

A Human-Translated FLORES+ Dataset for Chakma Language

Firoz Ahmed

Department of Linguistics
University of Florida
Gainesville, Florida, USA
firozahmed@ufl.edu

Dinalo Chakma

Department of English
University of Florida
Gainesville, Florida, USA
dinalo.chakma@ufl.edu

Abstract

We present a human-translated, native-script Chakma (ccp_Cakm) extension to FLORES+. The corpus comprises the 997-sentence dev and 1,012-sentence devtest English splits, for 2,009 records in total. We recruited three native Chakma speakers through a screening task, conducted six virtual training sessions, and required direct translation without machine-translation or language-model output. A native reviewer checked meaning, grammar, script, terminology, named entities, numbers, and punctuation, returning problematic items for revision. We describe the Bangladeshi Chakma variety used and our orthographic and Unicode decisions. We then report a completed release audit: FLORES+ identifiers were recovered from the canonical English order and renumbered to the 1-based scheme, all 2,009 records were verified one-to-one against the English source, text was normalised to Unicode NFC, a character inventory was produced, and all twenty-seven residual non-Chakma characters were located and resolved. The package passes every check with no open defects, is fixed by SHA-256 checksums, and has been submitted to FLORES+ through the Open Language Data Initiative under CC BY-SA 4.0. The work provides a native-script evaluation resource for a language whose existing computational resources frequently rely on Bengali-script transliteration.

1 Introduction

Machine-translation evaluation remains difficult for languages that lack trusted, independently produced test data. FLORES-101 established a professionally translated, many-to-many benchmark for 101 languages (Goyal et al., 2022), and the No Language Left Behind work expanded this framework to 200 languages (NLLB Team, 2022). FLORES+ continues that effort through the Open Language Data Initiative (OLDI), which supports community

contributions for languages and varieties not yet adequately represented (Burchell et al., 2024; Dale et al., 2025).

Chakma is a clear case for this kind of extension. It has its own Unicode-encoded writing system, but most computational work involving Chakma text has used Bengali-script transliteration or has focused on tasks other than native-script machine-translation evaluation. A benchmark written in the Chakma script is therefore necessary to distinguish translation ability from transliteration ability.

We contribute a human translation of the English FLORES+ dev and devtest material into the Chakma script, together with a documented workflow for translator selection, translation, adjudication, and release validation. This work was prepared for the OLDI open-data shared task at WMT 2026, and the dataset contribution has been submitted to the FLORES+ repository.¹ We make three contributions. First, we contribute native-script Chakma translations for the 2,009-sentence target corpus, aligned to the canonical FLORES+ identifiers and fixed by SHA-256 checksums. Second, we document a community-based workflow that excludes machine-generated translation output. Third, we describe the orthographic and quality-control decisions, and report the concrete defects that a systematic release audit uncovered, that must accompany a benchmark for a language with limited digital infrastructure.

This paper focuses on the resource and its preparation. We do not present a new MT model, and we do not use automatic model scores as a substitute for human review. Instead, we report the available evidence and state the limitations directly.

¹https://huggingface.co/datasets/openlanguedata/flores_plus/discussions/44

2 Chakma and Existing Resources

2.1 Language and writing system

Chakma is identified by ISO 639-3 code *ccp* and Glottocode *chak1266* (Glottolog, 2026). Bangladesh’s 2022 census reports 483,299 Chakma people in the country (Bangladesh Bureau of Statistics, 2022). Chakma communities are concentrated in the Chittagong Hill Tracts and also reside in parts of India and Myanmar. Sociolinguistic work documents sustained pressure from nationally dominant languages and limited institutional space for Indigenous languages in Bangladesh (Sultana, 2023).

The Chakma script is Brahmi-derived and occupies Unicode block U+11100–U+1114F (Unicode Consortium, 2025). We use ISO 15924 code *Cakm* and the OLDI dataset label *ccp_Cakm_chak1266*. The dedicated script matters methodologically: a system that produces Bengali-script Chakma or romanized Chakma has not demonstrated native-script generation, even if a reader can recover part of the intended content.

Our translations represent the Bangladeshi variety used by contributors from the Chittagong Hill Tracts. We do not claim that every spelling choice generalizes to Indian or Myanmar Chakma varieties. Where variants occurred, translators discussed them in a shared glossary and selected forms that were familiar in Bangladeshi Chakma writing. The release should therefore be interpreted as one documented variety rather than as a universal orthographic standard.

2.2 Related computational work

ChakmaNMT introduces Chakma–Bangla parallel and monolingual data and a Chakma–Bangla–English benchmark, using transliteration to exploit Bangla resources (Chakma et al., 2024). MELD provides sentence-level Chakma, Garo, and Marma material in Bengali script with Standard Bangla and English translations (Mahi et al., 2025). ChakmaBridge expands MELD into an 807-sentence, five-way corpus spanning English, Standard Bangla, Bengali-script Chakma, Romanized Bangla, and Romanized Chakma (Rahman et al., 2025). Khisa et al. (2025) use native-speaker-validated, Bangla-transliterated Chakma literature for masked-language-model adaptation. These resources demonstrate the utility of transliteration, but do not replace native-script evaluation.

Other work contributes complementary in-

Language	Chakma (<i>ccp</i> ; <i>chak1266</i>)
Script	Chakma (<i>Cakm</i>), U+11100–U+1114F
OLDI label	<i>ccp_Cakm_chak1266</i>
Source	English FLORES+
Splits	<i>dev</i> 997; <i>devtest</i> 1,012
Total	2,009 sentences
Contributors	3 translators, 1 native reviewer
MT/LLM output	Prohibited at every stage
License	CC BY-SA 4.0, via OLDI

Table 1: Dataset and workflow summary.

frastructure. DChakma develops part-of-speech tagging resources (Apu et al., 2025), and Self-ChakmaNet addresses handwritten-character recognition with a 47-class dataset (Podder et al., 2023). Recent language-identification studies recognize Chakma among several Bangladeshi languages from text or speech (Ahmod et al., 2025; Khan et al., 2025; Anik et al., 2026; Nirob et al., 2026). These tasks differ from translation evaluation: recognizing a language or its characters does not show that an MT system preserves sentence meaning in the native script. Ahmed et al. (2024) show how an OLDI language contribution can pair careful human data creation with transparent validation. Our work follows that resource-building tradition while focusing on native-script Chakma evaluation data.

3 Dataset Construction

3.1 Source data and scope

We translate from the English FLORES+ material rather than from Bengali or another intermediary language. The English source files contain 997 *dev* sentences and 1,012 *devtest* sentences, for a target total of 2,009. Direct English-to-Chakma translation avoids adding an uncontrolled pivot stage and retains compatibility with multilingual FLORES evaluation. Reviews of low-resource Indic parallel corpora likewise identify alignment quality, script variation, and domain representativeness as central data concerns (Raja and Vats, 2025). Table 1 summarises the resource.

3.2 Translator recruitment

We required native Chakma proficiency, literacy in the Chakma script, strong English proficiency, and prior translation or language-work experience. Four candidates completed a 20-sentence screening task. We assessed source fidelity, grammaticality, spelling consistency, and control of the script. We selected three contributors: two undergraduate

translators with approximately two years of experience and one senior Chakma language researcher with more than fifteen years of experience.

We conducted six virtual sessions. The first sessions introduced the purpose of evaluation data and the FLORES and OLDI requirements. Later sessions reviewed examples, named-entity handling, punctuation, numerals, and recurring spelling decisions. All three translators gave informed consent and were compensated for their translation work.

3.3 Translation protocol

We required direct human translation. Translators did not use, consult, or post-edit machine-translation or large-language-model output. This rule follows OLDI guidance for evaluation data, where model-generated text would contaminate the reference and compromise later evaluation (Open Language Data Initiative, 2026). The restriction applies to the dataset translations; our use of generative tools for manuscript editing is disclosed in the Acknowledgments.

Translators were instructed to preserve meaning, tone, register, units, and discourse function. Personal names were transliterated into Chakma script. Organizations were translated when a familiar Chakma form existed and transliterated otherwise. Translators could consult dictionaries and the shared glossary, but not machine-generated candidate translations. We used the Chakma Keyman keyboard for text entry.

Working documents used sequential labels written in Chakma digits to support discussion. These display labels are not release metadata. For the archival package, we discarded them and recovered alignment from the canonical English sequence, as described in Section 4.2.

3.4 Orthography and text representation

We use native Chakma Unicode characters and Bangladeshi spelling preferences established through contributor discussion. We retain conventional punctuation and permit non-Chakma characters where the source requires an acronym, symbol, scientific expression, or other conventional form. Consequently, script validation flags characters for review rather than deleting every character outside the Chakma block. This distinction proved to be productive: the audit in Section 4.2 separated 79 records whose Latin characters were justified acronyms from 19 records

Check	dev	devtest
Records	997	1,012
Aligned 1:1 to English	✓	✓
Identifiers 1-based, contiguous	✓	✓
Non-empty	✓	✓
Unicode NFC	✓	✓
No control characters	✓	✓
No duplicate targets	✓	✓
Metadata constant	✓	✓
Chakma-script characters	98.6%	98.5%
Characters mechanically repaired	10	11
Records revised by native reviewer	1	0
Records with open defects	0	0

Table 2: Release audit of the 2,009-record package. Script percentages are computed over non-space characters.

whose non-Chakma characters were genuine defects.

4 Quality Assurance and Validation

4.1 Native-speaker review

The second author, Dinalo Chakma, a PhD student in English who is proficient in both Chakma and English, reviewed the dataset against its English source. The review covered meaning preservation, omission and addition, grammar, spelling, punctuation, terminology, named entities, numerals, units, and register. Items that failed a criterion were returned to the original translator with an edit request and reviewed again after revision.

We performed this review over the full target corpus rather than relying only on spot checks. However, one reviewer conducted the adjudication, and we did not compute a FLORES-style sampled quality score or inter-annotator agreement. We therefore treat the evidence as procedural and human-reviewed, not as a quantified estimate of error rate.

4.2 Release audit

We ran a record-level audit over the complete package; Table 2 reports the outcome. The audit was scripted and re-runnable.

Three findings are worth reporting. First, the working export numbered records from zero, whereas FLORES+ identifiers are 1-based. Because the working files preserved the canonical English order, we recovered the correct identifiers by renumbering positionally against the English source and re-verified the result record by record; the sample in Section 5 carries its recovered identifier.

Second, the character inventory located 27 non-Chakma Indic characters in 19 records. Of these, 21 were mechanical defects with unambiguous repairs: a Devanagari DANDA (U+0964) used as a sentence terminator in place of the Chakma DANDA (U+11141), a Bengali letter RA WITH MIDDLE DIAGONAL (U+09F0) substituted for Chakma LETTER RAA (U+11122), and a spurious Bengali NUKTA (U+09BC) attached to Chakma vowel signs. All three are keyboard artefacts rather than translation choices, and we corrected them. The remaining six characters formed a single untranslated Bengali word in one dev record. We referred it back to the native reviewer rather than substituting a form on our own authority; he supplied a Chakma rendering, which the released package carries. No record in the corpus now holds an unresolved defect.

Third, the Latin characters that script validation flagged were overwhelmingly justified: they occur in retained acronyms such as *AOL*, *RSPCA*, *COVID*, *ISIL*, and *pH*, consistent with the policy in Section 3.4. This is the concrete reason a script validator should flag rather than strip characters outside the target block.

The audit also confirmed that the independent plain-text and JSONL copies of the corpus are byte-identical after normalisation, which closes out the divergence that had accumulated across collaborative editing. We computed SHA-256 checksums for the final files so that any later divergence is detectable.

4.3 Experimental validation

We do not report an MT-system baseline. Existing Chakma systems are centred on transliteration or Bengali-script resources, and a score from such a pipeline would not by itself validate native-script references. We also avoid presenting informal outputs from commercial systems as a reproducible experiment. The first post-release experiment should evaluate a documented baseline in both directions with chrF and spBLEU, alongside a script-validity diagnostic and native-speaker error analysis.

5 Data Sample

Table 3 shows record 6 of the dev split with its recovered FLORES+ identifier. Each released record is a JSON object carrying the twelve FLORES+ fields: `id`, `iso_639_3`, `iso_15924`, `glottocode`, `variant`, `text`, `url`, `domain`,

<code>id</code>	6
<code>English</code>	28-year-old Vidal had joined Barça three seasons ago, from Sevilla.
<code>Chakma</code>	২৮ বছর বয়সে ৩ বছর আগের সেভিলা থেকে বার্সেলোনাতে যোগ দিয়েছেন
<code>domain</code>	wikinews
<code>topic</code>	sports

Table 3: A released `ccp_Cakm` record, shown with its recovered FLORES+ identifier.

`topic`, `has_image`, `has_hyperlink`, and `last_updated`.

6 Discussion

A native-script benchmark changes what can be measured. It allows researchers to test whether a model produces Chakma characters, handles Chakma orthography, and preserves content without relying on a Bengali-script intermediary. It also makes errors more interpretable to literate community members who should participate in future evaluation and correction.

Our audit experience suggests a more general point for script-first contributions. The defects that survived careful human review were not translation errors; they were encoding artefacts introduced by input tooling, and they were invisible to a reader because the substituted characters are visually close to their Chakma counterparts. A character-level inventory over the whole corpus, rather than a sampled proofread, is what surfaced them. We would encourage other OLDI contributors working in less-supported scripts to treat that inventory as a required release step.

The resource should not be interpreted as training data or as a complete representation of Chakma variation. FLORES+ is an evaluation benchmark with a limited number of topic-diverse sentences. Its value lies in consistent multilingual alignment and careful reference translation. Future work should add independently reviewed Seed data, document Indian and Myanmar varieties, and compare native-script evaluation with transliteration-mediated pipelines.

7 Conclusion

We have described the preparation of a human-translated, native-script Chakma extension to FLORES+. Three native speakers translated the English dev and devtest material under a workflow that prohibited machine-generated translation out-

put, and a native reviewer adjudicated the translations. We document the language variety, writing system, orthographic decisions, and quality checks, and we report a completed release audit in which the canonical 2,009-record package was aligned to the FLORES+ identifiers, normalised, corrected, and fixed by SHA-256 checksums. The data provide a reproducible basis for evaluating machine translation into and out of native-script Chakma.

Limitations

Our quality evidence is based on one native reviewer and does not include inter-annotator agreement or a sampled numerical quality score. We have not evaluated a native-script MT baseline, so the paper does not establish the current performance of available systems. The contributors represent Bangladeshi Chakma, and the data should not be assumed to cover Indian or Myanmar orthographic preferences.

Our release audit is also mechanical in scope. It verifies alignment, encoding, identifiers, and script validity, and it caught defects that careful human proofreading had missed; it cannot establish that a sentence is semantically faithful or idiomatic. That judgement rests entirely on the native review described in Section 4.1, and therefore inherits the single-reviewer caveat above.

Ethics Statement

The translators and reviewer are members of the language community represented by the resource. The three translators received information about the dataset’s purpose, intended public release, and use for MT evaluation; they gave informed consent and were compensated for their work. We do not release private translator information with the sentences. The source material contains publicly available content and named entities inherited from FLORES+. The dataset contribution has been submitted to FLORES+ through OLDI under CC BY-SA 4.0. Because language technology can reproduce unequal access and representation, future work should retain community participation in decisions about spelling, correction, attribution, and downstream use.

Acknowledgments

We thank the Chakma translators who contributed their linguistic knowledge and time, and the OLDI organizers for maintaining an open contribution

pathway. Generative AI assistants were used for manuscript drafting, restructuring, citation checking, language editing, formatting, and for scripting the release audit described in Section 4.2. They were not used at any stage of the translation itself. We reviewed the final text and remain responsible for its claims, citations, translations, and release artifacts.

References

- Firoz Ahmed, Nitin Venkateswaran, and Sarah Moeller. 2024. [The Bangla/Bengali seed dataset submission to the WMT24 open language data initiative shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 556–566, Miami, Florida, USA. Association for Computational Linguistics.
- Rubel Ahmod, Musharof Ahmed, Munshaibur Rahman Mahin, Mahmudul Hasan Badhan, Nurul Islam Opu, and Md. Abu Naser Mojumder. 2025. [Sentence-level indigenous language identification in Bangladesh: A comparative study of attention-based RNNs and transformer models](#). In *2025 28th International Conference on Computer and Information Technology (ICCIT)*. IEEE.
- Saifullah Anik, Md. Obydul Hossain, Md. Shakib Ahammed, Gourab Kumar Pranta, and Abdus Sattar. 2026. [A speech driven framework for identifying low resourced ethnic languages of Bangladesh](#). In *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence and Networking (QPAIN)*. IEEE.
- S. M. Golam Sarwar Apu, Ahmed Imtiaz, Prosenjit Chakraborty, and U Mong Sing Marma Biplob. 2025. [DChakma: A deep learning framework for part-of-speech tagging in Chakma language](#). In *2025 IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS)*. IEEE.
- Bangladesh Bureau of Statistics. 2022. Population and housing census 2022.
- Laurie Burchell, Jean Maillard, Antonios Anastasopoulos, Christian Federmann, Philipp Koehn, and Skyler Wang. 2024. [Findings of the WMT 2024 shared task of the open language data initiative](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 110–117, Miami, Florida, USA. Association for Computational Linguistics.
- Aunabil Chakma, Aditya Chakma, Masum Hasan, Soham Khisa, Chumui Tripura, and Rifat Shahriyar. 2024. [ChakmaNMT: Machine translation for a low-resource and endangered language via transliteration](#). *arXiv preprint arXiv:2410.10219*.
- David Dale, Laurie Burchell, Jean Maillard, Idris Abdulmumin, Antonios Anastasopoulos, Isaac Caswell, and Philipp Koehn. 2025. [Findings of the WMT 2025 shared task of the open language data initiative](#).

- In *Proceedings of the Tenth Conference on Machine Translation*, pages 495–502. Association for Computational Linguistics.
- Glottolog. 2026. Chakma (chak1266). <https://glottolog.org/resource/language/id/chak1266>. Accessed 2026-09-15.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. [The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation](#). *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Anzir Rahman Khan, Mehraj Hossain Mahi, Mayen Uddin Mojumdar, Narayan Ranjan Chakraborty, Md. Alamgir Kabir, and Arpita Nath Boruah. 2025. [Transformer-based model for identifying ethnic languages in Bangladesh](#). In *2025 12th International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE.
- Adity Khisa, Nusrat Jahan Lia, Tasnim Mahfuz Nafis, Zarif Masud, Tanzir Pial, Shebuti Rayana, and Ahmedul Kabir. 2025. [Exploring cross-lingual knowledge transfer via transliteration-based MLM fine-tuning for critically low-resource Chakma language](#). In *Proceedings of the Second Workshop on Bangla Language Processing*, pages 280–291. Association for Computational Linguistics.
- Mehraj Hossain Mahi, Anzir Rahman Khan, Mobashsher Hasan Anik, Sheak Rashed Haider Noori, Arif Mahmud, and Mayen Uddin Mojumdar. 2025. [MELD: A multilingual ethnic dataset of Chakma, Garo, and Marma in Bengali script with English and standard Bengali translation](#). *Data in Brief*, 61:111745.
- Sowhanur Rahman Nirob, Priya Rani Das, Md. Tanzeem Rahat, Sazia Sharmin, and Kamruddin Nur. 2026. [CL-CNN: A low-resource character-level CNN for Bangladeshi ethnic language recognition](#). In *2026 IEEE 2nd International Conference on Quantum Photonics, Artificial Intelligence and Networking (QPAIN)*. IEEE.
- NLLB Team. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*.
- Open Language Data Initiative. 2026. Translation guidelines. <https://oldi.org/guidelines>. Accessed 2026-09-15.
- Kanchon Kanti Podder, Ludmila Emdad Khan, Jyoti Chakma, Muhammad E. H. Chowdhury, Proma Dutta, Khan Md Anwarus Salam, Amith Khandakar, Mohamed Arselene Ayari, Bikash Kumar Bhawmick, S. M. Arafin Islam, and Serkan Kiranyaz. 2023. [Self-ChakmaNet: A deep learning framework for indigenous language learning using hand-written characters](#). *Egyptian Informatics Journal*, 24(4):100413.
- Md. Abdur Rahman, Md. Tofael Ahmed Bhuiyan, and Abdul Kadar Muhammad Masum. 2025. [ChakmaBridge: A five-way parallel corpus for navigating the script divide in an endangered language](#). In *Proceedings of the Second Workshop on Bangla Language Processing*, pages 259–265. Association for Computational Linguistics.
- Rahul Raja and Arpita Vats. 2025. [Parallel corpora for machine translation in low-resource Indic languages: A comprehensive review](#). In *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages*, pages 129–143. Association for Computational Linguistics.
- Shaila Sultana. 2023. [Indigenous ethnic languages in Bangladesh: Paradoxes of the multilingual ecology](#). *Ethnicities*, 23(5):680–705.
- Unicode Consortium. 2025. The Unicode standard, version 17.0: Chakma, U+11100–U+1114F. <https://www.unicode.org/charts/PDF/U11100.pdf>.