Findings of the WMT 2025 Shared Task of the Open Language Data Initiative

David Dale[*] Meta FAIR

Laurie Burchell[*] Common Crawl Foundation

Jean Maillard Meta FAIR

Idris Abdulmumin University of Pretoria

Antonios Anastasopoulos George Mason University

Isaac Caswell Google

Philipp Koehn Johns Hopkins University

Correspondence: info@oldi.org

Abstract

We present the results of the WMT 2025 shared task of the Open Language Data Initiative. Participants were invited to contribute to the massively multilingual open datasets (FLORES+, MT Seed, WMT24++) or create new such resources. We accepted 8 submissions, including 7 extensions or revisions of the existing datasets and one submission with a new parallel training dataset, SMOL.

1 Introduction

Even if we assume that machine translation (MT) is already solved, this would be true only for the language pairs and domains where parallel training data is abundant, either as explicit parallel datasets, or as incidental bilingual signals in generic web data (Briakou et al., 2023). For the majority of the world's languages, we lack both parallel training data to train MT models and evaluation data to assess their translation capabilities. One way to address this bottleneck is creation of massively multilingual parallel datasets and their extension to new languages.

The second Shared task of Open Language Data Initiative (OLDI) at WMT25 invites the language communities to contribute to high-quality, massively parallel and open-source datasets by their extension with new languages, varieties or dialects, substantial improvements to existing datasets, or creation of new such datasets. These datasets include (but are not limited to) FLORES+ (Maillard et al., 2024), Seed (NLLB Team et al., 2024), and WMT24++ (Deutsch et al., 2025). OLDI itself is a community of researchers that maintains the former two datasets.

This year, we received 8 submissions, including 6 extensions of datasets to new language varieties, 2 revisions of existing translations, and one entirely new massively parallel dataset. All the data will be made available online under permissive licenses.[1].

2 Datasets

2.1 FLORES+

FLORES is a family of datasets designed to benchmark multilingual translation, with many-to-many alignment across over 200 languages. The first iteration of this dataset covered only three languages (Guzmán et al., 2019), but following iterations increased 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 et al., 2024). Finally, as part of the previous edition of this shared task, an additional 8 languages were included on top of several corrections to existing datasets (Maillard et al., 2024). This new, living version of the FLORES benchmark is released under the name FLORES+.

2.2 OLDI-Seed

The NLLB-Seed dataset of NLLB Team et al. (2024) was created as a source of starter data for languages without publicly-available high-quality bitext in sufficient quantity for training natural language processing (NLP) models. This dataset consists of around 6000 sentences sampled from the Wikipedia articles listed in English Wikimedia'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

[*] Equal contribution

[1] https://oldi.org and https://huggingface.co/collections/openlanguagedata

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

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 687–694
November 8–9, 2025. ©2025 Association for Computational Linguistics

Contributors	Type of contribution	Languages(s)
Caswell et al. (2025)	SMOL (new datasets)	123 languages
Frontull et al. (2025)	FLORES+ (new)	Ladin (2 varieties)
Jumashev et al. (2025)	MT Seed (new)	Kyrgyz
Mæhlum et al. (2025)	FLORES+ (corrections + new)	Norvegian Bokmål (+ a new variety)
Mamasaidov et al. (2025)	FLORES+ (new)	Southern Uzbek
Marmonier et al. (2025)	MT Seed (new)	French
Oktem et al. (2025)	FLORES+ & MT Seed (corrections)	Standard Moroccan Tamazight
Vamvas et al. (2025)	WMT24++ (new)	Romansh (6 varieties)

Table 1: A summary of all contributions to the WMT 2025 Shared Task of the Open Language Data Initiative.

this dataset (39 if including English), and experiments in (Maillard et al., 2023) demonstrated the gains of including these datasets in the training mix of MT models.

Participants to last year's edition of this shared task contributed three new languages (Maillard et al., 2024). To reflect the continuously updating nature of this dataset, and to distinguish it from prior iterations, it is released as OLDI-Seed.

2.3 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 domains (news is an overlapping domain, though) and in document sizes. Unlike the two previous datasets, WMT24++ is managed by Google Research and not by OLDI.

2.4 Other datasets

There are other massively parallel datasets that could have been potential targets for extension in the OLDI shared task. They include MT evaluation benchmarks such as NTREX-128 (Federmann et al., 2022) or BOUQuET (Andrews et al., 2025) and other parallel datasets that could be reused for MT, such as Global MMLU (Singh et al., 2025). FLEURS, a parallel datasets of speech (Conneau et al., 2023) and signed language (Tanzer, 2025; Costa-jussà et al., 2025) could also have been considered. Finally, there is the massively parallel GATITOS dataset (Jones et al., 2023) of 4000 frequently used words and phrases translated from English that served as a foundation for SMOL (Caswell et al., 2025), one of the contributions of the current shared task.

3 Shared task definition

The goal of this shared task was to expand high-quality, massively-parallel and open-source datasets: either FLORES+ and Seed managed by OLDI or any other dataset that matches the above criteria. Contributions could consist of either the addition of entirely new languages, varieties or dialects to the above datasets, or substantial improvements to existing datasets.

3.1 Contributing to FLORES+ and MT Seed

We encouraged the contributors of new languages to FLORES+ and Seed to start from the original English data; using a different pivot language was also possible, if clearly documented. We required the translations to be performed, wherever possible, by qualified, native speakers of the target language, and encouraged verification of the data by at least one additional native speaker. More recommendations were described in the OLDI contribution guidelines.[3]

For FLORES+ translation, we did not allow using or even referencing MT output was not allowed, including post-editing, to avoid introducing any machine bias in this evaluation dataset. For Seed data, the use of post-edited machine translated content was allowed, as long as all data was manually verified and the MT system allowed reusing their outputs to train other models (which is not the case for the major commercial LLMs).

We asked the participants to attach dataset cards to new data submissions, detailing precise language information and the translation workflow that was employed. In particular, we asked them to identify the language with both an ISO 639-3 individual language tag and a Glottocode, and identify the script with an ISO 15924 script code. For example, the Rumantsch Grischun variety was identified as `roh_Latn_ruma1247`.

[3] https://oldi.org/guidelines

Participants were encouraged to provide experimental validation of the quality of the data they were submitting.

3.2 Contributing other data

We also accepted extensions and improvements to other foundational multilingual datasets such as WMT24+ that are massively parallel, open source, and useful to under-served language communities. We suggested that contribution workflow should follow that for FLORES and Seed as closely as possible to ensure data quality and documentation. The contributed data was required to be released under an open license (allowing free research use as a minimum).

4 Submissions

4.1 Shared task submissions

Table 1 lists the contributions accepted as part of the shared task (more detailed list of contributions for each language can be found in Appendix A). Below, we briefly describe each submission.

Caswell et al. (2025) created the SMOL dataset: a multiway parallel training dataset with high lexical coverage. The first part of the dataset, SMOLSENT is based on 863 English sentences semi-manually selected from CommonCrawl to cover 5.5k of the most common English words (obtained by joining the GATITOS wordlist and the most frequent words in CommonCrawl). The second part, SMOLDOC is based on 584 English documents generated with LLMs using prompt templates that ensured diversity of topics and styles. The dataset was professionally translated from English into 115 languages, mostly under-resourced. Subsequently, additional volunteer translations were contributed, bringing the total number of languages to 123. To demonstrate the value of the dataset, the authors used it for in-context learning of several commercial LLMs and for fine-tuning of a GEMINI LLM for translation out of English into the 80 languages for which evaluation data were available. For most language subsets and models, in-context learning with SMOL examples was found to be superior to zero-shot translation. Fine-tuning demonstrated positive effect of both SMOL dataset parts and their combination with GATITOS.

Frontull et al. (2025) translated FLORES+ into two varieties of Ladin, a language spoken in Northern Italy: Val Badia and Gherdëina. The paper gives a detailed overview of Ladin and the resources available for it. The FLORES sentences were first manually translated into the Val Badia variety, using German, Italian, Friulian, and English references, then into the Gherdëina variant, using Val Badia as an additional reference. The authors additionally released training datasets for Gherdëina–Italian Val Badia–Gherdëina and used them to fine-tune an NLLB model to translate between the three languages. They used the newly translated FLORES dataset to benchmark the MT performance of this model and four LLMs (with and without retrieval of few-shot examples from the parallel training dataset). They found that even though retrieval helps, translation into Ladin variants remains a clear challenge for current LLMs.

Jumashev et al. (2025) expand Seed to Kyrgyz by post-editing LLM-based translations from English (using also Kazakh and Russian lexical resources) with a subsequent review to ensure term consistency throughout the dataset. Two post-edition techniques that the authors emphasize are fragmentation of a complex English sentence into two or more Kyrgyz sentences, more fluent under the Kyrguz SOV sentence structure, and a careful choice between native Kyrgyz words and Russian or English calques for scientific terms.

To demonstrate the effectiveness of the resulting parallel dataset, the authors finetuned four multilingual models on it and demonstrate gains in translation performance of each model on FLORES+ and X-WMT (Mirzakhalov et al., 2021).

Mæhlum et al. (2025) revise the FLORES+ dataset in Norvegian Bokmål and create a new version of it in Radical Bokmål, a sub-variety that is closer to spoken Norvegian dialects than the more Danish-like conservative Bokmål that dominates the formal discourse. The authors provide a detailed explanation of the difference between the varieties and the grammatical and lexical mistakes that had existed in the Bokmål FLORES+ dataset (such as anglicisms, word-by-word translations and problems in agreement) and had required correction (with the revisions affecting two thirds of the FLORES+ sentences). The authors demonstrate that the new version of the dataset, cleaned from anglicisms, serves as a more challenging reference set for English-Bokmål translation than the previous version.

Mamasaidov et al. (2025) extended FLORES+ to Southern Uzbek, a variety spoken in Afghanistan, written in Arabic script, and substantially different from Northern Uzbek spoken in Uzbekistan and written in Latin. The challenges of understanding and generating Southern Uzbek include the ambiguity of Arabic vowel characters and the use of a zero-width non-joiner character (U+200C) to separate the words' suffixes. Apart from the FLORES+ dev set translation into Southern Uzbek performed by a single native linguist, the paper contributes automatically aligned parallel dataset of the Southern and Northern Uzbek sentences, a NLLB model fine-tuned with this data and evaluated with FLORES+, and scripts for transliteration Southern Uzbek into Latin and for post-correction of missing U+200C characters. The newly finetuned model outperforms the strong LLM baselines on translation into Southern Uzbek, demonstrating their gap in supporting this language.

Marmonier et al. (2025) expand OLDI-Seed to French with the purpose of serving as a pivot language for the under-resourced regional languages of France. Each Seed sentence has been translated from English with 9 different MT systems, and two native French speakers selected and post-edited the most promising translation candidate from each such set. Finally, the translations were processed through a grammar checker. For validating the post-edited translations, the authors use MetricX-24 quality estimation system (Juraska et al., 2024), demonstrating that the human translations result in lower predicted error rates than any of the MT candidates. The paper emphasizes the terminological complexity of the Seed dataset and the challenges of producing fluent French translations despite the issues sometimes found in the English source sentences.

Oktem et al. (2025) revised FLORES+ and OLDI-Seed sentences in Standard Moroccan Tamazight as a part of the Awal initiative. The FLORES sentences were revised by two linguists using English as reference; Seed was revised by three professional Tamazight translators with English and Arabic references. 36% of FLORES and overall and 40% of Seed sentences required correction of spelling mistakes, transliteration errors, unnecessary or malformed loanwords, and mistranslations. The authors fine-tuned an NLLB-based model with the corrected Seed dataset and other Tamazight-English parallel datasets and evaluated it alongside with the original NLLB models and commercial LLMs on the original and corrected FLORES dataset. They found that the corrected FLORES dataset yields better MT evaluation metrics and that fine-tuning with the Seed data improves NLLB performance, making the model outperform the LLMs in the English-Tamazight direction.

Vamvas et al. (2025) expanded the WMT24++ benchmark with six varieties of the Romansh language: Rumantsch Grischun, a supra-regional variety, and five regional varieties: Sursilvan, Sutsilvan, Surmiran, Puter, and Vallader. The benchmark texts were translated from German by hired professionals who natively speak German and a Romansh variety and reviewed by two expert linguists. For automatic validation of the translations, the authors used language identification with a FastText model and cross-variety ChrF++ scores, demonstrating that the texts in the Romansh varieties are similar but distinguishable from each other. The resulting benchmark was used to assess the performance of MT system and LLMs on translation between German and Romansh, demonstrating that although some models already understand Romansh pretty well, translation into it is still challenging.

4.2 Other dataset extensions

It should be noted that not all contributors to the OLDI datasets submitted shared task papers. In the last year, FLORES+ has also received new translations in Chuvash, Dargwa, and Meadow Mari, regional languages in Russia, and incorporated the translations into Nko (Doumbouya et al., 2023) and five Indic languages (Gala et al., 2023): Bodo, Dogri, Konkani, Sindhi and Manipuri. Seed has been extended with the Nko language (Doumbouya et al., 2023).[4]

5 Discussion

Creating and maintaining language resources and technologies is hard, especially massively multilingual ones.

*Despite recent releases of state-of-the-art large-scale models (**?**) and the growing attention directed at speech and sign language translations (**??**Rust et al., 2024), the work on text-based MT remains ongoing. This is particularly true for many of the world's under-served languages, which compete with their higher-resource counterparts for research attention. Without sustained interest and contributions to key evaluation and seed data sets, the delta between high and low-resource*

[4] See the detailed list of changes and their attributions in the `CHANGELOG.md` files and dataset cards in the FLORES+ and OLD-Seed repositories.

languages will continue to expand, exacerbating already prominent technical divides.

Covering 16 languages spanning five continents, the papers in this shared task present a rigorous effort to improve the quality and scope of such data sets. Taken collectively, the authors developed protocols and tools to both refine and introduce new languages to existing FLORES+ and MT Seed data sets. Beyond their technical attributes, the work presented here also aligns with one of OLDI's core commitments: to be community-centric. Every paper in this shared task involves engaging with speakers of the languages of interest, with many authors being native speakers themselves. The linguistics expertise and cultural nuances these researchers brought, alongside the personal convictions many may have, culminated in a body of work that is both scientifically and socially meaningful. It is our hope that the papers showcased in this shared task are the first of a long series designed to consolidate the building blocks needed to advance language technologies for under-served linguistics communities across the world.

6 Conclusions

We presented the results of the WMT 2025 OLDI shared task. We accepted 8 submissions covering 16 languages, including the new SMOL dataset covering 123 languages, and extensions or revisions of the existing foundational datasets, FLORES+, OLDI-Seed, and WMT24++, in 14 language varieties. We are grateful to all the participants for their contributions and we hope that they would be soon adopted by the research community, enhancing a positive feedback loop between the developers of language technologies and the communities of speakers of all languages, including those who have been under-served by the modern tech.

References

Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R Costa-jussà, Joe Chuang, David Dale, Cynthia Gao, Jean Maillard, Alex Mourachko, Christophe Ropers, and 1 others. 2025. Bouquet: dataset, benchmark and open initiative for universal quality evaluation in translation. *arXiv preprint arXiv:2502.04314*.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Moussa Koulako Bala Doumbouya, Djibrila Diane, Baba Mamadi Diane, Solo Farabado, Edoardo Ferrante, Alessandro Guasoni, Mamadou K. Keita, Sudhamoy DebBarma, Ali Kuzhuget, David Anugraha, Muhammad Ravi Shulthan Habibi, and 3 others. 2025. Smol: Professionally translated parallel data for 115 under-represented languages. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 740–760, Suzhou, China. Association for Computational Linguistics.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Marta R. Costa-jussà, Bokai Yu, Pierre Andrews, Belen Alastruey, Necati Cihan Camgoz, Joe Chuang, Jean Maillard, Christophe Ropers, Arina Turkatenko, and Carleigh Wood. 2025. 2M-BELEBELE: Highly multilingual speech and American Sign Language comprehension dataset download PDF. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10893–10904, Vienna, Austria. Association for Computational Linguistics.

Daniel Deutsch, Eleftheria Briakou, Isaac 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. *Preprint*, arXiv:2502.12404.

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.

Christian Federmann, Tom Kocmi, and Ying Xin. 2022. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. 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 (WMT 2025)*, pages 698–708, 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, and 1 others. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *arXiv preprint arXiv:2305.16307*.

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.

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.

Murat Jumashev, Alina Tillabaeva, Aida Kasieva, Turgunbek Omurkanov, Akylai Musaeva, Meerim Emil kyzy, Gulaiym Chagataeva, and Jonathan North 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 (WMT 2025)*, pages 725–739, Suzhou, China. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. 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.

Jean Maillard, Laurie Burchell, 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.

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 (WMT 2025)*, pages 718–724, Suzhou, China. 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 (WMT 2025)*, pages 685–697, Suzhou, China. Association for Computational Linguistics.

Jamshidbek Mirzakhalov, Anoop Babu, Aigiz Kunafin, Ahsan Wahab, Bekhzodbek Moydinboyev, Sardana Ivanova, Mokhiyakhon Uzokova, Shaxnoza Pulatova, Duygu Ataman, Julia Kreutzer, Francis Tyers, Orhan Firat, John Licato, and Sriram Chellappan. 2021. Evaluating multiway multilingual NMT in the Turkic languages. In *Proceedings of the Sixth Conference on Machine Translation*, pages 518–530, Online. 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 (WMT 2025)*, pages 761–769, Suzhou, China. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 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 (WMT 2025)*, pages 709–717, Suzhou, China. Association for Computational Linguistics.

Phillip Rust, Bowen Shi, Skyler Wang, Necati Cihan Camgoz, and Jean Maillard. 2024. Towards privacy-aware sign language translation at scale. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8624–8641, Bangkok, Thailand. Association for Computational Linguistics.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Garrett Tanzer. 2025. FLEURS-ASL: Including American Sign Language in massively multilingual multitask evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6167–6191, Albuquerque, New Mexico. Association for Computational Linguistics.

Jannis Vamvas, Ignacio Pérez Prat, Not Battesta Soliva, Sandra Baltermia-Guetg, Andrina Beeli, Simona Beeli, Madlaina Capeder, Laura Decurtins, Gian Peder Gregori, Flavia Hobi, Gabriela Holderegger, Arina Lazzarini, Viviana Lazzarini, Walter Rosselli, Bettina Vital, Anna Rutkiewicz, and Rico Sennrich. 2025. Expanding the wmt24++ benchmark with rumantsch grischun, sursilvan, sutsilvan, surmiran, puter, and vallader. In *Proceedings of the Tenth Conference on Machine Translation (WMT 2025)*, pages 665–684, Suzhou, China. Association for Computational Linguistics.

A Full languages list

Table 2 describes the full list of language varieties contributed under the current shared tasks, except the 123 SMOL languages, for which we refer the reader to **?**.

actually fill the table

Language	Variety	ISO 639-3	ISO 15924	Glottocode	Contributor	Contributions
Ladin	Val Badia	lld	Latn	badi1244	Frontull et al. (2025)	FLORES+ new data
	Gherdëina	lld	Latn	gard1241	Frontull et al. (2025)	FLORES+ new data
Kyrgyz		kir	Cyrl	???	Jumashev et al. (2025)	Seed new data
Norvegian Bokmål	moderate	nob	Latn	???	Mæhlum et al. (2025)	FLORES+ revision
	radical	nob	Latn	???	Mæhlum et al. (2025)	FLORES+ revision
Southern Uzbek		uzs	Arab	???	Mamasaidov et al. (2025)	FLORES+ new data (??)
French		fra	Latn	???	Marmonier et al. (2025)	Seed new data
French		fra	Latn	???	Marmonier et al. (2025)	Seed new data

Table 2: A summary of all contributions to the WMT 2025 Shared Task of the Open Language Data Initiative.