Tokenization as Cultural Erasure: How Corpus Composition Shapes the Representation of Aymara Morphology in NLP Systems

Anonymous Authors[1]

Abstract

Tokenization is not a neutral operation for agglutinative languages whose morphology encodes culturally specific meaning. Aymara is spoken by roughly two million people in Bolivia, Peru, and Chile, and it grammatically encodes evidentiality, temporal orientation, and cosmological relationships through morpheme combinations with no equivalents in Spanish or English, and a tokenizer trained on complex surface forms can obscure these distinctions before downstream training begins. We present a controlled study of five SentencePiece tokenizers trained on linguistically stratified corpora of 17,856 Spanish-Aymara translation pairs, finding that the tokenizer trained exclusively on morphologically simple forms achieves the highest chrF at every evaluation level (17.01 ± 0.23 globally; 17.73 ± 0.40 on compositional structures) despite having the smallest vocabulary and highest fertility. We also show that a widely used morpheme integrity metric is systematically misleading for agglutinative languages, it assigns our best-performing tokenizer the lowest score precisely because correct boundary recovery breaks up the surface forms the metric rewards. We propose the Morphological Boundary Hypothesis and a design principle, morphological grounding as a concrete low cost intervention that protects compositional structure and improves the translation quality simultaneously.

1. Introduction

The decisions that go into building NLP infrastructure are not purely technical. Some of them carry cultural weight, even when they look routine. Tokenization is one of such decision.

For languages like English or Spanish, modest in morphology, enormous in corpus size, tokenization is invisible. A subword tokenizer trained on enough data tends to find reasonable boundaries, and nobody worries much about it. For Aymara, the situation is different in ways that matter. Consider *uñjatayna*. It encodes the act of seeing (*uñja-*) and, through the obligatory suffix *-tayna*, the speaker's epistemic stance: the event was not personally witnessed. This evidential system is not optional, it reflects a deep epistemological orientation in Aymara thought, one that grammatically distinguishes what you know from experience from what you heard from someone else (Hardman, 2001). A tokenizer that absorbs *-tayna* into a longer surface form removes it as a learnable unit. The distinction becomes substantially harder for the model to recover during training.

This paper argues that tokenizer corpus composition is an underappreciated site of cultural decision making for low-resource agglutinative languages. We frame this as an opportunity rather than just a critique. With modest effort, curating a morphologically simple subcorpus for tokenizer training, it is possible to build systems whose internal representations align with the compositional structure of the language, rather than overwriting it with the statistical regularities of surface frequency. And as we show, this also produces better translations. The tokenizer that better presents morphological structure also achieves the best downstream performance.

Contributions. We make three contributions. First, a controlled empirical study showing that a tokenizer trained on morphologically simple Aymara outperforms all alternatives at every linguistic level, with gains of $+1.63$ chrF globally and $+1.95$ on compositional structures over a full-corpus baseline. Second, and we think this deserves more attention than it usually gets, a methodological critique showing that the standard morpheme integrity metric actively rewards boundary erasure in agglutinative settings: our best performing tokenizer scores worst by this measure, an inversion that should give pause to anyone using this metric for similar lan-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

guages. Third, the Morphological Boundary Hypothesis and the design principle of morphological groundings, which reframe tokenizer construction as a culturally consequential act with a concrete, low-cost solution.

These results should be read as controlled evidence of a tokenization mechanism, not as a deployment ready system absolute quality is low, as we discuss in Section 6.

2. Aymara as a Cultural Technology

The workshop's framing of AI as an *interpretive technology* one engaged in acts of meaning-making, not just processing applies with unusual force to the preprocessing layer. Before a model interprets anything the tokenizer has already decided which units of meaning exist. For Aymara, those decisions carry consequences that go beyond benchmark accuracy.

2.1. Morphology as Cultural Encoding

Aymara belongs to the Aymaran language family, spoken primarily in Bolivia, Peru, and northern Chile. Its morphological structure encodes cultural content at three levels relevant to this study.

Agglutination. Words are formed by chaining suffixes onto roots in fixed positional classes. The word *munasiripuni* combines *munasi-* ('to want'), *-iri* (agentive nominalizer), and *-puni* (absolute certainty) three units with stable, well-defined boundaries. A tokenizer that collapses them into a single chunk makes it more difficult for downstream models to learn that *-iri* nominalizes agents or that *-puni* marks certainty as a general pattern. Those are learnable grammatical facts; erasing the boundaries makes them invisible.

Evidentiality. The suffix *-tayna* marks that the speaker reports an event they did not witness. Its use is grammatically obligatory in the relevant contexts not a stylistic choice (Hardman, 2001). Aymara grammar requires encoding the source of one's knowledge. An NLP system that cannot represent *-tayna* as an atomic unit may struggle to generalize this distinction across novel contexts.

Temporal cosmology. Núñez & Sweetser (2006) document that Aymara spatial metaphors for time reverse the dominant European convention: the known past is in front of the speaker, the unknown future behind. This is encoded grammatically through deictic morphology involving the root *pacha*, which participates in constructions encoding time, space, and world. *Pacha* is not a prefix of *pachamama* it is an independent, productive root. A tokenizer that fuses the two reduces their availability as independently reusable units.

2.2. Why Tokenization Is Particularly Problematic Here

SentencePiece (Kudo & Richardson, 2018) and BPE (Sennrich et al., 2016) learn vocabularies from corpus statistics. In large corpora, morphemes appear frequently enough as standalone forms to survive in the vocabulary. Aymara is largely absent from multilingual pretraining corpora (Joshi et al., 2020): with limited data, frequent surface forms which in Aymara are often multi-morpheme compounds dominate the learned vocabulary.

Fertility (average tokens per word) is the standard proxy for tokenizer quality (Rust et al., 2021), with lower values preferred. But fertility primarily reflects compression efficiency rather than linguistic structure, not linguistic structure. A tokenizer that represents *pachamama* as one token achieves low fertility at the cost of erasing the *pacha–mama* boundary. Our results show this tradeoff matters: the highest-fertility tokenizer produces the best translation, because its extra tokens correspond to actual morpheme boundaries, not arbitrary splits.

3. Experimental Design

3.1. Dataset

Our dataset consists of 17,856 Spanish–Aymara translation pairs constructed from three sources: pairs from the NLLB corpus (Team et al., 2022), back-translated pairs generated from NLLB using automatic translation, and pairs from the Global Voices NLP dataset (Nguyen & Daumé III, 2019). All sources are publicly available under open licenses. The dataset was annotated by linguistic complexity:

Table 1. Dataset composition by linguistic complexity level.

LEVEL	DESCRIPTION	N
L1	LEXEMES AND ROOT FORMS	4,416
L2	BASIC MORPHOLOGICAL INFLECTIONS	66
L3	COMPOSITIONAL MULTI-MORPHEME STRUCTURES	4,653
L4	FULL SENTENCE PAIRS	8,721

L2 yields only ≈6 test items under a 10% split, so we merge L1 and L2 into **L1+L2: Basic Morphology** ($n = 428$ test examples). All conditions and seeds share a single fixed split: 80% train (14,284 pairs), 10% validation, 10% test. We note that a portion of the dataset consists of back-translated pairs, which may introduce noise; we discuss this limitation in Section 6.

3.2. Tokenizer Conditions

We train five SentencePiece Unigram tokenizers, varying only training corpus composition. All other hyperparameters are fixed: character coverage 1.0, byte fallback, digit splitting, identity normalization.

Table 2. Tokenizer conditions. Vocabulary sizes scale with corpus character diversity to avoid training failures on small corpora.

ID	CORPUS COMPOSITION	VOCAB
T0	FULL CORPUS, UNWEIGHTED	8,100
T1	L1+L2 ONLY (MORPHOLOGICALLY SIMPLE)	3,580
T2	L3+L4 ONLY (MORPHOLOGICALLY COMPLEX)	8,100
T3	60% L1+L2, 25% L3, 15% L4	6,660
T4	15% L1+L2, 25% L3, 60% L4	8,100

3.3. Model and Evaluation

All conditions use an identical T5 model trained from scratch: 512d embeddings, 2,048 feed-forward width, 4 encoder and 4 decoder layers, 8 attention heads, 0.1 dropout. AdamW optimizer (lr $= 1\times10^{-4}$, cosine schedule, 8% warmup), batch size 8 with 4-step gradient accumulation, up to 30 epochs with early stopping (patience 5). We run three random seeds (42, 123, 777) per tokenizer, yielding 15 training runs total.

Primary metric: chrF (Popović, 2015). We prefer chrF over BLEU (Papineni et al., 2002) because its character-level matching is more sensitive to morphological variation in agglutinative output—a single incorrect suffix degrades chrF less catastrophically than a BLEU n-gram match.

4. Results

4.1. Downstream Translation Quality

Table 3. chrF (mean ± std, 3 seeds) by tokenizer and evaluation stratum. T1 leads at every level; gains reported over T0.

TOK.	GLOBAL	L1+L2	L3	L4
T0	15.38±0.18	9.47±0.54	15.78±0.26	15.42±0.13
T1	**17.01±0.23**	**10.83±0.93**	**17.73±0.40**	**16.94±0.18**
T2	15.56±0.54	10.11±0.84	15.69±0.51	15.67±0.58
T3	15.79±0.26	10.74±0.86	16.20±0.27	15.81±0.27
T4	15.65±0.22	10.08±0.60	15.68±0.48	15.81±0.34

T1 leads at every stratum across all three seeds: +1.63 chrF globally, +1.36 on basic morphology, +1.95 on compositional structures, +1.52 on full sentences. Standard deviations are small (0.18–0.40), suggesting the advantage is stable. T1 also achieves the lowest validation loss by a large margin (best val. loss ≈3.31–3.33 vs. 5.00–5.37 for all other tokenizers), suggesting systematic training differences across conditions.

T3 (60% simple forms) ranks second at every stratum, consistent with a gradient relationship between simple-form exposure and downstream quality. T2 (complex only) does not outperform T0 despite matching downstream data complexity—ruling out complexity-matching or corpus size as the explanation.

4.2. A Broken Metric: Why Standard Evaluation Misleads

Table 4. Tokenizer fertility and morpheme integrity. T1 ranks last on both standard metrics yet produces the best downstream results.

TOKENIZER	VOCAB	FERT.	MORPH. INT.
T0 BASELINE	8,100	2.33	92.71%
T1 EASY ONLY	3,580	3.66	82.91%
T2 HARD ONLY	8,100	2.34	92.25%
T3 CURR. PROG.	6,660	2.46	92.29%
T4 CURR. INV.	8,100	2.35	92.27%

By conventional evaluation, T1 should be the worst tokenizer: highest fertility, smallest vocabulary, lowest morpheme integrity. The downstream results contradict this at every stratum. The morpheme integrity figure is worth examining carefully. The metric computes the proportion of known Aymara morpheme strings that appear intact as substrings within a single token. When T1 correctly splits *pachamama* into [*pacha* | *mama*], both roots lose their status as intact substrings within one token the metric counts this as failure. T0, by fusing both into one token, scores well because *-ma*, *pa-*, and *-cha-* are all present inside that chunk.

The metric appears to reward boundary fusion more than boundary recovery. For agglutinative languages this may invert the intended evaluation singal, and it has a practical consequence: if researchers use morpheme integrity to select tokenizers for other low-resource agglutinative languages, they may be systematically choosing worse ones.

4.3. Qualitative Segmentation Analysis

Table 5 shows token counts for ten words selected for morphological density and cultural significance evidential suffixes (*-tayna*), purposive constructions (*-taki*), agentive nominalization (*-iri*), absolute certainty marking (*-puni*), and cosmological roots (*pacha*, *mama*).

Table 5. Tokens per word by tokenizer. The *Morph* column gives the linguistically correct morpheme count.

WORD	MORPH	T0	T1	T2	T3	T4
PACHAMAMARU	3	2	**3**	2	4	2
SARAÑATAKI	3	1	4	1	2	1
JIWASANAKARU	3	2	4	2	2	2
UÑJATAYNA	2	2	3	2	2	2
MUNASIRIPUNI	3	3	4	3	3	3
APSUWAYI	3	2	3	2	2	2
YATIQAÑKAMA	3	2	4	2	2	2
ARUSKIPASIÑA	3	2	2	2	2	2
JAKAÑATAKIWA	3	2	5	2	2	2
AYNACHARAYANI	3	3	3	3	3	3
DEMO FERT.		2.1	3.5	2.1	2.4	2.1

The clearest case is *pachamamaru*. T0, T2, and T4 produce [*pachamama* | *ru*]: both roots absorbed into an opaque unit, making *pacha* less directly available as a reusable unit. T3 recovers the *pacha–mama* boundary but incorrectly splits *mama* into [*ma* | *ma*]. T1 alone produces [*pacha* | *mama* | *ru*] the correct three-way segmentation.

The reason traces directly to corpus composition. T0, T2, and T4 trained on corpora where *pachamama* is a frequent surface form; the Unigram algorithm learns to compress it. T1 trained on L1+L2, where *pacha* and *mama* appear mostly as independent roots, so both enter the vocabulary as standalone units. When *pachamamaru* appears downstream, T1 segments by composition rather than lookup. T3's partial recovery reflects its 60% simple-form exposure: enough to stabilize *pacha*, not quite enough for *mama*.

One likely downstream consequence is *Pacha*, that participates in many productive constructions beyond *pachamama*. A model trained on T1-tokenized data can learn that *pacha* carries temporal-spatial meaning across contexts; a model trained on T0-tokenized primarly encounters *pacha* within larger fused units. For *sarañataki*, T0, T2, and T4 produce a single token, making the purposive suffix *-taki* and infinitive *-ña* invisible as morphological units. For *uñjatayna*, no tokenizer achieves a fully correct segmentation, but T1 at least recovers *uñja* as a recognizable root rather than producing [*uñj* | *atayna*], which fuses the evidential suffix to a root fragment.

4.4. The Morphological Boundary Hypothesis

Definition 4.1 (Morphological Boundary Hypothesis)**.** For agglutinative languages, a tokenizer trained on morphologically simple forms learns roots and suffixes as independent vocabulary units before encountering them in combination. When complex words appear in downstream data, this tokenizer segments by composing those units, placing boundaries at morphologically correct positions. The resulting representations expose the compositional structure of the language to the downstream model, enabling it to learn morphological regularities—evidential contrasts, purposive constructions, agentive patterns—that are invisible under vocabularies learned from morphologically mixed corpora.

Three clarifications prevent overreading. T1's advantage is a corpus composition effect, not a vocabulary size effect: T2 uses the same 8,100-token target and performs no better than T0. T1's high fertility reflects correct boundary placement, not arbitrary fragmentation: on words where all tokenizers agree (*aruskipasiña*, *aynacharayani*), T1 produces the same count. The hypothesis predicts the advantage should scale with morphological productivity—languages with sparser agglutination should show smaller effects. Testing across Quechua, Turkish, Finnish, and Nahuatl is a natural next step.

5. Tokenization as a Cultural Design Decision

The standard NLP practice is to train a tokenizer on all available data and treat this as a solved preprocessing step. For agglutinative languages with culturally encoded morphology, our results show this default has a specific failure mode: it tends to prioritize compression of frequent surface forms over explicit preservation of morphological structure. And it does so quietly, at a layer most practitioners never revisit.

Morphological grounding as a tokenizer design principle- The workshop asks how humanistic traditions of meaning-making can be embedded in AI design rather than applied after the fact. Morphological grounding is one concrete answer: it encodes, at the vocabulary level, a commitment to the compositionality that Aymara grammar actually expresses. The tokenizer becomes an interpretive technology that takes the structure of the language seriously one that reflects how Aymara speakers actually build meaning, rather than flattening that structure to serve compression efficiency.

Crucially, the cultural argument and the empirical argument converge. The tokenizer that better preserves Aymara morphological structure also produces better translations. This is not always the case when humanistic values are embedded into engineering decisions; here it is.

What gets erased under the default. The evidential suffix *-tayna*—grammatically obligatory when reporting non-witnessed events—is absorbed into longer tokens by T0, T2, T3, and T4. A model trained on this tokenization can at best memorize surface co-occurrences; it cannot generalize the evidential paradigm across novel roots. For speakers whose grammar encodes the source of knowledge as an obligatory distinction, this may affect the model's ability to generalize evidential distinctions.

The intervention is computationally inexpensive. Morphological grounding requires curating a morphologically simple subcorpus for tokenizer training only—the downstream model still trains on the full dataset. In our experiment, this means L1+L2 data (7,224 texts), available from existing lexical resources for many agglutinative languages or constructible at modest annotation cost. The gain is up to $+1.95$ chrF on the structures that most require morphological competence, across 15 controlled runs.

Tensions. Better NLP infrastructure for Aymara is not an unambiguous good, and we want to say that clearly. Improved translation systems can support information access, but they can also enable surveillance, commodify language data, or displace human translators whose work carries cultural authority that automated systems cannot replicate. Morphological grounding is infrastructure, not a solution. Meaningful NLP for Aymara ultimately requires Aymara-speaking communities as collaborators at every stage—not only as sources of data.

6. Limitations

Absolute quality is low. BLEU scores range from 1.82 to 2.01 and chrF from 15.38 to 17.01—consistent with comparable very low-resource translation settings (Anastasopoulos & Neubig, 2020; Gu et al., 2018; Neubig & Hu, 2018). These results are controlled evidence of a tokenization mechanism, not a deployment-ready system.

Three seeds. With $n = 3$, formal significance testing is not appropriate. Standard deviations (0.13–0.93 chrF) are consistently small relative to observed differences, supporting stable tendency claims.

Back-translated data. A portion of our dataset consists of back-translated pairs generated automatically from NLLB. This may introduce noise or systematic translation artifacts that could affect results. Future work should evaluate on datasets constructed entirely from human-translated pairs.

No pretrained baseline. Training from scratch isolates the tokenizer effect cleanly. Comparison with NLLB-200 or mBART-50 is planned for future work, particularly to see whether morphological grounding still matters when a pretrained model brings its own subword vocabulary.

Qualitative segmentation only. Formal boundary precision and recall require a gold-standard Aymara morphological lexicon that does not yet exist for this dataset. Human evaluation by Aymara morphologists is an essential next step.

Single language and architecture. Generalization to Quechua, Turkish, Finnish, and other agglutinative languages is an open empirical question.

7. Conclusion

Tokenization defines the initial representational units available to an NLP systems. For Aymara, we have shown that the corpus used to train the tokenizer determines whether culturally encoded morphological content—evidential suffixes, cosmological roots, purposive constructions—enters as learnable units or as substrings remain available as reusable units or appear primarly within fused surface forms. The difference is measurable and consistent: across 15 controlled training runs, the tokenizer built from morphologically simple forms produces gains of up to $+1.95$ chrF over a full-corpus baseline, and is the only one to correctly segment *pachamamaru* into its three constituent morphemes.

We have also shown that the standard morpheme integrity metric actively misleads in agglutinative settings, rewarding boundary erasure and penalizing recovery. This is a practical problem for anyone evaluating tokenizers in similar languages.

The Morphological Boundary Hypothesis and morphological grounding follow from these findings. The core claim is simple: for languages whose morphology encodes meaning that is culturally specific and linguistically productive, build the vocabulary from forms that expose that morphology, not compress it. Our experiments suggest that this design choice can be implemented with relatively low additional cost. What it cannot do, on its own, is make a system serve Aymara speakers well. That requires the language community's participation in defining what serving well even means. That is harder work and more important.

Acknowledgements

Impact Statement

This paper presents work on tokenization for Aymara, a low-resource indigenous language spoken primarily in Bolivia, Peru, and Chile. The primary goal is to improve NLP infrastructure for underrepresented language communities. However, we recognize that improved NLP tools for indigenous languages carry real risks: they can enable surveillance of speaker communities, commodify language data without community benefit, or displace human translators and language workers whose expertise carries cultural authority beyond what automated systems can replicate. The morphological grounding approach we propose is a technical contribution, not a substitute for community-centered language technology development. We strongly encourage future work in this space to be conducted in direct collaboration with Aymara-speaking communities as co-designers, not only as data sources.

References

Anastasopoulos, A. and Neubig, G. Should all cross-lingual embeddings speak english? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7214–7225, 2020.

Gu, J., Hassan, H., Devlin, J., and Li, V. O. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 344–354, 2018.

Hardman, M. J. *Aymara: Compendio de estructura fonólogica y gramatical*. Instituto de Lengua y Cultura Aymara, La Paz, 2001.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6282–6293, 2020.

Kudo, T. and Richardson, J. 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 (EMNLP)*, pp. 66–71, 2018.

Neubig, G. and Hu, J. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 875–880, 2018.

Nguyen, K. and Daumé III, H. Global voices: Crossing borders in automatic news summarization. *arXiv preprint arXiv:1910.00421*, 2019.

Núñez, R. and Sweetser, E. With the future behind them: Convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial constructs of time. *Cognitive Science*, 30(3):401–450, 2006.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 311–318, 2002.

Popović, M. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pp. 392–395, 2015.

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S., and Gurevych, I. 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 (ACL-IJCNLP)*, pp. 4866–4892, 2021.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1715–1725, 2016.

Team, N., Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

A. Full Segmentation Detail

Complete segmentations for the ten demo words, with · as boundary marker. Bold marks the only correct three-way segmentation of *pachamamaru*.

Table 6. Segmentations produced by each tokenizer for the ten demo words.

Word	T0	T1	T2	T3	T4
pachamamaru	pachamama·ru	**pacha·mama·ru**	pachamama·ru	pacha·ma·ma·ru	pachamama·ru
sarañataki	sarañataki	sara·ña·ta·ki	sarañataki	sar·añataki	sarañataki
jiwasanakaru	jiwasa·nakaru	jiwa·sa·naka·ru	jiwasa·nakaru	jiwasa·nakaru	jiwasa·nakaru
uñjatayna	uñj·atayna	uñja·tay·na	uñj·atayna	uñj·atayna	uñj·atayna
munasiripuni	munasi·ri·puni	mun·asiri·pu·ni	munasi·ri·puni	munasi·ri·puni	munasi·ri·puni
apsuwayi	apsu·wayi	apsu·way·i	apsu·wayi	apsu·wayi	apsu·wayi
yatiqañkama	yatiqañ·kama	yati·q·añ·kama	yatiqañ·kama	yatiqañ·kama	yatiqañ·kama
aruskipasiña	aruskip·asiña	aruskip·asiña	aruskip·asiña	aruskip·asiña	aruskip·asiña
jakañatakiwa	jakaña·takiwa	jak·aña·ta·ki·wa	jakaña·takiwa	jakaña·takiwa	jakaña·takiwa
aynacharayani	aynacha·r·ayani	aynach·araya·ni	aynacha·r·ayani	aynacha·r·ayani	aynacha·r·ayani

B. Raw Results by Seed

Table 7. Per-seed chrF for all tokenizer conditions. Means and standard deviations in Table 3 are computed from these values.

Tokenizer	Seed	chrF global	chrF L1+L2	chrF L3	chrF L4
T0	42	15.315	8.988	15.670	15.393
T0	123	15.244	9.360	15.588	15.312
T0	777	15.577	10.054	16.076	15.563
T1	42	17.195	9.756	18.106	17.087
T1	123	16.758	11.422	17.303	16.740
T1	777	17.077	11.325	17.773	16.999
T2	42	14.940	10.927	15.101	15.003
T2	123	15.904	10.165	16.009	16.042
T2	777	15.829	9.243	15.966	15.956
T3	42	16.042	9.746	16.385	16.088
T3	123	15.806	11.231	16.331	15.793
T3	777	15.518	11.251	15.889	15.543
T4	42	15.833	10.224	15.297	16.183
T4	123	15.717	9.427	16.214	15.727
T4	777	15.405	10.596	15.530	15.510