Bringing Mapudungun into the Modern MT Ecosystem: Morphology-Aware Tokenization for NLLB-200 Fine-Tuning

Isaac M. Thompson[1], **Brandon M. A. Rogers**[2], **Eric K. Ringger**[1]
[1]Department of Computer Science, Brigham Young University
[2]Department of Spanish & Portuguese, Brigham Young University
{it238, brandon.rogers, ringger}@byu.edu

Abstract

For Mapudungun arn→es translation, morphology-aware tokenization can substitute for a 5× increase in model parameters. We fine-tune three sizes of Meta's NLLB-200 on Mapudungun–Spanish translation across eight tokenization strategies, including our novel Morfessor-VC method, which constrains Morfessor morpheme segmentation to tokens already present in NLLB's pretrained vocabulary. Our 600M Morfessor-VC model is competitive with our own fine-tuned 3.3B Standard BPE model on arn→es (43.2 vs. 42.9 chrF++, $\Delta = +0.3$, $p = 0.039$, 95% CI [0.02, 0.60]) while using five times fewer parameters, and all fine-tuned conditions surpass frontier LLMs by over 27 chrF++. Mapudungun is an indigenous polysynthetic language spoken by 200,000+ Mapuche people in Chile and Argentina, absent from NLLB-200 and not supported by major commercial MT providers; prior work predates large-scale multilingual models and does not address the tokenization challenges posed by its agglutinative morphology. These results establish new state-of-the-art baselines for Mapudungun MT and provide a practical foundation for community language tools in pedagogy, social media, and language revitalization.

1 Introduction

Mapudungun (also *Mapuzugun*, *Mapudungu*, or *Mapuche*; historically *Araucanian*, the colonial designation reflected in ISO 639-3 code `arn`) is spoken by more than 200,000 Mapuche people across southern Chile and Argentina (Duan et al., 2020), making it one of the largest indigenous languages in South America by speaker count. Despite this, Mapudungun remains absent from nearly all commercial and open-source machine translation (MT) systems—a gap with direct consequences for community language vitality (Ahumada et al., 2022). The language's polysynthetic, agglutinative morphology—where a single verb form can encode tense, aspect, subject, object, evidentiality, and spatial deixis—poses well-documented challenges for standard subword tokenization (Rust et al., 2021; Mielke et al., 2021; Petrov et al., 2023). Byte-pair encoding (BPE) (Sennrich et al., 2016), the dominant tokenization strategy in multilingual models, is not designed to respect morpheme boundaries, and high fertility rates on agglutinative languages have been shown to correlate with degraded translation quality (Ahia et al., 2023; Banerjee and Bhattacharyya, 2018).

Prior work on Mapudungun NLP is sparse. Levin et al. (2002) collected early bilingual data and explored foundational language technologies for Mapudungun. Duan et al. (2020) introduced the AVENUE corpus and established sequence-to-sequence baselines. Pendas et al. (2023) and Lira et al. (2025) have since explored neural approaches, and Chandiá (2022) developed morphological analysis tools. None of this prior work has fine-tuned large-scale multilingual models on Mapudungun, nor systematically studied how tokenization choices interact with model capacity for this language.

We address both gaps. Our contributions are:

- The first fine-tuning study of Meta's NLLB-200 (NLLB Team et al., 2022) on Mapudungun–Spanish translation, across three model sizes (600M, 1.3B, 3.3B parameters) and both translation directions.
- A systematic ablation of eight tokenization strategies, ranging from standard NLLB BPE to Morfessor-based segmentation (Virpioja et al., 2013) and SentencePiece UnigramLM (Kudo, 2018).
- **Morfessor-VC**, a novel tokenization method that runs Morfessor segmentation and then constrains the resulting vocabulary to tokens

Proceedings of the Sixth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 173–185

already present in NLLB's pretrained vocabulary, preserving embedding alignment while improving morpheme boundary coverage.

- A comprehensive evaluation showing that Morfessor-VC with the 600M model achieves 43.2 chrF++ on Mapudungun→Spanish, matching our standard BPE 3.3B baseline (42.9 chrF++), and outperforming frontier LLMs (Aya Expanse 8B: 15.9 chrF++) by over 27 points.

Code and evaluation scripts are available at https://github.com/byu-matrix-lab/mapudungun-nllb; fine-tuned model weights are released as a HuggingFace collection at https://huggingface.co/collections/byumatrixlab/mapudungun-nllb.

2 Related Work

Mapudungun NLP. Levin et al. (2002) collected early bilingual data and explored foundational language technologies for Mapudungun at CMU. Duan et al. (2020) introduced the AVENUE corpus and established sequence-to-sequence baselines, reporting plain chrF (character n-gram F-score, 0–1 scale; not directly comparable to chrF++) of 0.50 arn→es and 0.40 es→arn on 220k training pairs using a custom Transformer. Ahumada et al. (2022) developed educational NLP tools for Mapudungun. Pendas et al. (2023) applied active learning to Mapudungun MT; their BLEU scores are not directly comparable to standard held-out evaluation due to test-train overlap in the active learning simulation. Lira et al. (2025) explored transfer learning from high-resource language pairs (Spanish–English, Spanish–Finnish) and reported 30.30 chrF on arn→es on a separate 1,250-pair test set. Chandiá (2022) developed a finite-state morphological analyser for Mapudungun, providing linguistic groundwork relevant to our tokenization study.

Morphology-aware tokenization. The interaction between subword tokenization and morphologically rich languages is well studied. Rust et al. (2021) showed that multilingual models produce unequal tokenization quality across languages, with agglutinative languages suffering highest fertility. Ahia et al. (2023) and Petrov et al. (2023) further quantified how tokenization inequity translates to downstream performance gaps. Ataman and Federico (2018) and Banerjee and Bhattacharyya (2018) proposed combining morphological segmentation with neural MT, motivating our Morfessor-BPE condition. Mager et al. (2022) conducted a direct comparison of BPE and morphological segmentation for four polysynthetic Amerindian languages, finding that morphological segmentation can outperform BPE in low-resource settings—consistent with our findings for Mapudungun. Gowda and May (2020) established that vocabulary size has diminishing returns beyond a language-specific threshold, informing our Mono BPE and Optuna BPE conditions. Morfessor 2.0 (Virpioja et al., 2013), the unsupervised morpheme segmenter underlying three of our conditions, uses minimum description length to learn morpheme boundaries without linguistic supervision. Subword regularization via UnigramLM (Kudo, 2018) has shown robustness benefits in low-resource settings, motivating our inclusion of that condition.

Low-resource MT with multilingual models. NLLB-200 (NLLB Team et al., 2022) supports 200 languages but does not include Mapudungun. Downey et al. (2023) studied methods for adapting multilingual vocabularies to new languages, directly relevant to our tokenization approach. Fine-tuning NLLB on new languages has shown strong results in AmericasNLP shared tasks (Ebrahimi et al., 2023, 2024; de Gibert et al., 2025). Gow-Smith and Sánchez Villegas (2023) and DeGenaro and Lupicki (2024) demonstrated competitive NLLB fine-tuning results across indigenous languages at the 2023 and 2024 shared tasks respectively; our work is the first to systematically study tokenization strategies—rather than data augmentation or model selection—for NLLB fine-tuning on a polysynthetic language. Mager et al. (2023) provides a comprehensive overview of MT for indigenous languages of the Americas.

3 Data

We use the AVENUE corpus (Duan et al., 2020), a Mapudungun–Spanish parallel resource derived from the Mapudungun Speech Corpus (Caniupil et al., 2019)—oral history interviews recorded by Mapuche community members and linguists, transcribed and aligned using ELAN (Wittenburg et al., 2006). The corpus was developed through an academic collaboration (CMU, Universidad de La Frontera, Chilean government partners) and uses

the Alfabeto Mapuche Unificado (AMU) orthography. The corpus covers domains such as traditional medicine, cultural practices, and personal narrative.

Extraction and cleaning. The AVENUE data is structured as parallel ELAN annotation blocks. We extract sentence-level pairs by treating each ELAN block as a translation unit. We then apply a cleaning pass that strips incomplete ELAN tags, CHAT transcription codes, overlap and uncertainty markers, and degenerate segments. Two standard defaults required language-specific overrides: the minimum word count was set to 1 (single-word Mapudungun clauses are grammatically complete in a polysynthetic language) and long-word filtering was disabled (Mapudungun compounds regularly exceed 30 characters). The resulting corpus is split into train / dev / test sets of 55,452 / 1,581 / 9,382 sentence pairs, respectively.

Code-switching. A notable property of the AVENUE corpus is widespread code-switching: approximately 24% of Mapudungun utterances contain embedded Spanish words or phrases, as marked by ELAN <SPA> tags in the original transcriptions. This reflects natural bilingual speech patterns in contemporary Mapuche communities rather than transcription artifacts. Our cleaning pipeline retains these mixed-language utterances; we analyze their effect on translation quality in Section 7.3.

Data statistics. Table 1 summarizes corpus statistics. The Mapudungun training side contains 656,969 whitespace-delimited tokens (avg. 11.8 tokens/sentence); the Spanish side contains 888,476 tokens (avg. 16.0 tokens/sentence). The higher Spanish token count reflects that Mapudungun's polysynthetic morphology encodes in a single word what Spanish expresses across multiple words.

Split	Lang	Sents	Tokens	Types	Avg len
Train	arn	55,452	656,969	105,375	11.8
	es	55,452	888,476	40,059	16.0
Dev	arn	1,581	23,519	6,863	14.9
	es	1,581	29,314	4,381	18.5
Test	arn	9,382	108,361	26,264	11.5
	es	9,382	144,870	13,624	15.4

Table 1: Corpus statistics after cleaning. Tokens and types are whitespace-delimited. Mapudungun has substantially more types per token than Spanish, reflecting its richer morphology.

4 Tokenization Methods

A core challenge in fine-tuning NLLB-200 on Mapudungun is the mismatch between NLLB's pretrained BPE vocabulary—optimized for 200 languages with Spanish heavily represented—and Mapudungun's polysynthetic morphology. Naive fine-tuning feeds the model unsegmented Mapudungun text, which NLLB's internal tokenizer fragments unpredictably. We compare eight tokenization strategies that differ in how they pre-segment Mapudungun before fine-tuning. All conditions use NLLB's standard tokenizer for the Spanish side; only the Mapudungun side is varied. Pre-segmented text uses the `@@` boundary marker convention (e.g., *mapu@ dun@ gun*) so that desegmentation is applied at inference time before scoring.

4.1 Standard BPE (NLLB)

The baseline condition feeds raw, unsegmented Mapudungun text directly to NLLB's built-in SentencePiece tokenizer without any pre-processing. This is the standard fine-tuning setup for NLLB and serves as our primary comparison point.

4.2 Joint BPE

Following Duan et al. (2020), we train a shared SentencePiece BPE model (Kudo and Richardson, 2018) on the concatenated Mapudungun and Spanish training data with a vocabulary size of 5,000. Joint training encourages shared subword representations across languages, which can aid cross-lingual transfer but may produce poor segmentations for the morphologically richer language.

4.3 Mono BPE

We train a Mapudungun-only SentencePiece BPE model, setting the vocabulary size via the 95%-character-coverage heuristic of Gowda and May (2020): we find the number of character types needed to cover 95% of character occurrences, then multiply by 10 to estimate the subword vocabulary size. This yields a vocabulary focused on Mapudungun morphology without interference from Spanish.

4.4 Optuna BPE

To address the objection that our results might be sensitive to vocabulary size selection, we use Optuna (Akiba et al., 2019) to tune the BPE vocabulary size over 15 trials, optimizing chrF++ on the development set using a 600M NLLB model

fine-tuned for 3 epochs per trial. The optimal vocabulary size found was 10,954. Notably, chrF++ was largely flat across a wide range of vocabulary sizes (1,431–14,517 all achieved 45.72 chrF++), suggesting BPE vocabulary size is not a critical hyperparameter for this language pair.

4.5 Morfessor

We apply Morfessor 2.0 (Virpioja et al., 2013), an unsupervised morpheme segmenter based on minimum description length, to the Mapudungun training corpus. Morfessor learns morpheme boundaries without linguistic supervision, producing segmentations that reflect statistical regularities in the character sequences. Boundaries are marked with `@@`. The Spanish side is left unsegmented.

4.6 Morfessor-VC (Our Method)

Standard Morfessor segmentation introduces boundaries that do not align with NLLB's pretrained vocabulary, causing *double tokenization*: NLLB's internal tokenizer re-splits the already-segmented morphemes, producing fragmented representations that were not seen during pretraining.

Morfessor-VC (Vocabulary-Constrained) is our proposed solution. Starting from Morfessor segmentation, we greedily merge adjacent morphemes whose concatenation corresponds to a single token in NLLB's SentencePiece vocabulary. The algorithm processes each word left-to-right:

1. Given morphemes $[m_0, m_1, \ldots, m_k]$ within a word, consider the candidate merge $m_i \oplus m_{i+1}$.
2. If $\_(m_i \oplus m_{i+1})$ or $(m_i \oplus m_{i+1})$ is a single piece in NLLB's vocabulary, merge and repeat from m_i.
3. Otherwise, retain the boundary and advance to m_{i+1}.

The result preserves morpheme boundaries only where the split is both linguistically motivated (by Morfessor) *and* corresponds to a genuine subword boundary in NLLB's vocabulary. Boundaries that span a single NLLB vocabulary item are removed. This reduces double-tokenization while retaining meaningful morphological signal. Unlike prior work applying standard Morfessor to polysynthetic languages (Mager et al., 2022), Morfessor-VC does not introduce segments that the pretrained model's tokenizer will re-split; the VC step is specific to fine-tuning scenarios where a fixed pretrained vocabulary must be respected.

Table 2 illustrates the difference on *kutrankautulay* ("does not become sick"; kutran = illness, kautu = to become, -lay = negation).

Method	Segmentation
Standard BPE	`_ku tran anka ut ulay`
Morfessor	`kutran@@ ka@@ u@@ tulay`
Morfessor-VC	`kutran@@ kau@@ tulay`

Table 2: Tokenization of *kutrankautulay* under three conditions. Standard BPE (top) shows NLLB's internal SentencePiece output on the raw word, ignoring morpheme structure entirely. Morfessor (middle) adds pre-segmentation boundaries, but splits *kau* into *ka* + *u*; when each pre-segment is fed to NLLB, *ka* and *u* receive word-initial _ sentinels—embeddings pretrained as independent Spanish function words, not as morpheme-internal pieces. Morfessor-VC (bottom) merges *ka* + *u* → *kau*, a single item in NLLB's vocabulary (id 22,381), eliminating this spurious split.

4.7 Morfessor-BPE

Following Banerjee and Bhattacharyya (2018), we apply BPE *within* Morfessor morphemes. First, Morfessor segments the corpus; then a BPE model (vocabulary size 4,000) is trained on the resulting morpheme tokens and applied within morpheme boundaries. This combines morphological segmentation with BPE's data-driven compression.

4.8 UnigramLM

We train a SentencePiece Unigram language model (Kudo, 2018) on the Mapudungun training data using the same 95%-character-coverage vocabulary size heuristic as Mono BPE. Unlike BPE, UnigramLM directly optimizes a probabilistic segmentation model and can produce multiple segmentations for the same word, which has been shown to improve robustness in low-resource settings (Kudo, 2018).

4.9 Fertility Comparison

Figure 1 shows the fertility (pre-segmented tokens per whitespace-delimited word) for each condition on the Mapudungun training side. Morfessor and Morfessor-VC have the lowest fertility (1.13), close to one-to-one morpheme–word mapping. BPE-based methods are more aggressive (1.55–2.04), and UnigramLM (1.98) is comparable to Mono BPE. The near-identical fertility of Morfessor and Morfessor-VC confirms that the VC merging step removes boundaries without adding

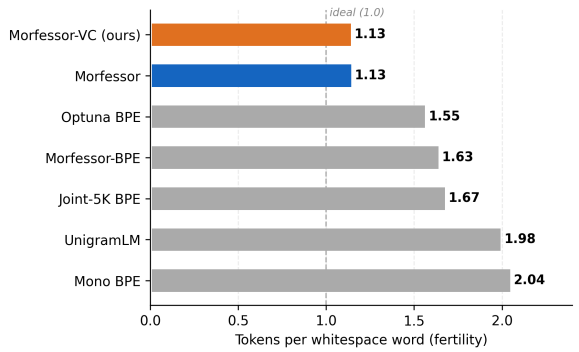

Figure 1: Pre-segmentation fertility (tokens per whitespace word) for each tokenization condition on the Mapudungun training set, sorted in ascending order. Morfessor and Morfessor-VC achieve the lowest fertility (1.13); BPE-based methods and UnigramLM are more aggressive (1.55–2.04).

new ones. Standard BPE is excluded from this comparison because its segmentation is applied internally by NLLB's SentencePiece tokenizer and produces incomparable fertility values.

5 Experimental Setup

Models. We fine-tune three sizes of NLLB-200-distilled (NLLB Team et al., 2022): 600M, 1.3B, and 3.3B parameters. All models are initialized from Meta's publicly released checkpoints. We fine-tune each model independently for each tokenization condition and both translation directions (arn→es and es→arn), for a total of 48 fine-tuning runs.

Training. All models are fine-tuned for up to 10 epochs with early stopping (patience 3, dev chrF++), AdamW ($lr = 3{\times}10^{-5}$, 1K warmup steps, weight decay 0.01), effective batch size 128, max length 128, beam size 4, on a single NVIDIA A100.

Evaluation. Our primary metric is chrF++ (Popović, 2017), a character n-gram F-score that has shown correlations with human judgments across language pairs. Our pilot human evaluation in Section 6.1 found no consistent preference between systems despite a ∼3-point chrF++ gap at 3.3B, though the study's single-annotator design limits the strength of that conclusion. We report BLEU (Papineni et al., 2002) as a secondary metric. Both metrics are computed with sacreBLEU (Post, 2018) against the held-out test set (9,382 sentence pairs). We also report COMET (wmt22-comet-da) in Appendix B for completeness; because Mapudungun is absent from XLM-R's training data, COMET scores on the Mapudungun side are unreliable and we do not use them to draw conclusions. For pre-segmented conditions, output is desegmented (removing `@@` markers) before scoring. We assess statistical significance for headline comparisons using paired bootstrap resampling (10,000 iterations) over per-sentence chrF++ scores (Koehn, 2004).

Human evaluation. We conduct a pilot preference ranking of Standard BPE (3.3B) vs. Morfessor-VC (3.3B) outputs on 50 stratified sentences per direction; see Section 6.1.

6 Results

Table 3 reports chrF++ for all 8 tokenization conditions across 3 model sizes and both translation directions, alongside zero-shot NLLB baselines and prompted LLM comparisons.

Fine-tuning vs. baselines. Zero-shot NLLB and prompted LLM scores serve as pre-fine-tuning references: both reflect performance without any Mapudungun-specific training, not directly comparable competitors to our fine-tuned systems. Fine-tuning yields dramatic gains over both in both directions (Figure 2). Even the smallest fine-tuned model under Standard BPE (34.9 chrF++) substantially outperforms Aya Expanse 8B (15.9 chrF++) on arn→es, and all morphology-aware conditions at 600M exceed 27 points above the best LLM baseline.

Tokenization effects. Figure 3 shows chrF++ by condition and model size. Standard BPE lags at 600M (34.9 arn→es) but catches up at 3.3B (42.9), while all morphology-aware methods cluster between 41.7–43.2 at 600M. Morfessor-VC leads arn→es at both 600M and 3.3B. For es→arn, tokenization effects are substantially smaller, with Standard BPE competitive at 3.3B. This asymmetry is important: es→arn produces Mapudungun output, the direction most relevant for community applications, and morphology-aware tokenization does not deliver a clear advantage there. This contrasts with Mager et al. (2022), who found the largest morphological segmentation gains on the polysynthetic-output direction; we attribute the difference to NLLB's fixed decoder-side tokenizer, which limits the benefit of pre-segmented targets. Claims about morphology-aware tokenization as

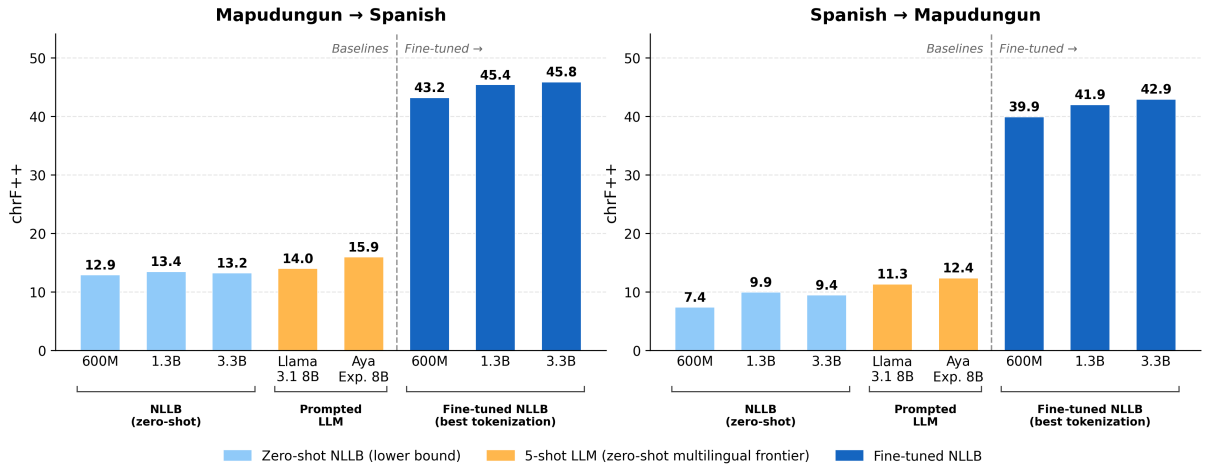

Figure 2: chrF++ by system and translation direction. Fine-tuned NLLB bars show the best-performing tokenization condition per direction (Morfessor-VC for arn→es, Standard BPE for es→arn). Dashed line separates baselines from fine-tuned systems. Zero-shot NLLB and prompted LLM bars establish lower bounds and the frontier of zero-shot multilingual capability before any Mapudungun-specific training; they are not fine-tuned competitors. Duan et al. (2020) and Pendas et al. (2023) are omitted: Duan reported plain chrF (0–1 scale), not directly comparable to chrF++; Pendas reported BLEU under an active learning simulation protocol. Lira et al. (2025) are omitted because their test set differs from ours; see Section 2.

Condition	arn→es				es→arn			
	600M	1.3B	3.3B	8B†	600M	1.3B	3.3B	8B†
Baselines (not fine-tuned)								
Zero-shot NLLB	12.86	13.43	13.19	—	7.37	9.91	9.42	—
Llama 3.1 8B	—	—	—	13.96	—	—	—	11.32
Aya Expanse 8B	—	—	—	15.92	—	—	—	12.36
Fine-tuned NLLB-200								
Standard BPE	34.90	40.64	42.85	—	39.87	41.94	**42.89**	—
Joint-5K BPE	42.36	44.68	45.25	—	38.17	41.06	42.30	—
Mono BPE	41.72	43.99	44.85	—	38.04	40.62	42.02	—
Optuna BPE	42.41	44.67	45.21	—	38.79	41.34	42.33	—
Morfessor	43.09	**45.40**	45.51	—	**39.90**	42.00	42.76	—
Morfessor-VC	**43.16**	45.37	**45.84**	—	39.89	**42.04**	42.77	—
Morfessor-BPE	42.80	44.97	45.56	—	38.74	41.25	42.40	—
UnigramLM	42.18	44.64	45.07	—	38.20	41.05	42.30	—

Table 3: chrF++ on the test set (9,382 pairs). Bold = best in column among fine-tuned conditions. †LLM baselines evaluated with 5-shot prompting.

an efficiency lever should therefore be understood as specific to the arn→es direction.

Scaling efficiency. Morfessor-VC at 600M (43.2 chrF++) matches Standard BPE at 3.3B (42.9 chrF++) on arn→es ($\Delta = +0.31$, $p = 0.039$, 95% CI $[0.02, 0.60]$; significant but with a narrow margin), using five times fewer parameters. Morfessor-VC 3.3B also significantly outperforms Standard BPE 3.3B ($\Delta = +2.99$, $p < 0.001$) and vanilla Morfessor 3.3B ($\Delta = +0.34$, $p < 0.001$, 95% CI $[0.14, 0.53]$).

6.1 Human Evaluation

We conducted a pilot preference ranking evaluation comparing Standard BPE (3.3B) and Morfessor-VC (3.3B) outputs across 50 sentences per direction, stratified by per-sentence chrF++ quartile. A single evaluator—a second-language learner of both Spanish and Mapudungun (native English speaker), not a native speaker of either language—indicated which system output was preferred for each sentence, or whether the two were of equal quality. No inter-annotator agreement was measured, as only one evaluator was available for this pilot study.

Results are shown in Table 4. Preferences are evenly split in both directions, with no clear advantage for either system. This is itself a finding: a 3-point chrF++ advantage for Morfessor-VC on arn→es does not translate to a perceptible quality

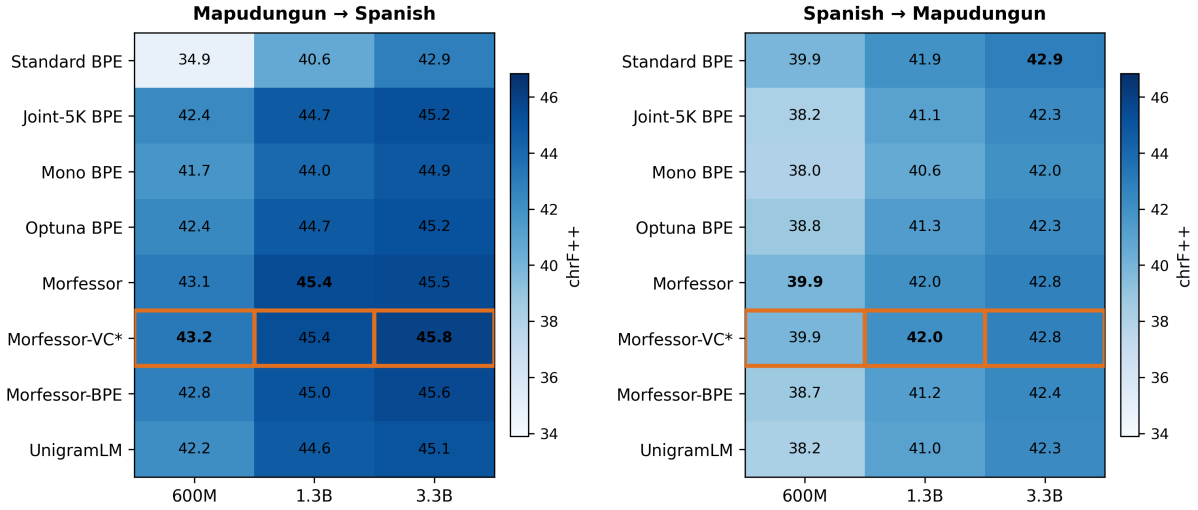

Figure 3: chrF++ by tokenization condition and model size for both translation directions. Darker cells indicate higher chrF++. Morfessor-VC achieves the highest arn→es scores at 600M and 3.3B; for es→arn, differences between conditions are smaller and Standard BPE is competitive at 3.3B.

Direction	Std BPE	Morfessor-VC	Tied
arn→es	19	19	12
es→arn	18	17	15

Table 4: Human preference rankings (out of 50 sentences per direction). Tied = annotator found the two outputs equally acceptable.

difference at the sentence level. The high tie rate (24% arn→es, 30% es→arn) further reflects that the two systems frequently produce equivalently adequate outputs. In this pilot study, chrF++ differences between systems did not translate to consistent human preference at the 3.3B scale, suggesting a possible ceiling effect or reference translation noise in the corpus; a larger study with multiple native-speaker annotators would be needed to draw firm conclusions about perceptual thresholds.

7 Analysis

7.1 Why Morfessor-VC Works

Figure 1 shows that Morfessor-VC has identical fertility to vanilla Morfessor (1.13), confirming that the VC step only removes boundaries—it never adds new ones.

The key mechanism is *double tokenization avoidance* (Table 2). When Morfessor splits two character sequences that together form a single NLLB vocabulary item, that boundary is spurious: NLLB's tokenizer will never produce it, so the corresponding embedding pair was never pretrained. By merging such adjacent morphemes, Morfessor-VC aligns pre-segmented tokens with embeddings NLLB actually learned.

Of the 87,382 boundaries introduced by Morfessor, the VC step removes 2,215 (2.5%)—confirming it is conservative. The VC advantage over vanilla Morfessor is 0.07 chrF++ at 600M ($p = 0.23$, n.s.) and 0.34 at 3.3B ($p < 0.001$). The primary driver at 600M is Morfessor-style morpheme segmentation; the VC refinement pays off more at larger scale where the model can exploit the improved embedding alignment. The broader finding—that any morphology-aware segmentation dramatically outperforms Standard BPE at 600M—is more robust than the VC-specific advantage.

Together with the Optuna BPE flatness across vocabulary sizes (1,431–14,517) and weak tokenization sensitivity in es→arn, these findings suggest that tokenization may not be the binding constraint at the current data scale—corpus size and domain narrowness may constitute harder ceilings.

7.2 Linguistic Analysis

A Mapudungun linguist annotated 20 Morfessor-VC outputs (10 per direction) for adequacy and naturalness in this pilot study. No systematic morphological assembly errors were observed in either direction, suggesting (on this small sample) that Morfessor-VC segmentation does not introduce boundary artifacts into the output.

arn→es. Outputs were generally adequate; in one case the model appeared to outperform the reference, retaining an adverb the transcriber omitted. A preliminary observation was that the model

tended toward narrow lexical interpretations of the polysemous noun *dungu* (language / thing / matter / way-of-being). Aspect errors (habitual vs. perfective) occurred occasionally.

es→arn. Several outputs were judged equivalent to or more concise than the reference, and the model sometimes preferred native vocabulary (e.g., *üytulafiñ*, "I will not name it") over calqued reference forms. One failure was severe: the model reproduced the Spanish source verbatim in Mapudungun orthography, a source-copy error likely triggered by high lexical overlap with mixed-language training sentences (Section 7.3). Repeated misuse of *chumuechi/chumngechi* ("how/in what way") suggests an overfit collocation pattern for this adverb.

7.3 Code-switching and es→arn Quality

Approximately 24% of Mapudungun utterances contain embedded Spanish words (ELAN <SPA> tags). This asymmetry explains why Standard BPE is relatively competitive for es→arn (Spanish input is well-covered by NLLB pretraining) and why the source-copy failure in Section 7.2 occurs only in that direction.

We used a word-level language identification model to correlate Spanish token proportion with per-sentence chrF++ from Morfessor-VC 3.3B on arn→es. The correlation is negligible (Pearson $r = +0.039$, $r^2 \approx 0.002$; Spearman $r = +0.051$); the $p < 0.001$ result reflects the large test set ($n = 9{,}382$) rather than practical effect size. Stratifying by code-switching level: no Spanish words averages 45.75 chrF++, light CS (1–20%) averages 45.11, heavy CS (>20%) averages 46.77. The uptick at heavy CS likely reflects Spanish loanwords and proper nouns that pass through unchanged, inflating character overlap with the reference. These results are post hoc and correlational; the model does not appear to substantially degrade on mixed-language arn→es input. A controlled future experiment—training conditions that tag or strip code-switched tokens—would directly measure the effect on es→arn quality, particularly the source-copy failure mode described in Section 7.2.

8 Conclusion

We presented the first systematic fine-tuning study of NLLB-200 on Mapudungun–Spanish translation, comparing eight tokenization strategies across three model sizes and both translation directions. Our proposed Morfessor-VC method—which constrains Morfessor segmentation to tokens present in NLLB's pretrained vocabulary—achieves the highest arn→es chrF++ at both 600M (43.2) and 3.3B (45.8), while matching Standard BPE at 3.3B using only a 600M model; the VC refinement over vanilla Morfessor is itself small ($\Delta \leq 0.34$), and the primary driver is morpheme-boundary presegmentation broadly. This 5× parameter efficiency gain—specific to the arn→es direction and significant but with a narrow margin ($p = 0.039$, 95% CI [0.02, 0.60])—suggests that morphology-aware tokenization can be a practical lever for compute-constrained deployment when translating from polysynthetic languages. For es→arn, tokenization choice has little impact and improving that direction—which produces Mapudungun output required for community use—is the direct priority for future work.

All fine-tuned conditions substantially outperform zero-shot NLLB and frontier LLMs, surpassing Aya Expanse 8B by over 27 chrF++ points and establishing new state-of-the-art baselines for Mapudungun MT. Linguistic analysis confirmed that Morfessor-VC does not introduce morpheme boundary artifacts, with the principal remaining error types being lexical (polysemy, collocation) rather than structural.

We release all models, tokenization code, and evaluation scripts to support future work on Mapudungun NLP and low-resource MT more broadly.

Several directions remain open. Improving es→arn—the direction that produces Mapudungun output, required for community use cases—is a direct priority; concrete next steps include back-translation augmentation, prefix-control for code-switching, and investigating whether corpus size or domain narrowness constitute harder ceilings than tokenization. Morfessor-VC is language-agnostic and applies to other polysynthetic languages in NLLB's coverage gap. Mapudungun is now included in the BOUQuET benchmark (Alastruey et al., 2026); we hope this work provides a technical foundation for future efforts pursued in coordination with Mapuche community priorities (Ahumada et al., 2022).

Limitations

Domain. All data comes from the AVENUE corpus, which consists of oral history interviews.

Translation quality on other domains (news, legal, educational texts) is unknown and may differ substantially.

Code-switching. The 24% code-switching rate in the AVENUE corpus is a property of this specific community's speech style. Our models may be poorly calibrated for monolingual Mapudungun text, and we have not evaluated on such data.

Reference quality. AVENUE translations were produced by human transcribers and may contain errors, as noted by our linguistic annotator in several cases. Automatic metrics evaluated against imperfect references will underestimate true translation quality.

Dialect variation. Mapudungun has significant dialectal variation across Chile and Argentina. The AVENUE corpus reflects a subset of Mapuche communities; performance on other dialects has not been assessed.

Evaluation metric. chrF++ is our primary metric, but it is a surface-level measure that does not capture morphological correctness or semantic adequacy directly.

Human and linguistic evaluation. Both our human preference evaluation (Section 6.1) and our linguistic analysis (Section 7.2) are pilot studies. The human evaluation was conducted by a single second-language learner of Spanish and Mapudungun (not a native speaker of either language); no inter-annotator agreement was measured. The linguistic annotation covers only 20 sentences. The divergence between chrF++ gains and human preferences may reflect a chrF++ ceiling effect at this performance level, reference translation noise in the AVENUE corpus, or both (see also the reference quality limitation above). A cleaner reference set would be needed to reliably discriminate between systems at the 3.3B scale. A stronger design would compare our best system against prior-work output, use multiple annotators including native Mapudungun speakers, and measure inter-annotator agreement. Preliminary observations in Section 7.2 (e.g., *dungu* polysemy errors, *chumuechi* overfitting) warrant confirmation on a larger annotated sample.

Missing ablations. We do not evaluate vocabulary-adaptation baselines such as WECHSEL (Minixhofer et al., 2022), which initializes new subword embeddings for cross-lingual transfer; this represents a complementary approach to Morfessor-VC and is a natural direction for future work. We also do not compare to parameter-efficient fine-tuning methods (e.g., LoRA) or to backtranslation-augmented training, both of which have shown benefits in indigenous-language MT (Ebrahimi et al., 2023).

Ethics Statement

This work develops machine translation technology for Mapudungun, an endangered indigenous language. We are attentive to the risk that MT systems can homogenize dialectal variation or be used to extract or commodify indigenous linguistic knowledge without community consent. The AVENUE corpus was collected through an academic collaboration involving Carnegie Mellon University, the Universidad de La Frontera (Temuco, Chile), and Chilean government partners. It is not missionary or religious data; it was collected with informed consent from Mapuche speakers and made publicly available by its creators (Duan et al., 2020).

Orthography. Mapudungun has multiple competing orthographic systems, including the Alfabeto Mapuche Unificado (AMU), Ragileo, and Azümchefe, each associated with different community and political positions. The AVENUE corpus uses the AMU orthography. Our models implicitly privilege AMU; community deployment should attend to which orthographic communities the models will serve and involve speakers of those communities in evaluation and adaptation.

Community engagement. We follow the CARE Principles for Indigenous Data Governance (Carroll et al., 2020) in spirit: data comes from a community-controlled corpus, we do not introduce new surveillance or extraction, and we release all models and code to support community-driven future work. We acknowledge that we did not establish direct partnerships with Mapuche governance bodies for this study, and encourage future work to do so. The ethical norms articulated by Mager et al. (2023) for indigenous-language MT research informed our approach.

Human evaluation. Human evaluation and linguistic annotation was conducted by a specialist collaborator who is a Mapudungun linguist.

We do not release any new primary data in this work.

Acknowledgments

We are grateful to the Mapuche speakers who contributed to the original corpus recordings, including Luis Caniupil, Flor Caniupil, Héctor Painequeo, Rosendo Huisca, and Hugo Carrasco. Compute resources were provided by the BYU Office of Research Computing. This work was supported by the BYU Computer Science Graduate Fellowship and generous donations to the BYU MAchine Translation Research and Interlingual eXperimentation (MATRIX) Lab.

References

Orevaoghene Ahia, Luke Bandarkar, Ife Henshaw, Sabrina Schneider, Sachin Kumar Singh, Antonios Anastasopoulos, David Mortensen, Graham Neubig, and Yulia Tsvetkov. 2023. All languages are NOT created (tokenized) equal. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14614–14627, Singapore. Association for Computational Linguistics.

Cristian Ahumada, Claudio Gutierrez, and Antonios Anastasopoulos. 2022. Educational tools for Mapuzugun. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 183–196, Seattle, Washington. Association for Computational Linguistics.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631. ACM.

Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, Artyom Kozhevnikov, Christophe Ropers, Eduardo Sánchez, Charles-Éric Saint-James, Ioannis Tsiamas, and 1 others. 2026. OmniLingual MT: Machine translation for 1,600 languages. *arXiv preprint arXiv:2603.16309*.

Duygu Ataman and Marcello Federico. 2018. Compositional representation of morphologically-rich input for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 305–311, Melbourne, Australia. Association for Computational Linguistics.

Sreejit Banerjee and Pushpak Bhattacharyya. 2018. Meaningless yet meaningful: Morphology grounded subword-level NMT. In *Proceedings of the 2nd Workshop on Subword/Character LEvels Models*, pages 55–60, New Orleans, Louisiana. Association for Computational Linguistics.

Luis Caniupil, Flor Caniupil, Héctor Painequeo, Rosendo Huisca, Hugo Carrasco, Rodolfo M. Vega, Lori Levin, and Jaime Carbonell. 2019. Mapudungun speech corpus.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE principles for indigenous data governance. *Data Science Journal*, 19(1):43.

Andrés Chandiá. 2022. A Mapudüngun FST morphological analyser and its web interface. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 6540–6547, Marseille, France. European Language Resources Association.

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.

Michael DeGenaro and Luke Lupicki. 2024. Low-resource machine translation for indigenous languages of the Americas. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 1–7, Mexico City, Mexico. Association for Computational Linguistics.

C.M. Downey, Terra Blevins, Nora Goldfine, and Shane Steinert-Threlkeld. 2023. Embedding structure matters: Comparing methods to adapt multilingual vocabularies to new languages. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 268–281, Singapore. Association for Computational Linguistics.

Mingjun Duan, Carlos Fasola, Sai Krishna Rallabandi, Rodolfo M. Vega, Antonios Anastasopoulos, Lori Levin, and Alan W. Black. 2020. A resource for computational experiments on Mapudungun. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2872–2877, Marseille, France. European Language Resources Association.

Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense, and Shruti Rijhwani. 2024. Findings of the AmericasNLP 2024 shared task on machine translation into indigenous languages.

In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 236–246, Mexico City, Mexico. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Peutêtre, Katharina Kann, and Alexis Palmer. 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 1–17, Toronto, Canada. Association for Computational Linguistics.

Edward Gow-Smith and Danaé Sánchez Villegas. 2023. Low-resource machine translation for low-resource languages: Corpus augmentation and pretrained model leverage. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 74–86, Toronto, Canada. Association for Computational Linguistics.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

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

Lori Levin, Rodolfo Vega, Jaime Carbonell, Ralf Brown, Alon Lavie, Eliseo Cañulef, and Carolina Huenchullan. 2002. Data collection and language technologies for Mapudungun. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Gran Canaria, Spain. European Language Resources Association.

Hernan Lira, Luis Martí, and Nayat Sanchez-Pi. 2025. Spanish–Mapudungun translation using transfer learning for low-resource languages. In *2025 15th IEEE International Conference on Pattern Recognition Systems (ICPRS)*, pages 1–7.

Manuel Mager, Rajat Bhatnagar, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2023. Neural machine translation for the indigenous languages of the Americas: An introduction. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 109–133, Toronto, Canada. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Elisabeth Mager, Katharina Kann, and Thang Vu. 2022. BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 961–971, Dublin, Ireland. Association for Computational Linguistics.

Sabrina J. Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *arXiv preprint arXiv:2112.10508*.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4007, Seattle, United States. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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.

María Begoña Pendas, Andrés Carvallo, and Carlos Aspillaga. 2023. Neural machine translation through active learning on low-resource languages: The case of Spanish to Mapudungun. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 6–11, Toronto, Canada. Association for Computational Linguistics.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. In *Advances in Neural Information Processing Systems*, volume 36.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Technical Report 25/2013 in Aalto University publication series SCIENCE + TECHNOLOGY, Aalto University.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, pages 1556–1559, Genoa, Italy. European Language Resources Association.

A Full BLEU Results

Table 5 reports BLEU (sacreBLEU, `13a` tokenizer) for all fine-tuned conditions across three model sizes and both translation directions.

B Full COMET Results

Table 6 reports COMET scores (wmt22-comet-da) for all fine-tuned conditions across three model sizes and both translation directions. *Note: the wmt22-comet-da model has not been validated on Mapudungun, which is absent from its training languages (based on XLM-R pretraining data). COMET scores for the Mapudungun side should be interpreted cautiously; we report them for completeness alongside the primary chrF++ results in the main text.*

Condition	arn→es			es→arn		
	600M	**1.3B**	**3.3B**	**600M**	**1.3B**	**3.3B**
Baselines (not fine-tuned)						
Zero-shot NLLB	1.68	1.94	2.16	0.32	0.64	0.68
Llama 3.1 8B		1.55			0.20	
Aya Expanse 8B		2.13			0.39	
Fine-tuned NLLB-200						
Standard BPE	10.24	15.83	16.78	8.53	**9.93**	**10.57**
Joint-5K BPE	14.20	15.91	17.20	7.92	9.02	9.73
Mono BPE	11.93	12.97	15.45	7.24	8.30	9.16
Optuna BPE	14.50	16.33	17.97	8.12	9.26	9.93
Morfessor	16.15	18.21	18.72	**8.61**	9.83	10.32
Morfessor-VC	**16.45**	**18.35**	**18.94**	8.50	9.84	10.32
Morfessor-BPE	14.74	15.98	17.79	8.00	9.12	9.93
UnigramLM	12.61	14.11	15.77	7.40	8.69	9.33

Table 5: BLEU on the test set (9,382 pairs). Bold = best in column among fine-tuned conditions. LLM baselines were 8B models evaluated with 5-shot prompting and are shown in a single merged column for clarity.

Condition	arn→es			es→arn		
	600M	**1.3B**	**3.3B**	**600M**	**1.3B**	**3.3B**
Baselines (not fine-tuned)						
Zero-shot NLLB	0.390	0.392	0.385	0.408	0.489	0.440
Llama 3.1 8B		0.483			0.484	
Aya Expanse 8B		0.509			0.497	
Fine-tuned NLLB-200						
Standard BPE	0.592	0.631	0.648	0.666	0.677	**0.681**
Joint-5K BPE	0.636	0.654	0.659	0.658	0.672	0.679
Mono BPE	0.628	0.644	0.653	0.658	0.670	0.677
Optuna BPE	0.638	0.654	0.662	0.660	0.672	0.679
Morfessor	0.644	0.660	0.664	0.666	0.677	0.680
Morfessor-VC	**0.644**	**0.661**	**0.665**	**0.667**	**0.677**	0.680
Morfessor-BPE	0.640	0.655	0.662	0.660	0.673	0.679
UnigramLM	0.633	0.650	0.657	0.657	0.672	0.677

Table 6: COMET (wmt22-comet-da) on the test set (9,382 pairs). Bold = best in column among fine-tuned conditions. LLM baselines were 8B models evaluated with 5-shot prompting and are shown in a single merged column for clarity.