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 (OLDI). Participants were invited to contribute to the existing massively multilingual open datasets supported by OLDI (FLORES+, OLDI-Seed, WMT24++), or create new resources in line with OLDI's aims. We accepted eight submissions: seven extensions or revisions of the existing datasets and one submission with a new massively parallel training dataset, SMOL. These contributions advance the coverage and quality of multilingual datasets, especially since for many languages, they are the first publicly-available training or evaluation data for machine translation. All contributions are released under permissive open-source licenses.

1 Introduction

Recent advances in machine translation (MT) have resulted in system output which is indistinguishable from that of human translators (Kocmi et al., 2024). However, even if we assume that the task of MT is 'solved', this would be true only for a small number of well-resourced language pairs. Achieving such high task performance requires abundant parallel training data specific to the language pair and domain of interest, 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 the parallel training data necessary to train MT models, as well as the 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. In this paper, we describe the second Shared Task for Open Language Data Initiative (OLDI) which invited language communities to contribute to high-quality, massively parallel and open-source datasets. Contributions could involve either extending existing datasets to new language varieties, making substantial improvements to existing datasets, or creating new massively multilingual parallel datasets. The datasets of interest to the shared task include (but are not limited to) FLORES+, OLDI-Seed (NLLB Team et al., 2024; Maillard et al., 2024), and WMT24++ (Deutsch et al., 2025). The OLDI itself is a community of researchers that maintains the former two datasets and promotes better language resources for under-served languages in general.[1]

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

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

*Equal contribution

[1] https://oldi.org/

[2] https://huggingface.co/collections/openlanguagedata

Proceedings of the Tenth Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, pages 495–502
November 8–9, 2025.

Language	Variety	ISO 639-3	ISO 15924	Glottocode	Contributors	Contributions
123 languages					Caswell et al. (2025)	SMOL (new dataset)
Ladin	Val Badia	lld	Latn	badi1244	Frontull et al. (2025)	FLORES+ new data
	Gherdëina			gard1241		
Kyrgyz		kir	Cyrl	kirg1245	Jumashev et al. (2025)	OLDI-Seed new data
Norvegian Bokmål	moderate	nob	Latn	under review	Mæhlum et al. (2025)	FLORES+ revision
	radical			under review		FLORES+ new data
Southern Uzbek		uzs	Arab	sout2699	Mamasaidov et al. (2025)	FLORES+ new data (dev split)
French		fra	Latn	stan1290	Marmonier et al. (2025)	OLDI-Seed new data
Standard Moroccan Tamazight		zgh	Tfng	stan1324	Oktem et al. (2025)	FLORES+ revision, OLDI-Seed revision
Romansh	Rumantsch Grischun	roh	Latn	ruma1247	Vamvas et al. (2025)	WMT24++ new data
	Sursilvan			surs1244		
	Surmiran			surm1243		
	Sutsilvan			suts1235		
	Puter			uppe1396		
	Vallader			lowe1386		

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

of the previous edition of this shared task, an additional 8 languages were included on top of several corrections to existing datasets (Abdulmumin et al., 2024; Ali et al., 2024; Gordeev et al., 2024; Kuzhuget et al., 2024; Mamasaidov and Shopulatov, 2024; Perez-Ortiz et al., 2024; Yu 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".[3] 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.

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

Participants to last year's edition of this shared task contributed three new languages (Ahmed et al., 2024; Cols, 2024; Ferrante, 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 document sizes and domains (though news is an overlapping domain). Unlike the two previous datasets, WMT24++ is managed by Google Research and not by OLDI, but in common with all datasets promoted by OLDI, it is released under a permissive license (Apache License 2.0).

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) and BOUQuET (Andrews et al., 2025), as well as other parallel datasets that could be reused for MT like Global MMLU (Singh et al., 2025) or MCS-350 (Agarwal et al., 2023). 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 the shared task was to expand high-quality, massively-parallel and open-source datasets to improve the resources available for multilingual applications like MT. Contributions could consist of the addition of new language varieties to existing datasets, substantial improvements to existing datasets, or novel datasets compatible with the aims of OLDI. Most contributions were to the datasets managed by OLDI: FLORES+ and OLDI-Seed.

3.1 Contributing to FLORES+ and OLDI-Seed

We encouraged the contributors of new languages to FLORES+ and OLDI-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.[4]

For FLORES+ translation, we did not allow using or even referencing MT output, including post-editing, to avoid introducing any machine bias in this evaluation dataset. For OLDI-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). This is because OLDI-Seed is intended as MT training data rather than evaluation data and so is subject to less strict requirements.

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`.

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 (e.g. 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 OLDI-Seed as closely as possible to ensure data quality and documentation. We required contributed data to be released under an open license (allowing free research use as a minimum).

4 Submissions

4.1 Shared task submissions

Table 1 summarises the contributions accepted as part of the shared task and the languages that were involved. In the rest of this section, 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 Common Crawl data[5] to cover 5.5k of the most common English words (obtained by joining the GATITOS wordlist and the most frequent words in Common Crawl). The second part of the dataset, 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.

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

[5] https://commoncrawl.org/

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 Val Badia and Gherdëina, two varieties of the Ladin language which is spoken in Northern Italy. 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 and Val Badia–Gherdëina pairs 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 the Ladin variants remains a clear challenge for current LLMs.

Jumashev et al. (2025) expand OLDI-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-editing techniques that the authors highlight are breaking a complex English sentence into two or more Kyrgyz sentences to be more fluent under the Kyrgyz 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 Norwegian dialects than the more Danish-like conservative Bokmål that dominates formal discourse. The authors provide a detailed explanation of the difference between the varieties, followed by an overview of the grammatical and lexical mistakes present in the original Bokmål FLORES+ dataset, such as anglicisms, word-by-word translations and problems in agreement. The necessary revisions affected two thirds of the FLORES+ sentences. The authors demonstrate that the new version of the dataset, cleaned from anglicisms and overly literal translations, 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 and written in Arabic script. It is substantially different from Northern Uzbek, which is 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 also contributes an 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 of 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 the lack of previous support for 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 OLDI-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 OLDI-Seed dataset and the challenges of producing fluent French translations as a result of 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 whilst OLDI-Seed was revised by three professional Tamazight translators with English and Arabic references. 36% of FLORES and overall and 40% of OLDI-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 OLDI-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 OLDI-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 are native speakers of both German and a Romansh variety. The translations were then 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 fairly 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. OLDI-Seed has been extended with the Nko language (Doumbouya et al., 2023).[6]

5 Discussion

Creating and maintaining language resources and technologies is hard, especially massively multilingual ones. There are tradeoffs and compromises: between the number of language varieties covered and the depth of the support of each variety, between the difficulty of benchmarks and the ease of translating them into new languages, between the naturalness of the translation in the target language and its faithfulness to the source content. Without the active interest of the communities actually speaking the language, advancing the NLP technologies for many of the world's under-served languages is hardly possible.

The contributions of this year's OLDI shared task highlight some of the issues with existing multilingual datasets and put forward suggestions as to how these might be solved. One issue which was highlighted repeatedly by the teams translating OLDI-Seed is its terminological complexity and the requirement of specialized knowledge for translating it. However, the emergence of SMOL as an alternative seed training dataset for MT helps circumvent this issue. An additional issue is that many popular multilingual datasets are English-centric. This is a barrier for extension into languages whose speakers use other languages as a lingua franca. Contributions like French OLDI-Seed by Marmonier et al. (2025) mitigate this. Finally, the work of Mæhlum et al. (2025) and Oktem et al. (2025) on revising the OLDI datasets in Norvegian Bokmål and Tamazight, respectively, show the need for continuous improvement of massive parallel datasets, especially with the direct involvement of the community of speakers.

Since the previous round of the OLDI shared task, contributions to FLORES+ and OLDI-Seed have already propagated to massively multilingual NLP benchmarks (e.g. Luo et al. (2025)) and to the extension of foundation models to new languages (e.g. Tsiamas et al. (2025)). We hope that the new datasets, languages and revisions contributed in the current shared task will similarly lead to further improvements in NLP resources and MT research for under-resourced languages.

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

6 Conclusion

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 truly grateful to all participants for their work and we hope that these contributions are soon adopted by the research community, enhancing a positive feedback loop between the developers of language technologies and the communities of language speakers.

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. AA also acknowledges support from the US National Science Foundation under award CIRC-2346334.

References

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo Putini, Miehleketo Mathebula, 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.

Milind Agarwal, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. LIMIT: Language identification, misidentification, and translation using hierarchical models in 350+ languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14496–14519, Singapore. 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.

Felermino 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.

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.

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.

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.

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, 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*.

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.

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.

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.

Hengyu Luo, Zihao Li, Joseph Attieh, Sawal Devkota, Ona de Gibert, Shaoxiong Ji, Peiqin Lin, Bhavani Sai Praneeth Varma Mantina, Ananda Sreenidhi, Raúl Vázquez, and 1 others. 2025. Gloteval: A test suite for massively multilingual evaluation of large language models. *arXiv preprint arXiv:2504.04155*.

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.

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, 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.

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, Jusèp 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.

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.

Ioannis Tsiamas, David Dale, and Marta R. Costa-jussà. 2025. Improving language and modality transfer in translation by character-level modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20171–20187, Vienna, Austria. 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.

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.