Using Encipherment to Isolate Conditions for the Successful Fine-tuning of Massively Multilingual Translation Models

Carter Louchheim, Denis Sotnichenko, Yukina Yamaguchi and Mark Hopkins
Williams College
Williamstown, MA

Abstract

When fine-tuning massively multilingual translation models for low-resource languages, practitioners often include auxiliary languages to improve performance, but factors determining successful auxiliary language selection remain unclear. This paper investigates whether syntactic similarity or lexical overlap is more important for effective multilingual fine-tuning. We use encipherment to create controlled experimental conditions that disentangle these confounded factors, generating novel languages with identical syntax but no lexical overlap, and conversely, languages that preserve lexical overlap. Through extensive NLLB-200 fine-tuning experiments across Europarl and AmericasNLP datasets, we demonstrate that lexical overlap is the dominant factor. Syntactically identical auxiliary languages provide negligible benefits (< 1.0 ChrF), while languages with significant lexical overlap provide substantial improvements (> 5.0 ChrF), with effectiveness strongly correlated to KL-divergence between token distributions ($r = -0.47$, $p < .001$). Our findings provide clear guidance: when selecting auxiliary languages for multilingual fine-tuning, prioritize lexical overlap over syntactic similarity.

1 Introduction

A popular modern approach to low-resource machine translation is the fine-tuning of massively multilingual encoder-decoder transformers (Vaswani et al., 2017). For example, the top two entrants in the AmericasNLP 2023[1] Shared Task on Machine Translation into Indigenous Languages (which solicits systems that translate Spanish into Indigenous American languages) both used this technique (Ebrahimi et al., 2023). The teams from the University of Sheffield (Gow-Smith and Sánchez Villegas, 2023) and the University of Helsinki (De Gibert et al., 2023) fine-tuned distillations of Meta's NLLB-200 model (Costa-Jussà et al., 2022) simultaneously on all eleven language pairs of the shared task.

Simultaneous training on a set of related language pairs (as opposed to training a separate system per language pair) has frequently been reported to yield performance benefits (Aharoni et al., 2019; Maillard et al., 2023). One survey (Ranathunga et al., 2023) on low-resource machine translation claims: "This is mainly due to the capability of the model to learn an interlingua (shared semantic representation between languages)". But when the parent model (as in the case of NLLB-200) has already been pre-trained on 200 language pairs, when (and why) does it remain beneficial to simultaneously fine-tune on multiple language pairs? What do the fine-tuning languages learn from each other?

In the context of multilingual language modeling, investigators have focused on two candidates: syntax and lexicon. According to a recent review (Philippy et al., 2023): "In previous research, syntax has been suggested as potentially the most important linguistic contributor for better cross-lingual transfer."[2] The same article also reports that "lexical overlap is particularly important when the pre-training corpus for the source language is small or when the word order between the source and target languages is dissimilar," but concludes that "lexical overlap is not a sufficient standalone factor to explain cross-lingual transfer."

An obstacle to drawing definitive conclusions is the difficulty of isolating the confounding factors of shared syntax and lexicon – related languages typically share both. In this work, we use **encipherment** to disentangle these factors. Encipher-

[1] These two systems were the baselines for the 2024 edition of the shared task, and continued to be the top systems for most language pairs. The Sheffield system was used as the baseline for the 2025 shared task, and maintained its superiority.

[2] The review, however, hypothesizes that the impact of syntax "may be overestimated" due to shortcomings in the research methods.

Proceedings of the Tenth Conference on Machine Translation, pages 241–252
November 8-9, 2025

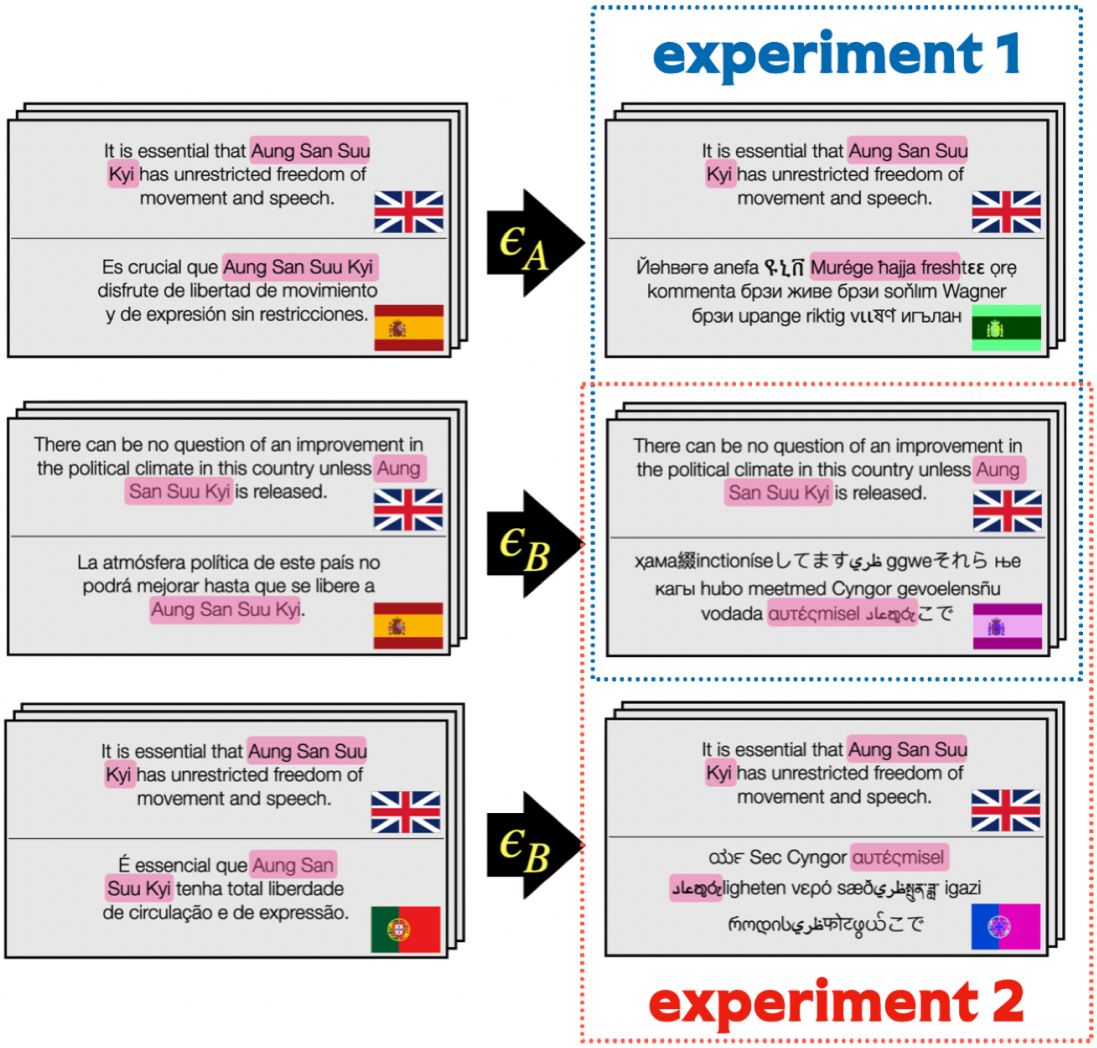


Figure 1: We use encipherment to create statistically realistic languages that are unseen by pre-trained translation models, affording experimental control over the often-confounded factors of syntactic and lexical overlap. For the top two rows, we apply different encipherments to disjoint subsets of English-Spanish Europarl, creating two novel languages with identical syntax but no lexical overlap. For the bottom two rows, we apply the same encipherment to disjoint subsets of English-Spanish and English-Portuguese Europarl, creating two novel languages that preserve the lexical overlap between Spanish and Portuguese.

ment allows us to generate statistically realistic languages that have not been previously included in translation model pre-training, while also providing experimental control:

1. We can produce syntactically identical languages with no lexical overlap. Figure 1 (top two rows) shows an example where two English-Spanish corpora are enciphered using different encipherments (ϵ_A and ϵ_B), producing two languages that are syntactically identical but lexically distinct.

2. We can produce novel languages that preserve cross-lingual lexical overlap. Figure 1 (bottom two rows) also shows an example where an English-Spanish corpus and an English-Portuguese corpus are enciphered using the same encipherment (ϵ_B), producing two languages that preserve the lexical overlap between Spanish and Portuguese.

The goal of this paper is to provide practical guidance to those seeking to build translation engines for low-resource languages. Specifically, when fine-tuning a massively multilingual translation model like NLLB-200, how should one select auxiliary languages to include in the fine-tuning (or should one include them at all)? We ultimately arrive at the following recommendations:

- **Recommendation 1:** Lexical overlap is the

most important factor to consider. If there is a low relative entropy (KL-divergence) between the token distribution of two source or target languages, then you can get considerable performance benefits from multilingual fine-tuning.

- **Recommendation 2:** Even if you can find an auxiliary language with extreme grammatical similarity to your low-resource language of interest, the performance benefits of multilingual fine-tuning (attributable to common syntax) are liable to be negligible.

Bottom line: When choosing auxiliary languages for the multilingual fine-tuning of massively multilingual translation models, focus on languages with high lexical overlap with your low-resource language of interest.

2 Related Work

Neural Machine Translation

Zoph et al. (2016) approached low-resource neural machine translation by leveraging a "parent" model (pre-trained on a high-resource language pair) to train a "child" model (for a low-resource language pair). Nguyen and Chiang (2017) and Kocmi and Bojar (2018) streamlined this process so that it could be succinctly described as follows (Kocmi and Bojar, 2018): "We train the parent language pair for a number of iterations and switch the training corpus to the child language pair for the rest of the training, without resetting any of the training (hyper)parameters."

Ensuing work studied conditions resulting in successful transfer from a parent to a child model. Among this work, Dabre et al. (2017) explored several parent-child combinations and reported that "transfer learning done on a X-Y language pair to [a] Z-Y language pair has maximum impact when Z-Y is resource-scarce and when X and Z fall in the same or linguistically similar language family." Lin et al. (2019) trained gradient-boosted decision tree models to predict synergistic parent/child language pairs, and observed that dataset size and word overlap were the most common splitting features. Aji et al. (2020) determined that the "inner" transformer layers were more crucial to transfer than the embedding layer, and noted that even using a simple copy model as the parent had performance benefits over training from scratch. This earlier work focused on parent models that were trained on a single language pair.

Over the past few years, the trend has been to pre-train massively multilingual translation models (Aharoni et al., 2019; Costa-Jussà et al., 2022) by simultaneously training on many language pairs. Focusing on the pragmatics of this training paradigm, Shaham et al. (2023) studied "interference," i.e. when multilingual pre-training underperforms bilingual pre-training. They concluded that the main cause of interference is when the model size is too small relative to the available training data. Tan and Monz (2023) used a linear regression model to determine factors that predict the zero-shot X→Y translation performance (where neither X nor Y is English) of multilingual models trained exclusively on English-centered language pairs. Our focus is on fine-tuning massively multilingual parent models for a low-resource language pair, specifically the factors that promote synergistic multilingual fine-tuning. We also believe we are the first work in this space to use encipherment as an instrument to de-confound the factors of syntactic similarity and lexical overlap.

Multilingual Language Modeling

It was quickly observed (Lin et al., 2019; Pires et al., 2019) that pre-trained multilingual encoder-only language models like mBERT (Devlin et al., 2019) could be fine-tuned on certain tasks (like named entity recognition or textual entailment) using monolingual supervision (typically English supervision), and only suffer minor performance degradation when applied zero-shot to other languages from the pre-training corpus. Several papers have studied this phenomenon (Dufter and Schütze, 2020; K et al., 2020; Lauscher et al., 2020; Ahuja et al., 2022; Deshpande et al., 2022; de Vries et al., 2022; Wu et al., 2023) – enough to have merited a survey paper (Philippy et al., 2023). Among these papers, K et al. (2020) used the most similar methodology to ours. To determine the impact of lexical overlap on cross-lingual transfer, they pre-trained two versions of BERT: one on English and Hindi[3] and another on "fake English" and Hindi, where "fake English" was derived from English by shifting the Unicode encoding of each character by a fixed constant. They concluded that lexical overlap plays only a minor role in cross-lingual transfer for textual entailment and named entity

[3]They also conducted this experiment with Russian instead of Hindi.

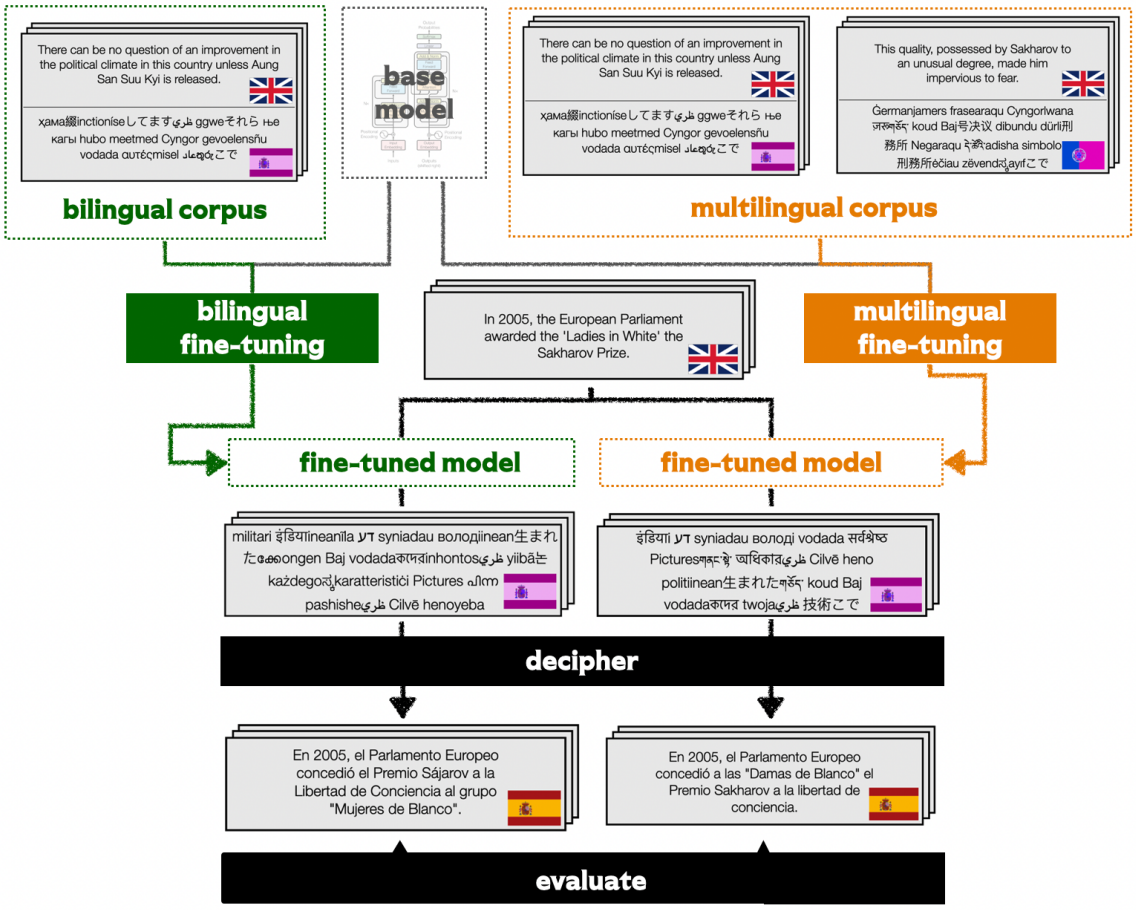


Figure 2: Overview of an experiment trial: a base model is fine-tuned using one or more enciphered bitexts. The resulting model is evaluated by translating held-out test sets, then deciphering and scoring the translations.

recognition. Also relevant to our approach is the work of Wu et al. (2023), who used "controlled studies" to investigate the factors contributing to the success of cross-lingual transfer learning – for instance, they perform artificial syntactic manipulations before fine-tuning on the GLUE dataset (Wang et al., 2018).

3 Preliminaries

This paper uses the following formalisms to describe enciphered parallel corpora.

Let $\mathcal{T}$ and $\mathcal{L}$ be finite alphabets that respectively correspond to a token vocabulary[4] and a set of language ids (e.g. `eng_Latn`, `rus_Cyrl`, etc.). Let $\mathcal{T}^*$ be the set of all sequences of tokens from $\mathcal{T}$.

Define a *parallel corpus* as a function $\pi : D \mapsto \mathcal{T}^*$, where $D \subset \mathcal{L} \times \mathbb{Z}^+$ and $\forall (l_1, i), (l_2, i) \in D$, $\pi(l_1, i)$ and $\pi(l_2, i)$ have the same meaning (i.e. they are translations of one another).

Define an *encipherment* ϵ as a permutation[5] of token vocabulary $\mathcal{T}$, i.e. a bijection $\epsilon : \mathcal{T} \mapsto \mathcal{T}$. For a token sequence $\pi(l, i) = \langle t_1, ..., t_k \rangle$ from parallel corpus π, denote the ϵ-enciphered sequence as $\pi_\epsilon(l, i) = \langle \epsilon(t_1), ..., \epsilon(t_k) \rangle$.

We extract a *bitext* from parallel corpus π using the following notation:

$$\hat{\pi}(l, l', \epsilon, \epsilon', I) = \{(\pi_\epsilon(l, i), \pi_{\epsilon'}(l', i) \mid i \in I\}$$

where $l, l' \in \mathcal{L}$ are language ids, ϵ, ϵ' are encipherments, and $I \subset \mathbb{Z}^+$ is a finite set of indices.

4 Experiment Design

Figure 2 provides an overview of our experiment design. Each experiment involves K bitexts ex-

[4]Throughout this paper, we assume a fixed token vocabulary. Namely, the token vocabulary we use in all our experiments is the NLLB-200 (Costa-Jussà et al., 2022) token vocabulary.

[5]One might worry that a translation model could simply learn to invert the permutation, but previous experimental work (Aji et al., 2020) suggests "that the [transformer] model is incapable of untangling [an] embedding permutation."

tracted from parallel corpus π:

$$\hat{\pi}(l_1, l'_1, \epsilon_1, \epsilon'_1, I_1)$$
$$\hat{\pi}(l_2, l'_2, \epsilon_2, \epsilon'_2, I_2)$$
$$\vdots$$
$$\hat{\pi}(l_K, l'_K, \epsilon_K, \epsilon'_K, I_K)$$

The index sets $I_1, \ldots, I_K$ are pairwise disjoint.

We use these bitexts to train several translation models:

- **bilingual fine-tuning**: For each bitext $\hat{\pi}(l_k, l'_k, \epsilon_k, \epsilon'_k, I_k)$, we fine-tune model M_k from pre-trained model M_{base}.
- **multilingual fine-tuning**: We use the entire collection of bitexts to simultaneously fine-tune a single model M_{multi} from pre-trained model M_{base}. During training, we sample evenly from the bitexts.

We fine-tune each model using a batch size of 64 for a maximum of 60,000 training steps. Validation is performed every 500 steps, and training is terminated early if the validation loss does not decrease for five consecutive evaluations. We use the Adafactor optimizer (Shazeer and Stern, 2018), as implemented in the Transformers library (Wolf et al., 2020), with a fixed learning rate of 1×10^{-4}, disabling both parameter scaling and relative step sizing. Gradient clipping is applied with a threshold of 1.0, and a weight decay of 1×10^{-3} is used for regularization. We adopt a constant learning rate schedule with warm-up, increasing the learning rate linearly over the first 1,000 steps.

We evaluate the resulting models by translating held-out test sets, then deciphering the translations and scoring them using standard machine translation metrics (e.g. BLEU (Papineni et al., 2002) and ChrF (Popović, 2015)). Because one of the languages in our experiments is polysynthetic, we report results using ChrF, but the choice of metric does not affect the experimental conclusions. We run 5 trials for each of the following training bitext sizes: 1024, 2048, 4096, 8192, and 16392 (so 25 trials in total). A bilingual fine-tuning trial consists of a fine-tuning for each bitext – the resulting systems are then evaluated and their scores are averaged. A multilingual fine-tuning trial is a single fine-tuning over all bitexts – the resulting system is then evaluated on each language pair, and these scores are averaged.

5 Base Models

We focus on fine-tuning NLLB-200 models (Costa-Jussà et al., 2022). These models were trained on a large-scale multilingual corpus covering 200 languages, using a transformer-based encoder-decoder architecture, following the general design of the M2M-100 model (Fan et al., 2021). Several distillations of this model are provided, including a 600M parameter model (6 encoder/decoder layers, 768 hidden size, 12 attention heads) and a 1.3B parameter model (12 encoder/decoder layers, 1024 hidden size, 16 attention heads). To increase experimental throughput, our experiments focus on the 600M parameter model.

6 Datasets

This section describes the parallel corpora that we use in our experiments.

Europarl

The Europarl Parallel Corpus (Koehn, 2005) is a set of sentence-aligned proceedings from the European Parliament covering sessions from 1996 to 2011. It spans 21 European languages, with each language contributing approximately 60 million words across 30 million aligned sentence fragments. We preprocess the corpus to eliminate repeated sentences.

AmericasNLP

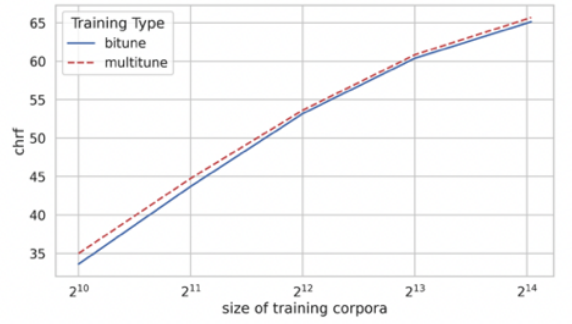
Referenced in the introduction, the AmericasNLP Workshop solicits systems that translate Spanish into Indigenous American languages. They provide official training corpora for these language pairs. In 2025, the workshop introduced Spanish → Wayuunaiki as a new language pair (Prieto et al., 2024). Wayuunaiki is "an Arawakan language spoken in northern Colombia and Venezuela, primarily by the Wayuu community, with about 420,000 speakers. It is an agglutinative language with a predominant SOV word order." (De Gibert et al., 2025)

7 Experiment 1: The Impact of Syntactic Similarity on the Success of Multilingual Fine-tuning

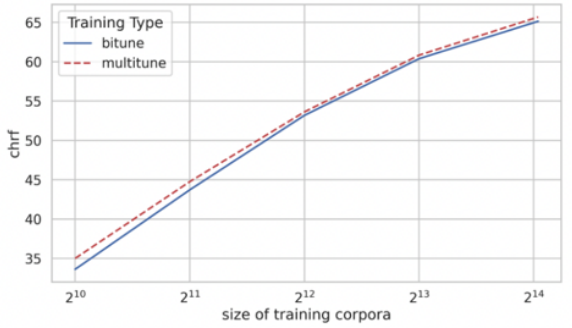
7.1 Syntactic Similarity of Target Languages

For our first set of experiments, we construct a scenario in which we have target languages with no lexical overlap but identical syntax (i.e. the top two rows of Figure 1). Specifically, we extract the

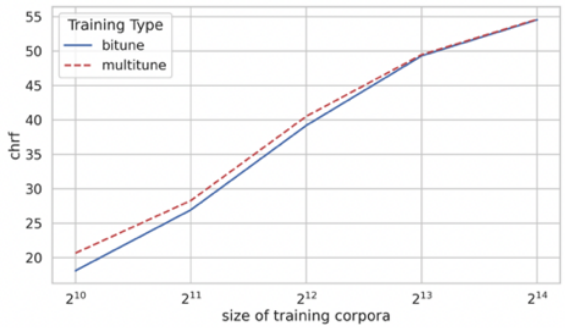
base model	corpus	source	target (enciphered)
nllb-200(600M)	europarl	english	spanish-ε1 spanish-ε2



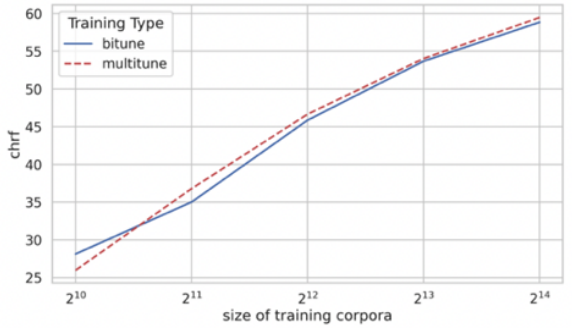
base model	corpus	source (enciphered)	target
nllb-200(600M)	europarl	spanish-ε1 spanish-ε2	english



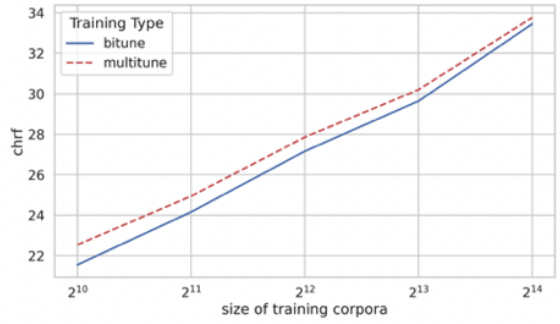
base model	corpus	source	target (enciphered)
nllb-200(600M)	europarl	english	czech-ε1 czech-ε2



base model	corpus	source (enciphered)	target
nllb-200(600M)	europarl	czech-ε1 czech-ε2	english



base model	corpus	source	target (enciphered)
nllb-200(600M)	anlp	spanish	wayuunaiki-ε1 wayuunaiki-ε2



base model	corpus	source (enciphered)	target
nllb-200(600M)	anlp	wayuunaiki-ε1 wayuunaiki-ε2	spanish

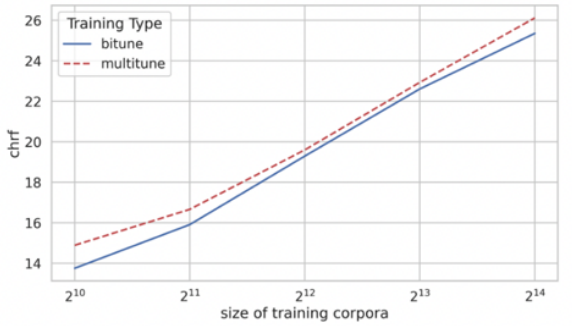


Figure 3: Experiment 1 results. Multilingual fine-tuning using two lexically distinct but syntactically identical languages provides only marginal improvement over bilingual fine-tuning of each language independently. While more pronounced for Wayuunaiki (whose unenciphered analogue is not part of the NLLB-200 pre-training corpus), the benefits are still minor (< 1.0 ChrF).

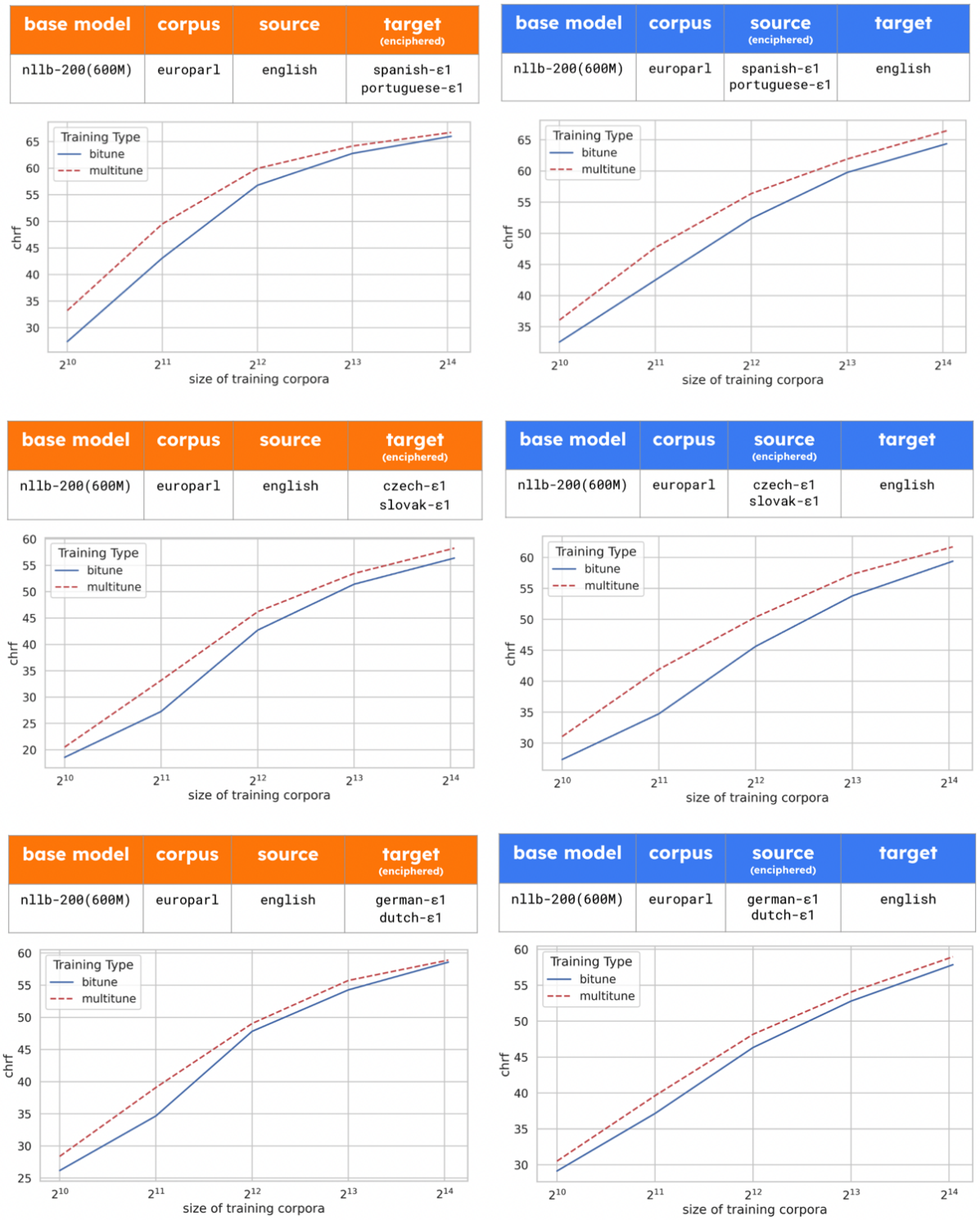


Figure 4: Experiment 2 results. Multilingual fine-tuning using two enciphered languages that preserve lexical overlap can provide significant improvement ($>$ 5.0 ChrF at certain data sizes) over bilingual fine-tuning of each language independently.

following k bitexts from parallel corpus π:

$$\hat{\pi}(l, l', \epsilon^0, \epsilon'_1, I_1)$$
$$\hat{\pi}(l, l', \epsilon^0, \epsilon'_2, I_2)$$
$$\vdots$$
$$\hat{\pi}(l, l', \epsilon^0, \epsilon'_k, I_k)$$

Here (and henceforth), ϵ^0 denotes the identity function (i.e. no encipherment occurs, since the encipherment maps each token to itself). Note that we have constructed k target languages with identical syntax but distinct lexicons. During fine-tuning, we freeze the encoder to eliminate the confounding impact of encoder domain adaptation to the source language.

7.2 Syntactic Similarity of Source Languages

Analogously, we construct a scenario in which we have **source** languages with no lexical overlap but identical syntax, i.e., we extract the following k bitexts from parallel corpus π:

$$\hat{\pi}(l, l', \epsilon_1, \epsilon^0, I_1)$$
$$\hat{\pi}(l, l', \epsilon_2, \epsilon^0, I_2)$$
$$\vdots$$
$$\hat{\pi}(l, l', \epsilon_k, \epsilon^0, I_k)$$

This time, we have constructed k source languages with identical syntax but distinct lexicons. During fine-tuning, we freeze the **decoder** to eliminate the confounding impact of decoder domain adaptation to the target language.

7.3 Results

We conducted these experiments using the following language pairs:

- Target Side Encryption: English $\rightarrow$ {Spanish, Czech, Wayuunaiki}
- Source Side Encryption: {Spanish, Czech, Wayuunaiki} $\rightarrow$ English

Figure 3 shows results from this set of experiments. Multilingual fine-tuning generally shows little advantage over bilingual fine-tuning, even though the two target languages are syntactically identical (they are both encipherments of the same language). Only in the case of Wayuunaiki, a language whose family (Arawak) was not represented in the original NLLB-200 training set, do we observe a small benefit from multilingual fine-tuning. This suggests that even in the best-case scenario – where we can find an auxiliary language with nearly identical syntax to our low-resource language of interest – the benefits of multilingual training (in the absence of lexical overlap) is minor.

Conclusion: Syntactic similarity of the source or target languages appears to have little impact on the effectiveness of multilingual fine-tuning.

8 Experiment 2: The Impact of Lexical Overlap on the Success of Multilingual Fine-tuning

8.1 Lexical Overlap of Target Languages

The only difference between this experiment and Experiment 1 is that we extract the following k bitexts from parallel corpus π:

$$\hat{\pi}(l, l'_1, \epsilon^0, \epsilon', I_1)$$
$$\hat{\pi}(l, l'_2, \epsilon^0, \epsilon', I_2)$$
$$\vdots$$
$$\hat{\pi}(l, l'_k, \epsilon^0, \epsilon', I_k)$$

We use different languages but the same encipherment, so that shared tokens remain shared after encipherment (see the bottom two rows of Figure 1). This produces unseen languages that have the same amount of lexical overlap as their unenciphered analogues. Again, we freeze the encoder during fine-tuning to eliminate the confounding impact of encoder domain adaptation to the source language.

8.2 Lexical Overlap of Source Languages

We also construct a scenario in which we have unseen **source** languages that have the same amount of lexical overlap as their unenciphered analogues, i.e., we extract the following k bitexts from parallel corpus π:

$$\hat{\pi}(l'_1, l, \epsilon', \epsilon^0, I_1)$$
$$\hat{\pi}(l'_2, l, \epsilon', \epsilon^0, I_2)$$
$$\vdots$$
$$\hat{\pi}(l'_k, l, \epsilon', \epsilon^0, I_k)$$

During fine-tuning, we freeze the **decoder** to eliminate the confounding impact of decoder domain adaptation to the target language.

8.3 Results

For these experiments, we used English as the unenciphered language l. As the enciphered

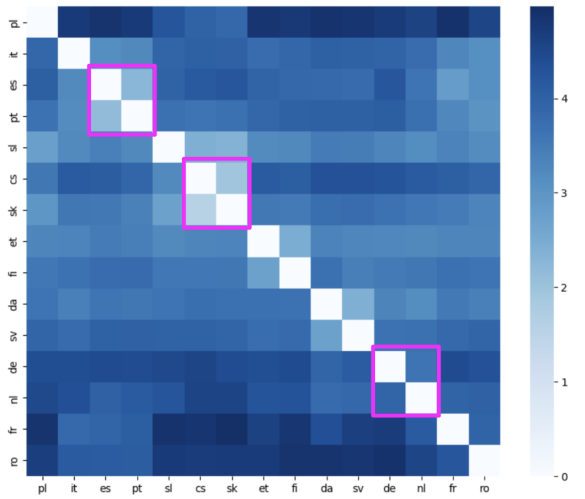


Figure 5: KL-divergence heatmap between Europarl languages. The heatmap shows the KL-divergence between token distributions for all pairs of Latin script languages in the Europarl corpus. Lighter colors indicate lower KL-divergence (greater lexical overlap). The magenta boxes highlight language pairs used in Experiment 2: Spanish-Portuguese (KL $\approx$ 2.26), Czech-Slovak (KL $\approx$ 1.96), and German-Dutch (KL $\approx$ 3.63). This visualization helps explain why Spanish-Portuguese and Czech-Slovak multilingual fine-tuning show greater benefits than German-Dutch, as their lower KL-divergence values indicate higher lexical overlap.

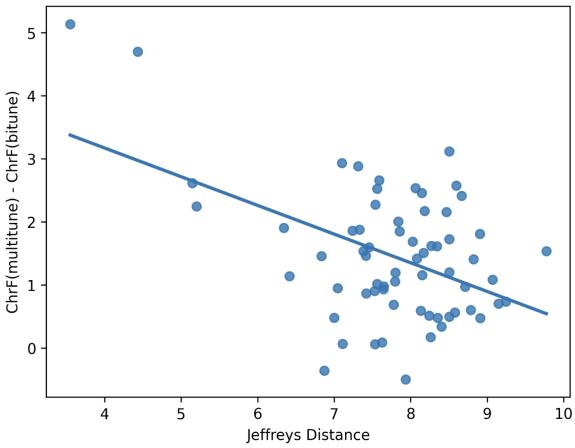


Figure 6: Relationship between lexical overlap and multilingual fine-tuning effectiveness. The scatter plot shows the ChrF improvement (multilingual minus bilingual fine-tuning) versus the Jeffreys distance between token distributions for all pairs of non-English Latin script languages in Europarl. Each point represents a single trial of Experiment 2 with bitext size 4096. The negative correlation (r = -0.47, p < .001) demonstrates that languages with greater lexical overlap (lower Jeffreys distance) benefit more from multilingual fine-tuning, supporting the conclusion that lexical overlap is the primary factor driving successful auxiliary language selection.

language combinations, we used three pairs of geographically-proximate European languages (which we assumed would have significant lexical overlap):

- l'_1 = Spanish and l'_2 = Portuguese
- l'_1 = Czech and l'_2 = Slovak
- l'_1 = German and l'_2 = Dutch

Figure 4 shows the results of these experiments. Under these conditions, multilingual fine-tuning substantially outperforms bilingual fine-tuning, often by more than five ChrF points. Given that Experiment 1 showed little benefit to incorporating syntactically similar languages during multilingual fine-tuning, it would appear that the observed benefits are mainly attributable to the lexical overlap.

However, the multilingual fine-tuning of English $\leftrightarrow$ German and English $\leftrightarrow$ Dutch is notably less effective than the others. To explain this difference, we computed the KL-divergence between the token distributions of all Europarl languages that use Latin script (see Figure 5). The KL-divergence from Spanish to Portuguese (approximately 2.26) and from Czech to Slovak (approximately 1.96) is considerably smaller than the KL-divergence from German to Dutch (approximately 3.63).

To assess the general impact of lexical overlap on the effectiveness of multilingual training, we conducted a single trial of Experiment 2 (using bitext size 4096) for every pair (l'_1, l'_2) of non-English Latin script languages in Europarl. Figure 6 plots the ChrF delta between multilingual and bilingual fine-tuning, versus the Jeffreys distance (additive symmetrization of KL-divergence) between languages l'_1 and l'_2. There is a moderate, statistically significant negative correlation ($r(63) = -0.47$, $p < .001$), suggesting that lexical overlap is a significant determining factor in the effectiveness of multilingual fine-tuning.

9 Conclusion

This work addresses a fundamental question in low-resource machine translation: when fine-tuning massively multilingual models like NLLB-200, which factors determine the success of multilingual fine-tuning with auxiliary languages? Through controlled experiments using encipherment to disen-

tangle syntactic similarity and lexical overlap, we provide empirical evidence that lexical overlap is the primary driver of performance improvements. Our key findings are:

- Syntactic similarity provides minimal benefit: Even when auxiliary languages share identical syntax with the target language, multilingual fine-tuning shows little advantage over bilingual approaches. The benefits are most pronounced (but still minor) only for languages from families not represented in the pre-training corpus.
- Lexical overlap drives substantial improvements: Languages that share vocabulary can provide significant performance gains (> 5.0 ChrF in many cases), with effectiveness inversely correlated to the KL-divergence between token distributions.

The encipherment methodology introduced here also provides an experimental framework for future research on cross-lingual transfer, allowing researchers to control for confounding factors that typically make it difficult to isolate the impact of different linguistic properties.

Future work should explore whether these findings generalize to other massively multilingual models, investigate optimal methods for measuring and maximizing lexical overlap, and examine whether the relative importance of syntax versus lexicon changes with different model architectures or pre-training objectives.

References

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

Kabir Ahuja, Shanu Kumar, Sandipan Dandapat, and Monojit Choudhury. 2022. Multi task learning for zero shot performance prediction of multilingual models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5454–5467, Dublin, Ireland. Association for Computational Linguistics.

Alham Fikri Aji, Nikolay Bogoychev, Kenneth Heafield, and Rico Sennrich. 2020. In neural machine translation, what does transfer learning transfer? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7701–7710, Online. Association for Computational Linguistics.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Raj Dabre, Tetsuji Nakagawa, and Hideto Kazawa. 2017. An empirical study of language relatedness for transfer learning in neural machine translation. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 282–286. The National University (Phillippines).

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

Ona De Gibert, Raúl Vázquez, Mikko Aulamo, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2023. Four approaches to low-resource multilingual NMT: The Helsinki submission to the AmericasNLP 2023 shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 177–191, Toronto, Canada. Association for Computational Linguistics.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's submission to the AmericasNLP shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 192–199, Toronto, Canada. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small data, big impact: Leveraging minimal data for effective machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Juan Prieto, Cristian Martinez, Melissa Robles, Alberto Moreno, Sara Palacios, and Rubén Manrique. 2024. Translation systems for low-resource colombian indigenous languages, a first step towards cultural preservation. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages

7–14, Mexico City, Mexico. Association for Computational Linguistics.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Uri Shaham, Maha Elbayad, Vedanuj Goswami, Omer Levy, and Shruti Bhosale. 2023. Causes and cures for interference in multilingual translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15849–15863, Toronto, Canada. Association for Computational Linguistics.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*.

Shaomu Tan and Christof Monz. 2023. Towards a better understanding of variations in zero-shot neural machine translation performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13553–13568, Singapore. Association for Computational Linguistics.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45.

Zhengxuan Wu, Alex Tamkin, and Isabel Papadimitriou. 2023. Oolong: Investigating what makes transfer learning hard with controlled studies. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3280–3289, Singapore. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.