Findings of the WMT 2024 Shared Task of the Open Language Data Initiative

Laurie Burchell[*]
University of Edinburgh

Jean Maillard[*]
Meta FAIR

Antonios Anastasopoulos
George Mason University

Christian Federman
Microsoft

Philipp Koehn
Johns Hopkins University

Skyler Wang
McGill University

Correspondence: info@oldi.org

Abstract

We present the results of the WMT 2024 shared task of the Open Language Data Initiative. Participants were invited to contribute to the FLORES+ and MT Seed multilingual datasets, two foundational open resources that facilitate the organic expansion of language technology's reach. We accepted ten submissions covering 16 languages, which extended the range of languages included in the datasets and improved the quality of existing data.

1 Introduction

Machine translation research has advanced at breakneck speed in recent years (Kocmi et al., 2023). That said, progress made in translation quality has largely been directed at high-resource languages, leaving many languages behind. More recently, the focus has shifted towards under-served languages (also called low-resource) (Haddow et al., 2022). Foundational, high-coverage datasets have made it easier to develop and evaluate language technologies for a growing number of languages. Given the high impact of these components, extending such datasets becomes imperative.

The aim of the WMT 2024 shared task of the Open Language Data Initiative (OLDI) is to empower language communities to contribute such key datasets. In particular, we solicited contributions to the MT evaluation dataset FLORES+ and the MT Seed dataset. Additionally, we also solicited other high-quality, human-verified monolingual text datasets in under-resource languages. This builds on previous work to create these datasets and extend machine translation (MT) models and evaluation tools to more languages (Guzmán et al., 2019; Goyal et al., 2022; NLLB Team et al., 2024; Maillard et al., 2023).

We accepted ten submissions to the task, and the data contributed covered 16 languages. We required all contributions to be released under open licenses so that they can be useful to as many community members as possible. We make the data available online and encourage future work to build on these foundational datasets even further.[1]

2 Related Work

In recent years, there has been a growing recognition of the need for high-quality, representative datasets to broaden access to language technologies across a more diverse range of languages. Several initiatives have emerged to address this need.

In machine translation, the FLORES family of datasets (Guzmán et al., 2019; Goyal et al., 2022; NLLB Team et al., 2024) and NTREX-128 (Federmann et al., 2022) have provided the research community with massively multilingual, professionally translated benchmark data that is open source; while NLLB-Seed (Maillard et al., 2023; NLLB Team et al., 2024) played a similar role but focused on training data. Since the release of these resources, several authors have provided coverage for new languages (Gala et al., 2023; Doumbouya et al., 2023; Aepli et al., 2023) or even extended the datasets to the speech modality (Conneau et al., 2022).

Thanks to the availability of higher-quality data for an increasingly larger number of languages, recent language identification models have been able to expand coverage. Projects such as AfroLID (Adebara et al., 2022) and OpenLID (Burchell et al., 2023) improved upon pre-existing models by a careful curation and auditing of existing data sources; while LIMIT (Agarwal et al., 2023) further expanded data coverage and performance by creating and releasing a new high-quality corpus.

Several crowdsourced projects have proven invaluable as a source of knowledge for under-served

[*]Equal contribution

[1]https://huggingface.co/openlanguagedata

Proceedings of the Ninth Conference on Machine Translation, pages 110–117

languages. The Tatoeba project,[2] not designed explicitly for language technologies but as a language learning aid, provides a large database of aligned multilingual sentences. Mozilla Common Voice (Ardila et al., 2020) has enabled communities to build open-source ASR corpora for their own language and counts over 160 languages to date. The Aya initiative (Singh et al., 2024) has created the largest instruction finetuning dataset for large language models.

3 Datasets: FLORES+ and MT Seed

3.1 FLORES+

One of the biggest challenges in extending effective natural language processing (NLP) to underserved languages is a lack of high-quality, high-coverage evaluation benchmarks. In particular, few benchmarks are suitable for evaluating multilingual translation, since this requires many-to-many alignment between different languages in the evaluation dataset.

The FLORES family of datasets was released to address this problem. While the first iteration of this dataset covered only three languages (Guzmán et al., 2019), following iterations increased coverage to 101 languages (FLORES-101, Goyal et al., 2022) and finally to over 200 languages as part of the "No Language Left Behind" project (FLORES-200, NLLB Team et al., 2024). The current iteration of this dataset set is managed by OLDI, and we refer to it as FLORES+ to disambiguate between the original datasets and this new actively developed version.

FLORES+ consists of sentences extracted from English Wikinews, Wikijunior, and Wikivoyage: 997 for the dev split and 1012 for the devtest split.[3] These were then professionally translated into each language (almost universally from English) and underwent quality assessment and adjustment as necessary. The fact that all sentences in all languages are translations of each other means that they can be used for any-way multilingual evaluation.

3.2 MT Seed

The MT Seed dataset (previously NLLB Seed) was created as a source of "starter data" for languages without publicly-available high-quality bitext in sufficient quantity for training NLP models (NLLB Team et al., 2024, p.23). Previous work showed that employing the relatively small amount of high-quality data in MT Seed for training models had a significant impact on performance even when larger but lower quality corpora are used (Maillard et al., 2023). By extending MT Seed, OLDI aims to improve the quality of NLP applications for underserved languages by providing an initial source of reliable training data.

MT Seed consists of around 6000 sentences sampled from the Wikipedia articles listed in English Wikimedia's "List of articles every Wikipedia should have". These were professionally translated into each of the 38 languages covered by the first iteration of this dataset (39 if including English). Since this dataset is intended as a source of training data rather than evaluation, it did not undergo the quality assurance as the FLORES family of datasets.

4 Shared Task Definition

The goal of this shared task was to expand the open datasets managed by OLDI. Primarily, we solicited contributions to FLORES+ and MT Seed (described in Section 3), which could be either fixes to existing data or entirely new translations. It also accepted other high-quality, human-verified monolingual text datasets in under-resource languages.

4.1 Contributing to FLORES+ and MT Seed

To contribute to FLORES+ and MT Seed, we encouraged participants to translate from English into the target language so as to follow the original standard FLORES-200 workflow (NLLB Team et al., 2024, p.21). We required that translations were performed by qualified, native speakers of the target language and that translators acknowledged our translation guidelines (Appendix A). We strongly encouraged the verification of the data by at least one additional native speaker.

The acceptability of machine-translated content varied between the two datasets. Since the FLORES+ dataset is used to evaluate MT systems, new contributions must be entirely human-translated. Using or even referencing MT output was not allowed, including post-editing. However, post-edited MT content was allowed for contributions to MT Seed, provided all content was verified manually. This was done because MT Seed is intended for training rather than evaluation and, therefore, has less stringent translation requirements.

[2] https://www.tatoeba.org

[3] The separate blind test set, originally developed by Meta, is not managed by OLDI and is not part of FLORES+.

Participants were encouraged to provide experimental validation to demonstrate the quality of their submitted datasets. Due to the heterogeneous nature of submissions, we left the exact nature of the experimental validation up to the participants, though we gave some suggestions. For example, MT Seed data contributions could train a simple MT model and evaluate it on FLORES+.

All submissions were labeled with the same standardized language codes used throughout OLDI. These are made up of three parts, separated by underscores:

- An ISO 639-3 language code. Macrolanguage codes must not be used if a more specific code is possible: e.g., cmn, yue, wuu, etc., rather than zho.
- An ISO 15924 script code
- A Glottocode identifying the specific language variety.

For example, `apc_Arab_sout3123` indicates South Levantine Arabic written in the Arabic script.

All submissions were accompanied by a dataset card summarizing key facts about the data and how it was created. This is critical to foster informed and responsible use of the submitted data (Pushkarna et al., 2022). Submitted datasets were required to be released under the open CC BY-SA 4.0 license to match FLORES+ and MT Seed.

4.2 Contributing other monolingual data

Contributions of monolingual data had similar requirements to those for FLORES+ or MT Seed. The aim was to collect high-quality, human-verified monolingual text in multiple under-served languages for training NLP tools and systems. Synthetic data of any kind was not allowed. Parallel datasets were excluded from the scope of the shared task to not conflict with existing corpus-building efforts like Opus (Tiedemann, 2009).

For FLORES+ and MT Seed, submissions were encouraged to be manually verified by native speakers of the target language. All submissions needed to be accompanied by a data card and released under an open license (allowing free research use as a minimum).

5 Submissions

There were 24 expressions of interest in the shared task, and we ultimately accepted 10 papers. Table 1 summarizes the data submitted. We describe each submission in the following section.

Abdulmumin et al. (2024) contributed an improved version of the FLORES+ datasets for Hausa, Northern Sotho (Sepedi), Xitsonga, and isiZulu. They carried out error analysis of the datasets for the four languages and found problems such as poor translation of named entities, incorrect handling of morphological changes, a lack of consistency in vocabulary, and poor handling of borrowed terms. The Hausa dataset was particularly weak, with evidence that it was built upon Google Translate outputs. The participants corrected the translations following the guidelines in the shared task description and evaluated the alterations to the dataset using a range of metrics.

Ahmed et al. (2024) contributed a translation of MT Seed into the Bangla variety of Bangla/Bengali, an Indo-Aryan language that is the official language of Bangladesh and the state of West Bengal in India (as well as others). The dataset was translated by a native speaker with translation experience, per the OLDI translation guidelines. They validated the quality of their dataset by fine-tuning a range of pre-trained multilingual models on their generated translations and compared performance with the same pre-trained models fine-tuned on different but comparable datasets. They found that the models pre-trained on their translation of MT Seed showed the best performance after controlling for dataset size.

Ali et al. (2024) produced a translation of the FLORES+ dataset into the Central variety of Emakhuwa, a Bantu language spoken primarily in Mozambique. They verified their translation by using a second translator to revise the work of the first, followed by quality assessment involving three raters using a Direct Assessment pipeline. The participants conducted several experiments to benchmark current progress in Emakhuwa–Portuguese MT. They found that a lack of standardized orthography remains a challenge for Emakhuwa MT, though multiple reference translations can help with this issue.

Cols (2024) released Seed-CAT, an open-source web application specifically designed to assist human translators in translating MT Seed dataset files.[4] Using Seed-CAT, they produced a trans-

[4] https://github.com/josecols/seed-cat

Contributors	Type of contribution	Languages(s)
Abdulmumin et al. (2024)	FLORES+ (corrected)	Hausa, Northern Sotho (Sepedi), Xitsonga, isiZulu.
Ahmed et al. (2024)	MT Seed	Bangla/Bengali
Ali et al. (2024)	FLORES+ (new)	Emakhuwa
Cols (2024)	MT Seed (new) and CAT tool	Spanish (Latin American)
Ferrante (2024)	MT Seed (new)	Italian
Gordeev et al. (2024)	FLORES+ (new)	Erzya
Kuzhuget et al. (2024)	FLORES+ (new)	Tuvan
Mamasaidov and Shopulatov (2024)	FLORES+ devtest (new)	Karakalpak
Perez-Ortiz et al. (2024)	FLORES+ (new and corrected)	Aragonese, Aranese, Asturian, Valencian
Yu et al. (2024)	FLORES+ (new)	Wu Chinese

Table 1: A summary of all accepted contributions to the WMT 2024 Shared Task of the Open Language Data Initiative.

lation of MT Seed into Latin American Spanish. To validate their dataset's quality, they trained an English–Spanish MT model using the MT Seed data and compared its performance to models trained to translate between English and three Italic languages using existing MT Seed data. They found similar performance, suggesting that quality was similar to existing data in MT Seed.

Ferrante (2024) contributed a translation of MT Seed into Italian, building on a previous translation by Haberland et al. (2024). For this submission, the existing post-edited machine translation was reviewed and amended by two native speakers. The dataset was verified by training an Italian–Ligurian MT system and finding comparable results to those of Haberland et al. (2024).

Gordeev et al. (2024) contributed a translation of FLORES+ into Erzya, a Finno-Ugric language spoken primarily in Russia. As part of their work, they created a set of neologisms to aid future translators working in the digital space. They used their FLORES+ translation to evaluate the quality of existing English–Erzya and Russian–Erzya MT systems and train new competitive models for translating these language pairs.

Kuzhuget et al. (2024) translated the FLORES+ dataset from Russian into the Central dialect of Tuvan, a Turkic language primarily spoken in the Republic of Tuva in South Central Siberia, Russia. The team of translators worked from guidelines prepared in Russian to ensure consistent and high-quality translation.

Mamasaidov and Shopulatov (2024) contributed a translation of FLORES+ devtest split into Karakalpak, a Turkic language primarily spoken in the Republic of Karakalpakstan, which is an autonomous region within Uzbekistan. In addition, they also released a training dataset containing 100,000 sentence pairs for each of the language pairs: Uzbek–Karakalpak, Russian–Karakalpak, and English–Karakalpak. They carried out MT experiments using their datasets, releasing the trained models for further research.

Perez-Ortiz et al. (2024) contributed translations of FLORES+ into four Romance languages spoken in Spain: specifically new datasets for Aragonese, Aranese, and Valencian, and a corrected dataset for Asturian. The datasets were used as part of the evaluation of a shared task on MT from Spanish to low-resource languages of Spain (Sánchez-Martínez et al., 2024). Even though post-edited MT was used in the creation of these datasets, they were exceptionally accepted due to their use in a major shared task with the use of post-editing flagged in the metadata.

Yu et al. (2024) contributed a translation of FLORES+ into the Chongming dialect of Wu Chinese. The translation was done by two native speakers and checked by a third. Since Wu Chinese is typically colloquial while FLORES+ contains relatively formal text, the translators examined online written content and asked for community guidance about translations on fora to arrive at the best translations. To validate their dataset, the participants ran a three-way language identification task between Wu Chinese, Mandarin Chinese, and Yue Chinese. Their language identification model could distinguish between the three language varieties with high accuracy, though there was some confusion between Mandarin and Wu Chinese.

6 Discussion

Despite recent releases of state-of-the-art large-scale models (NLLB Team et al., 2024) and the growing attention directed at speech and sign language translations (Seamless Communication et al., 2023a,b; 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 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.

7 Conclusion

We presented the results of the WMT 2024 OLDI shared task. We accepted ten submissions covering 16 languages, which extend the range of languages included in the foundational datasets FLORES+ and MT Seed. We thank all participants for their contributions and hope that this shared task encourages further efforts towards improved language technologies for more language varieties.

References

Idris Abdulmumin, Sthembiso Mkhwanazi, Mahlatse S. Mbooi, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Neo N. 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*, Miami, USA. Association for Computational Linguistics.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022. AfroLID: A neural language identification tool for African languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1958–1981, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Noëmi Aepli, Chantal Amrhein, Florian Schottmann, and Rico Sennrich. 2023. A benchmark for evaluating machine translation metrics on dialects without standard orthography. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1045–1065, Singapore. 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*, Miami, 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*, Miami, USA. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. 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*, Miami, USA. Association for Computational Linguistics.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *Preprint*, arXiv:2205.12446.

Moussa Doumbouya, Baba Mamadi Diané, Solo Farabado Cissé, Djibrila Diané, Abdoulaye Sow, Séré Moussa Doumbouya, Daouda Bangoura, Fodé Moriba Bayo, Ibrahima Sory 2. Condé, 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*, Miami, USA. Association for Computational Linguistics.

Jay Gala, Pranjal A Chitale, A K Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar M, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Transactions on Machine Learning Research*.

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*, Miami, 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.

Christopher R. Haberland, Jean Maillard, and Stefano Lusito. 2024. Italian-Ligurian machine translation in its cultural context. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 168–176, Torino, Italia. ELRA and ICCL.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. 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*, Miami, 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 and Abror Shopulatov. 2024. Open Language Data Initiative: Advancing low-resource machine translation for Karakalpak. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. 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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers,

Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

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, Jusép 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*, Miami, USA. Association for Computational Linguistics.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826.

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.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023a. Seamlessm4t-massively multilingual & multimodal machine translation.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023b. Seamless: Multilingual expressive and streaming speech translation. *Preprint*, arXiv:2312.05187.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Felipe Sánchez-Martínez, Juan Antonio Perez-Ortiz, Aaron Galiano Jimenez, and Antoni Oliver. 2024. Findings of the WMT 2024 shared task translation into low-resource languages of Spain: Blending rule-based and neural systems. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. Association for Computational Linguistics.

Jörg Tiedemann. 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

Hongjian Yu, Yiming Shi, Zherui Zhou, and Christopher Haberland. 2024. Machine translation evaluation benchmark for Wu. In *Proceedings of the Ninth Conference on Machine Translation*, Miami, USA. 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 for evaluation purposes, such as for FLORES+, using or even referencing machine translation output is not allowed (this includes post-editing).
- **Note** that some machine translation services – including DeepL, Google Translate, and ChatGPT – prohibit the use of their output for training other translation or AI models, so their use is not permitted.

General Guidelines

1. You will be translating sentences coming from different sources. Please refer to the source document if available.
2. Do not convert any units of measurement. Translate them exactly as noted in the source content.
3. 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.
4. Provide fluent translations without deviating too much from the source structure. Only allow necessary changes.
5. 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.
6. Do not ignore any meaningful text present in the source.
7. 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.).
8. 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.
9. For proper nouns and common abbreviations, please see the guidelines on Named Entities below.
10. 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.
11. 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.
12. Foreign words and phrases used in the text should be kept in their original language when necessary to preserve the meaning of the sentence (e.g., if given as an example of a foreign word).

Named entities

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

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