A French Version of the SMOLSENT Corpus

Malik Marmonier **Rachel Bawden** **Benoît Sagot**
Inria, Paris, France
{firstname.lastname}@inria.fr

Abstract

We present a French version of the SMOLSENT corpus, a contribution to the WMT 2026 Open Language Data Initiative (OLDI) shared task. Originally curated to maximize unique-word coverage across 863 English sentences, SMOLSENT constitutes a valuable, token-efficient training set for the machine translation of low-resource languages. We describe our translation workflow, which relied on a custom-built interface to select and refine translation hypotheses generated by a traditional encoder-decoder model, alongside two state-of-the-art reasoning-enabled large language models (LLMs). Experimental validation based on the MetricX-24-XXL model for quality estimation, supported by bootstrap resampling significance tests, indicates that human post-editing results in significant quality improvements over all raw model outputs, even as reasoning models like Gemma-4-31B-it achieve a competitive 39.51% zero-edit rate. This French partition is not an end in itself, but a pivot resource meant to facilitate ongoing data collection efforts for under-resourced regional languages of France, such as Savoyard and Gallo.

1 Introduction

While machine translation (MT) has advanced tremendously in recent years, its progress remains overwhelmingly concentrated on a small set of high-resource languages (Joshi et al., 2020; Haddow et al., 2022). For most language varieties, the development of state-of-the-art MT systems is hampered by a lack of quality training and test data. The WMT Open Language Data Initiative (OLDI) shared tasks (Burchell et al., 2024; Dale et al., 2025) aim to redress this imbalance by empowering language communities to expand key foundational datasets. We contribute to this effort by compiling a French version of the SMOLSENT dataset (Caswell et al., 2025). Originally curated to maximize unique-word coverage across 863 English sentences, SMOLSENT constitutes a valuable, token-efficient training set for low-resource MT.

Since French is a high-resource language, this new partition is naturally not an end in itself; rather, it is meant to serve as a pivot for the under-resourced languages of France, such as Savoyard, Réunion Creole, or Gallo, languages which typically exist in a diglossic relationship with French and whose translators are predominantly native French speakers likely to benefit from a French source text or to need it as a prerequisite.

In the following sections, in compliance with shared task requirements, we present the French language, our data collection process, which followed a machine-translation-post-editing (MTPE) approach, and the results of validation experiments conducted through quality estimation.

2 Linguistic Overview

French is a Romance language.[1] It evolved from Vulgar Latin, which was brought into Northern Gaul by the Roman conquest and gradually replaced the local Celtic languages, taking hold first in urban centers before spreading across the wider territory (Rickard, 2014). This spoken Latin was markedly influenced by the Germanic languages of the fifth-century invaders who toppled the Roman Empire, and particularly by that of the Franks, which caused the loss of final unstressed syllables,[2] initiating a process of phonetic erosion that starkly distinguishes French from its southern Romance counterparts to this day. By the ninth century, this spoken vernacular had diverged so much from Latin that it was recognized as a distinct language, with the Oaths of Strasbourg of 842 considered the earliest French written document. In the following centuries, the dialect of the Île-de-France region

[1] Some parts of this section draw closely on material and language from our previous work (Marmonier et al., 2025).

[2] This point is now debated, however; see (Marchello-Nizia, 1995; Noske, 2009).

Proceedings of the Eleventh Conference on Machine Translation, pages 2517–2526
October 28-29, 2026 ©2026 Association for Computational Linguistics

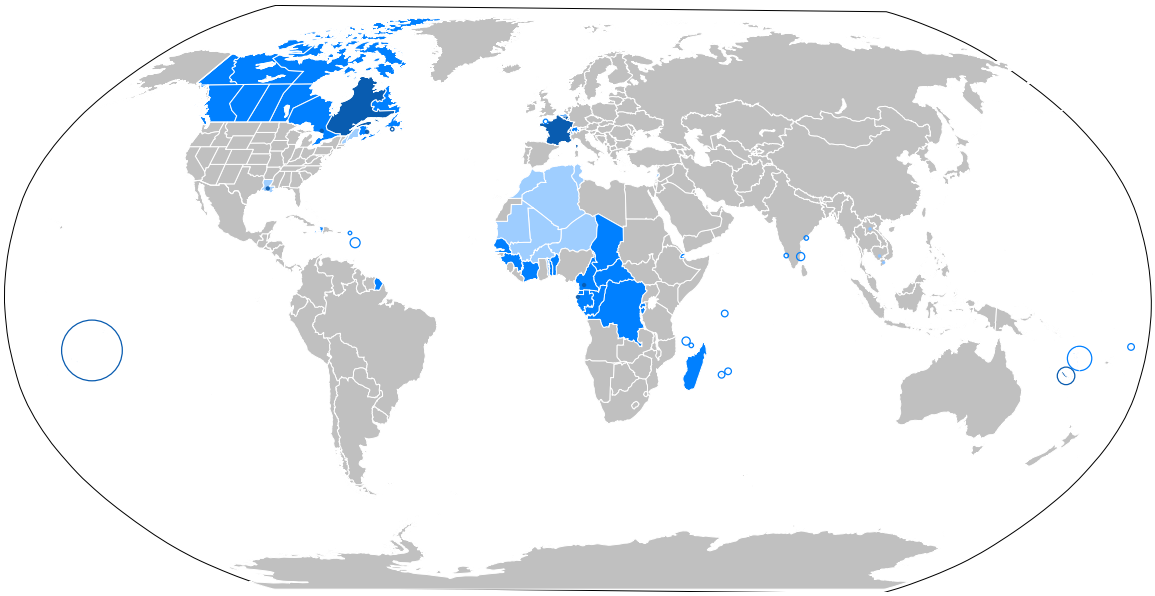

Figure 1: The French language in the world. States where French is the majority native language; states where it is a *de jure* official language; states where it is an administrative language or is culturally important without holding official status. Map by Wikimedia Commons contributors.

(Francien) began to emerge as a standard due to the political and cultural prominence of Paris.

This Old French still maintained a two-case declensional system and was written according to heterogeneous spelling conventions. The Middle French period—spanning the fourteenth to early seventeenth centuries—saw the final collapse of this two-case system and significant phonological changes, such as the muting of final consonants, making French spelling increasingly complex and etymological—a property for which French is faulted to this day by second-language learners and native speakers alike. The 1539 Ordinance of Villers-Cotterêts established French as the exclusive language of legal and judicial proceedings in the Kingdom of France, while scholars and poets like Joachim du Bellay fiercely defended it as a literary medium capable of rivaling both Latin and Greek (Du Bellay 2008 [1549]). During this time, French vocabulary expanded massively through deliberate borrowings from Latin, Greek, and Italian, at times replacing older Germanic loanwords. In the seventeenth century, the French language underwent rigorous codification and standardization, spearheaded by grammarians like de Malherbe (1862, [1606]) and Favre de Vaugelas (2009, [1647]), as well as the newly founded *Académie Française*.

This classical era established a rigid linguistic standard that remained the preserve of a privileged elite until the profound social upheavals of the French Revolution at the end of the eighteenth century. The revolutionary government viewed the linguistic fragmentation of France as a direct threat to national unity and republican ideals (Grégoire, 1794). This ideological stance drove the First French Republic to actively repress and attempt to destroy minority languages and regional dialects, an aggressive nation-building effort systematically enforced through a highly centralized administration, universal military conscription, and later through compulsory state education. Concurrently, the French language was exported globally through France's expanding colonial empire, initially to the Americas and the Caribbean, where colonization and the slave trade gave rise to French-based creoles, and later to territories across Africa, the Far East, and Oceania, and French consequently became a major international language.

Today, the latest report of the *Organisation Internationale de la Francophonie* (OIF) estimates the global Francophone population at 396 million, ranking French as the fourth most spoken language worldwide (Organisation internationale de la Francophonie, 2026). However, this demographic optimism relies heavily on a new counting method that includes millions of young African schoolchildren and masks a contested geopolitical reality. French is experiencing significant pushback in its former colonies, losing ground to both indigenous

languages and English. Following similar constitutional downgrades by Mali and Burkina Faso, Niger officially demoted French to a mere "working language" in early 2025, elevating Hausa to the status of national language. Underscoring this anti-colonial sentiment, all three Sahelian nations formally withdrew from the OIF in March 2025. French also faces fierce competition from English in North African educational systems and a continuous decline in second-language learner enrollments internationally (Lusin et al., 2023; Collen and Duff, 2025). English pressure, moreover, is not confined to France's former colonies: within France itself, English also dominates high-prestige domains like scientific publishing (OST, 2024) and higher education, and increasingly encroaches on the domestic sphere, with some scholars arguing that a dramatic surge of anglicisms across media, everyday speech, and corporate branding is reshaping the French language itself, in effect turning English into a linguistic superstratum within France (Meney, 2024)—a claim other linguists dispute (Les Linguistes atterrées, 2023).

Despite these trends, French still retains substantial international standing, making it a valuable pivot for future data collection efforts targeting the under-resourced languages of France and its former colonies. France French is one of several standardized varieties of French; owing to its extensive documentation, we adopted it as the target variety for our corpus.

3 Data Collection

In order to produce the French partition of SMOLSENT, we adopted an MTPE workflow, a *de facto* norm in the translation industry, particularly for the English-French language pair (International Organization for Standardization, 2017; ELIS Research, 2026). We obtained a set of three candidate translations per source segment, using three distinct systems comprising a "traditional" NMT model—NLLB-200 3.3B (NLLB Team, 2024)—and two reasoning LLMs—Gemma-4-31B-it (Google DeepMind, 2026) and DeepSeek V4 Pro (DeepSeek-AI et al., 2026).

For NLLB-200 3.3B, inference was conducted on the Jean Zay HPC cluster in FP16 precision via the `transformers` library (Wolf et al., 2020), using beam search decoding with a beam size of 4.

For the LLMs, we crafted zero-shot prompts—one per source segment—giving the models the persona of an expert English-French translator, asking for an accurate and fluent French translation of the English source segment, targeting the France French locale. The prompts encouraged the models to reflect on the task prior to outputting the final target-language segment between specified `<translation>...</translation>` XML tags. We provide a prompt sample in Appendix A.

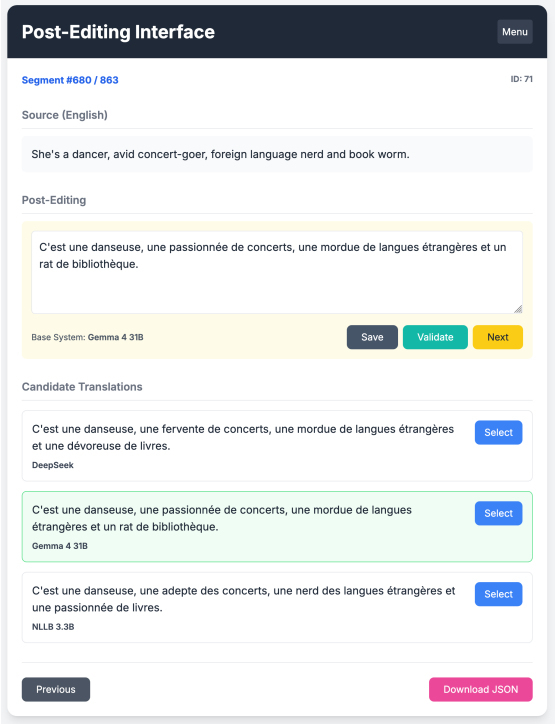

Figure 2: Our custom-built post-editing interface, adapted from Marmonier et al. (2025).

Gemma-4-31B-it's reasoning mode was triggered through the addition of a <|think|> token at the beginning of its system prompt, and inference was conducted on the Jean Zay HPC cluster via the vLLM framework (Kwon et al., 2023), with a sampling temperature of 0.3.

For DeepSeek V4 Pro, candidate translations were obtained via the official DeepSeek API with the model's reasoning effort set to "high", the lowest of the two settings available via the API at the time.[3]

The candidate translations were extracted from LLM output through regular expressions targeting the specified XML tags and loaded into an online post-editing interface we adapted from previous work (Marmonier et al., 2025). The interface presented each source segment with the three previ-

[3]The DeepSeek API silently ignores any temperature parameter passed while reasoning mode is enabled; the value used internally in this mode is not documented.

ously generated candidate translations and allowed the annotator to select the most promising translation hypothesis as a starting point and refine it in an editable text field (see Figure 2). The post-editors were the first and third authors of this work, both native French speakers with C2 command of English and translation experience.[4]

The SMOLSENT corpus was pleasant to translate, with minimal terminological challenges (compared to OLDI Seed, for instance) and no insurmountable complications resulting from disfluencies in the crawled source segments.[5] Out of the 863 segments that make up SMOLSENT, 495 (57.36%) required no alteration in post-editing, which is not unexpected for the language pair at hand.[6] A breakdown of these zero-edit matches reveals a clear advantage for the reasoning LLMs, with Gemma-4-31B-it and DeepSeek V4 Pro producing the final, unaltered translation for 39.51% (341 segments) and 25.26% (218 segments) of the total corpus, respectively. Conversely, the traditional NLLB-200 model lagged significantly behind its larger counterparts, generating a zero-edit candidate for a mere 3.13% (27 segments) of the compiled dataset.

The final version of the dataset was reviewed manually in the Grammalecte interface,[7] and all its quotation marks and apostrophes were normalized to their straight forms (" / " and '), rather than French guillemets (« / ») and curly apostrophes (’), against official stylistic conventions but in keeping with actual contemporary French digital writing practices.[8]

We provide a random sample of the compiled data in Appendix B.

[4]The second author, a native English speaker with C2 command of French, was consulted on occasion to dispel ambiguities in the source segments.

[5]Segment #743 “The entire castle of every Disney movie has featured a logo in the shape of an ‘s’!” was the only one whose meaning was not immediately transparent.

[6]We note that a segment being left unaltered does not reflect the silent post-editing work involved in verifying the correctness of its terminology, for instance.

[7]https://grammalecte.net/

[8]We note that this normalization operation was deliberately performed after the reported zero-edit matches were computed rather than during the post-editing work so as to not artificially distort these figures with trivial alterations. Should users require strict adherence to official typographical norms, restoring standard French guillemets and curly apostrophes from their straight counterparts is a trivial operation using regular expressions.

4 Experimental Validation

To assess the quality of the compiled French partition and measure the added value of human post-editing over raw system outputs, we evaluated all candidate translations alongside the final post-edited dataset using MetricX-24-XXL[9] (Juraska et al., 2024) in a reference-free Quality Estimation (QE) configuration. MetricX-24 scores predict translation error on a continuous scale from 0 to 25, where lower values denote higher translation quality.

Table 1 presents the mean MetricX-24-XXL scores across all 863 segments. To determine whether the differences in translation quality between “systems” were statistically significant, we conducted paired bootstrap resampling tests (Koehn, 2004), with $10,000$ resamples across all pairwise system combinations (see Figure 3).

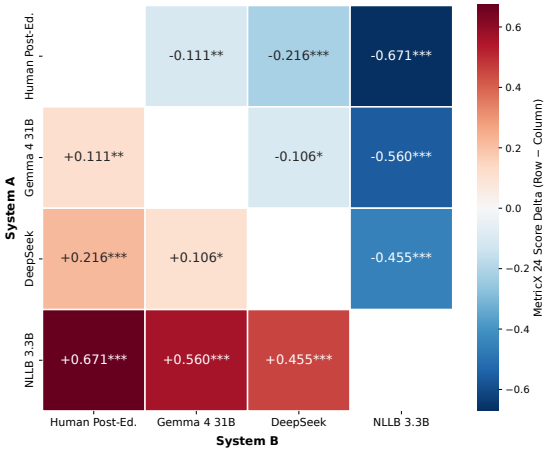

Figure 3: Heatmap of MetricX-24-XXL score differences between raw system outputs and human post-edits. Negative deltas (blue) indicate that the row system outperforms the column system. Statistical significance was computed via paired bootstrap resampling ($10,000$ iterations); asterisks denote significance level ($^{*}: p \leq 0.05$, $^{**}: p < 0.01$, $^{***}: p < 0.001$).

System	MetricX-24-XXL (↓)
NLLB-200 3.3B	3.1143
DeepSeek V4 Pro	2.6595
Gemma-4-31B-it	2.5540
Human Post-Ed.	**2.4434**

Table 1: Mean MetricX-24-XXL Quality Estimation scores across all 863 segments. Lower is better.

The validation results highlight three key empirical observations.

[9]metricx-24-hybrid-xxl-v2p6

Human post-editing results in statistically superior quality The final post-edited dataset achieved the best overall score (2.4434), demonstrating significant quality gains over all raw model outputs—including Gemma-4-31B-it ($\Delta = -0.111, p < 0.01$), DeepSeek V4 Pro ($\Delta = -0.216, p < 0.001$), and NLLB-200 3.3B ($\Delta = -0.671, p < 0.001$).[10]

Reasoning LLMs outperform traditional NMT Both reasoning-enabled LLMs generated significantly higher quality candidate translations than the traditional NLLB-200 3.3B baseline (3.1143). Among the raw candidate systems, Gemma-4-31B-it achieved the strongest performance (2.5540), followed by DeepSeek V4 Pro (2.6595).

Strict system hierarchy Pairwise significance tests reveal a clear ranking among automated systems. Gemma-4-31B-it significantly outperformed DeepSeek V4 Pro ($\Delta = -0.106, p < 0.05$), while both LLMs demonstrated a massive, statistically significant lead over NLLB-200 3.3B ($\Delta = -0.560$ and $\Delta = -0.455$, respectively; $p < 0.001$).

5 Related Work

Data-driven approaches to MT have been explored as early as the 1950s (Edmundson and Hays, 1958), but only came into their own in the early 1990s thanks to the pioneering work on statistical MT (SMT) of IBM researchers (Brown et al. 1990; Berger et al. 1994).[11] Anecdotally, the French-English language pair played a central role in these seminal efforts, largely thanks to the Canadian Hansard, a bilingual record of parliamentary proceedings. This role was so central, in fact, that a lasting convention in SMT literature consisted in using the letters f and e as the standard variables denoting source ($f$rench) and target ($e$nglish) language strings in equations, regardless of the actual languages considered.

Caswell et al. (2025) compiled SMOLSENT together with the larger SMOLDOC corpus as the two sentence- and document-level components of SMOL, an initiative that also incorporates the older GATITOS corpus of parallel wordlists (Jones et al., 2023) and is continuously extended through volunteer efforts.[12]

SMOL is one of several open datasets targeted by the OLDI shared task (Burchell et al., 2024; Dale et al., 2025) in order to grow foundational MT resources through community contributions. Among these are also the OLDI Seed corpus (originally the NLLB Seed corpus; NLLB Team et al., 2022; Maillard et al., 2023; Cols, 2024; Ferrante, 2024; Haberland et al., 2024; Ahmed et al., 2024; Marmonier et al., 2025; Oktem et al., 2025), the FLORES+ benchmark (Goyal et al., 2022; NLLB Team, 2024; Gordeev et al., 2024; Kuzhuget et al., 2024; Ali et al., 2024; Abdulmumin et al., 2024; Perez-Ortiz et al., 2024; Mæhlum et al., 2025; Oktem et al., 2025; Frontull et al., 2025; Lou et al., 2025; Samil and Adelani, 2026), and test sets such as NTREX-128 (Federmann et al., 2022), WMT24++ (Deutsch et al., 2025; Vamvas et al., 2025) and BOUQuET (Andrews et al., 2025). The present submission is a small step in this effort—fittingly, on the SMOL side.

6 Conclusion

We have presented a French partition of the SMOLSENT corpus, produced through an MTPE workflow in which candidate translations from NLLB-200 3.3B, Gemma-4-31B-it, and DeepSeek V4 Pro were selected and refined by native French speakers in a dedicated interface. Of the 863 source segments, 57.36% had at least one candidate translation that required no alteration on the part of the post-editors, with the reasoning-enabled LLMs accounting for the vast majority of these zero-edit matches. Quality estimation with MetricX-24-XXL nevertheless shows that human post-editing yields statistically significant gains over every raw system output, indicating that strong state-of-the-art LLMs still benefit from human revision, even for a high-resource language pair such as English-French.

This partition is intended primarily as a pivot resource for the regional languages of France, whose translators are typically native speakers of French. It has already served a similar purpose in a companion submission to the present shared task, where its French segments were used as the source side of a synthetic data pipeline for French→Savoyard translation.

[10] We further note that our choice to normalize the quotation marks and apostrophes to their straight forms might slightly penalize the human post-edited translations in the eyes of MetricX-24.

[11] Elements of this section draw closely on language from our previous work (Marmonier et al., 2025).

[12] See https://huggingface.co/datasets/google/smol.

Limitations

We acknowledge two related primary limitations arising from this work.

First, because the English source segments are isolated sentences drawn from a broader web crawl, their original document-level context was not easily recoverable. To account for the resulting ambiguity, our post-editing strategy favored translations that preserved the widest plausible range of meanings, rather than committing to a single, context-dependent reading. This cautious approach occasionally came at the cost of stylistic fluency and naturalness in the target language.

Second, our reliance on a machine translation post-editing (MTPE) workflow introduces the well-documented risk of translationese (Zhang and Toral, 2019; Graham et al., 2020). Even with careful human revision, target segments generated via MTPE may retain lexical or syntactic traces of their English source material—a phenomenon sometimes called "interference" by translation scholars (Toury, 1995) and generally considered to be a universal property of translated texts, but one that is likely more pronounced in an MTPE setting (Toral, 2019).

Acknowledgments

This work was funded by the French *Agence Nationale de la Recherche* (ANR) under the project TraLaLaM ("ANR-23-IAS1-0006"), as well as by Inria under the "*Défi*"-type project COLaF. It was also partly funded by Benoît Sagot's chair in the PRAIRIE-PSAI institute, funded by the ANR as part of the "France 2030" strategy under the reference ANR-23-IACL-0008.

This work was also granted access to the HPC resources of IDRIS under allocations 2025-AD011015117R2, 2025-AD011012254R5 and 2025-AD011017228 made by GENCI. The authors acknowledge the use of LLMs as coding assistants during the implementation of this work.

The authors wish to thank Isaac Caswell for providing valuable assistance during the initial phase of this project.

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.

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, Mark Duppenthaler, Nathanial Paul Ekberg, Cynthia Gao, Daniel Edward Licht, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, and Shireen Yates. 2025. BOUQuET : dataset, benchmark and open initiative for universal quality evaluation in translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27515–27535, Suzhou, China. Association for Computational Linguistics.

Adam L. Berger, Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, John R. Gillett, John D. Lafferty, Robert L. Mercer, Harry Printz, and Lubos Ures. 1994. The Candide System for Machine Translation. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

Laurie Burchell, Jean Maillard, Antonios Anastasopoulos, Christian Federmann, Philipp Koehn, and Skyler Wang. 2024. Findings of the WMT 2024 shared task of the open language data initiative. In *Proceedings of the Ninth Conference on Machine Translation*, pages 110–117, Miami, Florida, USA. Association for Computational Linguistics.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Djibrila Diane, Solo Farabado Cissé, Koulako Moussa Doumbouya, Edoardo Ferrante, Alessandro Guasoni, Christopher Homan, Mamadou K. Keita, Sudhamoy DebBarma, Ali Kuzhuget, David Anugraha, and 5 others. 2025. SMOL: Professionally translated parallel data for 115 under-represented languages. In

Proceedings of the Tenth Conference on Machine Translation, pages 1103–1123, Suzhou, China. Association for Computational Linguistics.

Ian Collen and Jayne Duff. 2025. Language Trends England 2025: Language teaching in primary, secondary and independent schools in England. Survey report, British Council.

Jose Cols. 2024. Spanish Corpus and Provenance with Computer-Aided Translation for the WMT24 OLDI Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 624–635, Miami, Florida, USA. Association for Computational Linguistics.

David Dale, Laurie Burchell, Jean Maillard, Idris Abdulmumin, Antonios Anastasopoulos, Isaac Caswell, and Philipp Koehn. 2025. Findings of the WMT 2025 shared task of the open language data initiative. In *Proceedings of the Tenth Conference on Machine Translation*, pages 495–502, Suzhou, China. Association for Computational Linguistics.

François de Malherbe. 1862. Commentaire sur Desportes. In Ludovic Lalanne, editor, *Œuvres de Malherbe*, volume 4, pages 249–474. Hachette, Paris.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, and 300 others. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence. *Preprint*, arXiv:2606.19348.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.

Joachim Du Bellay. 2008. *La Deffence et Illustration de la Langue françoyse*. Number DXLIII in Textes Littéraires Français. Librairie Droz, Genève.

H. P. Edmundson and D. G. Hays. 1958. Research methodology for machine translation. *Mechanical Translation*, 5(1):8–15.

ELIS Research. 2026. European language industry survey 2026: Trends, expectations and concerns of the european language industry. Technical report, ELIA, EMT, EUATC, FIT Europe, GALA, LIND, Women in Localization.

Claude Favre de Vaugelas. 2009. *Remarques sur la langue françoise*. Société de Publications Romanes et Françaises. Librairie Droz, Genève. First published 1647.

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*, pages 1061–1071, Suzhou, China. Association for Computational Linguistics.

Google DeepMind. 2026. Gemma 4 model card. `https://ai.google.dev/gemma/docs/core/model_card_4`.

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.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Henri Grégoire. 1794. Rapport sur la nécessité et les moyens d'anéantir les patois et d'universaliser l'usage de la langue française. Rapport, Convention nationale. Available on Wikisource.

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 Underresourced Languages @ LREC-COLING 2024*, pages 168–176.

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.

International Organization for Standardization. 2017. *ISO 18587:2017 Translation Services — Post-editing of Machine Translation Output — Requirements*. ISO, Geneva.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. GATITOS: Using a new multilingual lexicon for low-resource machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. 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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Les Linguistes atterrées. 2023. *Le français va très bien, merci*. Number 49 in Tracts. Gallimard, Paris.

Andrés Lou, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Miquel Esplà-Gomis, and Víctor M. Sánchez-Cartagena. 2025. FLORES+ mayas: Generating textual resources to foster the development of language technologies for Mayan languages. In *Proceedings of Machine Translation Summit XX: Volume 2*, pages 89–90, Geneva, Switzerland. European Association for Machine Translation.

Natalia Lusin, Terri Peterson, Christine Sulewski, and Rizwana Zafer. 2023. Enrollments in Languages Other Than English in US Institutions of Higher Education, Fall 2021. Report, Modern Language Association of America, New York, NY.

Petter Mæhlum, Anders Næss Evensen, and Yves Scherrer. 2025. Improved Norwegian Bokmål translations for FLORES. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1124–1132, Suzhou, China. Association for Computational Linguistics.

Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. Small Data, Big Impact: Leveraging Minimal Data for Effective Machine Translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.

Christiane Marchello-Nizia. 1995. *L'évolution du français: ordre des mots, démonstratifs, accent tonique*. Armand Colin, Paris.

Malik Marmonier, Benoît Sagot, and Rachel Bawden. 2025. A French version of the OLDI seed corpus. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1048–1060, Suzhou, China. Association for Computational Linguistics.

Lionel Meney. 2024. *Le naufrage du français, le triomphe de l'anglais: enquête*. L'espace public. Presses de l'Université Laval, Québec. Diffusion: Hermann (Paris).

NLLB Team. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

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. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *Preprint*, arXiv:2207.04672.

Roland Noske. 2009. Autonomous typological prosodic evolution versus the Germanic superstrate in diachronic French phonology. In Edward J. Rubin and 1 others, editors, *Romance Languages and Linguistic Theory 2007*, pages 231–252. John Benjamins, Amsterdam.

Alp Oktem, Mohamed Aymane Farhi, Brahim Essaidi, Naceur Jabouja, and Farida Boudichat. 2025. Correcting the tamazight portions of FLORES+ and OLDI seed datasets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1072–1080, Suzhou, China. Association for Computational Linguistics.

Organisation internationale de la Francophonie. 2026. *La Langue française dans le monde 2023–2026*. Éditions Gallimard.

OST. 2024. The scientific position of France in the world and in Europe: Analysis of various corpora of publications and European projects. Prepared under

the direction of Frédérique Sachwald; with contributions by Mounir Amdaoud, Agénor Lahatte, Esther Lardreau, and Françoise Laville.

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.

Peter Rickard. 2014. *A History of the French Language*. Routledge. Google-Books-ID: kHm3oAEACAAJ.

Hadia Mohmmedosman Ahmed Samil and David Ifeoluwa Adelani. 2026. Sudanese-Flores: Extending FLORES+ to Sudanese Arabic dialect. In *Proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026)*, pages 243–247, Rabat, Morocco. Association for Computational Linguistics.

Antonio Toral. 2019. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 273–281, Dublin, Ireland. European Association for Machine Translation.

Gideon Toury. 1995. *Descriptive Translation Studies – and Beyond*, volume 4 of *Benjamins Translation Library*. John Benjamins, Amsterdam.

Jannis Vamvas, Ignacio Pérez Prat, Not 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*, pages 1028–1047, Suzhou, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

A Prompt Sample

The following are populated examples of the system and user prompts used to obtain translation hypotheses from Gemma-4-31B-it and DeepSeek V4 Pro.

System Prompt

You are an expert translator. Your working languages are English and French.

User Prompt

Translate the following English segment into French (for the France locale). Provide a high-quality translation that is both accurate and fluent. You can reflect on the task, prior to producing your final translation. Your final translation must, however, be wrapped inside XML tags thusly: `<translation>your translation here</translation>`

Here is the English segment you must translate:
Processes outgoing mail, maintains files, orders supplies, and ships packages.

B Data Sample

We provide a random sample of the compiled dataset in Table 2.

ID	English Source	French Translation
53	The injured boy lost consciousness and fell to the pavement and his blood imbued the road dust.	*Le garçon blessé perdit connaissance et tomba sur la chaussée, et son sang imprégna la poussière de la route.*
274	Drop by and say hello; stay strong and live awesome, my friend.	*Passe dire bonjour ; reste fort et vis ta vie à fond, mon ami.*
613	Madam, your glasses are fogged up!	*Madame, vos lunettes sont embuées !*
787	He tells me he originally tried the injection technique and, as he expected, it was too fiddly and gave him zero success.	*Il me dit qu'il a d'abord essayé la technique de l'injection et que, comme il s'y attendait, c'était trop fastidieux et ne donnait aucun résultat.*
417	What I mean is, you can't buy stuff cheap and sell it dear.	*Ce que je veux dire, c'est qu'on ne peut pas acheter des trucs bon marché et les revendre cher.*
286	You can arrange in rows, in a circular pattern, whatever takes your fancy.	*Vous pouvez les disposer en rangées, en cercle, ou comme bon vous semble.*
806	Sets were moved, lighting was adjusted, and actors made costume changes.	*Les décors ont été déplacés, l'éclairage a été ajusté et les acteurs ont changé de costume.*
372	If you missed the signs of the first stage, you will not have a clue that the cow should have already delivered a calf.	*Si vous avez manqué les signes du premier stade, vous ne vous rendrez pas compte que la vache aurait déjà dû vêler.*
32	So, considering the ratio of net value added to gross output does not yield a complete picture.	*Ainsi, le rapport entre la valeur ajoutée nette et la production brute ne permet pas d'obtenir une vision complète.*
667	Use fresh basil, flax oil, or virgin coconut oil and replace the cheese with yeast.	*Utilisez du basilic frais, de l'huile de lin ou de l'huile de coco vierge, et remplacez le fromage par de la levure.*

Table 2: Random sample of ten sentence pairs from the French partition of the SMOLSENT dataset.