 94.2  | 76.9 |
| sl     | 59M   | Slovenian          | 2M     | Europe | Latn   | sl     | 97.9  | 72.7 |
| lo     | 59M   | Lao                | 20M    | Asia   | Laoo   | lo     | 100   | 72.0 |
| gd     | 57M   | Scottish Gaelic    | 50K    | Europe | Latn   | gd     | 99.4  | 74.5 |
| so     | 56M   | Somali             | 24M    | Africa | Latn   | so     | 99.7  | 67.4 |
| si     | 56M   | Sinhala            | 16M    | Asia   | Sinh   | si     | 100   | 61.3 |
| uz     | 53M   | Uzbek              | 57M    | Europe | Latn   | uz     | 99.8  | 62.2 |
| sw     | 52M   | Swahili            | 24M    | Africa | Latn   | sw     | 99.3  | 78.1 |
| km     | 51M   | Central Khmer      | 16M    | Asia   | Khmr   | km     | 99.9  | 69.4 |
| ha     | 51M   | Hausa              | 70M    | Africa | Latn   | ha     | 99.4  | 70.5 |
| wry    | 50M   | Merwari            | 8M     | Asia   | Deva   | wry    | 68.2  | -    |
| ky     | 49M   | Kirghiz            | 8M     | Europe | Cyril  | ky     | 99.8  | 56.7 |
| am     | 47M   | Amharic            | 22M    | Africa | Ethii  | am     | 95.1  | 58.0 |
| ca     | 44M   | Catalan            | 4M     | Europe | Latn   | ca     | 99.5  | 75.0 |
| ceb    | 44M   | Cebuano            | 16M    | Asia   | Latn   | ceb    | 99.1  | 73.9 |
| yo     | 43M   | Yoruba             | 21M    | Africa | Latn   | yo     | 99.6  | 73.7 |
| sd     | 42M   | Sindhi             | 68M    | Asia   | Arab   | sd     | 99.0  | 67.3 |
| giu    | 41M   | Gujarati           | 1M     | Asia   | Deva   | giu    | 68.6  | -    |
| co     | 41M   | Corsican           | 50K    | Europe | Latn   | co     | 97.5  | 75.2 |
| mg     | 40M   | Malagasy           | 18M    | Africa | Latn   | mg     | 99.9  | 65.4 |
| pa     | 37M   | Eastern Punjabi    | 28M    | Asia   | Guru   | pa     | 99.7  | 71.9 |
| ar-MA  | 36M   | Arabic             | 180M   | Africa | Arab   | ar-MA  | 46.1  | 69.0 |
| bew    | 33M   | Betawi             | 5M     | Asia   | Latn   | bew    | 93.8  | 67.2 |
| ny     | 33M   | Chichewa           | 12M    | Africa | Latn   | ny     | 98.8  | 66.1 |
| sq     | 32M   | Albanian           | 5M     | Europe | Latn   | sq     | 99.3  | 79.0 |
| ka     | 30M   | Georgian           | 4M     | Europe | Geor   | ka     | 99.9  | 71.9 |
| haw    | 30M   | Hawaiian           | 50K    | Americas| Latn  | haw    | 99.7  | 58.3 |
| hmm    | 28M   | Hmong              | 3M     | Europe | Latn   | hmm    | 99.7  | 71.5 |
| yi     | 28M   | Yiddish            | 2M     | Europe | Hebr   | yi     | 100   | 84.8 |
| ig     | 27M   | Igbo               | 18M    | Africa | Latn   | ig     | 99.0  | 68.6 |
| ckb    | 25M   | Central Kurdish    | 7M     | Europe | Arab   | ckb    | 99.3  | 60.0 |
| tt     | 25M   | Tatar              | 4M     | Europe | Cyril  | tt     | 97.0  | 52.5 |
| gsw    | 25M   | Central Alemannic  | 5M     | Europe | Latn   | gsw    | 93.2  | 73.3 |
| arq    | 24M   | Algerian Arabic    | 27M    | Africa | Arab   | arq    | 79.0  | 66.0 |
| sm     | 24M   | Samoan             | 500K   | Oceania| Latn   | sm     | 99.9  | 67.6 |
| eo     | 24M   | Esperanto          | 2M     | Europe | Latn   | eo     | 99.7  | 78.4 |
| zu     | 23M   | Zulu               | 11M    | Africa | Latn   | zu     | 96.3  | 71.3 |
| st     | 23M   | Southern Sotho     | 14M    | Africa | Latn   | st     | 98.8  | 69.2 |
| ru-Latin| 23M  | Russian            | 150M   | Europe | Latn   | ru-Latin| 97.2  | 51.4 |
| cy     | 22M   | Welsh              | 700K   | Europe | Latn   | cy     | 99.7  | 79.4 |
| la     | 22M   | Latin              | 0K     | Europe | Latn   | la     | 93.2  | 67.9 |
| or     | 21M   | Odia               | 35M    | Asia   | Orya   | or     | 100   | 63.2 |
| nt     | 21M   | Maltese            | 500K   | Europe | Latn   | nt     | 99.5  | 84.2 |
| rkt    | 20M   | Rangpuri           | 15M    | Asia   | Beng   | rkt    | 98.8  | 59.2 |
| ht     | 20M   | Haitian            | 10M    | Americas| Latn  | ht     | 99.9  | 79.3 |
| sn     | 18M   | Shona              | 9M     | Africa | Latn   | sn     | 99.6  | 68.6 |
| bo     | 18M   | Tibetan            | 1M     | Asia   | Tibet | bo     | 99.9  | 58.5 |
| ps     | 18M   | Nuclear Pashto     | 13M    | Asia   | Arab   | ps     | 98.9  | 59.7 |
| ga     | 17M   | Irish              | 200K   | Europe | Latn   | ga     | 98.8  | 80.9 |
| shu    | 16M   | Chadian Arabic     | 1M     | Africa | Arab   | shu    | 89.5  | -    |
| xh     | 16M   | Xhosa              | 19M    | Africa | Latn   | xh     | 97.8  | 69.0 |
| lb     | 15M   | Moselle Franconian | 400K   | Europe | Latn   | lb     | 96.7  | 63.7 |
| ku     | 15M   | Kurdish            | 32M    | Europe | Latn   | ku     | 98.2  | 57.7 |
| 58     |
| BCP-47 | Mono | Language Name | Speak. | Cont. | Script | Clust. | F1 | RTT |
|--------|------|---------------|--------|-------|--------|--------|----|-----|
| tg     | 14M  | Tajik         | 8M     | Europe| Cyril  | tg     | 99.9| 64.9|
| hi-Latn| 14M  | Hindi         | 320M   | Asia  | Latin  | bgh-13l| 98.1| 71.7|
| jv     | 13M  | Javanese      | 85M    | Asia  | Latin  | jv     | 99.4| 77.0|
| apd-SD | 13M  | Sudanese Arabic| 17M | Africa | Arab   | acm-9l | 74.2| 69.1|
| su     | 12M  | Sundanese     | 39M    | Asia  | Latin  | su     | 99.1| 71.1|
| mzn    | 12M  | Mazanderani   | 6M     | Asia  | Arab   | apr-8l | 52.4| 64.8|
| meo    | 11M  | Kedah Malay   | 3M     | Asia  | Latin  | ace-12l| 94.4| 66.7|
| fy     | 11M  | Western Frisian| 800K | Europe| Latin  | fy     | 99.7| 71.7|
| mrw    | 10M  | Maranao       | 800K   | Asia  | Latin  | agn-14l| 98.0| 69.2|
| rw     | 10M  | Kinyarwanda   | 11M    | Africa| Latin  | rw     | 98.6| 47.7|
| apc    | 10M  | North Levantine| 15M | Europe| Arabic | abv-10l| 40.4| 67.7|
| as     | 9M   | Assamese      | 13M    | Asia  | Beng   | as     | 99.7| 63.7|
| aec    | 9M   | Saidi Arabic  | 19M    | Africa| Arabic | abv-10l| 78.0| 64.3|
| fo     | 9M   | Faroese       | 50K    | Europe| Latin  | bn-Latn-6l | 98.0| 71.6|
| pap    | 9M   | Papiamento    | 600K   | Americas| Latin  | ach-10l| 94.8| 61.4|
| sah    | 9M   | Sakha         | 500K   | Europe| Cyril  | az-RU-13l| 97.7| 43.4|
| acm    | 9M   | Gilit         | 26M    | Europe| Arabic | acm-9l | 35.2| 64.2|
| lus    | 8M   | Mizo          | 700K   | Asia  | Latin  | bgr-6l | 100 | 35.9|
| grc    | 8M   | Ancient Greek (to 1453) | 0K | Europe| Greek  | grc | 99.9 | - |
| dv     | 8M   | Dhivehi       | 300K   | Asia  | Thaa   | bpy-7l | 100 | 53.6|
| oc     | 7M   | Occitan       | 500K   | Europe| Latin  | acf-10l| 96.7| 73.4|
| luz    | 7M   | Southern Luri | 900K   | Asia  | Arabic | fay-14l| 68.4| 64.2|
| sa     | 6M   | Sanskrit      | 100K   | Asia  | Deva   | bho-13l| 97.8| 57.6|
| bgp-Ab| 6M   | Eastern Balochi| 2M | Asia  | Arabic | bal-13l| 55.1 | - |
| kg     | 6M   | Uyghur        | 10M    | Asia  | Arabic | ckb-2l | 100 | 50.5|
| ba     | 6M   | Bashkir       | 1M     | Europe| Cyril  | az-RU-13l| 98.0| 58.2|
| om     | 6M   | Oromo         | 24M    | Africa| Latin  | amf-6l | 99.7| 56.0|
| zza    | 5M   | Zaza          | 2M     | Europe| Latin  | aem-Latn-11l| 71.7| - |
| zh-Latn| 5M   | Chinese       | 1000M  | Asia  | Latin  | cqd-8l | 99.2| 22.2|
| vas-Gujr| 4M | Vassavini | 1M | Asia  | Gujr  | bbb-Gujr-3l | 78.7| 68.5|
| vec    | 4M   | Venetian      | 4M     | Europe| Latin  | bi-6l | 94.5| 73.5|
| tk     | 4M   | Turkmen       | 8M     | Europe| Latin  | tk     | 99.9| 57.5|
| hil    | 4M   | Hiligaynon    | 6M     | Asia  | Latin  | abs-11l| 99.3| 70.8|
| ti     | 4M   | Tigrinya      | 9M     | Africa| Ethnic | ethi-gez-8l | 99.9| 57.0|
| syl    | 4M   | Sylheti       | 23M    | Asia  | Beng   | bpy-7l | 86.9| 65.8|
| phr    | 4M   | Pahari Potwari| 2M | Asia  | Arabic | bal-13l| 54.5| 73.6|
| ayl    | 4M   | Libyan Arabic | 5M    | Africa| Arabic | acm-9l | 86.3| 65.0|
| el-Latn| 3M   | Modern Greek  | 13M    | Europe| Latin  | bg-Latn-5l| 98.9| 26.0|
| kaa    | 3M   | Kara-Kalpak   | 600K   | Europe| Cyril  | az-RU-13l| 99.3| 61.8|
| bs     | 3M   | Bosnian Standard| 9M | Europe| Cyril  | bs     | 70.8| 74.1|
| bgp    | 3M   | Eastern Balochi| 2M | Asia  | Latin  | bgp-13l| 91.1| 70.0|
| br     | 3M   | Breton        | 200K   | Europe| Latin  | br-4l | 82.8| 42.8|
| gdq    | 3M   | Mehri         | 100K   | Europe| Arabic | acm-9l | 51.7 | - |
| tet    | 3M   | Tetum         | 500K   | Asia  | Latin  | ceg-5l | 98.4| 53.5|
| to     | 3M   | Tonga (Tonga Islands) | 200K | Oceania| Latin | mi-4l | 100 | 57.5|
| cv     | 3M   | Chuvenh       | 1M     | Europe| Cyril  | az-RU-13l| 92.9| 48.6|
| kfr    | 3M   | Kachchi       | 900K   | Asia  | Gujr  | bbb-Gujr-3l| 82.5| 63.4|
| ilo    | 3M   | Ilocano       | 9M     | Asia  | Latin  | agn-14l| 93.4| 47.6|
| bg-Latn| 3M   | Bulgarian     | 7M     | Europe| Latin  | bg-Latn-5l| 97.2| 65.0|
| bal    | 3M   | Balochi       | 8M     | Asia  | Arabic | bal-13l| 23.0| 61.8|
| hif    | 2M   | Fiji Hindi    | 500K   | Oceania| Latin  | bgp-13l| 99.6| 69.4|
| nes    | 2M   | Hassassiyaw   | 3M     | Africa| Arabic | acm-9l | 69.5 | - |
| dz     | 2M   | Dzongkha      | 200K   | Asia  | Tibet | bo-2l | 98.2| 39.5|
| tn     | 2M   | Tswana        | 8M     | Africa| Latin  | cce-8l | 90.5| 66.5|
| bn-Latn| 2M   | Bengali       | 260M   | Asia  | Latin  | bn-Latn-6l| 98.5| 57.5|
| lg     | 2M   | Ganda         | 4M     | Africa| Latin  | cgg-15l| 98.7| 58.2|
| te-Latn| 2M   | Telugu        | 82M    | Asia  | Latin  | bgp-13l| 98.6| 59.3|
| mde    | 2M   | Maba (Chad)   | 300K   | Africa| Arabic | acm-9l | 48.5 | - |
| BCP-47 | Mono | Language Name | Speak. | Cont. | Script | Clust. | F1   | RTT  |
|--------|------|---------------|--------|-------|--------|--------|------|------|
| nso    | 2M   | Pedi          | 14M    | Africa| Latn   | cee-8l | 98.6 | 68.0 |
| rm     | 2M   | Romanes       | 50K    | Europe| Latn   | rm     | 99.9 | 61.7 |
| za     | 2M   | Northern Daic | 15M    | Asia  | Latn   | eq8-8l | 98.7 | 27.3 |
| tpi    | 2M   | Tok Pisin     | 4M     | Oceania| Latn   | bi-6l  | 99.3 | 61.3 |
| pcm    | 2M   | Nigerian Pidgin| 30M   | Africa| Latn   | pcm    | 97.9 | 71.1 |
| mtq    | 2M   | Haka Chin     | 100K   | Asia  | Latn   | bgr-6l | 99.9 | 37.2 |
| iba    | 2M   | Iban          | 2M     | Asia  | Latn   | amo-7l | 99.8 | 54.8 |
| ja-Latn| 2M   | Japanese      | 130M   | Asia  | Latn   | dgr-5l | 98.6 | 20.9 |
| ium    | 2M   | Iu Mien       | 800K   | Asia  | Latn   | eq8-8l | 98.9 | 18.9 |
| crs    | 2M   | Seselwa Creole| 50K    | Africa| Latn   | acf-10l| 99.8 | 72.9 |
| se     | 1.5M | North Saami   | 50K    | Europe| Latn   | bn-Latn-6l | 90.2 | -    |
| kha    | 1.4M | Khasi         | 800K   | Asia  | Latn   | ace-12l| 99.9 | 38.3 |
| ln     | 1.4M | Kinshasa Lingala| 15M   | Africa| Latn   | jbu-8l | 96.4 | 56.1 |
| tyv    | 1.4M | Tuvinian      | 300K   | Europe| Cyrl   | sty-2l | 99.4 | 47.4 |
| ce     | 1.4M | Chechen       | 1M     | Europe| Cyrl   | ce-3l  | 99.9 | 41.6 |
| mai    | 1.3M | Maithili      | 17M    | Asia  | Deva   | awa-14l| 98.6 | 60.8 |
| nr     | 1.3M | South Transvaal| 1M    | Africa| Latn   | gsp-8l | 86.3 | -    |
| dty    | 1.3M | Dotyali       | 800K   | Asia  | Deva   | bho-13l| 77.6 | 57.8 |
| ts     | 1.3M | Tsonga        | 13M    | Africa| Latn   | cjk-14l| 72.9 | 55.0 |
| fj     | 1.3M | Fijian        | 300K   | Oceania| Latn   | fj-4l  | 98.8 | 37.5 |
| ak     | 1.3M | Akan          | 13M    | Africa| Latn   | abr-5l | 93.8 | 49.5 |
| kaa-Latn| 1.3M | Kara-Kalpak   | 600K   | Europe| Latn   | bqcq-Latn-12l | 86.7 | 55.9 |
| os     | 1.3M | Iron Ossetian | 600K   | Europe| Cyrl   | gaw-8l | 93.5 | 55.2 |
| bik    | 1.2M | Bikol         | 2M     | Asia  | Latn   | abr-11l| 94.4 | 68.7 |
| ber-Latn| 1.2M | Berber        | 2M     | Africa| Latn   | abq-4l | 99.7 | 32.9 |
| ctd-Latn| 1.2M | Tedim Chin    | 300K   | Asia  | Latn   | bgs-9l | 99.5 | 23.5 |
| tly-IR | 1.2M | North-Central | 900K   | Europe| Arab   | fay-14l| 88.7 | 63.1 |
| plk    | 1.1M | Kohistani Shina| 200K  | Asia  | Arab   | bal-13l| 65.1 | -    |
| tks    | 1.1M | Takestani     | 200K   | Asia  | Arab   | fay-14l| 88.9 | -    |
| kmz    | 1.1M | Khorasan Turkic| 400K  | Asia  | Arab   | fay-14l| 60.7 | 64.4 |
| mnw    | 1.1M | Mon           | 900K   | Asia  | Mynmr  | ksw-3l | 99.9 | 24.4 |
| scl    | 1M   | Shina         | 400K   | Asia  | Arab   | bal-13l| 30.1 | -    |
| ml-Latn| 1M   | Malayalam     | 35M    | Asia  | Latn   | bgp-13l| 98.7 | 58.2 |
| el-CY  | 1M   | Modern Greek  | 13M    | Europe| Grek   | el-CY  | 99.6 | 71.0 |
| skr    | 1M   | Saraiki       | 26M    | Asia  | Arab   | doi-Arab-51 | 99.8 | 65.1 |
| udm    | 1M   | Udmurt        | 300K   | Europe| Cyrl   | ady-9l | 98.1 | 45.6 |
| bbc    | 900K | Batak Toba   | 2M     | Asia  | Latn   | ace-12l| 98.3 | 47.9 |
| war    | 900K | Waray (Philip-| 5M    | Asia  | Latn   | bgs-9l | 98.3 | 70.8 |
| mki    | 900K | Dhatki        | 200K   | Asia  | Arab   | bal-13l| 26.8 | -    |
| sm     | 900K | Sranan Tongo | 100K   | Americas| Latn| djk-3l  | 98.4 | 56.9 |
| pck    | 900K | Paite Chin    | 50K    | Asia  | Latn   | bgs-9l | 99.5 | 32.3 |
| gn     | 900K | Guarani       | 5M     | Americas| Latn| gn-6l  | 98.8 | 50.3 |
| qu     | 800K | Quechua       | 9M     | Americas| Latn| ae-Latn-10l | 98.3 | 48.8 |
| kj     | 800K | Kuanyama      | 600K   | Africa| Latn   | kj-5l  | 98.5 | 39.9 |
| bug    | 800K | Buginese      | 5M     | Asia  | Latn   | amo-7l | 83.5 | -    |
| ee     | 800K | Ewe           | 4M     | Africa| Latn   | avn-3l | 99.5 | 49.6 |
| ltg    | 800K | East Latvian  | 200K   | Europe| Latn   | gaw-8l | 94.9 | 63.7 |
| hsb    | 800K | Upper Sorbian | 5K     | Europe| Latn   | ach-10l| 59.6 | -    |
| kl     | 700K | Kalaallisut   | 50K    | Americas| Latn| bi-6l  | 87.7 | 34.1 |
| bho    | 700K | Bhojpuri     | 51M    | Asia  | Deva   | bho-13l| 96.6 | 56.2 |
| ta-Latn| 700K | Tamil         | 76M    | Asia  | Latn   | bgr-13l| 99.1 | 53.8 |
| ve     | 700K | Venda        | 2M     | Africa| Latn   | cce-8l | 99.9 | 52.9 |
| bi     | 700K | Bislama      | 200K   | Oceania| Latn| bi-6l  | 99.1 | -    |
| az-RU  | 700K | Azerbajiani  | 18M    | Europe| Cyrl   | az-RU-13l | 99.7 | 58.1 |
| kbd    | 700K | Kabardian     | 2M     | Europe| Cyrl   | ady-9l | 87.7 | 46.1 |
| nut    | 600K | Nung (Viet Nam)| 1M    | Asia  | Latn   | bhcq-Latn-4l | 97.9 | -    |
| hif-Deva| 600K | Fiji Hindi   | 500K   | Oceania| Deva | awa-14l| 89.5 | 69.9 |
| BCP-47 | Mono | Language Name | Speak. | Cont. | Script | Clust. | F1  | RTT |
|--------|------|-------------|--------|------|--------|-------|-----|-----|
| ar-Latn | 600K | Arabic | 180M | Europe | Latin | aeb-Latin-111 | 95.2 | 29.5 |
| krc | 600K | Karachay-Balkar | 300K | Europe | Cyril | az-RU-131 | 98.5 | 52.7 |
| chn | 600K | Mari | 500K | Europe | Cyril | az-RU-131 | 99.5 | 44.5 |
| pag | 600K | Pangasinan | 1M | Asia | Latin | agn-141 | 92.0 | 48.9 |
| mui | 600K | Musi | 3M | Asia | Latin | abs-131 | 90.4 | 62.5 |
| shn | 600K | Shan | 3M | Asia | Latin | Mymr | 100 | 40.3 |
| gom-Latn | 500K | Goan Konkani | 2M | Asia | Latin | bgp-131 | 99.0 | 56.0 |
| noe | 500K | Nimadi | 2M | Asia | Deva | ahr-131 | 99.2 | 62.8 |
| nh | 500K | Marshallese | 50K | Oceania | Latin | kjp-5l | 100 | 31.4 |
| min | 500K | Minangkabau | 6M | Asia | Latin | ace-121 | 100 | 59.1 |
| yue | 500K | Yue Chinese | 84M | Asia | Hant | yue | 99.1 | 55.4 |
| new | 500K | Kathmandu Valley | 900K | Asia | Deva | ahr-131 | 98.3 | 51.4 |
| cjk | 500K | Chokwe | 1M | Africa | Latin | cjk-141 | 98.7 | 43.7 |
| jax | 500K | Jambi Malay | 1M | Asia | Latin | abs-131 | 85.8 | 69.1 |
| bft | 500K | Balti | 300K | Asia | Arab | bal-131 | 13.7 | - |
| dov | 500K | Dombe | 2M | Africa | Latin | dov-9l | 100 | 44.8 |
| kv | 500K | Komi | 200K | Europe | Cyril | ady-9l | 95.2 | 44.0 |
| cfm | 500K | Falam Chin | 100K | Asia | Latin | bgr-6l | 99.9 | 36.7 |
| fay | 500K | Fars Dialects | 50K | Asia | Arab | fay-141 | 60.9 | 62.3 |
| mr-Latn | 500K | Marathi | 83M | Asia | Latin | bgp-131 | 98.0 | 64.5 |
| lu | 400K | Luba-Katanga | 2M | Africa | Latin | jbu-8l | 98.3 | 39.1 |
| wa | 400K | Walloon | 600K | Europe | Latin | acf-10l | 91.2 | 54.1 |
| yua | 400K | Yucatec Maya | 800K | Americas | Latin | chd-6l | 99.3 | 57.9 |
| abv | 400K | Baharna Arabic | 300K | Europe | Arab | abv-10l | 8.1 | - |
| sv | 400K | Sango | 2M | Africa | Latin | exn-4l | 99.7 | 31.0 |
| iso | 400K | Isoko | 400K | Africa | Latin | agg-8l | 99.9 | 35.1 |
| kac | 400K | Southern Jinghpaw | 900K | Asia | Latin | ace-121 | 100 | 38.4 |
| syr | 400K | Syriac | 1M | Europe | Syrc | ach-10l | 99.9 | 40.4 |
| pam | 400K | Pampanga | 2M | Asia | Latin | agrn-141 | 99.5 | 52.5 |
| afb | 400K | Gulf Arabic | 5M | Europe | Arab | abv-10l | 4.1 | - |
| qxm | 400K | Qashqä’i | 2M | Asia | Arab | fay-14l | 61.8 | 61.0 |
| ada | 400K | Adangme | 800K | Africa | Latin | ada-131 | 99.7 | 44.1 |
| ctg | 400K | Chittagonian | 15M | Asia | Beng | bry-7l | 90.9 | 65.2 |
| kg | 400K | Kongo | 6M | Africa | Latin | jbu-8l | 93.8 | - |
| mtr | 400K | Mewari | 5M | Asia | Deva | bfrz-151 | 89.2 | 44.8 |
| io | 400K | Ido | 5K | Europe | Latin | acf-10l | 84.6 | - |
| pmy | 400K | Papuan Malay | 500K | Asia | Latin | abs-131 | 88.3 | 68.3 |
| ss-SZ | 400K | Swati | 3M | Africa | Latin | gup-8l | 87.3 | 64.7 |
| mwr | 400K | Marwari | 8M | Asia | Deva | bfrz-151 | 97.3 | - |
| ktu | 400K | Kituba (DRC) | 5M | Africa | Latin | jbu-8l | 99.8 | 43.6 |
| nd | 300K | Zimbabwean Ndebele | 2M | Africa | Latin | gup-8l | 98.6 | 69.1 |
| kek | 300K | Kekchí | 800K | Americas | Latin | kekc-2l | 100 | 35.1 |
| mfe | 300K | Morisjen | 1M | Africa | Latin | acf-10l | 100 | 66.0 |
| bua | 300K | Buriat | 400K | Europe | Cyril | bua-2l | 96.9 | 44.5 |
| tvl | 300K | Tuvaluan | 50K | Oceania | Latin | mi-4l | 99.6 | - |
| chk | 300K | Chuukese | 50K | Oceania | Latin | chk-2l | 100 | 34.4 |
| dcr | 300K | Decan | 13M | Asia | Arab | bal-131 | 89.2 | 61.2 |
| gom | 300K | Goan Konkani | 2M | Asia | Deva | awa-141l | 85.6 | 59.2 |
| pon | 300K | Pohnpeian | 50K | Oceania | Latin | bgs-9l | 99.7 | 39.9 |
| av | 300K | Avar | 800K | Europe | Cyril | ady-9l | 99.1 | 45.8 |
| mfb | 300K | Bangka | 300K | Asia | Latin | abs-131 | 91.5 | 66.6 |
| sjp | 300K | Surjapuri | 1M | Asia | Deva | bho-131l | 96.3 | 55.4 |
| tiv | 300K | Tiv | 2M | Africa | Latin | bmm-3l | 100 | 27.1 |
| ady | 300K | Adyge | 600K | Europe | Cyril | ady-9l | 89.2 | 43.2 |
| btx | 300K | Batak Karo | 600K | Asia | Latin | ban-3l | 100 | 39.1 |
| wo | 300K | Wolof | 4M | Africa | Latin | aeb-Latin-111 | 99.3 | 50.1 |
| tsc | 300K | Tswana | 1M | Africa | Latin | cce-8l | 98.6 | 66.9 |
| hne | 300K | Chhattisgarhi | 18M | Asia | Deva | bho-131l | 98.9 | 56.9 |
| ay | 300K | Central-Southern Aymara | 3M | Americas | Latin | auc-6d | 99.6 | 31.0 |
| BCP-47 | Mono | Language Name       | Speak. | Cont. | Script | Clust.       | F1    | RTT |
|--------|------|---------------------|--------|-------|--------|--------------|-------|-----|
| pa-Arab| 300K | Eastern Punjabi     | 28M    | Asia  | Arab   | don-Arab-5t  | 96.4  | 64.3|
| gug    | 300K | Gagauz              | 600K   | Europe| Latn   | aeb-Latn-111 | 99.0  | 61.9|
| raj    | 300K | Rajasthani          | 18M    | Asia  | Deva   | bzf-15t      | 88.0  | 47.9|
| quc    | 200K | K’iche’             | 2M     | Americas| Latn | acr-4t      | 99.9  | 38.8|
| zyj    | 200K | Youjiang Zhuang     | 900K   | Asia  | Latn   | cqpl-8t      | 99.1  | 26.4|
| bhh-Gujr| 5l  | Bhili               | 3M     | Asia  | Gujr   | bhh-Gujr-3t  | 98.2  | -   |
| ho     | 300K | Hiri Motu           | 100K   | Oceania| Latn | gah-11t      | 99.4  | 32.6|
| zap    | 200K | Zapotec             | 400K   | Americas| Latn | zaa-5t      | 100   | 32.1|
| crh    | 200K | Crimean Tatar       | 500K   | Europe| Cyr   | az-RU-13t    | 99.7  | 53.1|
| mam    | 200K | Mam                 | 500K   | Americas| Latn | chd-6t      | 99.8  | 28.3|
| rhg-Latn| 6l | Rohingya            | 2M     | Asia  | Latn   | bn-Latn-6t   | 98.7  | 50.0|
| rom    | 200K | Romani              | 4M     | Europe| Latn   | rmc-3t      | 99.1  | 44.0|
| mgh    | 200K | Makhua-Meetto       | 7M     | Africa| Latn   | ahu-13t    | 99.6  | 30.0|
| lrc    | 200K | Northern Luri       | 2M     | Asia  | Arab   | apr-8t      | 67.2  | -   |
| kmz-Latn| 8l | Khorasan Luri       | 400K   | Asia  | Latn   | aeb-Latn-111 | 93.0  | 54.9|
| myv    | 200K | Erzya               | 400K   | Europe| Cyr   | ady-9t      | 98.0  | 36.3|
| ace    | 200K | Acehnese            | 4M     | Asia  | Latn   | ace-12t     | 99.7  | 50.1|
| bns    | 200K | Bundeli             | 3M     | Asia  | Deva   | awa-14t     | 87.8  | -   |
| acw    | 200K | Hijazi Arabic       | 14M    | Europe| Arab   | abv-10t     | 6.4   | -   |
| nan-Latn-TW | 200K | Min Nan Chinese | 59M | Asia | Latn | cdo-5t | 100 | 27.1 |
| quh    | 200K | South Bolivian Quechua | 3M | Americas| Latn | aeb-Latn-10t | 99.9 | 39.2 |
| gbm    | 200K | Garhwali            | 3M     | Asia  | Deva   | bzf-15t     | 95.1  | 53.9|
| ctu    | 200K | Chol                | 200K   | Americas| Latn | chr-9t      | 100   | 29.8|
| acq    | 200K | Ta’izz-i-Adeni Arab- | 7M | Europe| Arab | abv-10t | 12.3 | -   |
| bfy    | 200K | Bagheli             | 8M     | Asia  | Deva   | awa-14t     | 91.5  | 62.0|
| gv     | 200K | Manx                | 5K     | Europe| Latn   | gv-3t       | 98.3  | 36.6|
| xmm    | 200K | Manado Malay        | 800K   | Asia  | Latn   | abs-13t     | 94.5  | 59.7|
| mel    | 200K | Central Melanau     | 100K   | Asia  | Latn   | abs-111     | 89.7  | 74.2|
| ksw    | 200K | S’gaw Karen        | 1M     | Asia  | Mymr   | ksw-3t      | 100   | 38.0|
| seh    | 200K | Sena                | 2M     | Africa| Latn   | cjk-14t     | 99.5  | 52.5|
| msb    | 200K | Masbatenyo          | 400K   | Asia  | Latn   | abs-111     | 99.4  | 57.3|
| tsg    | 200K | Tausug             | 1M     | Asia  | Latn   | abs-111     | 98.4  | 54.0|
| dtp    | 200K | Central Dusun      | 100K   | Asia  | Latn   | abs-111     | 100   | 23.9|
| yap    | 200K | Yapese             | 5K     | Oceania| Latn | bru-6t     | 99.3  | -   |
| kum    | 200K | Kumyk              | 400K   | Europe| Cyr   | az-RU-13t    | 99.9  | 47.9|
| fon    | 200K | Fon                | 1M     | Africa| Latn   | ada-13t     | 99.9  | 40.9|
| alt    | 200K | Southern Altai     | 50K    | Europe| Cyr   | alt       | 100   | 43.4|
| wal    | 200K | Wolaytta           | 2M     | Africa| Latn   | dta-Latn-8t | 99.7  | 29.6|
| en-Arab| 200K | English            | 550M   | Europe| Arab   | abv-10t    | 97.4  | 56.5|
| nzi    | 200K | Nzima              | 300K   | Africa| Latn   | bow-4t      | 100   | 40.6|
| cak    | 200K | Kaqchikel          | 400K   | Americas| Latn | aeb-Latn-111 | 100   | 24.1|
| ban    | 200K | Balinese           | 3M     | Asia  | Latn   | ban-3t      | 100   | 51.7|
| zne    | 200K | Zande              | 1M     | Africa| Latn   | agt-6t      | 99.5  | 16.9|
| bm     | 200K | Bambara           | 14M    | Africa| Latn   | ada-131     | 99.1  | 42.1|
| mdh    | 200K | Maguindanao        | 1M     | Asia  | Latn   | agn-14t     | 96.4  | 43.7|
| xal    | 200K | Oirad-Kalmyk-Darkhat | 400K | Europe| Cyr   | az-RU-13t | 91.5  | 42.8|
| abs    | 200K | Ambonese Malay     | 2M     | Asia  | Latn   | abs-131     | 88.8  | 63.7|
| tzo    | 200K | Tzotzil            | 200K   | Americas| Latn | aeb-Latn-111 | 99.9 | 29.5|
| bge    | 200K | Haryanvi           | 13M    | Asia  | Deva   | xar-13t     | 91.7  | 56.9|
| vkt    | 200K | Tenggarong Kutai Malai | 200K | Asia | Latn | abs-131 | 93.6 | 64.8|
| doi    | 200K | Kangri-Dogri       | 2M     | Asia  | Deva   | bzf-15t     | 93.5  | -   |
| gez    | 200K | Geez              | 0K     | Africa| Ethi   | gez-8t      | 87.6  | -   |
| rwr    | 200K | Marwari (India)    | 8M     | Asia  | Deva   | bzf-15t     | 84.9  | 44.9|
| tum    | 200K | Tumbuka           | 2M     | Africa| Latn   | cjk-14t     | 47.9  | -   |
| ang    | 200K | Old English (ca. 450-1100) | 0K | Europe| Latn | ang-7t | 97.6  | -   |
| BCP-47 | Mono | Language Name                  | Speak. | Cont. | Script | Clust. | F1   | RTT  |
|--------|------|--------------------------------|--------|------|--------|--------|------|------|
| nia    | 200K | Nias                           | 800K   | Asia | Latn   | ceg-5l | 100  | 23.1 |
| bts    | 200K | Batak Simalungun               | 1M     | Asia | Latn   | agn-14l| 100  | 34.1 |
| ng     | 200K | Ndonda                         | 800K   | Africa| Latn  | kj-5l  | 95.0 | 35.0 |
| ach    | 200K | Acoli                          | 1M     | Africa| Latn  | ach-108| 99.6 | 21.0 |
| rcf    | 200K | Réunion Creole French         | 600K   | Africa| Latn  | acf-108| 99.3 | 66.9 |
| kos    | 200K | Kosraean                       | 5K     | Oceania| Latn | ceg-5l | 98.8 | -    |
| skg    | 200K | West Malagasy Sakalava         | 1M     | Africa| Latn  | apy-7l| 93.4 | 57.6 |
| mrj    | 200K | Western Mari                   | 5K     | Europe| Cyrl  | az-RU-13l| 94.3 | 45.8 |
| bci    | 200K | Baoulé                         | 2M     | Africa| Latn  | ada-13l| 98.9 | 32.1 |
| ch     | 200K | Chamorro                       | 50K    | Oceania| Latn | ayo-4l| 98.9 | 31.9 |
| qvi    | 200K | Imbabura Highland Quichua      | 300K   | Americas| Latn | qsw-5l| 100  | 38.7 |
| gym    | 200K | Ngâbere                        | 200K   | Americas| Latn | gym   | 100  | 19.5 |
| cyo    | 150K | Cuyomonon                      | 100K   | Asia | Latn   | agn-14l| 85.0 | 69.1 |
| sqq    | 150K | Ems-Weser Frisian             | 5K     | Europe| Latn  | ap-8l  | 89.6 | -    |
| tcy    | 150K | Tulu                           | 2M     | Asia | Knda  | lmu-Knda-2l| 100  | 49.4 |
| meu    | 150K | Motu                           | 50K    | Oceania| Latn | gah-11l| 99.1 | 27.4 |
| pau    | 150K | Palauan                        | 50K    | Oceania| Latn | pau   | 100  | -    |
| bas    | 150K | Basa (Cameroon)               | 300K   | Africa| Latn  | bas-11l| 99.4 | 21.0 |
| gac    | 140K | Wayuu                          | 300K   | Americas| Latn | gac   | 100  | 18.5 |
| quy    | 140K | Ayacucho                       | 900K   | Americas| Latn | ae-Latn-10l| 99.7 | 29.8 |
| gu-Latn| 140K | Gujarati                       | 56M    | Asia | Latn   | bgr-13l| 98.1 | 60.9 |
| mad    | 140K | Madurese                       | 7M     | Asia | Latn   | ade-4l | 99.8 | 44.7 |
| chk    | 140K | Chavacano                      | 700K   | Asia | Latn   | ebk   | 100  | -    |
| awn    | 140K | Awadhi                         | 4M     | Asia | Deva  | awa-14l| 98.1 | 61.4 |
| brx    | 130K | Bodo-Mech                      | 1M     | Asia | Deva  | ahr-13l| 100  | 32.8 |
| koi    | 130K | Komi-Permyak                   | 50K    | Europe| Cyrl  | ady-9l| 93.7 | 40.6 |
| arn    | 130K | Mapudungun                     | 300K   | Americas| Latn | arn-6l| 99.6 | 33.6 |
| nv     | 130K | Navajo                         | 200K   | Americas| Latn | nv-2l | 99.2 | 24.6 |
| nhe    | 130K | Eastern Huasteca English       | 400K   | Americas| Latn | azz-7l| 99.7 | 29.4 |
| fip    | 130K | Fipa                           | 200K   | Africa| Latn | dov-9l | 100  | 32.0 |
| dyu    | 130K | Dyula                          | 16M    | Africa| Latn   | ada-13l| 99.8 | 27.8 |
| xnr    | 130K | Kangri                         | 2M     | Asia | Deva  | bfr-15l| 96.2 | 57.7 |
| kbp    | 130K | Kabiyyé                        | 1M     | Africa| Latn  | kbp-2l | 100  | 32.1 |
| kw     | 130K | Cornish                        | 5K     | Europe| Latn  | br-4l  | 92.6 | -    |
| kri    | 130K | Krio                           | 4M     | Africa| Latn  | cjk-14l| 99.6 | 61.0 |
| mn     | 130K | Rundi                          | 9M     | Africa| Latn  | cgg-15l| 97.5 | -    |
| mdf    | 120K | Moksha                         | 5K     | Europe| Cyrl  | ady-9l | 84.8 | -    |
| tzh    | 120K | Tzeltal                        | 400K   | Americas| Latn | arn-6l| 99.4 | 30.9 |
| cab    | 120K | Garifuna                       | 200K   | Americas| Latn | bch-10l| 100  | 18.8 |
| trp    | 120K | Kok Borok                      | 800K   | Asia | Latn   | ace-12l| 99.8 | 38.3 |
| mkn    | 120K | Kupang Malay                   | 200K   | Asia | Latn   | abs-13l| 98.6 | 53.7 |
| rki    | 120K | Rakhine                        | 2M     | Asia | Mynr  | rki   | 99.9 | 49.2 |
| gyn    | 120K | Guyanese Creole English        | 600K   | Americas| Latn | brj-9l| 99.4 | 71.8 |
| syl-Latn| 120K | Sylheti                        | 23M    | Asia | Latn   | bn-Latn-6l| 99.1 | -    |
| max    | 120K | North Moluccan Malay           | 700K   | Asia | Latn   | abs-13l| 91.3 | -    |
| izz    | 120K | Izi                            | 500K   | Africa| Latn  | anw-8l | 99.7 | 26.6 |
| dje    | 110K | Zarma-Kaado                    | 2M     | Africa| Latn  | ang-7l | 99.6 | 22.2 |
| smt    | 110K | Simte                          | 50K    | Asia | Latn  | bgs-9l | 99.9 | -    |
| ii     | 110K | Inukitut                       | 50K    | Americas| Cans | crn-3l| 99.8 | 26.9 |
| bmm    | 110K | Northern Beisimisarakam Malagasy|      | Asia | Africa| Latn | apy-7l| 91.8 | -    |
| lhu    | 110K | Lahu                           | 500K   | Asia | Latn  | hni-8l | 99.9 | 29.3 |
| dkj    | 110K | Aukan                          | 50K    | Americas| Latn | djk-3l | 100  | 35.3 |
| cuk    | 110K | San Blas Kuna                  | 50K    | Americas| Latn | bnj-6l| 100  | 12.3 |
| mni    | 110K | Manipuri                       | 2M     | Asia | Beng  | mni   | 99.9 | 37.9 |
| mak    | 100K | Makasar                        | 5M     | Asia | Latn  | mak   | 100  | 19.2 |
| BCP-47 | Mono | Language Name | Speak. | Cont. | Script | Clust. | F1 | RTT |
|--------|------|---------------|--------|-------|--------|--------|----|-----|
| bto    | 100K | Igira Bicolano | 200K   | Asia  | Latn   | agn-14l| 95.2| 64.6|
| tab    | 100K | Tabasaran     | 100K   | Europe| Cyril  | ce-3l  | 100 | 33.4|
| mps    | 100K | Dadibi        | 50K    | Oceania| Latn   | mps-3l | 99.8| -   |
| mbt    | 100K | Matigsalug Manobo | 50K | Asia | Latn | mbt | 100 | 21.5|
| gui    | 100K | Eastern Bolivian Guaraní | 50K | Americas | Latn | gui | 100 | 20.6|
| fan    | 100K | Fang (Equatorial Guinea) | 1M | Africa | Latn | bas-11l | 99.1 | 31.0|
| ngu    | 100K | Central Guererro Nahuatl | 200K | Americas | Latn | azz-7l | 99.7 | 28.3|
| srm    | 100K | Saramaccan   | 50K    | Americas | Latn | djk-3l | 100 | -   |
| ms-Arab-BN | 100K | Malay | 80M | Asia | Arab | mfa-3l | 92.7 | 53.9|
| ndc-ZW | 100K | Ndua       | 2M     | Africa | Latn | cjk-14l | 99.0 | 60.5|
| nmb    | 100K | Nande       | 900K   | Africa | Latn | cgg-15l | 99.7 | 35.7|
| ber    | 100K | Berber      | 2M     | Africa | Tiflag | ber | 100 | 26.5|
| ibb    | 100K | Ibibio      | 2M     | Africa | Latn | answ-8l | 98.6 | 48.3|
| ntu    | 90K | Nyungwe     | 400K   | Africa | Latn | cjk-14l | 98.0 | -   |
| gub    | 90K | Guajajáwe   | 50K    | Americas | Latn | abr-5l | 100 | -   |
| kmb    | 90K | Kimbundu   | 4M     | Africa | Latn | dov-9l | 100 | 25.0|
| njj    | 90K | Torvaldi   | 50K    | Asia | Arab | bal-13l | 88.4 | 43.8|
| cgg    | 90K | Chiga       | 2M     | Africa | Latn | cgg-15l | 99.3 | 50.5|
| dhw    | 90K | Termamu    | 50K    | Asia | Arab | bhw-9l | 100 | 19.9|
| kn-Latn | 90K | Kannada    | 43M    | Asia | Latn | bpd-13l | 98.6 | 46.0|
| ms-Arab | 90K | Malay       | 80M    | Asia | Arab | mfa-3l | 88.9 | 55.7|
| oj     | 90K | Ojibwe     | 50K    | Americas | Latn | agr-3l | 99.9 | -   |
| acf    | 90K | Saint Lucian Creole French | 200K | Americas | Latn | acf-10l | 100 | 48.9|
| cac    | 90K | Chuj        | 50K    | Americas | Latn | cac | 100 | 13.8|
| jiv    | 90K | Shuar       | 50K    | Americas | Latn | bs-Latn-6l | 100 | 12.0|
| alz    | 90K | Alur        | 2M     | Africa | Latn | adh-5l | 99.6 | 21.3|
| akb    | 90K | Batak Angkola | 800K | Asia | Latn | akb-4l | 100 | 15.4|
| ff     | 90K | Fula       | 12M    | Africa | Latn | aa-13l | 90.6 | 40.6|
| aa     | 80K | Afar        | 2M     | Africa | Latn | aa-13l | 99.7 | 19.8|
| miq    | 80K | Miskito     | 200K   | Americas | Latn | miq-2l | 100 | 19.1|
| bgz    | 80K | Banggai    | 100K   | Asia | Latn | amo-7l | 92.7 | 56.2|
| nyn    | 80K | Nyankole   | 2M     | Africa | Latn | cgg-15l | 99.7 | 37.7|
| adh    | 80K | Adhola     | 400K   | Africa | Latn | adh-5l | 100 | 18.5|
| bhz    | 80K | Belize Kriol English | 200K | Americas | Latn | bzh-9l | 100 | 57.9|
| nyo    | 80K | Nyoro      | 700K   | Africa | Latn | cgg-15l | 99.3 | 46.7|
| kfy    | 80K | Kumaoni    | 2M     | Asia | Devanagari | bfr-15l | 93.8 | 50.7|
| kjb    | 80K | Q’anjob’al | 50K    | Americas | Latn | cbr-6l | 96.7 | 14.3|
| frp    | 80K | Arptian    | 100K   | Europe | Latn | acf-10l | 90.4 | -   |
| jvn    | 80K | Caribbean Japanese | 50K | Americas | Latn | jvn | 100 | 26.9|
| mdy    | 80K | Male (Ethiopia) | 100K | Africa | Ethi | mdy | 100 | 22.7|
| hae    | 80K | Eastern Oromo | 4M | Africa | Latn | amf-6l | 99.5 | 48.6|
| qxr    | 80K | Cañar Highland Quichua | 200K | Americas | Latn | qw-5-6l | 99.7 | 37.7|
| noa    | 80K | Woun Meu   | 50K    | Americas | Latn | cbv-6l | 99.9 | -   |
| krt    | 70K | Kinaray-a  | 400K   | Asia | Latn | abx-11l | 99.5 | 58.9|
| hui    | 70K | Huli       | 200K   | Oceania | Latn | hui-3l | 99.9 | 12.0|
| enq    | 70K | Enga       | 200K   | Oceania | Latn | avl-12l | 99.5 | 12.1|
| sgj    | 70K | Surugia    | 1M     | Asia | Devanagari | bho-13l | 97.5 | 56.8|
| lrl    | 70K | Larestani  | 200K   | Asia | Arab | fay-14l | 56.8 | -   |
| gos    | 70K | Gofa       | 400K   | Africa | Ethi | drs-Latn-8l | 99.6 | 24.1|
| snx    | 70K | Sangir     | 200K   | Asia | Latn | snx | 100 | 31.9|
| hus    | 70K | Huastec    | 100K   | Americas | Latn | chd-6l | 100 | 19.6|
| tsc    | 70K | Mutu       | 5K     | Oceania | Latn | bmu-6l | 100 | -   |
| maj    | 70K | Central Mazahua | 50K | Americas | Latn | jic-3l | 100 | 15.7|
| hix    | 70K | Hixkaryána | 5K     | Americas | Latn | hix | 99.6 | -   |
| Code | Mono | Language Name | Speak. | Cont. | Script | Clust. | F1  | RTT |
|------|------|---------------|--------|------|--------|--------|-----|-----|
| ks   | 70K  | Kashmiri     | 6M     | Asia | Arab   | don-Arab-5l | 96.5 | 47.0 |
| clu  | 70K  | Caluyanun    | 50K    | Asia | Latn   | abh-1l   | 98.9 | -    |
| aex  | 70K  | Omani Arabic | 1M    | Europe | Arab | abv-10l | 4.9 | -    |
| prk  | 70K  | South Wa     | 800K  | Asia | Latn | ace-12l | 100 | 26.3 |
| mas  | 70K  | Masai        | 1M    | Africa | Latn | big-5l | 99.8 | 16.2 |
| pis  | 70K  | Pijin        | 400K  | Oceania | Latn | bi-6l | 100 | 21.3 |
| amu  | 70K  | Guerrero Amuzgo | 50K | Americas | Latn | amu-10l | 100 | 12.0 |
| myu  | 70K  | Mundurukú    | 5K    | Americas | Latn | bch-10l | 99.9 | -    |
| kig  | 70K  | Khmu         | 700K  | Asia | Laao | kgs-2l | 100 | 37.1 |
| sda  | 70K  | Toraja-Sa’dan | 800K | Asia | Latn | jke-6l | 98.4 | 20.1 |
| kpg  | 70K  | Kapingamarangi | 5K | Oceania | Latn | kpg-1l | 100 | -    |
| bus  | 70K  | Bokobaru     | 50K   | Africa | Latn | akp-4l | 99.5 | 16.5 |
| mag  | 70K  | Magahi       | 28M   | Asia | Deva | bho-13l | 99.6 | 57.7 |
| dwr  | 70K  | Dawro        | 500K  | Africa | Latn | dvs-Latin-8l | 100 | 19.8 |
| sjv  | 70K  | Epēna        | 5K    | Americas | Latn | mmp-1l | 99.1 | -    |
| ahk  | 60K  | Akha         | 600K  | Asia | Latn | akh-1l | 100 | 26.2 |
| ffm  | 60K  | Maasina Fulfulde | 1M | Africa | Latn | aa-13l | 98.6 | 29.0 |
| xav  | 60K  | Xavante      | 5K    | Americas | Latn | xav-1l | 87.5 | -    |
| chr  | 60K  | Cherokee      | 50K   | Americas | Cher | chr-1l | 100 | 19.7 |
| kyz  | 60K  | Kabáyí       | 5K    | Americas | Latn | avt-1l | 71.1 | -    |
| tll  | 60K  | Tetela       | 800K  | Africa | Latn | kj-5l | 99.9 | 24.8 |
| tee  | 60K  | Hueheueta Tepehua | 5K | Americas | Latn | tee-2l | 99.9 | -    |
| niq  | 60K  | Nandi        | 1M    | Africa | Latn | awk-4l | 98.8 | 20.0 |
| din  | 60K  | Dinka        | 1M    | Africa | Latn | dm-3l | 99.7 | 27.6 |
| nog  | 60K  | Nogai        | 50K   | Europe | Cyril | bcc-Latin-12l | 100 | 51.6 |
| tca  | 60K  | Ticuna       | 50K   | Americas | Latn | tca-1l | 100 | 8.6  |
| abt  | 60K  | Ambulas      | 50K   | Oceania | Latn | abt-3l | 99.8 | 10.3 |
| bbw  | 60K  | Biak         | 50K   | Asia | Latn | bbw-4l | 100 | 20.5 |
| dln  | 60K  | Darlong      | 5K    | Asia | Latn | bla-7l | 99.9 | -    |
| xog  | 60K  | Soga         | 3M    | Africa | Latn | cgg-15l | 99.4 | 56.6 |
| cht  | 60K  | Shawai       | 5K    | Americas | Latn | cbt-2l | 100 | -    |
| vep  | 60K  | Veps         | 5K    | Europe | Cyril | gaw-8l | 90.9 | -    |
| gil  | 60K  | Gilbertese   | 100K  | Oceania | Latn | aaz-6l | 100 | 14.3 |
| tyz  | 60K  | Tây          | 3M    | Asia | Latn | bll-Latin-4l | 99.5 | 22.8 |
| bqc  | 60K  | Boko (Benin) | 200K  | Africa | Latn | akp-4l | 99.8 | 19.0 |
| ksd  | 60K  | Kuanua       | 50K   | Oceania | Latn | apb-13l | 99.2 | 15.4 |
| tob  | 60K  | Toba         | 50K   | Americas | Latn | hni-8l | 100 | 13.3 |
| bru  | 60K  | Eastern Bru  | 50K   | Asia | Latn | bru-6l | 100 | 21.4 |
| sat-Latn | 60K | Santali   | 6M    | Asia | Latn | aar-6l | 98.9 | 40.0 |
| msy  | 60K  | Araamu       | 5K    | Oceania | Latn | msy-2l | 99.6 | -    |
| cbu  | 60K  | Candoshi-Shapra | 5K | Americas | Latn | apu-13l | 99.1 | -    |
| shp  | 60K  | Shipibo-Conibo | 50K | Americas | Latn | apu-13l | 100 | -    |
| qub  | 60K  | Huallaga Huánuco Quechua | 50K | Americas | Latn | queb-7l | 99.5 | 26.0 |
| agr  | 60K  | Aguaruna     | 50K   | Americas | Latn | agr-3l | 100 | 14.6 |
| emp  | 60K  | Northern Emberá | 50K | Americas | Latn | cbv-6l | 99.6 | 12.3 |
| suz  | 60K  | Sunwar       | 50K   | Asia | Deva | suz-1l | 100 | 17.2 |
| maa  | 60K  | San Jerónimo Tecóatl Mazatec | 50K | Americas | Latn | kcg-9l | 100 | -    |
| ifk  | 60K  | Tuvali Ifugao | 50K | Asia | Latn | bnc-8l | 100 | 19.9 |
| arc  | 50K  | Official Aramaic (700-300 BCE) | 0K | Europe | Syrc | ach-10l | 80.0 | -    |
| bim  | 50K  | Bimoba       | 400K  | Africa | Latn | bim-3l | 100 | 15.6 |
| oiq  | 50K  | Querétaro Otomi | 50K | Americas | Latn | amz-3l | 99.9 | 47.0 |
| tdx  | 50K  | Tandroy Malagasy | 1M | Africa | Latn | apy-7l | 99.6 | 48.2 |
| tay  | 50K  | Tayuyu       | 5K    | Americas | Latn | des-4l | 100 | -    |
| pir  | 50K  | Wá’ikhana    | 5K    | Americas | Latn | des-4l | 99.6 | -    |
| ded  | 50K  | Dedua        | 5K    | Oceania | Latn | bru-6l | 99.7 | -    |
| txu  | 50K  | Kayapó       | 5K    | Americas | Latn | apu-3l | 99.2 | -    |
| qvh  | 50K  | Huamalíes-Dos de Mayo Huánuco Quechua | 50K | Americas | Latn | queb-7l | 100 | 19.7 |
| BCP-47 | Mono | Language Name | Speak. | Cont. | Script | Clust. | F1 | RTT |
|--------|------|---------------|--------|-------|--------|--------|----|-----|
| mav    | 50K  | Sateré-Mawé   | 5K     | Americas | Latn   | bwa-4l | 100 |     |
| ubu    | 50K  | Umbu-Ungu     | 5K     | Oceania | Latn   | gbi-8l | 99.9 | 11.8|
| toj    | 50K  | Triojolabal    | 5K     | Americas | Latn   | toj-4l | 100 | 27.2|
| inb    | 50K  | Inga          | 5K     | Americas | Latn   | inb   | 100 |     |
| aoj    | 50K  | Muñian        | 5K     | Oceania | Latn   | aoj   | 99.6 |     |
| mbb    | 50K  | Western Bukidnon | Manobo | Asia | Latn | mbb-3l | 85.9 |     |
| acr    | 50K  | Achi          | 5K     | Americas | Latn   | acr-4l | 99.9 | 21.1|
| xtd    | 50K  | Mixtec        | 5K     | Americas | Latn   | mprn-4l | 99.9 |     |
| kjh    | 50K  | Khakas        | 5K     | Europe | Cyril  | kjh   | 100 | 37.9|
| dyi    | 50K  | Djimini Senoufo | Africa | Latn | dhi    | 100 | 18.5|
| mpm    | 50K  | Yosondúa Mixtec | Americas | Latn | mprn-4l | 99.7 |     |
| kwi    | 50K  | Awa-Cuaíquer  | 5K     | Americas | Latn   | bla-7l | 100 |     |
| bjj    | 50K  | Kanaúji      | 10M    | Asia    | Deva   | awa-14l | 97.8 | 55.4|
| ann    | 50K  | Amanab      | 5K     | Oceania | Latn   | ann-2l | 99.9 |     |
| plu    | 50K  | Palikúr      | 5K     | Americas | Latn   | plu   | 98.7 |     |
| ood    | 50K  | Tobono O’odham | Americas | Latn | aq-4l | 100 |     |
| czt    | 50K  | Zotung Chin  | 5K     | Asia    | Latn   | czt   | 100 | 19.9|
| tbg    | 50K  | North Taíraora | 5K     | Oceania | Latn   | amnp-12l | 100 |     |
| gik    | 50K  | Kachi Koli   | 600K   | Asia    | Arab   | bal-13l | 77.3 | 59.6|
| crn    | 50K  | El Nayar Cora | 5K     | Americas | Latn   | cok-2l | 99.7 |     |
| cni    | 50K  | Asháninka    | 5K     | Americas | Latn   | apo-13l | 100 | 13.0|
| qup    | 50K  | Southern Pastaza Quichua | 1M | Americas | Latn | qup-2l | 99.7 |     |
| ify    | 50K  | Keley-i Kallahan | 5K | Asia | Latn | ify | 99.9 |     |
| sqq    | 50K  | Sirionó      | 5K     | Americas | Latn   | sqq-3l | 98.9 |     |
| tbz    | 50K  | Ditungamari  | 200K   | Africa  | Latn   | gev-4l | 100 | 16.1|
| aeb    | 50K  | Tunisian Arabic | 11M | Africa | Arab | abv-10l | 13.1 |     |
| sus    | 50K  | Susu         | 1M     | Africa  | Latn   | bas-11l | 100 | 20.5|
| itv    | 50K  | Itawit       | 100K   | Asia    | Latn   | bnj-6l | 99.9 | 28.9|
| cce    | 50K  | Chopi        | 800K   | Africa  | Latn   | cce-8l | 99.4 | 33.3|
| qwh    | 50K  | Huaylas Ancash Quichua | 300K | Americas | Latn | nhg-3l | 100 | 25.4|
| teo    | 50K  | Teso         | 2M     | Africa  | Latn   | blaj-7l | 100 | 14.4|
| lon    | 50K  | Malawi Lomwe | 2M     | Africa  | Latn   | cjk-14l | 99.5 |     |
| gun    | 50K  | Mbyá Guaraní | 50K    | Americas | Latn   | gn-6l  | 99.4 |     |
| ppp    | 50K  | Uma          | 50K    | Asia    | Latn   | pmnf-2l | 100 |     |
| ann    | 50K  | Obolo        | 200K   | Africa  | Latn   | ann-3l | 100 | 26.3|
| laj    | 50K  | Lango (Uganda) | 2M | Africa | Latn | adh-5l | 99.7 | 25.7|
| eza    | 50K  | Ezāa         | 600K   | Africa  | Latn   | anw-8l | 99.7 | 29.2|
| caf    | 50K  | Southern Carrier | 5K | Americas | Latn | bpr-7l | 99.9 |     |
| lex    | 50K  | Luang        | 50K    | Asia    | Latn   | kje-6l | 100 |     |
| msn    | 50K  | Agusan Manobo | 50K | Asia | Latn | ait-4l | 100 | 24.6|
| tuf    | 50K  | Central Tunebo | 5K | Americas | Latn | tuf   | 100 |     |
| quf    | 50K  | Lambayeque   | 50K    | Americas | Latn   | aer-Latn-10f | 99.5 |     |
| bzc    | 50K  | Southern Betsimis-araka Malagasy | 2M | Africa | Latn | apy-7l | 96.4 | 61.4|
| jbo    | 50K  | Lojban       | 0K     | Europe  | Latn   | bi-6l  | 98.0 |     |
| yss    | 50K  | Yessan-Mayo  | 5K     | Oceania | Latn   | bom-5l | 99.9 |     |
| tzj    | 50K  | Tz’utujil    | 50K    | Americas | Latn   | tzj   | 100 | 13.4|
| qvc    | 40K  | Cajamarca    | 50K    | Americas | Latn   | qvc   | 100 | 16.4|
| stp    | 40K  | Southeastern Tepehuan | 50K | Americas | Latn | cbr-6l | 100 |     |
| wsk    | 40K  | Waskia       | 50K    | Oceania | Latn   | aom-10f | 97.7 |     |
| knj    | 40K  | Akateko      | 50K    | Americas | Latn   | knj-8l | 99.3 | 15.8|
| qvz    | 40K  | Northern Pastaza Quichua | 50K | Americas | Latn | quv-5l | 100 |     |
| pad    | 40K  | Paumari      | 5K     | Americas | Latn   | anw-8l | 100 |     |
| cbi    | 40K  | Cha’palaa    | 5K     | Americas | Latn   | cbi   | 100 |     |
| tac    | 40K  | Lowland Tarahumara | 50K | Americas | Latn | alp-2l | 99.9 |     |
| BCP-47 | Mono | Language Name          | Speak. | Cont.     | Script                   | Clust. | F1  | RTT |
|--------|------|------------------------|--------|-----------|--------------------------|--------|-----|-----|
| ngl    | 40K  | Mozambique             | 2M     | Africa    | lat                        | cjk-14l| 96.2| -   |
| kaq    | 40K  | Capanaíha              | 50K    | Americas  | lat                        | bgs-9l | 100 | -   |
| jum    | 40K  | Jamaican Creole        | 3M     | Americas  | lat                        | bjt-9l | 99.9| 57.9|
| cpy    | 40K  | South Ucayali Ashéninka| 5K     | Americas  | lat                        | apu-13l| 99.2| 15.0|
| en-Cyr | 40K  | English                | 550M   | Europe    | Cyril                      | az-RU-13l| 98.0| 87.3|
| car    | 40K  | Galibi Carib           | 5K     | Americas  | lat                        | car    | 100 | -   |
| gju    | 40K  | Gujarí                 | 1M     | Asia      | Arab                      | doi-Arab-5l| 98.9| 51.3|
| pab    | 40K  | Parecis                | 5K     | Americas  | lat                        | pab-2l | 99.9| -   |
| yuj    | 40K  | Karkar-Yuri            | 5K     | Oceanía   | lat                        | amp-12l| 98.0| -   |
| jac    | 40K  | Poptí                 | 5K     | Americas  | lat                        | ayo-4l | 98.1| -   |
| mjc    | 40K  | San Juan Colorado      | 5K     | Americas  | lat                        | cta-4l | 99.6| -   |
| med    | 40K  | Melpa                  | 100K   | Oceania   | lat                        | cqvi-4l| 100 | 13.6|
| qvs    | 40K  | San Martín             | 50K    | Americas  | lat                        | qyu-2l | 99.7| -   |
| urb    | 40K  | Urubí-Kaapor           | 5K     | Americas  | lat                        | bww-9l | 99.7| -   |
| bnp    | 40K  | Bola                   | 50K    | Oceania   | lat                        | bch-10l| 99.0| -   |
| yle    | 40K  | Yele                   | 5K     | Oceania   | lat                        | bbb-2l | 100 | -   |
| acu    | 40K  | Achuar-Shiwiar         | 5K     | Americas  | lat                        | acu-8l | 99.1| -   |
| cof    | 40K  | Tsafiki                | 5K     | Americas  | lat                        | bmr-3l | 100 | -   |
| poh    | 40K  | Poqomchi’              | 50K    | Americas  | lat                        | hni-8l | 100 | 12.6|
| gbi    | 40K  | Galela                 | 50K    | Asia      | lat                        | gbi-8l | 100 | 13.6|
| kto    | 40K  | Kuot                   | 5K     | Oceania   | lat                        | bkv-4l | 99.9| -   |
| bfr    | 40K  | Mahasu Pahari          | 1M     | Asia      | Deva                      | bfr-15l| 94.4| -   |
| cao    | 40K  | Chácobo                | 5K     | Americas  | lat                        | auc-6l | 99.7| -   |
| myy    | 40K  | Macuna                 | 5K     | Americas  | lat                        | cub-2l | 100 | -   |
| azz    | 40K  | Amarasi                | 50K    | Asia      | lat                        | aaz-6l | 100 | 15.3|
| nch    | 40K  | Central Huasteca       | 200K   | Americas  | lat                        | aaz-7l | 95.7| 19.2|
| ixl    | 40K  | Ixil                   | 100K   | Americas  | lat                        | cip-4l | 100 | 15.1|
| rwo    | 40K  | Rawa                   | 50K    | Oceania   | lat                        | gbi-8l | 99.9| -   |
| azg    | 40K  | San Pedro Amuzgos Amuzgo| 50K  | Americas  | lat                        | aaz-4l | 100| -   |
| yaq    | 40K  | Yaqui                  | 50K    | Americas  | lat                        | yaq    | 99.6| -   |
| iow    | 40K  | Tuma-Irumu             | 5K     | Oceanía   | lat                        | awx-2l | 97.0| -   |
| qvw    | 40K  | Huaylla Wanca          | 300K   | Americas  | lat                        | qub-7l | 100 | 16.7|
| snn    | 40K  | Siona-Tetete           | 5K     | Americas  | lat                        | azg-4l | 100| -   |
| sxu    | 40K  | Central East Middle German| 2M   | Europe    | lat                        | ang-7l | 99.9| 64.3|
| bch    | 40K  | Barrai                 | 5K     | Oceania   | lat                        | bch-10l| 98.8| -   |
| bud    | 40K  | Ntcham                 | 300K   | Africa    | lat                        | amf-6l | 100 | 17.2|
| tih    | 40K  | Timugon Murut          | 5K     | Asia      | lat                        | thh-2l | 99.8| -   |
| mcd    | 40K  | Sharanaíha             | 5K     | Americas  | lat                        | ccb-2l | 100 | -   |
| apn    | 40K  | Apinayé                | 5K     | Americas  | lat                        | apn-3l | 91.4| -   |
| yaa    | 40K  | Yaminahua              | 5K     | Americas  | lat                        | yaa    | 100 | -   |
| mcy    | 40K  | Manggarai              | 900K   | Asia      | lat                        | abs-13l| 97.8| 48.9|
| yka    | 40K  | Yakan                  | 100K   | Asia      | lat                        | bpr-7l | 99.8| 23.6|
| rmc    | 40K  | Central Romani         | 100K   | Europe    | lat                        | rmc-3l | 100 | 25.7|
| tfr    | 40K  | Teribe                 | 5K     | Americas  | lat                        | psm-2l | 100 | -   |
| tna    | 40K  | Tacana                 | 5K     | Americas  | lat                        | nhe-3l | 100 | -   |
| nhy    | 40K  | Northern Oaxaca        | 50K    | Americas  | lat                        | nhy-3l | 99.7| -   |
| nfa    | 40K  | Dhaó                   | 5K     | Asia      | lat                        | bww-9l | 99.1| -   |
| zos    | 40K  | Francisco León         | 50K    | Americas  | lat                        | ayo-4l | 98.4| -   |
| zpx    | 40K  | Texmelucan             | 5K     | Americas  | lat                        | srq-3l | 100| -   |
| mjq    | 40K  | Mamasa                 | 50K    | Asia      | lat                        | kje-6l | 97.9| 14.5|
| xsi    | 40K  | Sio                    | 5K     | Oceanía   | lat                        | kji-2l | 99.9| -   |
| srr    | 40K  | Serere                 | 1M     | Africa    | lat                        | aa-13l | 81.6| 21.4|
| BCP-47 | Mono | Language Name       | Speak. | Cont.  | Script | Clust. | F1   | RTT  |
|--------|------|---------------------|--------|--------|--------|--------|------|------|
| lac    | 40K  | Gurani              | 40K    | Asia   | Arab   | fay-14l| 68.6 | 52.1 |
| yal    | 40K  | Yalunka             | 100K   | Africa  | Latn   | bas-11l| 100  | 16.4 |
| aey    | 40K  | Amele               | 5K     | Oceania | Latn   | aey-6l | 97.3 | -    |
| lep-Tibt | 40K | Lepcha              | 50K    | Asia   | Tibt   | lep-Tibt| 100  | 9.9  |
| kia    | 40K  | Kim                 | 50K    | Africa  | Latn   | cut-5l | 99.9 | -    |
| sef    | 40K  | Cebaara Senoufo     | 3M     | Africa  | Latn   | ada-13l| 98.7 | 41.1 |
| lac    | 40K  | Lacandon            | 5K     | Americas| Latn   | lac   | 99.8 | -    |
| ro     | 40K  | Waima               | 50K    | Oceania | Latn   | apb-13l| 97.7 | -    |
| cav    | 40K  | Cavineña            | 5K     | Americas| Latn   | aak-3l| 99.9 | -    |
| bvf    | 40K  | Bauzi               | 5K     | Asia   | Latn   | adz-3l | 100  | -    |
| qxf    | 40K  | Panao Huánuco       | 50K    | Americas| Latn   | quh-7l| 99.5 | 17.3 |
| mcf    | 40K  | Matsés              | 5K     | Americas| Latn   | con-3l| 100  | -    |
| gvl    | 40K  | Guayabero           | 200K   | Africa  | Latn   | dgr-5l | 92.6 | -    |
| alq    | 40K  | Algonquin           | 5K     | Americas| Latn   | alq-4l| 99.9 | -    |
| gun    | 40K  | Guambiano           | 50K    | Americas| Latn   | cjr-4l| 100  | -    |
| cr-Latn|      | Cooee               | 200K   | Americas| Latn   | ach-10l| 24.2 | -    |
| guh    | 40K  | Guahibo             | 50K    | Americas| Latn   | aca-8l| 100  | 10.8 |
| sri    | 40K  | Siriano             | 5K     | Americas| Latn   | cbv-6l| 99.2 | -    |
| kvn    | 40K  | Border Kuna         | 50K    | Americas| Latn   | guo-2l| 100  | -    |
| acn    | 40K  | Longchuan           | 50K    | Asia   | Latn   | acn-11l| 100  | 14.4 |
| mxb    | 40K  | Tezoatlán Mixtec    | 50K    | Americas| Latn   | kcg-9l| 98.9 | -    |
| ppo    | 40K  | Polopa              | 5K     | Oceania | Latn   | pab-2l| 98.2 | -    |
| atd    | 40K  | Ata Manobo          | 50K    | Asia   | Latn   | atd-4l| 70.9 | -    |
| zam    | 40K  | Cuixe-Xitla Zapotec | 5K     | Americas| Latn   | sqx-3l| 100  | -    |
| mti    | 40K  | Matwa (Papua New Guinea) | 5K | Oceania | Latn | dgr-5l | 99.8 | - |
| jic    | 40K  | Tol                 | 5K     | Americas| Latn   | jic-3l| 100  | -    |
| cui    | 40K  | Cuiba               | 5K     | Americas| Latn   | aca-8l| 100  | -    |
| ata    | 40K  | Pele-Ata            | 5K     | Oceania | Latn   | ata-5l| 98.8 | -    |
| enk    | 40K  | Khumi Chin          | 50K    | Asia   | Latn   | cdo-5l| 100  | 16.9 |
| bnc    | 40K  | Bontok              | 50K    | Asia   | Latn   | bnc-8l| 99.6 | 15.5 |
| con    | 40K  | Cofán               | 5K     | Americas| Latn   | con-3l| 100  | -    |
| guo    | 40K  | Guayabero           | 5K     | Americas| Latn   | guo-2l| 100  | -    |
| cca    | 40K  | Chortí              | 50K    | Americas| Latn   | cca   | 100  | 12.5 |
| toh    | 30K  | Gitonga             | 400K   | Africa  | Latn   | cce-8l| 99.9 | 29.9 |
| xon    | 30K  | Konkomba            | 900K   | Africa  | Latn   | daa-3l| 98.6 | 17.3 |
| rug    | 30K  | Roviana             | 5K     | Oceania | Latn   | akh-4l| 99.9 | -    |
| bbb    | 30K  | Barai               | 5K     | Oceania | Latn   | asg-10l| 99.6 | -    |
| chf    | 30K  | Tabasco Chontal     | 50K    | Americas| Latn   | chd-6l| 100  | 20.3 |
| cub    | 30K  | Cuboe               | 5K     | Americas| Latn   | cub-2l| 100  | -    |
| tke    | 30K  | Takwane             | 2M     | Africa  | Latn   | tke   | 97.3 | -    |
| ape    | 30K  | Bukiyip             | 50K    | Oceania | Latn   | aon-13l| 69.9 | -    |
| tuo    | 30K  | Tucano              | 5K     | Americas| Latn   | meh-3l| 99.3 | -    |
| npl    | 30K  | Nahuatl, South-eastern Puebla | 300K | Americas | Latn | nhy-3l | 94.5 | - |
| mmx    | 30K  | Madak               | 5K     | Oceania | Latn   | avt-12l| 98.4 | -    |
| ifa    | 30K  | Angamad Ifugao      | 50K    | Asia   | Latn   | bnc-8l| 99.7 | -    |
| boj    | 30K  | Anjam               | 5K     | Oceania | Latn   | boj-4l| 100  | -    |
| ter    | 30K  | Terena-Kiniknao-Chane | 50K | Americas | Latn | enx-4l | 100  | -    |
| mpx    | 30K  | Misima-Paneati      | 50K    | Oceania | Latn   | awi-14l| 97.5 | -    |
| msk    | 30K  | Mansaka             | 50K    | Asia   | Latn   | agn-14l| 100  | 34.2 |
| sdh    | 30K  | Southern Kurdish    | 3M     | Asia   | Arab   | fay-14l| 70.2 | -    |
| ber-Arab | 30K | Berber             | 2M     | Africa  | Arab   | abv-10l| 99.5 | 21.1 |
| zia    | 30K  | Zia                 | 5K     | Oceania | Latn   | avt-12l| 96.2 | -    |
| mna    | 30K  | Mbula               | 5K     | Oceania | Latn   | bmo-6l| 99.0 | -    |
| apy    | 30K  | Apalai              | 5K     | Americas| Latn   | apy-7l| 99.2 | -    |
| kdi    | 30K  | Kumam               | 200K   | Africa  | Latn   | adl-5l| 99.9 | 15.8 |
| aon    | 30K  | Bumbita Arapesh     | 5K     | Oceania | Latn   | aon-13l| 76.9 | -    |
| mpg    | 30K  | Marba               | 200K   | Africa  | Latn   | agt-6l| 100  | 13.1 |
| mlp    | 30K  | Bargam              | 5K     | Oceania | Latn   | hnz-81| 98.4 | -    |
| BCP-47 | Mono | Language Name | Speak. | Cont. | Script | Clust. | F1 | RTT |
|--------|------|---------------|--------|-------|--------|--------|----|-----|
| kyc    | 30K  | Kyaka         | 50K    | Oceania | Latn   | avt-121 | 99.2 | -   |
| huu    | 30K  | Murui Huitoto | 5K     | Americas | Latn   | aak-31  | 100  | -   |
| sl     | 30K  | Salt-Yai      | 5K     | Oceania | Latn   | aen-131 | 97.6 | -   |
| qwu    | 30K  | Tena Lowland  | 1M     | Americas | Latn   | qwu-51  | 99.9 | -   |
| bkq    | 30K  | Bakairí       | 5K     | Americas | Latn   | abr-51  | 99.2 | -   |
| yaf    | 30K  | Yaka-Pelende-Lonzo | 900K | Africa | Latn   | jbu-81  | 99.6 | 27.3 |
| ncl    | 30K  | Michoacán Nahuatl | 5K | Americas | Latn   | azz-71  | 98.8 | -   |
| sgb    | 30K  | Mag-Anchi Ayta | 5K     | Asia    | Latn   | kgp-41  | 100  | -   |
| rop    | 30K  | Kriol         | 5K     | Oceania | Latn   | bg-Latn-51 | 99.7 | -   |
| rkb    | 30K  | Rikbaktsa     | 5K     | Americas | Latn   | apy-71  | 99.2 | -   |
| kqc    | 30K  | Doromu-Koki   | 5K     | Oceania | Latn   | adz-31  | 99.5 | -   |
| kew    | 30K  | West Kewa     | 50K    | Oceania | Latn   | aau-141 | 95.4 | 11.5 |
| snp    | 30K  | Siane         | 50K    | Oceania | Latn   | knj-81  | 100  | -   |
| cjp    | 30K  | Cabécar       | 5K     | Americas | Latn   | cjg-41  | 100  | -   |
| viv    | 30K  | Iduna         | 5K     | Oceania | Latn   | auu-141 | 99.5 | -   |
| sim    | 30K  | Mende (Papua New Guinea) | 5K | Oceania | Latn   | apb-131 | 98.2 | -   |
| xsm    | 30K  | Kasem         | 200K   | Africa  | Latn   | lua-31  | 100  | 12.2 |
| iws    | 30K  | Sepik Iwam    | 5K     | Oceania | Latn   | avt-121 | 98.8 | -   |
| amm    | 30K  | Ama (Papua New Guinea) | 5K | Oceania | Latn   | amm-94  | 99.4 | -   |
| opm    | 30K  | Oksapmin      | 5K     | Oceania | Latn   | agg-51  | 98.8 | -   |
| kgk    | 30K  | Kiawi         | 50K    | Americas | Latn   | gn-6l   | 98.8 | -   |
| mta    | 30K  | Cotabato Manobo | 50K | Asia    | Latn   | mta-100 | 17.0 | -   |
| bef    | 30K  | Benabena      | 50K    | Oceania | Latn   | aau-121 | 99.7 | 12.2 |
| ura    | 30K  | Urarina       | 5K     | Americas | Latn   | bmr-31  | 100  | -   |
| tim    | 30K  | Timbe         | 50K    | Oceania | Latn   | bom-51  | 99.9 | -   |
| ife    | 30K  | Ife           | 100K   | Africa  | Latn   | ife-99.9 | 18.7 | -   |
| gqr    | 30K  | Gor           | 50K    | Africa  | Latn   | dgr-51  | 93.0 | -   |
| haz    | 30K  | Hazaragi      | 2M     | Asia    | Arab   | fay-141 | 13.9 | 57.8 |
| row    | 30K  | Dela-Oenale   | 5K     | Asia    | Latn   | bsw-91  | 98.6 | -   |
| lcm    | 30K  | Tungang       | 50K    | Oceania | Latn   | apb-131 | 97.0 | -   |
| wos    | 30K  | Hanga Hundi   | 5K     | Oceania | Latn   | aau-141 | 99.5 | -   |
| agn    | 30K  | Agutaynen     | 50K    | Asia    | Latn   | agn-141 | 99.9 | -   |
| lew    | 30K  | Ledo Kaili    | 400K   | Asia    | Latn   | kzw-31  | 100  | 18.0 |
| mbc    | 30K  | Macushi       | 50K    | Americas | Latn   | mbc-100 | -   | -   |
| byx    | 30K  | Qaqet         | 5K     | Oceania | Latn   | acn-111 | 97.7 | -   |
| apz    | 30K  | Safeyoka      | 5K     | Oceania | Latn   | amp-121 | 99.2 | -   |
| qxn    | 30K  | Northern Conchucos Ancash Quechua | 200K | Americas | Latn   | qub-71  | 99.1 | 18.2 |
| rgu    | 30K  | Ringgou       | 50K    | Asia    | Latn   | bsw-91  | 98.0 | -   |
| for    | 30K  | Fore          | 50K    | Oceania | Latn   | asw-41  | 99.7 | -   |
| sny    | 30K  | Saniyo-Hiyewe | 5K     | Oceania | Latn   | amp-121 | 97.1 | -   |
| mop    | 30K  | Mopán Maya    | 5K     | Americas | Latn   | chd-6l  | 100  | -   |
| men    | 30K  | Mende (Sierra Leone) | 3M | Africa | Latn    | men-100 | 20.8 | -   |
| arl    | 30K  | Arabela       | 5K     | Americas | Latn   | arl-100 | 100  | -   |
| hvn    | 30K  | Hawu          | 100K   | Asia    | Latn   | aom-101 | 100  | 15.8 |
| bba    | 30K  | Baatonum      | 600K   | Africa  | Latn   | bas-111 | 100  | 17.0 |
| des    | 30K  | Desano        | 5K     | Americas | Latn   | des-41  | 99.9 | -   |
| tue    | 30K  | Tuyuca        | 5K     | Americas | Latn   | cbc-21  | 100  | -   |
| ino    | 30K  | Inoke-Yate    | 50K    | Oceania | Latn   | ata-51  | 100  | -   |
| tod    | 30K  | Toma          | 200K   | Africa  | Latn   | moa-21  | 100  | 16.9 |
| ots    | 30K  | Estado de México Otomi | 50K | Americas | Latn   | kcg-91  | 99.9 | -   |
| bpr    | 30K  | Koromadal Blaan | 200K | Asia    | Latn   | bpr-71  | 97.3 | 13.6 |
| aau    | 30K  | Abau          | 5K     | Oceania | Latn   | aau-141 | 98.2 | -   |
| khz    | 30K  | Keapara       | 50K    | Oceania | Latn   | gab-111 | 97.6 | -   |
| bum    | 30K  | Bulu (Cameroon) | 900K | Africa | Latn   | bas-111 | 98.6 | 29.9 |
| aer    | 30K  | Eastern Arrernte | 5K | Oceania | Latn   | aer-100 | 100  | -   |
| BCP-47 | Mono | Language Name | Speak. | Cont. | Script | Clust. | F1  | RTT |
|--------|------|---------------|--------|-------|--------|--------|-----|-----|
| azz    | 30K  | Highland Puebla Nahual | 200K   | Americas | Latin | azz-7l | 99.7 | 13.5 |
| dah    | 30K  | Gwahatike       5K | Oceania | Latin | aem-13l | 97.2   | -   |
| yml    | 30K  | Jamalele        5K | Oceania | Latin | fj-4l  | 99.7   | -   |
| bzh    | 30K  | Mapos Buang     50K | Oceania | Latin | apb-13l | 97.5   | -   |
| kgp    | 30K  | Kaingang        50K | Americas | Latin | bzh-9l | 100    | -   |
| tpz    | 30K  | Tinputz         5K | Oceania | Latin | ach-10l | 99.3   | -   |
| bex    | 30K  | Jur             50K | Africa | Latin | bbo-3l | 100    | 11.0 |
| mek    | 30K  | Mekeo           50K | Oceania | Latin | fj-4l  | 99.1   | -   |
| sbl    | 30K  | Botolan         50K | Asia    | Latin | kgp-4l | 100    | 23.0 |
| uvl    | 30K  | Lote           5K | Oceania | Latin | ata-5l | 98.4   | -   |
| abx    | 30K  | Inabaknon       50K | Asia    | Latin | abx-11l | 96.4   | -   |
| ese    | 30K  | Ese Eija        5K | Americas | Latin | ese-2l | 100    | -   |
| mcb    | 30K  | Machiguenga     5K | Americas | Latin | apo-13l | 100    | -   |
| pma    | 30K  | Paama           5K | Oceania | Latin | ktm-4l | 99.4   | -   |
| qvm    | 30K  | M-Y-L Quechua   50K | Americas | Latin | qub-7l | 99.6   | 15.6 |
| mbj    | 30K  | Nadëb           5K | Americas | Latin | bgp-13l | 99.2   | -   |
| gso    | 30K  | Southwest Ghaya | 200K   | Africa | bas-11l | 100    | 16.8 |
| chd    | 30K  | Highland Oaxaca Chontal | 5K  | Americas | latin | chd-6l | 99.6   | -   |
| sxx    | 30K  | Samberigi       5K | Oceania | Latin | aui-14l | 99.9   | -   |
| zaa    | 30K  | Sierra de Juárez Zapotec | 5K  | Americas | Latin | zaa-5l | 100    | -   |
| bqp    | 30K  | Busa            50K | Africa | Latin | akp-4l | 99.7   | -   |
| kgf    | 30K  | Kulungtfu- Yuanggeng-Tobo | 5K | Oceania | Latin | kgf-4l | 99.6   | -   |
| tkv    | 30K  | Southeast Ambrym | 5K | Oceania | Latin | gaw-8l | 99.1   | -   |
| pah    | 30K  | Tenharim- Parintintin-Diahoi | 5K  | Americas | Latin | gn-6l  | 99.7   | -   |
| mog    | 30K  | Mongondow       200K | Asia    | Latin | awb-4l | 100    | 18.6 |
| zas    | 30K  | Santo Domingo Al- barradas Zapotec | 5K  | Americas | Latin | jic-3l | 100    | -   |
| zpl    | 30K  | Lachixió Zapotec | 5K | Americas | Latin | cbr-6l | 99.4   | -   |
| pib    | 30K  | Yine            5K | Americas | Latin | pib    | 71.2   | -   |
| omw    | 30K  | South Taiwora   50K | Oceania | Latin | amp-12l | 99.9   | -   |
| faa    | 30K  | Fasu            5K | Oceania | Latin | apr-8l | 99.6   | -   |
| dob    | 30K  | Dobu            100K | Oceania | Latin | aui-14l | 99.1   | 13.2 |
| dwf    | 30K  | Western Dani    200K | Asia    | Latin | bmu-6l | 100    | 13.5 |
| gor    | 30K  | Gorontalo       1M | Asia    | Latin | bco-8l | 100    | 18.4 |
| hak    | 30K  | Hakka Chinese   30M | Asia    | Latin | cdo-5l | 91.6   | 28.0 |
| poe    | 30K  | San Juan Atzingo Popoloca | 5K  | Americas | Latin | mix-4l | 99.5   | -   |
| gwi    | 30K  | Gwich’in        5K | Americas | Latin | agr-3l | 99.9   | -   |
| npy    | 30K  | Napu            5K | Asia    | Latin | gah-11l | 100    | -   |
| kpr    | 30K  | Korafe-Yegha    5K | Oceania | Latin | aom-10l | 97.6   | -   |
| vid    | 30K  | Vidunda         50K | Africa   | Latin | cwe-10l | 100    | 17.1 |
| yva    | 30K  | Yawa            50K | Asia    | Latin | wnc-3l | 100    | -   |
| llg    | 30K  | Lole            50K | Asia    | Latin | bbw-9l | 94.6   | -   |
| ssp    | 30K  | Saposa          5K | Oceania | Latin | apb-13l | 97.3   | -   |
| wmm    | 30K  | Mwani           100K | Africa  | Latin | bep-4l | 99.5   | 19.9 |
| mif    | 30K  | Mofu-Gudur      50K | Africa  | Latin | meq-3l | 100    | 16.1 |
| fuf    | 30K  | Pular           3M | Africa   | Latin | aa-13l | 99.9   | 20.9 |
| kbg    | 30K  | Kamano          50K | Oceania | Latin | aar-14l | 98.8   | 11.8 |
| fuv    | 30K  | Nigerian Fulfulde 12M | Africa  | Latin | aa-13l | 98.1   | -   |
| ptu    | 30K  | Bambam          50K | Asia    | Latin | atq-2l | 99.7   | -   |
| pri    | 30K  | Paiçã           5K | Oceania | Latin | ae-Latn-10l | 97.2 | -   |
| en-Deva| 30K  | English         550M | Europe  | Latin | awa-14l | 97.1   | 75.9 |
| ymm    | 30K  | Maay            3M | Africa   | Latin | aa-13l | 99.8   | -   |
| bnh    | 30K  | Kein            5K | Oceania | Latin | acm-11l | 98.1   | -   |
| zpo    | 30K  | Amatlán Zapotec | 50K | Americas | Latin | zaa-5l | 88.4   | -   |
| smk    | 30K  | Bolinao         50K | Asia    | Latin | bbr-4l | 99.0   | 28.6 |
| amp    | 30K  | Alamblak        5K | Oceania | Latin | amp-12l | 97.6   | -   |
| mic    | 30K  | Mi’kmaq         5K | Americas | Latin | hni-8l | 100    | -   |
| BCP-47 | Mono | Language Name       | Speak. | Cont.   | Script       | Clust. | F1  | RTT |
|--------|------|---------------------|--------|---------|--------------|--------|-----|-----|
| gkn    | 30K  | Gõkana              | 50K    | Africa  | Latin        | bsg-Latn-12l | 100 | 22.4 |
| cbs    | 30K  | Cashinahua          | 5K     | Americas| Latin        | cbs    | 100 | -   |
| gvc    | 30K  | Kotiria             | 5K     | Americas| Latin        | des-4l | 100 | -   |
| xla    | 30K  | Kamula              | 5K     | Oceania | Latin        | ktm-4l | 99.6| -   |
| sab    | 30K  | Bugleré             | 50K    | Americas| Latin        | sab    | 100 | -   |
| spp    | 30K  | Supyire Senoufo     | 500K   | Africa  | Latin        | spp    | 100 | 12.0|
| gvn    | 30K  | Kuku-Yalanji        | 5K     | Oceania | Latin        | gvn-6l | 99.6| -   |
| trc    | 30K  | Copula Triqui       | 50K    | Americas| Latin        | cut-5l | 88.0| -   |
| thk    | 30K  | Tharakaa            | 200K   | Africa  | Latin        | thk    | 100 | 20.9|
| poy    | 30K  | Pogolo              | 200K   | Africa  | Latin        | dov-9l | 100 | 15.7|
| gej    | 30K  | Gen                 | 300K   | Africa  | Latin        | gej-4l | 100 | 24.1|
| aak    | 30K  | Ankaive             | 5K     | Oceania | Latin        | aak-3l | 100 | -   |
| ml     | 30K  | Halam               | 5K     | Asia    | Latin        | akb-4l | 100 | -   |
| kpf    | 30K  | Komba               | 50K    | Oceania | Latin        | avf-12l| 98.6| -   |
| ghs    | 30K  | Guhu-Samane         | 50K    | Oceania | Latin        | agg-5l | 99.2| -   |
| agi    | 30K  | Aja (Benin)         | 600K   | Africa  | Latin        | agi-8l | 99.8| 29.7|
| bmu    | 30K  | Burum-Mindik        | 5K     | Oceania | Latin        | bmu-6l | 97.4| -   |
| fuq    | 30K  | Central-Eastern     | 500K   | Africa  | Latin        | aa-13l | 100 | 16.3|
| gvf    | 30K  | Golin               | 50K    | Oceania | Latin        | aen-13l| 98.2| 12.6|
| cbv    | 30K  | Kakua               | 5K     | Americas| Latin        | cbv-6l | 100 | -   |
| ake    | 30K  | Akawaio             | 50K    | Americas| Latin        | ake    | 100 | -   |
| usp    | 30K  | Usupanteco          | 5K     | Americas| Latin        | acr-4l | 100 | -   |
| kpj    | 30K  | Karajá              | 5K     | Americas| Latin        | kpj-5l | 98.6| -   |
| mtp    | 30K  | Wichí Lhamtés       | 5K     | Americas| Latin        | kqg-4l | 100 | -   |
| tku    | 30K  | Upper Necaxa To-    | 5K     | Americas| Latin        | tku-2l | 99.6| -   |
| tonac  |      |                     |        |         |              |        |     |     |
| ttc    | 30K  | Wantoat             | 5K     | Oceania | Latin        | tnc-3l | 99.8| -   |
| gwv    | 30K  | Sebat Bet Gurance   | 1M     | Africa  | Latin        | gwi-8l | 99.2| 35.6|
| ppg    | 20K  | Gapapaiwa           | 5K     | Oceania | Latin        | asg-10l| 97.7| -   |
| huv    | 20K  | San Mateo del Mar   | 50K    | Americas| Latin        | huv    | 100 | -   |
| tly-Cyr-| 20K  | North-Central      | 900K   | Europe  | Cyril        | az-RU-13l| 99.8| 28.7|
| RU     |      | Tálыш                |        |         |              |        |     |     |
| kmu    | 20K  | Kanite              | 5K     | Oceania | Latin        | gah-11l| 99.9| -   |
| bhh    | 20K  | Bhili               | 5M     | Asia    | Deva         | awa-14l| 100 | 44.2|
| ljp    | 20K  | lampung Api         | 800K   | Asia    | Latin        | hui-3l | 100 | 25.3|
| wrc    | 20K  | Waris               | 5K     | Oceania | Latin        | acm-11l| 98.7| -   |
| tec    | 20K  | Tekiteko            | 5K     | Americas| Latin        | azz-7l | 100 | -   |
| myb    | 20K  | Mbay                | 50K    | Africa  | Latin        | kpj-5l | 99.1| 14.2|
| mhx    | 20K  | Maru                | 50K    | Asia    | Latin        | isu-2l | 100 | 15.9|
| knv    | 20K  | Tabo                | 5K     | Oceania | Latin        | bcc-8l | 99.9| -   |
| nbl    | 20K  | Maxakalí            | 5K     | Americas| Latin        | arm-6l | 99.3| -   |
| nhi    | 20K  | Zacatlan-           | 50K    | Americas| Latin        | atr-4l | 98.7| -   |
|        |      | Ahuacatlán-        |        |         |              |        |     |     |
|        |      | Tepetzintla-Nahuatl |        |         |              |        |     |     |
| tby    | 20K  | Tabaru              | 50K    | Asia    | Latin        | gbi-8l | 100 | -   |
| soq    | 20K  | Kanasi              | 5K     | Oceania | Latin        | acm-11l| 98.8| -   |
| agg    | 20K  | Angor               | 5K     | Oceania | Latin        | agg-5l | 99.0| -   |
| kwd    | 20K  | Kwaio               | 50K    | Oceania | Latin        | knj-8l | 99.9| -   |
| ssg    | 20K  | Seimat              | 5K     | Oceania | Latin        | gaw-8l | 96.3| -   |
| zaj    | 20K  | Zaramo              | 5K     | Africa  | Latin        | cwe-10l| 99.2| -   |
| ahr    | 20K  | Ahirani             | 2M     | Asia    | Deva         | ahr-3l | 99.1| 32.3|
| pmf    | 20K  | Pamona              | 100K   | Asia    | Latin        | pmf-2l | 100 | 14.2|
| nco    | 20K  | Coatlan Mixe        | 50K    | Americas| Latin        | dys-3l | 100 | -   |
| txq    | 20K  | Tii                 | 50K    | Asia    | Latin        | bbw-9l | 94.5| -   |
| lif    | 20K  | Limbu               | 800K   | Asia    | Deva         | lif    | 99.9| -   |
| ycn    | 20K  | Yucuna              | 5K     | Americas| Latin        | acu-8l | 100 | -   |
| kln    | 20K  | Kelenjin            | 9M     | Africa  | Latin        | cro-3l | 99.9| 12.5|
| xsu    | 20K  | Sanumá              | 5K     | Americas| Latin        | xsu    | 100 | -   |
| nus    | 20K  | Nuer                | 900K   | Africa  | Latin        | dm-3l  | 100 | 22.8|
| BCP-47 | Mono | Language Name          | Speak. | Cont. | Script   | Clust.    | F1  | RTT |
|--------|------|------------------------|--------|-------|----------|-----------|-----|-----|
| nak    | 20K  | Nakain                 | 50K    | Oceania | Latn     | bch-10l   | 98.9| -   |
| lid    | 20K  | Nyindrou               | 5K     | Oceania | Latn     | gam-8l    | 99.2| -   |
| spl    | 20K  | Selepet                | 5K     | Oceania | Latn     | gab-11l   | 99.5| -   |
| aom    | 20K  | Ómie                   | 5K     | Oceania | Latn     | aom-10l   | 99.9| -   |
| nhg    | 20K  | Tetelcingo Nahuatl     | 5K     | Americas| Latn     | nhg-3l    | 100 | -   |
| cpu    | 20K  | Pichis Ashéninka       | 50K    | Americas| Latn     | apa-13l   | 96.9| 13.3|
| muh    | 20K  | Mündü                  | 50K    | Africa  | Latn     | muh       | 100 | -   |
| nmz    | 20K  | Nawdm                  | 200K   | Africa  | Latn     | amf-6l    | 99.9| 14.7|
| taj    | 20K  | Eastern Tamang         | 300K   | Asia    | Deva     | taj-2l    | 100 | 18.3|
| nca    | 20K  | Iyo                    | 5K     | Oceania | Latn     | kgf-4l    | 99.3| -   |
| ik     | 20K  | Alaskan Inupiaq        | 5K     | Americas| Latn     | bla-7l    | 99.6| -   |
| gam    | 20K  | Kandawo                | 5K     | Oceania | Latn     | gam-3l    | 99.1| -   |
| kue    | 20K  | Kuman                  | 100K   | Oceania | Latn     | agt-6l    | 99.1| 13.1|
| bbr    | 20K  | Girawa                 | 5K     | Oceania | Latn     | acn-11l   | 96.0| -   |
| roo    | 20K  | Rotokas                | 5K     | Oceania | Latn     | amp-12l   | 91.7| -   |
| big    | 20K  | Biangai                | 50K    | Oceania | Latn     | big-5l    | 100 | -   |
| dip    | 20K  | Northeastern           | 5M     | Africa  | Latn     | din-3l    | 100 | 14.9|
| aeb-Latn | 20K | Tunisian Arabic        | 11M    | Africa  | Latn     | aeb-Latn-11l | 76.9| 15.4|
| lee    | 20K  | Lyélé                  | 100K   | Africa  | Latn     | lee       | 99.6| 12.4|
| bgs    | 20K  | Tagabawa               | 50K    | Asia    | Latn     | bgs-9l    | 92.3| -   |
| apb    | 20K  | Sa’a                   | 50K    | Oceania | Latn     | apb-13l   | 99.5| -   |
| kmh    | 20K  | Kalam                  | 50K    | Oceania | Latn     | dth-4l    | 99.5| -   |
| soy    | 20K  | Miyobe                 | 50K    | Africa  | Latn     | soy       | 100 | -   |
| gfk    | 20K  | Patpatar              | 5K     | Oceania | Latn     | acn-11l   | 98.4| -   |
| nvm    | 20K  | Namiae                 | 5K     | Oceania | Latn     | dgr-5l    | 99.1| -   |
| bgr    | 20K  | Bawm Chin              | 50K    | Asia    | Latn     | bgr-6l    | 100 | -   |
| gnn    | 20K  | Gumatj                 | 5K     | Oceania | Latn     | bch-10l   | 100 | -   |
| qxo    | 20K  | Southern Conchucos     | 900K   | Americas| Latn     | qub-7l    | 98.9| 17.1|
| cbr    | 20K  | Cashibo-Cacataibo      | 5K     | Americas| Latn     | cbr-6l    | 100 | -   |
| bhl    | 20K  | Bimin                  | 5K     | Oceania | Latn     | aey-6l    | 99.6| -   |
| amr    | 20K  | Amarakaeri             | 5K     | Americas| Latn     | amr       | 100 | -   |
| kde    | 20K  | Makonde                | 1M     | Africa  | Latn     | cwe-10l   | 98.3| 24.4|
| zad    | 20K  | Cajonos Zapotec        | 5K     | Americas| Latn     | cip-4l    | 100 | -   |
| bps    | 20K  | Sarangani Blaan        | 200K   | Asia    | Latn     | bpr-7l    | 97.4| 15.3|
| mks    | 20K  | Silacayoapan Mixtec    | 50K    | Americas| Latn     | mks       | 99.6| -   |
| muy    | 20K  | Muyang                | 50K    | Africa  | Latn     | dgr-6l    | 100 | 12.8|
| ncu    | 20K  | Chumburung             | 50K    | Africa  | Latn     | ncu       | 100 | 15.0|
| stn    | 20K  | Owa                    | 5K     | Oceania | Latn     | apr-8l    | 100 | -   |
| sd-Deva| 20K | Sindhi                 | 68M    | Asia    | Deva     | bfr-15l   | 98.3| 34.4|
| hub    | 20K  | Huambisa               | 5K     | Americas| Latn     | acu-8l    | 98.8| -   |
| cpb    | 20K  | Ucayali-Yurúa Ashéninka| 5K    | Americas| Latn     | apu-13l   | 97.0 | 14.2|
| mio    | 20K  | Pinotepa Nacional Mixtec| 50K | Americas| Latn     | cta-4l    | 100 | -   |
| chz    | 20K  | Ozumacín Chinantepec   | 5K     | Americas| Latn     | chz-9l    | 100 | -   |
| aia    | 20K  | Arosi                  | 5K     | Oceania | Latn     | aia       | 100 | -   |
| ian    | 20K  | Iatmul                 | 5K     | Oceania | Latn     | agg-5l    | 99.9| -   |
| way    | 20K  | Wayana                 | 5K     | Americas| Latn     | way       | 100 | -   |
| ndz    | 20K  | Nlogo                  | 50K    | Africa  | Latn     | beq-Latn-12l | 100 | -   |
| agd    | 20K  | Agarabi                | 50K    | Oceania | Latn     | acn-11l   | 98.5| -   |
| mcp    | 20K  | Makaa                  | 50K    | Africa  | Latn     | mcp       | 100 | 15.0|
| wbp    | 20K  | Warlpiri               | 5K     | Oceania | Latn     | gvn-6l    | 99.9| -   |
| mox    | 20K  | Molima                 | 5K     | Oceania | Latn     | au-14l    | 96.2| -   |
| tbc    | 20K  | Takia                  | 50K    | Oceania | Latn     | aan-14l   | 97.7| 12.4|
| sxb    | 20K  | Suba                   | 100K   | Africa  | Latn     | cgg-15l   | 99.9| 25.0|
| kbn    | 20K  | Iwal                   | 5K     | Oceania | Latn     | gam-3l    | 99.8| -   |
| BCP-47 | Mono | Language Name | Speak. | Cont. | Script | Clust. | F1 | RTT |
|--------|------|---------------|--------|------|--------|--------|----|-----|
| bon    | 20K  | Bine          | 5K     | Oceania | Latn | bon | 99.7 | -   |
| pls    | 20K  | San Marcos Tlalcoyolco Popoloca | 50K | Americas | Latn | knj-8l | 100 | -   |
| vmy    | 20K  | Ayauatl Mazatec | 5K | Americas | Latn | nhr-2l | 98.6 | -   |
| nop    | 20K  | Numanggang     | 5K     | Oceania | Latn | aav-14l | 98.7 | -   |
| qvn    | 20K  | North Junín Quechua | 50K | Americas | Latn | qvn | 100 | 17.1 |
| cgc    | 20K  | Kagarayanen    | 50K     | Asia | Latn | cgc-2l | 100 | 22.4 |
| kwj    | 20K  | Kwanga         | 50K     | Americas | Latn | avt-12l | 97.3 | -   |
| kcg    | 20K  | Tyap           | 100K    | Africa | Latn | kcg-9l | 100 | 13.9 |
| enx    | 20K  | Southern Lengua | 5K | Americas | Latn | enx-4l | 99.0 | -   |
| mfy    | 20K  | Mayo           | 50K     | Americas | Latn | ese-2l | 99.9 | -   |
| aby    | 20K  | Aneme Wake     | 5K     | Oceania | Latn | aby-4l | 100 | -   |
| cco    | 20K  | Comaltepec Chinantec | 50K | Americas | Latn | cco | 100 | -   |
| boa    | 20K  | Bora           | 5K     | Americas | Latn | boa | 100 | -   |
| ikw    | 20K  | Ikwere         | 200K    | Africa | Latn | aaw-8l | 99.9 | 20.1 |
| lmn    | 20K  | Lambadi        | 6M      | Asia | Knda | lmn-Knda-2l | 100 | 33.3 |
| rai    | 20K  | Ramoaaina      | 50K     | Oceania | Latn | apb-13l | 95.9 | -   |
| leu    | 20K  | Kara (Papua New Guinea) | 5K | Oceania | Latn | apb-13l | 99.1 | -   |
| sue    | 20K  | Suena          | 5K     | Oceania | Latn | avt-12l | 99.6 | -   |
| auc    | 20K  | Waorani        | 5K     | Americas | Latn | auc-6l | 100 | -   |
| kdc    | 20K  | Kutu           | 50K     | Africa | Latn | cwe-10l | 97.7 | 19.1 |
| awb    | 20K  | Awa (Papua New Guinea) | 5K | Oceania | Latn | awb-4l | 99.9 | -   |
| kpw    | 20K  | Kobon          | 50K     | Oceania | Latn | aon-13l | 99.6 | -   |
| mux    | 20K  | Bo-Ung         | 50K     | Oceania | Latn | gbi-8l | 99.5 | 11.2 |
| mqb    | 20K  | Mbuko          | 50K     | Africa | Latn | bbl-2l | 100 | -   |
| knc    | 20K  | Kankanacay     | 200K    | Asia | Latn | bnc-8l | 100 | 25.7 |
| hlz    | 20K  | Balantak       | 50K     | Asia | Latn | big-4l | 100 | 14.2 |
| vmw    | 20K  | Makhusuwa      | 3M      | Africa | Latn | apu-13l | 99.9 | 16.2 |
| tcs    | 20K  | Torres Strait Creole | 5K | Oceania | Latn | bg-Latn-5l | 99.4 | -   |
| mie    | 20K  | Ocotop Mixtec | 50K     | Americas | Latn | mie | 99.9 | -   |
| mhy    | 20K  | Ma’anyan       | 200K    | Africa | Latn | amo-7l | 99.5 | 34.7 |
| btd    | 20K  | Batak Dairi    | 1M      | Asia | Latn | aom-10l | 100 | 21.5 |
| lem    | 20K  | Nomaande       | 5K     | Africa | Latn | box-3l | 100 | -   |
| hop    | 20K  | Hopi           | 5K     | Americas | Latn | hop | 99.9 | -   |
| ken    | 20K  | Kenyang        | 50K     | Africa | Latn | ken-3l | 97.3 | -   |
| ssd    | 20K  | Siroi          | 5K     | Oceania | Latn | beh-10l | 99.1 | -   |
| kms    | 20K  | Kamasau        | 5K     | Oceania | Latn | cwe-4l | 97.4 | -   |
| yuw    | 20K  | Yau (Morobe Province) | 5K | Oceania | Latn | amp-12l | 97.6 | -   |
| agt    | 20K  | Central Cagayan Agta | 5K | Asia | Latn | agt-6l | 97.6 | -   |
| maj    | 20K  | Jalapa De Díaz Mazatec | 50K | Americas | Latn | kcg-9l | 100 | -   |
| tpm    | 20K  | Tampulma       | 50K     | Africa | Latn | sil-3l | 100 | -   |
| khs    | 20K  | Kasua          | 5K     | Oceania | Latn | aui-14l | 99.9 | -   |
| kzf    | 20K  | Da’a Kaili     | 400K    | Asia | Latn | kzf-3l | 100 | 16.6 |
| ign    | 20K  | Ignaciano      | 50K     | Americas | Latn | aca-8l | 100 | 11.3 |
| mwv    | 20K  | Mentawai       | 50K     | Asia | Latn | bjp-4l | 100 | 14.7 |
| cul    | 20K  | Culina         | 5K     | Americas | Latn | cul | 100 | -   |
| apu    | 20K  | Apuriná        | 5K     | Americas | Latn | apu-13l | 99.9 | -   |
| ncj    | 20K  | Northern Puebla Nahuatl | 50K | Americas | Latn | bjr-4l | 100 | 18.6 |
| xnn    | 20K  | Northern Kankanay | 50K | Asia | Latn | bnc-8l | 99.2 | 24.4 |
| bvd    | 20K  | Buegu          | 5K     | Oceania | Latn | bvc-4l | 99.1 | -   |
| nko    | 20K  | Nkonya         | 50K     | Africa | Latn | kjt-6l | 100 | -   |
| cwe    | 20K  | Kwere          | 50K     | Africa | Latn | cwe-10l | 97.6 | 21.7 |
| tap    | 20K  | Taabwa         | 300K    | Africa | Latn | tap | 100 | 23.5 |
| BCP-47 | Mono | Language Name              | Speak. | Cont. | Script | Clust. | F1  | RTT |
|--------|------|-----------------------------|--------|-------|--------|--------|-----|-----|
| wrk    | 20K  | Garrwa                      | 5K     | Oceania | Latn   | wrk    | 100 | -   |
| gah    | 20K  | Alekano                     | 50K    | Oceania | Latn   | gah-11l | 97.6 | -   |
| yam    | 20K  | Yamba                       | 50K    | Africa  | Latn   | pry-2i  | 100 | 7.0 |
| geb    | 20K  | Kire                        | 5K     | Oceania | Latn   | bma-6l  | 100 | -   |
| ctp    | 20K  | Western Highland Chatino    | 50K    | Americas| Latn   | ctp-4l  | 83.9 | -   |
| mbs    | 20K  | Sarangani Manobo            | 50K    | Asia    | Latn   | mbb-3l  | 99.6 | 16.0|
| nlg    | 20K  | Gela                        | 50K    | Oceanía | Latn   | bgt-4l  | 99.0 | -   |
| nsn    | 20K  | Nehan                       | 5K     | Oceanía | Latn   | bgt-4l  | 99.4 | -   |
| bao    | 20K  | Waimaha                     | 5K     | América  | Latn   | bao     | 99.7 | -   |
| cle    | 20K  | Lealao Chinantec            | 5K     | América  | Latn   | chz-9l  | 100 | -   |
| mvn    | 20K  | Minaveha                    | 5K     | Oceanía | Latn   | gah-11l | 98.2 | -   |
| tpt    | 20K  | Tlachichilco Tepehua        | 5K     | América  | Latn   | gam-3l  | 100 | -   |
| zyp    | 20K  | Zyphoe                      | 50K    | Asia    | Latn   | arm-6d  | 99.9 | -   |
| pio    | 20K  | Piapoco                     | 5K     | América  | Latn   | pio     | 100 | -   |
| bgx-Latn| 20K | Rumelian Turkish            | 300K   | Europe  | Latn   | dch-Latin-11 | 95.6 | 59.5|
| myx    | 20K  | Masaaba                     | 3M     | Africa  | Latn   | cgg-15l | 99.8 | 25.0|
| tsz    | 20K  | Purepecha                   | 50K    | América  | Latn   | apa-13l | 100 | 15.2|
| dww    | 20K  | Dawawa                      | 5K     | Oceanía | Latn   | abt-3l  | 98.3 | -   |
| gnw    | 20K  | Western Bolivian Guarani    | 5K     | América  | Latn   | gnw     | 100 | -   |
| ksr    | 20K  | Borong                      | 5K     | Oceanía | Latn   | kpi-5l  | 98.5 | -   |
| otn    | 20K  | Tenango Otomi               | 50K    | América  | Latn   | kcg-9l  | 100 | 15.1|
| gdn    | 20K  | Umanakaina                  | 5K     | Oceanía | Latn   | asg-10l | 99.5 | -   |
| kmo    | 20K  | Kwoma                       | 5K     | Oceanía | Latn   | abt-3l  | 100 | -   |
| nii    | 20K  | Nii                         | 50K    | Oceanía | Latn   | nii     | 100 | -   |
| tsw    | 20K  | Salka-Tsishingini           | 200K   | Africa  | Latn   | asg-10l | 98.1 | 15.9|
| not    | 20K  | Nomatsiguenga               | 5K     | América  | Latn   | cot-2l  | 100 | -   |
| ury    | 20K  | Orya                        | 5K     | Asia    | Latn   | bzh-9l  | 100 | -   |
| anv    | 20K  | Denya                       | 50K    | América  | Latn   | anv     | 100 | -   |
| pbc    | 20K  | Patamona                    | 50K    | América  | Latn   | pbc     | 100 | -   |
| bilt-Latn| 20K | Tai Dam                     | 800K   | Asia    | Latn   | bilt-Latin-4l | 99.7 | 24.4|
| aii    | 20K  | Assyrian Neo-Aramaic        | 800K   | Europe  | Syrc   | ach-10l | 100 | 35.7|
| mhl    | 20K  | Mauwake                     | 5K     | Oceanía | Latn   | aon-13l | 65.5 | -   |
| myk    | 20K  | Mamara Senoufo              | 700K   | Africa  | Latn   | myk     | 100 | 13.2|
| mee    | 20K  | Mengen                      | 5K     | Oceanía | Latn   | ata-5l  | 99.8 | -   |
| kpx    | 20K  | Mountain Koialai            | 5K     | Oceanía | Latn   | apr-8l  | 99.1 | -   |
| udu    | 20K  | Uduk                        | 50K    | Africa  | Latn   | knj-8l  | 100 | -   |
| mnp    | 20K  | Min Bei Chinese             | 10M    | Asia    | Latn   | cdo-5l  | 99.8 | 16.2|
| alp    | 20K  | Alune                       | 50K    | Asia    | Latn   | alp-2i  | 100 | -   |
| wiu    | 20K  | Wiru                        | 50K    | Oceanía | Latn   | gbi-8l  | 99.0 | -   |
| gyr    | 20K  | Guarayu                     | 5K     | América  | Latn   | gn-6l   | 100 | -   |
| meq    | 20K  | Merey                       | 50K    | Africa  | Latn   | meq-3l  | 100 | -   |
| kjs    | 20K  | East Kewa                   | 50K    | Oceanía | Latn   | aau-14l | 95.3 | 11.4|
| mcq    | 20K  | Ese                         | 50K    | Oceanía | Latn   | acn-11l | 96.7 | -   |
| ong    | 20K  | Olo                         | 50K    | Oceanía | Latn   | kpz-2l  | 99.9 | -   |
| tnn    | 20K  | North Tanna                 | 5K     | Oceanía | Latn   | nwi-5l  | 99.9 | -   |
| coe    | 20K  | Koreguaje                   | 5K     | América  | Latn   | coe     | 99.9 | -   |
| hrx    | 20K  | Hunsrik                     | 3M     | América  | Latn   | hrx-2l  | 99.6 | 48.8|
| tte    | 20K  | Bwanabwana                  | 5K     | Oceanía | Latn   | aui-14l | 98.3 | -   |
| yut    | 20K  | Yopno                       | 5K     | Oceanía | Latn   | dbh-4l  | 99.2 | -   |
| sgz    | 20K  | Sursurunga                  | 5K     | Oceanía | Latn   | apb-13l | 97.7 | -   |
| ceg    | 20K  | Chamaroco                   | 5K     | América  | Latn   | ceg-5l  | 100 | -   |
| qul    | 20K  | North Bolivian Quechua      | 100K   | América  | Latn   | ae-Latin-10 | 99.6 | 20.9|
| mvp    | 20K  | Duri                        | 100K   | Asia    | Latn   | kje-6l  | 99.7 | 14.4|
| xed    | 20K  | Hdi                         | 50K    | Africa  | Latn   | cko-5l  | 100 | -   |
| bqj    | 20K  | Bandial                     | 50K    | Africa  | Latn   | bkv-4l  | 99.6 | -   |
| duo    | 20K  | Dupaninan Agta              | 5K     | Asia    | Latn   | dgc-6l  | 99.5 | -   |
| mza    | 20K  | Santa María Zacatepec Mixtec| 5K     | América  | Latn   | mpm-4l  | 99.6 | -   |

74
| BCP-47 | Mono | Language Name | Speak. | Cont. | Script | Clust. | F1 | RTT |
|--------|------|---------------|--------|-------|--------|--------|----|-----|
| mva    | 20K  | Manam         | 5K     | Oceania | Latn   | gal-11l | 97.7 | -   |
| akh    | 20K  | Angal Heneng  | 50K    | Oceania | Latn   | aar-14l | 97.8 | 12.0|
| blw    | 20K  | Balangao      | 50K    | Asia   | Latn   | blw-2l  | 92.5 | -   |
| luy    | 20K  | Luhya         | 1M     | Africa  | Latn   | cgg-15l | 99.7 | 20.6|
| mnb    | 20K  | Muna          | 300K   | Asia   | Latn   | ata-5l  | 100  | 13.5|
| nou    | 20K  | Ewage-Notu    | 50K    | Oceania | Latn   | aom-10l | 99.9 | -   |
| ebk    | 20K  | Eastern Bontok| 50K    | Asia   | Latn   | bnc-8l  | 97.7 | -   |
| tpp    | 20K  | Pisallores Tepehua | 5K | Americas | Latn | tee-2l | 100 | -   |
| wim    | 20K  | Wik-Mungkan   | 5K     | Oceania | Latn   | beu-2l  | 99.9 | -   |
| avt    | 20K  | Au            | 5K     | Oceania | Latn   | avt-12l | 98.3 | -   |
| blh    | 20K  | Kuwaa         | 50K    | Africa  | Latn   | bcq-Latn-12l | 100 | - |
| mfi    | 20K  | Wandala       | 50K    | Africa  | Latn   | acu-8l  | 100  | 13.5|
| eri    | 20K  | Ogea          | 5K     | Oceania | Latn   | agl-6l  | 98.3 | -   |
| npt    | 20K  | Mian          | 5K     | Oceania | Latn   | amp-12l | 98.1 | -   |
| swp    | 20K  | Suau          | 5K     | Oceania | Latn   | apb-13l | 99.4 | -   |
| nho    | 20K  | Takuu         | 5K     | Oceania | Latn   | bgt-4l  | 96.0 | -   |
| aso    | 20K  | Dano          | 50K    | Oceania | Latn   | aso     | 100  | 11.6|
| wnu    | 20K  | Usan          | 5K     | Oceania | Latn   | amp-12l | 96.7 | -   |
| zav    | 20K  | Yatzachi Zapotec | 5K | Americas | Latn | cut-5l | 100 | -|
| bcw    | 20K  | Bana          | 50K    | Africa  | Latn   | bcw-3l  | 100  | -   |
| dgz    | 20K  | Daga          | 5K     | Oceania | Latn   | dgz-5l  | 99.8 | -   |
| ipi    | 20K  | Ipili         | 50K    | Oceania | Latn   | gbi-8l  | 99.5 | -   |
| nnu    | 20K  | Sierra Negra Nahuelatl | 50K | Americas | Latn | nhy-3l | 99.5 | -|
| urt    | 20K  | Urat          | 5K     | Oceania | Latn   | aey-6l  | 99.1 | -   |
| obo    | 20K  | Obo Manobo    | 50K    | Asia    | Latn   | aat-4l  | 100  | 16.1|
| zpi    | 20K  | Santa María Quiegolani Zapotec | 5K | Americas | Latn | zaa-5l | 100 | -|
| sat-Deva | 20K | Santali      | 6M     | Asia    | Deva   | bbo-13l | 98.1 | -   |
| cvj    | 20K  | Chuave        | 50K    | Oceania | Latn   | bom-5l  | 97.5 | -   |
| ifu    | 20K  | Mayoyao Ifugao| 50K | Asia | Latn | bnc-8l | 100 | - |
| atg    | 20K  | Ivbie North-Opkela-Arhe | 50K | Africa | Latn | aig-8l | 100 | - |
| icer   | 20K  | San Andres Creole English | 50K | Americas | Latn | bzz-9l | 100 | - |
| neb    | 20K  | Toura (Côte d'Ivoire) | 50K | Africa | Latn | avn-3l | 100  | 8.9|
| apr    | 20K  | Arop-Lokep    | 5K     | Oceania | Latn   | apr-8l  | 99.2 | -   |
| aoz    | 20K  | Uab Meto      | 800K   | Asia    | Latn   | aar-6l  | 100  | 15.5|
| mca    | 20K  | Maca          | 5K     | Americas | Latn   | hxc-2l  | 100  | -   |
| tiy    | 20K  | Tiraray       | 50K    | Asia    | Latn   | cip-4l  | 100  | 13.2|
| gdr    | 20K  | Wipi          | 5K     | Oceania | Latn   | aom-13l | 98.4 | -   |
| kwk    | 20K  | Kandas        | 5K     | Oceania | Latn   | apb-13l | 93.8 | -   |
| zaw    | 20K  | Milta Zapotec | 50K | Americas | Latn | toj-4l | 100 | -|
| atq    | 20K  | Aralle-Tabulahan | 50K | Asia | Latn | atq-2l | 99.9 | -|
| ruf    | 20K  | Luguru        | 700K   | Africa  | Latn   | cwe-10l | 100  | 17.3|
| kje    | 20K  | Kisar         | 50K    | Asia    | Latn   | kje-6l  | 100  | -   |
| fae    | 20K  | Füiwol        | 5K     | Oceania | Latn   | bnj-6l  | 99.8 | -   |
| zpq    | 20K  | Zoogocho Zapotec | 5K | Americas | Latn | cut-5l | 100 | -|
| mcn    | 20K  | Masana        | 200K   | Africa  | Latn   | dzg-5l  | 50.7 | 47.6|
| nle    | 20K  | Nalca         | 50K    | Asia    | Latn   | mps-3l  | 100  | -   |
| yby    | 20K  | Yuweyuha      | 5K     | Oceania | Latn   | aar-14l | 97.6 | -   |
| pao    | 20K  | Northern Paiute| 5K | Americas | Latn | chq-3l | 99.9 | -|
| mnc    | 20K  | Donga Mambila | 50K    | Africa  | Latn   | mcn-2l  | 100  | 7.6 |
| kbe    | 20K  | Kadiwéu       | 5K     | Americas | Latn   | etn-2l  | 100  | -   |
| dje    | 20K  | Djambarrpuyngu| 5K | Oceania | Latn | bch-10l | 100 | -|
| zae    | 20K  | Yareni Zapotec | 5K | Americas | Latn | mnc-3l | 100 | -|
| wer    | 20K  | Weri          | 50K    | Oceania | Latn   | acn-11l | 99.2 | -   |
| maw    | 20K  | Mamprulei     | 200K   | Africa  | Latn   | cko-5l  | 100  | 14.1|
| kud    | 20K  | Auhelawa      | 5K     | Oceania | Latn   | aui-14l | 98.0 | -   |
| dad    | 20K  | Marik         | 5K     | Oceania | Latn   | amp-12l | 94.4 | -   |
| BCP-47 | Mono | Language Name | Speak. | Cont. | Script | Clust. | F1   | RTT |
|--------|------|----------------|--------|-------|--------|--------|------|-----|
| cax    | 20K  | Lomerianno-Ignaciano Chiquitano | 5K     | Americas | Latin | aca-8l | 100  | -   |
| ons    | 20K  | Ono            | 50K    | Oceania | Latin | kgf-4l | 100  | -   |
| wuw    | 20K  | Wuwulu-Aua     | 5K     | Oceania | Latin | aom-10l| 98.6 | -   |
| gay    | 20K  | Gayo           | 300K   | Asia   | Latvia | abs-13| 97.2 | 31.9|
| gng    | 20K  | Ngangam        | 50K    | Africa | Latin | ada-13| 100  | 15.4|
| nim    | 20K  | Nilamba        | 700K   | Africa | Latin | esx-4l| 100  | 16.5|
| snc    | 20K  | Sinaugoro      | 50K    | Oceania| Latin | gih-11| 97.6 | -   |
| nhu    | 20K  | Noone          | 50K    | Africa | Latin | cme-5l| 100  | 10.9|
| daa    | 20K  | Dangaleat      | 50K    | Africa | Latin | daa-3l| 100  | 12.3|
| yli    | 20K  | Angguruk Yali  | 50K    | Asia   | Latin | yli   | 100  | -   |
| bgq-Arab| 20K  | Bagri          | 2M     | Asia   | Arab  | bgq-Arab-4l | 99.7 | -   |
| tkn    | 20K  | Kwamera        | 5K     | Oceania| Latin | nwi-5l| 99.4 | -   |
| knf    | 20K  | Mankanya       | 50K    | Africa | Latin | knf-2l| 100  | 13.5|
| mbb    | 20K  | Mangseng       | 5K     | Oceania| Latin | acn-11| 98.3 | -   |
| cme    | 20K  | Cerma          | 50K    | Africa | Latin | cme-5l| 100  | 10.3|
| dgr    | 20K  | Dogrib         | 5K     | Americas| Latin | bgq-Latn-12l| 99.8 | -   |
| kup    | 20K  | Kunimaipa      | 50K    | Oceania| Latin | brj-4l| 99.5 | -   |
| lww    | 20K  | Lewo           | 5K     | Oceania| Latin | big-5l| 99.5 | -   |
| ktb    | 20K  | Kambaata       | 900K   | Africa | Ethio | gez-8l| 99.9 | 13.5|
| dop    | 20K  | Lukpa          | 50K    | Africa | Latin | dop   | 100  | -   |
| kki    | 20K  | Kagulu         | 200K   | Africa | Latin | dov-9l| 100  | 17.3|
| mpp    | 20K  | Migabac        | 5K     | Oceania| Latin | bru-6l| 99.3 | -   |
| sua    | 20K  | Sulka          | 5K     | Oceania| Latin | bwa-4l| 98.7 | -   |
| yon    | 20K  | Yonggom        | 5K     | Oceania| Latin | wnc-3l| 100  | -   |
| eka    | 15K  | Ekajuk         | 50K    | Africa | Latin | apn-3l| 100  | 12.6|
| mib    | 15K  | Chayuco Mixtec | 50K    | Americas| Latin | gaw-8l| 99.5 | -   |
| cux    | 15K  | Tepeuxila Cuicatec | 50K    | Americas| Latin | cux   | 100  | -   |
| tbl    | 15K  | Tobelo         | 50K    | Asia   | Latin | gbi-8l| 100  | -   |
| klv    | 15K  | Maskelynes     | 5K     | Oceania| Latin | bkv-4l| 99.9 | -   |
| due    | 15K  | Umiray Dumaget Agta | 5K     | Asia   | Latin | dgc-6l| 100  | -   |
| lmm-Deva| 15K  | Lambadi        | 6M     | Asia   | Deva  | awa-14l| 98.4 | 24.8|
| xuo    | 15K  | Kuo            | 50K    | Africa | Latin | xuo   | 100  | -   |
| shk    | 15K  | Shilluk        | 200K   | Africa | Latin | shk   | 100  | -   |
| xsb    | 15K  | Tiná Sambal    | 50K    | Asia   | Latin | dgc-6l| 100  | 21.8|
| amk    | 15K  | Ambai          | 50K    | Asia   | Latin | aby-4l| 100  | -   |
| zab    | 14K  | San Juan Guelavia Zapotec | 50K    | Americas| Latin | toj-4l| 99.7 | -   |
| mig    | 14K  | San Miguel El Grande Mixtec | 50K    | Americas| Latin | mig   | 60.8 | -   |
| hla    | 14K  | Halia          | 50K    | Oceania| Latin | aom-13| 97.6 | -   |
| dgc    | 14K  | Casiguran-Nagtipunan Agta | 5K     | Asia   | Latin | dgc-6l| 99.1 | -   |
| iry    | 14K  | Iraya          | 50K    | Asia   | Latin | agn-14l| 96.8 | -   |
| vun    | 14K  | Vunjo          | 900K   | Africa | Latin | nq-4l | 99.9 | 15.9|
| hot    | 14K  | Hote           | 5K     | Oceania| Latin | aey-6l| 93.1 | -   |
| otd    | 14K  | Ot Danum       | 50K    | Asia   | Latin | bw-9l | 100  | -   |
| ade    | 14K  | Adele          | 50K    | Africa | Latin | ade-4l| 100  | -   |
| nin    | 14K  | Ninzo          | 50K    | Africa | Latin | aig-8l| 100  | 13.4|
| gaw    | 14K  | Nobobob        | 5K     | Oceania| Latin | gaw-8l| 100  | -   |
| nwi    | 14K  | Southwest Tanna | 5K    | Oceania| Latin | nwi-5l| 100  | -   |
| naf    | 14K  | Nabak          | 50K    | Oceania| Latin | mig-2l| 98.1 | -   |
| ztg    | 14K  | Quioquitani-Quier Zapotec | 5K    | Americas| Latin | gv-3l| 99.7 | -   |
| bkd    | 14K  | Talaandig-Binukid | 50K    | Asia   | Latin | abh-11| 93.9 | -   |
| mit    | 14K  | Southern Puebla | 5K    | Americas| Latin | mit   | 100  | -   |
| kbr    | 14K  | Kafa           | 800K   | Africa | Latin | drc-Latin-8l| 100  | 12.8|
| cnt    | 14K  | Tepetotula Chichimec | 5K    | Americas| Latin | chr-9l| 100  | -   |
| BCP-47 | Mono | Language Name                  | Speak. | Cont.   | Script | Clust.  | F1   | RTT |
|--------|------|--------------------------------|--------|---------|--------|---------|------|-----|
| bdd    | 14K  | Bunama                         | 5K     | Oceania | Latn   | au-14l  | 97.3 | -   |
| bm-Nkoo| 14K  | Bambara                        | 14M    | Africa  | Nkoo   | bn-Nkoo-2l | 78.3 | 12.1|
| myw    | 14K  | Muyuw                          | 5K     | Oceania | Latn   | aom-13l | 97.9 | -   |
| box    | 14K  | Buamu                          | 200K   | Africa  | Latn   | box-3l   | 100  | 13.1|
| prf    | 13K  | Paranan                        | 50K    | Asia    | Latn   | dgc-6l   | 99.6 | -   |
| ziw    | 13K  | Ziguula-                       | 400K   | Africa  | Latn   | cwe-10l  | 98.3 | 17.6|
| maq    | 13K  | Chiquihuitlán                  | 5K     | Americas| Latn   | chr-9l   | 99.9 | -   |
| yrb    | 13K  | Yareba                         | 5K     | Oceania | Latn   | aby-4l   | 100  | -   |
| lwo    | 13K  | Luwo                           | 50K    | Africa  | Latn   | chq-3l   | 100  | 13.2|
| ozm    | 13K  | Koonzime                       | 50K    | Africa  | Latn   | mib-2l   | 100  | 9.5  |
| csk    | 13K  | Jola-Esulalu                   | 50K    | Africa  | Latn   | csk-3l   | 100  | 13.0 |
| kdl    | 13K  | Tsikimba                       | 50K    | Africa  | Latn   | asg-10l  | 99.0 | 16.3|
| hno    | 13K  | Northern Hindko                | 4M     | Asia    | Arab   | bal-13l  | 40.5 | 32.3|
| waj    | 13K  | Waffa                          | 5K     | Oceania | Latn   | waj     | 99.8 | -   |
| kkc    | 13K  | Odoodee                        | 5K     | Oceania | Latn   | adh-5l   | 99.8 | -   |
| nil    | 12K  | Peñoles Mixtec                 | 50K    | Americas| Latn   | mb-2l    | 100  | -   |
| ndj    | 12K  | Ndamba                         | 50K    | Africa  | Latn   | dov-9l   | 100  | 15.9|
| ngp    | 12K  | Ngulu                          | 100K   | Africa  | Latn   | cwe-10l  | 98.3 | 16.3|
| okv    | 11K  | Oronaiva                       | 50K    | Oceania | Latn   | aom-10l  | 99.1 | 14.6|
| nnq    | 11K  | Ngindo                         | 200K   | Africa  | Latn   | nnq-4l   | 99.9 | 13.9|
| grt    | 11K  | Garo                           | 1M     | Asia    | Beng   | bpy-7l   | 100  | -   |
| psh    | 11K  | Southwest Pashayi              | 50K    | Asia    | Arab   | fay-14l  | 15.5 | 38.0|
| hnd    | 11K  | Southern Hindko                | 600K   | Asia    | Arab   | bal-13l  | 17.3 | -   |
| tgp    | 11K  | Movono                         | 5K     | Oceania | Latn   | akb-4l   | 100  | -   |
| tbo    | 10K  | Tawala                         | 50K    | Oceania | Latn   | aui-14l  | 98.5 | -   |
| bxh    | 9K   | Buhutu                         | 5K     | Oceania | Latn   | aui-14l  | 93.2 | -   |
| mfa    | 7K   | Kelantan-Pattani               | 1M     | Asia    | Arab   | mfa-3l   | 95.9 | 23.2|

77