

Findings of the WMT 2026 Shared Task of the Open Language Data Initiative

Laurie Burchell
Common Crawl Foundation

David Dale
Meta FAIR

Jean Maillard
Meta FAIR

Idris Abdulmumin
University of Pretoria

Antonios Anastasopoulos
George Mason University

Isaac Caswell
Google Deepmind

Philipp Koehn
Johns Hopkins University

Correspondence: info@oldi.org

Abstract

We present the results of the WMT 2026 shared task of the Open Language Data Initiative (OLDI). Participants were invited to contribute to the massively multilingual open datasets supported by OLDI, FLORES+ and OLDI Seed, or to create new resources in line with OLDI’s ethos. We accepted ten submissions: five extending FLORES+, and five extending other open and massively-parallel datasets (BOU-QuET, Seed, SMOL, and WMT24+). These contributions advance the coverage and quality of multilingual datasets, especially for under-served language varieties. All contributions are released under permissive open-source licenses.

1 Introduction

Recent advances in machine translation (MT) and other multilingual natural language processing (NLP) tasks depend on a sufficient quantity of training data in the languages of interest, as well as high-quality evaluation data to measure progress in a meaningful way (Kocmi et al., 2025). However, such data is only available for a small number of the world’s language varieties, limiting high-performance NLP applications to the well-resourced few (Joshi et al., 2020). The question then is how to expand data coverage, and consequently NLP access, to language communities that are missing out?

The Open Language Data Initiative (OLDI) aims to empower language communities to contribute to the key datasets underlying NLP technologies. It champions open data so that more researchers and communities can benefit from multilingual resources. As part of this, OLDI manages two open and massively-multilingual datasets, FLORES+ and Seed (described in Section 2), as well as

supporting work to create and expand other open corpora in line with its objectives.

This paper describes the third shared task of OLDI, run as part of the Eleventh Conference on Machine Translation (WMT 2026). Like previous iterations (Burchell et al., 2024; Dale et al., 2025), the primary goal of the shared task was to expand the high-quality, massively-parallel and open-source datasets underpinning the development of multilingual applications. Contributions could consist of adding new language varieties to existing datasets, substantial improvements to existing datasets, or novel datasets compatible with the aims of OLDI. This year, the shared task received ten submissions covering thirteen language varieties, expanding five existing datasets. All contributions are publicly available under permissive licenses.

2 Datasets

In the third iteration of the OLDI shared task, participants contributed to five datasets: FLORES+ and Seed, which are managed by OLDI, as well as the independently-owned datasets BOUQuET, Seed, SMOL, and WMT24+.

2.1 FLORES+

FLORES is a family of datasets designed to benchmark multilingual translation, with many-to-many alignment across over 200 languages. It consists of translations of English sentences sampled in equal amounts from Wikinews (an international news source), Wikijunior (a collection of age-appropriate non-fiction books), and Wikivoyage (a travel guide). There are two public splits: dev (997 sentences) and devtest (1012). The first iteration of this dataset covered only three languages (Guzmán et al., 2019), but following iterations in-

creased coverage first to 101 languages (FLORES-101, Goyal et al., 2022) and then to over 200 languages as part of the “No Language Left Behind” project (NLLB Team, 2024).

Previous iterations of this task and ongoing volunteer effort have added further languages as well as correcting existing data (Burchell et al., 2024; Dale et al., 2025; Doumbouya et al., 2023; Gala et al., 2023; Perez-Ortiz et al., 2024; Abdulmumin et al., 2024; Ali et al., 2024; Kuzhuget et al., 2024; Yu et al., 2024; Mamasaidov and Shopulatov, 2024; Gordeev et al., 2024; Frontull et al., 2025; Oktem et al., 2025; Mamasaidov et al., 2025; Mæhlum et al., 2025; Rajcoomar, 2025; Zhang et al., 2025; Samil and Adelani, 2026). This living version of the benchmark is released as FLORES+¹ under the CC BY-SA 4.0 license.

2.2 OLDI Seed

The NLLB-Seed dataset of NLLB Team (2024) was created as a source of starter data for languages without publicly-available high-quality bitext in sufficient quantity for training NLP models. This dataset consists of around 6000 sentences sampled from the Wikipedia articles listed in English Wikipedia’s “List of articles every Wikipedia should have”.² These were professionally translated into each of the 38 languages covered by the first iteration of this dataset (39 if including English), and experiments by Maillard et al. (2023) demonstrated the gains of including these datasets in the training mix of MT models.

Like FLORES+, previous iterations of this shared task and ongoing volunteer effort have expanded the original data (Doumbouya et al., 2023; Ahmed et al., 2024; Ferrante, 2024; Cols, 2024; Marmonier et al., 2025; Jumashev et al., 2025; Oktem et al., 2025). The living version of this benchmark is released as OLDI Seed³ under the CC BY-SA 4.0 license.

2.3 BOUQuET

BOUQuET is a multi-way parallel, multi-centric and multi-register/domain dataset and benchmark for MT quality (Andrews et al., 2025). It consists of short paragraphs, fully parallel in all languages at the sentence level, in two public splits: dev (504

unique sentences, 120 paragraphs) and test (854 unique sentences, 198 paragraphs). The underlying texts have been handcrafted by linguists in eight diverse languages (Egyptian Arabic, French, German, Hindi, Indonesian, Mandarin Chinese, Russian, and Spanish) and translated to English and 266 other language + script combinations. The dataset is intended to be extensible to virtually any other written language and is open to volunteer contributions. BOUQuET is managed by Meta FAIR and released under the CC BY-SA 4.0 license.⁴

2.4 SMOL

SMOL is a multi-way parallel training dataset with high lexical coverage (Caswell et al., 2025). The first part of the dataset, SMOLSENT, is based on 863 English sentences semi-manually selected from Common Crawl data⁵ to cover 5.5k of the most common English words (obtained by joining the GATITOS wordlist (Jones et al., 2023) and the most frequent words in Common Crawl). These sentences were then professionally translated into 88 languages. The second part of the dataset, SMOLDOC, is based on 584 English documents generated by a large language model (LLM) using prompt templates to ensure diversity of topics and styles. These were professionally translated from English into 115 languages, then volunteers contributed additional translations to bring the total number of languages to 123. SMOL is managed by Google Research and released under the CC BY 4.0 license.⁶

2.5 WMT24++

The WMT24+ dataset (Deutsch et al., 2025) was created by translating the test dataset from the WMT24 General MT shared task (Kocmi et al., 2024) from English to 54 other languages. The 998 paragraph-sized English source documents come from four different domains: literary, news, social, and transcribed speech. Thus, WMT24+ is mostly complementary to FLORES+ in document sizes and domains (though news is an overlapping domain). WMT24+ is managed by Google Research and released under the permissive Apache License 2.0.⁷

¹https://huggingface.co/datasets/openlanguage/flores_plus

²https://meta.wikimedia.org/wiki/List_of_articles_every_Wikipedia_should_have

³https://huggingface.co/datasets/openlanguage/oldi_seed

⁴<https://huggingface.co/datasets/facebook/bouquet>

⁵<https://commoncrawl.org/>

⁶<https://huggingface.co/datasets/google/smol>

⁷<https://huggingface.co/datasets/google/wmt24pp>

3 Shared task description

As described in the introduction, we invited three main categories of contribution: new language varieties for existing datasets, substantial improvements to existing datasets, or novel datasets compatible with the aims of OLDI. We were open to contributions of new or extended datasets of translation quality annotations for under-served languages, but we received no submissions of this type. We describe the standard contribution process below.

3.1 Contributing to OLDI’s datasets

When contributing to the datasets managed by OLDI, FLORES+ and Seed, we encouraged participants to translate from the English data, so as to minimise translation drift. Participants could translate from a different language if necessary, but we asked them to document this choice clearly. Where possible, we asked participants to use translators who were qualified, native speakers of the target language, and to have at least one additional native speaker verify the data. For FLORES+ translation, we did not allow any use of MT output, including post-editing, to avoid introducing any machine bias in this evaluation dataset. For Seed, post-editing was allowed as long as all outputs were verified manually. Full translation guidelines are provided on the OLDI website⁸ and copied to Appendix A.

We asked the participants to provide dataset cards as part of their submissions, detailing precise language information and the translation workflow that was employed. In particular, we asked them to identify the language with an ISO 639-3 individual language tag, the precise variety with a Glottocode, and the script with an ISO 15924 script code. For example, the Val Badia variety of Lombard is tagged as `lld_Latn_badi1244`.

Participants were encouraged to provide experimental validation of the quality of the data they were submitting. All contributions had to be released under open licenses matching the original datasets e.g. CC BY-SA 4.0 for FLORES+.

3.2 Other contributions

We also accepted extensions and improvements to other foundational, open, and multilingual datasets in keeping with OLDI’s principles. We recommended that the contribution workflow for these submissions should follow OLDI’s as closely as possible to ensure data quality and documentation.

⁸<https://oldi.org/guidelines>

4 Submissions

4.1 Shared task submissions

Table 1 summarises the contributions received as part of the shared task. We received ten papers describing contributions to five datasets: FLORES+ (7 language varieties), BOUQuET (3 language varieties), Seed (1 variety), SMOL (3 language varieties), and WMT24+ (1 language variety). Overall, participants made submissions for 13 different language varieties. In the rest of this section, we give a brief description of each submission.

Ahmed and Chakma (2026) translate the dev and devtest part of FLORES+ into Chakma, a language spoken in India and written in its own script (part of Unicode). There are variants of the language. The contribution represents the Bangladeshi variety used by contributors from the Chittagong Hill Tracts. The contribution is fully manual and did not use any machine translation or large language model. The source language is English. Translations were reviewed by one of the authors. There is no experimental validation of the data.

Ali et al. (2026) contribute three FLORES+ evaluation sets for Mozambican Bantu varieties: Xichangana, Mozambican Nyanja, and Sena (the first FLORES+ entry for this 2M speaker language), each translated from Portuguese by professional native translators and independently assessed. Since South African Tsonga and Malawian Chichewa share ISO 639-3 codes with varieties already in FLORES+, the authors score identical system outputs against both references, showing that existing sets overestimate quality for Mozambican users by 7 to 9 spBLEU, with the gap driven by borrowing source, orthography, and register rather than ordinary translation errors. Fine-tuning NLLB-200 on around 30k Mozambican Bible verses recovers most of the gap while degrading scores on the sibling sets, confirming that the two references measure genuinely different targets and motivating variety-aware language tags (e.g., `nya_Latn_MZ`) for cross-border Bantu languages.

Ibrahim (2026) extend the SMOL dataset with two varieties of the Tigre language (a language spoken in Eritrea and eastern Sudan): Modern Standardized Tigre and its Mansa dialect (a classical literary norm). The translations from English were pre-generated by an NLLB-based model and post-edited and reviewed by volunteer contributors. Their validity was confirmed by fine-tuning an NLLB-200 model with this data and observing

Language	Variety	ISO 639-3	ISO 15924	Glottocode	Contributors	Contribution
Chakma	Bangladeshi (Chit-tagong Hill Tracts)	ccp	Cakm	chak1266	Ahmed and Chakma (2026)	FLORES+
Xichangana	Gaza and Maputu (Mozambican)	tso	Latn	chan1318	Ali et al. (2026)	FLORES+
Nyanja	Tete and Niassa (Mozambican)	nya	Latn	nyan1308	Ali et al. (2026)	FLORES+
Sena	Central Lower-Zambezi (Mozambican)	seh	Latn	nuc11396	Ali et al. (2026)	FLORES+
Ladin	Gherdëina	lld	Latn	gard1241	Frontull et al. (2026)	BOUQuET
Ladin	Val Badia	lld	Latn	badi1244	Frontull et al. (2026)	BOUQuET
Tigre	Modern Standardized Tigre & Mansa literary dialect	tig	Ethi	tigr1270	Ibrahim (2026)	SMOL
Franco-Provençal	Savoyard	frp	Latn	savo1253	Marmonier et al. (2026b)	FLORES+
French	-	fra	Latn	stan1290	Marmonier et al. (2026a)	SMOLSENT
Lombard	Eastern	lmo	Latn	east2276	Signoroni and Rychlý (2026)	FLORES+
Manipuri (Meitei)	-	mni	Mtei	mani1292	Singh (2026)	FLORES+ devtest
Chechen	-	che	Cyrl	chec1245	Umishov and Grigorian (2026)	SMOL, BOUQuET, WMT24++
Croatian	-	hrv	Latn	croal245	Zundanović et al. (2026)	Seed

Table 1: A summary of all dataset contributions to the WMT 2026 Shared Task of the Open Language Data Initiative.

its translation quality on a held-out subset increase.

[Frontull et al. \(2026\)](#) contribute BOUQuET translations in two varieties of Ladin, Val Badia and Gherdëina. Four native speakers, two per variety, translated BOUQuET from the English and German references via the official web interface, using variety-specific spellcheckers. Translations for the two varieties were then aligned side-by-side to harmonise the terminology and the outputs were checked by two reviewers. Additional validation relied on automated metrics such as source/target ratio checks, checking back-translations into English against the English reference, and blind ranking the Ladin references against automatic translations.

[Marmonier et al. \(2026a\)](#) translate the SMOL sentence-level dataset from English into French by manually post-editing translations from NLLB-200, Gemma-4-31B, and DeepSeek V4 Pro. The candidates were compared using reference-free MetricX scores, with that the post-edited texts scoring significantly better than the MT candidates. The authors intend to use the French SMOL translation as a pivot for the regional languages of France, including the experiments with Savoyard described below.

[Marmonier et al. \(2026b\)](#) translate FLORES+ from French into Savoyard, an endangered Arpitan dialect spoken around the French–Italian–Swiss alpine borders. The draft translations for the dev

split were generated with a French-Arpitan rule-based Apertium system and post-edited by two proficient speakers; the devtest split was translated from scratch. The authors use the dataset to evaluate several systems for translation between French and Savoyard: the copy-source baseline, Apertium, a fine-tuned NLLB-200 model, and several versions of Gemma 4. According to automatic and human evaluations, Apertium leads for translation into Savoyard, while Gemma 4 is the best for translating Savoyard into French.

[Signoroni and Rychlý \(2026\)](#) translate FLORES+ from Italian into Eastern Lombard, an under-represented variety of the Lombard language continuum (FLORES+ already had Western Lombard translation). They use these translations to evaluate several LLMs and MT models for translation of English and Italian into Western and Eastern Lombard. They find that the models tend to generate the more represented Western variety even when instructed otherwise, and that translation into both Lombard varieties with any system does not significantly outperform the copy-the-Italian-source baseline.

[Singh \(2026\)](#) create a four-way parallel benchmark between English, Hindi, Meitei script Manipuri, and Bengali script Manipuri by transliterating the FLORES-devtest Bengali script Manipuri

into Meitei script. This new split was produced by two professional translators: one to transliterate, and a second to review independently. They use their benchmark to evaluate two dedicated MT systems and five general purpose LLMs. They find that evidence that differences in performance across scripts reflects training data curation, meaning that different systems vary in ability and failure modes for the different scripts. They recommend that script not be taken as constant across a language’s benchmark entries, since conflating “Manipuri” scores across scripts without qualification would obscure the models’ real-world performance. They also highlight that failure modes are script-specific, not just magnitude differences, and the observed failures indicate data scarcity interacting with architecture rather than uniform degradation.

Umishov and Grigorian (2026) add Chechen translations to the BOUQuET, WMT24+, and SMOL datasets. For all 3 datasets, Russian serves as a pivot language to translate from. The authors validate their datasets by fine-tuning a MADLAD-400 model using their SMOL translation and evaluating it with WMT24+ and BOUQuET. On both benchmarks, the quality of translation between Russian and Chechen language improves.

Zundanović et al. (2026) post-edit machine translations of Seed from English into Croatian. They validate data quality by fine-tuning several models from NLLB and Gemma with the post-edited translations to translate between English and Croatian. On the evaluation datasets, they observe improvements in all configurations in ChrF++ and COMET scores, in all cases except English-Croatian LLM-based translation. The authors ablate the effect of human post-edition by fine-tuning the NLLB-based models using the original machine-translated Croatian Seed sentences, finding that the post-edited data always wins.

4.2 Other contributions

Not all contributors to the OLDI datasets submitted shared task papers. In the past year, Seed incorporated new translations for Mauritian Creole (**Rajcoomar, 2025**), Mongolian in traditional Mongolian script (**Zhang et al., 2025**), Khakas, Sudanese Arabic (**Samil and Adelani, 2026**), Haryanvi, Bodo, Garhwali, Konkani, Rajasthani, Malvi and Marwari (**Singh et al., 2024**). FLORES+ also received corrections for Amharic and Turkish. SMOL received translations into Abaza, Brahui, Tigre, Juhuri, Nko, and Khakas; and Gatitos received 1,052 transla-

tions of Cantonese food items. For a detailed list of changes and attributions, see the CHANGELOG.md files and dataset cards in the **FLORES+**, **OLDI-Seed**, and **SMOL** repositories.

5 Discussion

The third iteration of the Open Language Data shared task turned out as fruitful as the previous two, with 10 submissions, 13 involved language varieties, and 5 datasets that the contributions targeted.

An important theme of this year is fine-grained variation of a language: **Ali et al. (2026)**, **Frontull et al. (2026)**, and **Signoroni and Rychlý (2026)** all explicitly contrast several dialects of the same language, while **Singh (2026)** compares two writing systems. Contributions such as these add more depth to the coverage of the datasets, as they allow developers to assess variation in language use.

This year, more submissions than before (**Ibrahim, 2026**; **Marmonier et al., 2026a,b**; **Zundanović et al., 2026**) chose not to translate the training dataset contributions from scratch, but rather to post-edit their machine translations and then validate the contribution of the human post-edits to the resulting data quality. We also observe a growing degree of rigour in validating the overall quality of the resulting translations, using a combination of automatic quality estimation metrics, comparison of machine translation outputs, and human evaluation. This is an encouraging development, as it reflects the increasing appreciation in the community for the impact of high-quality multilingual benchmarks.

With the recent progress of LLMs, the sentiment is sometimes expressed that machine translation is a "solved problem". This is not the case, as ablation results from these submissions demonstrate. Even for such a high-resourced language as French, human post-edits are quantifiably better than the best MT outputs; for translation into some lower-resourced languages, LLMs appear to be inferior to rule-based systems (Savoyard) or even struggle to produce the correct target variety (Manipuri in Meitei script). The newly contributed dataset extensions provide a leverage for improving the way modern technologies handle the world’s linguistic diversity.

6 Conclusion

We presented the results of the WMT 2026 OLDI shared task. We accepted ten submissions covering thirteen language varieties, extending five existing foundational datasets: FLORES+, BOUQuET, Seed, SMOL, and WMT24+. We are truly grateful to all participants for their work, and we hope that these contributions are soon adopted by the research community and used to enhance user-led language technologies.

Acknowledgments

OLDI functions on volunteer time and community contributions. We are grateful to the language communities, researchers, and reviewers that contributed to this shared task with new resources for under-served languages.

References

- Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathabula, Matimba Shingange, Tajuddeen Gwadabe, and Vukosi Marivate. 2024. [Correcting FLORES evaluation dataset for four African languages](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 570–578, Miami, Florida, USA. Association for Computational Linguistics.
- Firoz Ahmed and Dinalo Chakma. 2026. A human-translated flores+ dataset for chakma language. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- 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.
- Felermينو Dario Mario Ali, Henrique Lopes Cardoso, and Rui Sousa-Silva. 2024. [Expanding FLORES+ benchmark for more low-resource settings: Portuguese-emakhuwa machine translation evaluation](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 579–592, Miami, Florida, USA. Association for Computational Linguistics.
- Felermينو Dario Mario Ali, Delfina Lázaro Mateus, and Manuel Valente Mangué. 2026. Measuring the cost of variety conflation in multilingual mt evaluation: Adding mozambican xichangana, nyanja and sena to flores+. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Mark Duppenthaler, Nathaniel Paul Ekberg, Cynthia Gao, Daniel Edward Licht, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, and Shireen Yates. 2025. [BOUQuET : dataset, benchmark and open initiative for universal quality evaluation in translation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27515–27535, Suzhou, China. Association for Computational Linguistics.
- 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.
- Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Djibrila Diane, Solo Farabado Cissé, Koulako Moussa Doumbouya, Edoardo Ferrante, Alessandro Guasoni, Christopher Homan, Mamadou K. Keita, Sudhamoy DebBarma, Ali Kuzhuget, David Anugraha, and 5 others. 2025. [SMOL: Professionally translated parallel data for 115 under-represented languages](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1103–1123, Suzhou, China. Association for Computational Linguistics.
- Jose Cols. 2024. [Spanish corpus and provenance with computer-aided translation for the WMT24 OLDI shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 624–635, Miami, Florida, USA. Association for Computational Linguistics.
- 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, Suzhou, China. Association for Computational Linguistics.
- Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. [WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.
- Moussa Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye

- Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory Conde, Kalo Mory Diané, Chris Piech, and Christopher Manning. 2023. [Machine translation for Nko: Tools, corpora, and baseline results](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 312–343, Singapore. Association for Computational Linguistics.
- Edoardo Ferrante. 2024. [A high-quality seed dataset for Italian machine translation](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 567–569, Miami, Florida, USA. Association for Computational Linguistics.
- Samuel Frontull, Lara Castlunger, Maximilian Oberholzenzer, Karin Comploj, Ruth Videsott, Robert Sama, Gabriel Perathoner, Carlo Zoli, and Paul Videsott. 2026. Adding *ciofs* and *ciüfs* to bouquet: Benchmark extension to ladin. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Samuel Frontull, Thomas Ströhle, Carlo Zoli, Werner Pescosta, Ulrike Frenademez, Matteo Ruggeri, Daria Valentin, Karin Comploj, Gabriel Perathoner, Silvia Liotto, and Paolo Anvidalfarei. 2025. [Bringing Ladin to FLORES+](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1061–1071, Suzhou, China. Association for Computational Linguistics.
- Jay Gala, Pranjal A Chitale, Raghavan Ak, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, et al. 2023. [Indictans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages](#). *arXiv preprint arXiv:2305.16307*.
- Isai Gordeev, Sergey Kuldin, and David Dale. 2024. [FLORES+ translation and machine translation evaluation for the Erzya language](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 614–623, Miami, Florida, USA. Association for Computational Linguistics.
- 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.
- Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc’Aurelio Ranzato. 2019. [The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.
- Beshir Ibrahim. 2026. Closing the resource gap for tigre: A diaspora-sourced english–tigre parallel dataset for smol. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. [GATITOS: Using a new multilingual lexicon for low-resource machine translation](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [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*, pages 6282–6293, Online. Association for Computational Linguistics.
- Murat Jumashev, Alina Tillabaeva, Aida Kasieva, Turgunbek Omurkanov, Akylai Musaeva, Meerim Emil Kyzy, Gulaiym Chagataeva, and Jonathan Washington. 2025. [The Kyrgyz seed dataset submission to the WMT25 open language data initiative shared task](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1088–1102, Suzhou, China. Association for Computational Linguistics.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakouga, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. [Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.
- Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. [Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.
- Ali Kuzhuget, Airana Mongush, and Nachyn-Enkhedorzhu Oorzhak. 2024. [Enhancing tuvan language resources through the FLORES dataset](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 593–599, Miami, Florida, USA. Association for Computational Linguistics.

- Petter Mæhlum, Anders Næss Evensen, and Yves Scherrer. 2025. [Improved Norwegian Bokmål translations for FLORES](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1124–1132, Suzhou, China. Association for Computational Linguistics.
- Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. [Small data, big impact: Leveraging minimal data for effective machine translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.
- Mukhammadsaid Mamasaidov, Azizullah Aral, Abror Shopulatov, and Mironshoh Inomjonov. 2025. [Filling the gap for Uzbek: Creating translation resources for Southern Uzbek](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1081–1087, Suzhou, China. Association for Computational Linguistics.
- Mukhammadsaid Mamasaidov and Abror Shopulatov. 2024. [Open language data initiative: Advancing low-resource machine translation for Karakalpak](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 606–613, Miami, Florida, USA. Association for Computational Linguistics.
- Malik Marmonier, Rachel Bawden, and Benoît Sagot. 2026a. [A french version of the smolsent corpus](#). In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Malik Marmonier, Alain Favre, Benoît Sagot, and Rachel Bawden. 2026b. [A savoyard version of the flores+ corpus](#). In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Malik Marmonier, Benoît Sagot, and Rachel Bawden. 2025. [A French version of the OLDI seed corpus](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1048–1060, Suzhou, China. Association for Computational Linguistics.
- NLLB Team. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630(8018):841–846.
- Alp Oktem, Mohamed Aymane Farhi, Brahim Essaidi, Naceur Jabouja, and Farida Boudichat. 2025. [Correcting the tamazight portions of FLORES+ and OLDI seed datasets](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1072–1080, Suzhou, China. Association for Computational Linguistics.
- Juan Antonio Perez-Ortiz, Felipe Sánchez-Martínez, Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Aaron Galiano Jimenez, Antoni Oliver, Claudi Aventín-Boya, Alejandro Pardos, Cristina Valdés, Jusep Loís Sans Socasau, and Juan Pablo Martínez. 2024. [Expanding the FLORES+ multilingual benchmark with translations for Aragonese, aranese, Asturian, and Valencian](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 547–555, Miami, Florida, USA. Association for Computational Linguistics.
- Yush Rajcoomar. 2025. [KozKreolMRU WMT 2025 CreoleMT system description: Koz kreol: Multi-stage training for English–mauritian creole MT](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1183–1190, Suzhou, China. Association for Computational Linguistics.
- Hadia Mohmmadosman Ahmed Samil and David Ifeoluwa Adelani. 2026. [Sudanese-Flores: Extending FLORES+ to Sudanese Arabic dialect](#). In *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)*, pages 243–247, Rabat, Morocco. Association for Computational Linguistics.
- Edoardo Signoroni and Pavel Rychlý. 2026. [FiÛr: A benchmark dataset for eastern lombard machine translation](#). In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Harman Singh, Nitish Gupta, Shikhar Bharadwaj, Dinesh Tewari, and Partha Talukdar. 2024. [IndicGenBench: A multilingual benchmark to evaluate generation capabilities of LLMs on Indic languages](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11047–11073, Bangkok, Thailand. Association for Computational Linguistics.
- Kshetrimayum Boynao Singh. 2026. [Script matters: Benchmarking open-source machine translation for manipuri in meitei mayek and bengali script](#). In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Abu-Viskhan A. Umishov and Vladislav A. Grigorian. 2026. [Expanding chechen machine translation resources: Smol, bouquet, and wmt24++ extensions](#). In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Hongjian Yu, Yiming Shi, Zherui Zhou, and Christopher Haberland. 2024. [Machine translation evaluation benchmark for Wu Chinese: Workflow and analysis](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 600–605, Miami, Florida, USA. Association for Computational Linguistics.
- Chen Zhang, Mingxu Tao, Zhiyuan Liao, and Yansong Feng. 2025. [MiLiC-eval: Benchmarking multilingual LLMs for China’s minority languages](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11086–11102, Vienna, Austria. Association for Computational Linguistics.

Dragana Zundanovać, Hrvoje Leventić, Dragana Božić Lenard, and Krešimir Romić. 2026. The croatian dataset seed submission to the wmt26 open language data initiative shared task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

A Translation guidelines

These translation guidelines must be acknowledged by all translators who will be contributing data.

Important note

Your translations will be used to help train or evaluate machine translation engines. For this reason, this project requires human translation.

- If you are translating data to be used for evaluation purposes (e.g. FLORES+), using or even referencing machine translation output is usually not allowed (this includes post-editing). In exceptional circumstances, use of CAT (computer-assisted translation) tools and translation frameworks such as Apertium may be admissible. This is decided on a case by case basis: please contact the organisers to discuss.
- If you are translating data to be used for training purposes, such as Seed, the use of post-edited machine translated content is allowed, provided all data is manually verified and edited where necessary.
- CAUTION: Some commercial machine translation or LLM services – including DeepL, Google Translate, ChatGPT, Claude and Gemini – prohibit the use of their output for training other translation or AI models. Their use is not permitted.

General guidelines

- You will be translating sentences coming from different sources. Please refer to the source document if available.
- Do not convert any units of measurement. Translate them exactly as noted in the source content.
- When translating, please maintain the same tone used in the source document. For example, encyclopedic content coming from sources like Wikipedia should be translated using a formal tone.
- Provide fluent translations without deviating too much from the source structure. Only allow necessary changes.
- Do not expand or replace information compared to what is present in the source documents. Do not add any explanatory or parenthetical information, definitions, etc.
- Do not ignore any meaningful text that was present in the source.

- In case of multiple possible translations, please pick the one that makes the most sense (e.g., for gender concordance, cultural fit in the target language, level of formality, etc.).
 - Translations must be faithful to the source in terms of pragmatics such as (if applicable) level of hedging/modality, sentiment and its intensity, negation, speech effects (disfluencies), etc.
 - For proper nouns and common abbreviations, please see the guidelines on Named Entities below.
 - Idiomatic expressions should not be translated word for word. Use an equivalent idiom, if one exists. If no equivalent idiom exists, use an idiom of similar meaning. If no similar expressions exist in the target language, paraphrase the idiom such that the meaning is retained in the target language.
 - When a pronoun to be translated is ambiguous (for instance, when it could be interpreted as either *him/her* or *he/she*), opt for gender neutral pronouns (such as *them/they*) if those exist in the target language. However, when a pronoun to be translated is clearly marked for gender, you should follow the source material and continue to mark for gender.
 - Foreign words and phrases used in the text should be kept in their original language when this is necessary to preserve the meaning of the sentence (e.g. if given as an example of a foreign word).
- If a transliteration would not be commonly understood in the context, and the source term would be more acceptable, you may retain the original term.

Named entities

Named entities are people, places, organisations, etc., that are commonly referred to using a proper noun. This section provides guidance on how to handle named entities. Please review the following guidelines carefully:

- If there is a commonly used term in the target language for the Named Entity:
 - If the most commonly used term is the same as in the source language, then keep it as it is.
 - If the most commonly used term is a translation or a transliteration, then use that.
- If there is no commonly used term:
 - If possible, a transliteration of the original term should be used.