The MaTOS Pipeline for the Translation of Scientific Abstracts on the HAL Platform

Panagiotis Tsolakis[1] **Ziqian Peng**[1,2] **Laurent Romary**[1]
François Yvon[2] **Rachel Bawden**[1]
[1]Inria, Paris, France
[2]Sorbonne Université, CNRS, ISIR, Paris, France
firstname.lastname@inria.fr
lastname@isir.upmc.fr

Abstract

English dominates scientific publishing, which disadvantages researchers who are not native English speakers, especially those in the earlier stages of their careers. Being able to write and engage with scientific content written in their own language would clearly facilitate scientific production. The MaTOS project (Machine Translation for Open Science) seeks to reduce these barriers by developing machine translation tools for scientific documents in English and French. This article presents the design of the MaTOS pipeline for the HAL platform to automatically translate article abstracts, with author validation, to increase the number of bilingual abstracts on the platform. We also report preliminary experiments comparing translation of sentence, three-sentence chunks, and whole abstracts, evaluated using quality estimation metrics. We release all the code of the different stages of the pipeline.[1]

1 Introduction

A vast majority of academic publications are published in English (Gordin, 2015). While having a lingua franca has certain advantages, it also creates linguistic barriers for non-native speakers of English; they may feel less at ease both reading the literature and writing their own articles. In a world of ever-increasing publication, it is important to create a more inclusive research ecosystem that does not disadvantage non-native speakers and, on the contrary, embraces the diversity that comes with a multilingual community. Initiatives such as the *Helsinki Initiative on Multilingualism in Scholarly Communication*[2] recommend just that, encouraging the community to promote and facilitate writing in their own language, and for the Natural Language Processing (NLP) domain, the ACL 60-60 initiative[3] sought to break down linguistic barriers with their ambition to translate the ACL anthology into 60 different languages.

Machine translation (MT) is one way of both facilitating access to publications in English and enabling researchers to write in their own language. Progress in MT has been remarkable, thanks to neural architectures (Bahdanau et al., 2015), large-scale language models trained on large amounts of data (Lewis et al., 2020; Raffel et al., 2020), and, in more recent years, the use of large language models (LLMs) (Hendy et al., 2023; Zhu et al., 2024; Bawden and Yvon, 2023), which have provided new opportunities for longer document translation (beyond the sentence level) (Guo et al., 2025; O'Brien et al., 2025; Peng et al., 2025b) and the integration of domain-specific translation guidelines and terminologies (Lu et al., 2024a; Oncevay et al., 2025).

The MaTOS project[4] (Machine Translation for Open Science), a project funded by the French National Research agency, seeks to improve translation of academic documents between French and English (Bénard et al., 2023; Bawden et al., 2025). As part of the project, we have developed a pipeline to translate monolingual abstracts

[1]https://github.com/ANR-MaTOS/HAL_pipeline

[2]https://www.helsinki-initiative.org

[3]https://2022.aclweb.org/dispecialinitiative.html

[4]https://anr-matos.github.io

of scholarly publications submitted to the HAL platform from English into French and vice versa. The HAL platform is the French national scientific publication archive, hosting publications of a wide range of scientific domains. Its use is recommended, sometimes even mandatory, across all French public academic institutions.

As of February 2026, HAL hosts 4.5M references. Out of these, 1.7M references have a declared English abstract and no French abstract,[5] while 582k references have a declared French abstract and no English abstract. Through our translation pipeline, we aim to provide HAL users with an automatic translation of their monolingual abstracts in French or English, depending on the original source language, which the authors can then post-edit before it is associated with their publication. In this article, we present the design of the pipeline and also compare and analyse the translation quality of an LLM integrated into this pipeline when translating segments of varying lengths (isolated sentences, blocks of consecutive sentences and the whole abstract).

2 Related work

Traditionally, MT models have been trained to perform sentence-level translation, for computational reasons as well as the availability of large sentence-level parallel datasets. However, translating at the sentence level means that certain aspects of the translation cannot be ensured, notably those that span across several sentences or even whole documents such as lexical cohesion, which requires extra-sentential processing (Fernandes et al., 2023). Context-aware and document-level translation has therefore been an important area of study in MT from as far back as statistical MT (Hardmeier, 2012), and the task has been made more flexible with the use of neural MT models (Miculicich et al., 2018; Lopes et al., 2020) and nowadays with LLMs (Wang et al., 2023). The advent of LLMs that can handle large contexts, sometimes of thousands of tokens, has enabled the research community to envisage holistic document translation that can potentially ensure document-level consistency and coherence. Although theoretically whole-document translation would be ideal, it is less obvious that this is the optimal granularity for current models, particularly for longer documents. Despite the theoretical context window being big enough to accommodate the whole document and prompt, in practice it has been shown that additional problems can arise for the translation of longer documents, including duplicated segments, missing or hallucinated content, resulting in worse overall quality (O'Brien et al., 2025; Peng et al., 2025a). It has even been shown that sentences that appear far from the beginning of the input text are more affected by this decrease in translation quality than those at the beginning (Peng et al., 2025a). We therefore choose to test several granularities of translation in our pipeline to find the right balance. An additional advantage of LLMs for translation is that they can easily integrate additional information sources such as terminologies (Moslem et al., 2023; Lu et al., 2024a; Oncevay et al., 2025), style guidelines and domain information (Hu et al., 2024), as well as examples of the type of translation to be carried out (in the form of few-shot examples) (Tan et al., 2024) through prompting. We conduct experiments with both 0-shot and few-shot prompting for document-level translation.

A previous project had a similar aim to our own. The ANR COSMAT project (Lambert et al., 2012) aimed to translate entire PDF documents submitted to HAL, with the goal of having the authors post-edit, validate or reject the suggested translation. However, this pipeline was not integrated in the platform. Much has changed since, notably the MT technologies available for translation (and therefore the quality and usefulness of translation). Our ambition to translation abstracts initially rather than whole documents is partly due to it being more straightforward and more likely to be adopted by authors in this initial implementation of the pipeline, and abstracts are part of the publications' metadata and therefore belong to the public domain, making it possible to translate them for all articles, even those without permissive licences.

In a user-oriented setup, it is important to evaluate the quality of MT output before submitting it to the user for validation. While traditional MT evaluation scores like BLEU (Papineni et al., 2002) and chrF (Popović, 2015) require a reference translation (which we do not have in our setting, except in rare cases), it is possible to carry out reference-less evaluation, known as quality estimation (QE). Several neural QE metrics have been made available in the last years. COMET-

[5] HAL users may submit an abstract for their publication without specifying its language.

QE (Rei et al., 2021) and CometKiwi (Rei et al., 2022a), like their reference-based version, COMET (Rei et al., 2020), are based on a neural language model fine-tuned on human judgments of MT quality. We choose to use CometKiwi in this article as a measure of MT quality, and we also use it within the pipeline (alongside heuristic measures such as translation length ratio and language identification tools) to identify poor quality translations. More recently, Kocmi and Federmann (2023b) introduced GEMBA, an LLM-powered metric for assessing MT quality with or without a reference translation. GEMBA relies on zero-shot prompting of GPT-models to assess a translation and return a quality score. There have also been approaches to obtain fine-grained assessment of translation quality, including specific errors (Kocmi and Federmann, 2023a; Lu et al., 2024b). The one we use in this article for the assessment of MT quality is GEMBA-MQM (Kocmi and Federmann, 2023a), which emulates the multidimensional quality metrics (MQM) framework (Lommel et al., 2013; Freitag et al., 2021).

3 The MaTOS pipeline

The MaTOS pipeline involves the extraction and translation of abstracts from HAL and submitting the translated abstracts to the authors for validation and/or post-editing. Abstracts belong to the metadata of articles, and therefore we are able to process all submitted abstracts in compliance with the intellectual property code.[6] While it would be interesting in the future to work on the translation of whole articles, we believe that this initial first step (i.e. translating abstracts) is important.

As shown in Figure 1, the pipeline is composed of five steps, which are described in more detail below: (i) Extraction of publication metadata, (ii) language identification and filtering, (iii) translation using an LLM, (iv) filtering of the abstracts based on the translation to avoid insufficiently correct translations, and finally (v) submission of the translation to the authors for validation.

3.1 Extraction of submissions from HAL

This first step consists in extracting metadata of publications submitted to HAL. We are planning to initially deploy the pipeline only for publications submitted to the Inria portal of HAL (in computer science). However, in this paper we present results both from Inria portal publications (mostly computer science-related) and publications from all of HAL (from different institutions and scientific fields). We download the submissions on a daily basis using the HAL API.[7] Extracted metadata fields include the document ID, title, keywords, abstracts, author IDs and names, the day of submission, and the scientific field.[8] All kinds of publications are taken into account, including articles, theses, reports, images, videos, etc.

3.2 Language identification and source-side filtering/processing

This step has two aims: (i) identify English and French abstracts that are to be translated by the pipeline (i.e. those for which a translation does not already exist), and, as a by-product of the pipeline (ii) identify abstracts for which the translation is provided by the author(s), with a view to creating a parallel corpus of scientific publication abstracts.

The metadata fields specify the language of the abstract (i.e. English or French), but we carry out language identification to verify that the declared language matches the actual language of the abstract. We use an off-the-shelf fastText model (Joulin et al., 2016; Joulin et al., 2017) without a confidence threshold to predict the language. Any abstracts for which the predicted language does not match are excluded. We also exclude abstracts that contain fewer than 15 words or that end in an ellipsis (indicating an incomplete abstract).

We also collapse multiple spaces, replace subscripts and superscripts by their Unicode equivalents and remove HTML tags which are not part of the abstract content and posed problems for MT in preliminary experiments.

3.3 Machine Translation

We choose to use EuroLLM-9B-Instruct (Martins et al., 2025) for translation, as preliminary experiments showed the results to be good whilst minimising computational resources required. The model is an instruction-following model based on the 9B-parameter EuroLLM model. It has the advantage of being open-weight and trained on publicly available data in 35 languages, including parallel corpus to increase its translation abilities.

[6] https://doc.hal.science/en/legal-aspects/

[7] https://api.archives-ouvertes.fr/docs

[8] Although we do not currently use all extracted fields to help translation (e.g. keywords and the scientific field), we believe this could be useful information to include in the future.

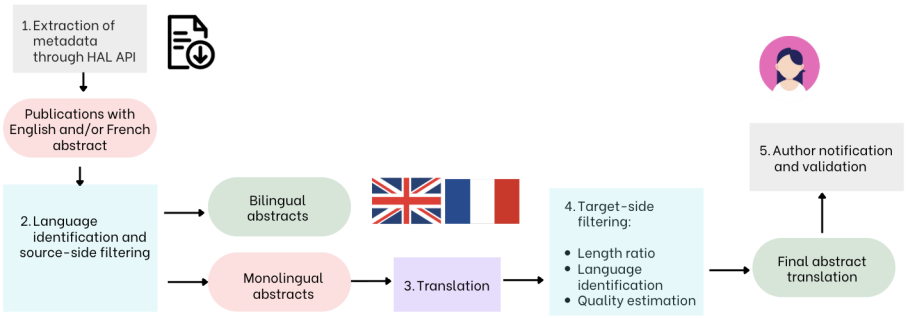

Figure 1: Illustration of the stages of the MaTOS pipeline.

We batch the translation requests to accelerate inference, and translate using vLLM (Kwon et al., 2023). The average time for processing a single request on an NVIDIA A100 GPU is 5.95 seconds. We only translate those abstracts for which the prompt and source sequence do not exceed the model's maximum context length of 4096 tokens.

In the experiments presented in this article, we test zero-shot, one-shot and two-shot MT. A zero-shot prompt example is displayed in Figure 2. We take few-shot examples from the ACAD-TRAIN and ACAD-BENCH subsets of the ACADATA parallel dataset of academic abstract (Lacunza et al., 2026) and use BM25 (Lù, 2024) to select the most similar texts, after all texts are lemmatised with SpaCy (Honnibal et al., 2020).

```
<|im_start|>system
You are a professional translator of scientific
documents. Translate the following text from
English into French. The target text must have the
same number of paragraphs as the source text.
Reply only with the translated text.
<|im_end|>

<|im_start|>user
English: Geological data show that, early in its
history, the Earth had a large-scale magnetic
field with an amplitude comparable to the one of
the present geomagnetic field.
French:
<|im_end|>

<|im_start|>assistant
```

Figure 2: Example prompt (English-French, sentence-level).

For the translation of scientific abstracts, which can be considered to be small documents, the size of the segment to be translated is important for the overall consistency of the output. In Section 4, we compare different translation granularities (isolated sentences, chunks of 3 sentences and whole abstracts), to test the trade-off between maximal access to context to aid translation and minimising potential deterioration in quality linked to having to translate longer sequences. As shown in Section 4, we find that document-level translation leads to fewer errors overall, notably concerning terminology and style. We therefore choose to translate at the document-level, but still envisage to split very long documents into smaller chunks to avoid processing issues and the degraded quality that can come with MT of very long texts.

3.4 Target-side filtering

It is important to ensure that the translated abstracts submitted to the authors are of sufficiently good quality to maximise the chances that the authors will accept to validate (or post-edit) the abstracts and to avoid any potential disillusion with the tools provided. We therefore apply the following filters on the output: (i) language identification to check that the output is in the expected language, (ii) verification of the length ratios between the source and target texts (rejecting any that differ by more than 35%)[9] and finally (iii) carrying out QE using CometKiwi (Rei et al., 2022b).

3.5 Author notification and validation

Only the translations of the abstracts that pass the previously described filters are submitted to authors. It is important for authors to maintain control over the information associated with their article deposits. We therefore implement a mechanism by which authors are notified of the translation of their abstract and are encouraged to validate and post-edit it, after which the translation will be added to the metadata of their article.

[9]This is based on the length ratios of the bilingual abstracts extracted from HAL in February 2026; French abstracts are 29% longer than their English translations, while English abstracts are 23% shorter than their French translations.

4 Experimenting with the translation segment size

As described above, we attempt to find the optimal granularity of the translation segment. In theory, more context should be useful (i.e. translating abstracts as a whole). However, in practice, it has been shown that this can have adverse effects, especially with longer texts, such as hallucinations and deleted sentences (Karpinska and Iyyer, 2023).

We compare three levels of granularity for translation: sentence-level, chunks of three sentences and whole abstracts. For the first two methods, we split abstracts into individual sentences using Trankit (Nguyen et al., 2021). 3-sentence chunks are the concatenation of three sentences, although there may be fewer than three sentences at the end of an abstract. For the document level, we also experiment with one-shot and two-shot prompting.

4.1 Data and evaluation

We experiment with two datasets collected from HAL and filtered as described in Section 3.4. The HAL-Inria corpus contains the first 500 abstracts submitted to the Inria collection of HAL starting from 1st February 2026,[10] while the HAL-all corpus contains the first 500 abstracts submitted to HAL (every discipline combined) starting from 23rd February 2026. Detailed statistics about the evaluation data can be found in Table 1.

	#docs	#sents	#aligned segs	#toks
HAL-Inria				
en–fr	500	3,775	3,429	121,692
fr–en	45	299	287	10,694
HAL-all				
en–fr	500	5,105	4,701	180,339
fr–en	500	2,831	2,479	120,210

Table 1: Statistics for the evaluation data.

For evaluation, we calculate QE scores with CometKiwi and with GEMBA-MQM to provide a finer-grained analysis of translation errors. We apply GEMBA-MQM to the complete document translations, i.e. where the sentence-level and 3-sentence-level translations are concatenated to form documents. However, CometKiwi's ability to faithfully evaluate longer text segments is subject to discussion, as the model is trained mainly on sentence-level judgments (Dahan et al., 2026). Therefore, we decide to apply it to shorter (sentence-like) segments. The same sentence alignments are not guaranteed for all granularities, so we create segments by aligning input and output sentences for each granularity using Bertalign (Liu and Zhu, 2022) and calculate the alignment that is compatible with all three, concatenating sentences in the case of many-to-1 alignments. We then apply CometKiwi to each aligned input-output pair.

[10]There are way fewer monolingual French abstracts in the INRIA collection, therewefore we only gathered 45 abstracts, submitted in all of February and March 2026

4.2 Quality estimation results and analysis

QE results can be found in Table 2 for HAL-all and in Table 3 for HAL-Inria.

Segment	#shots	CometKiwi	GEMBA-MQM	LR
		en–fr		
Sent	0	85.71	-6.29	1.419
3-sents	0	85.65	-4.87	1.370
Doc	0	85.24	-3.87	1.356
Doc	1	85.22	-3.98	1.349
Doc	2	85.05	-4.63	1.341
		fr–en		
Sent	0	83.31	-6.51	0.847
3-sents	0	83.39	-4.86	0.828
Doc	0	82.92	-4.44	0.822
Doc	1	82.74	-3.66	0.817
Doc	2	82.16	-4.15	0.816

Table 2: Comparison of MT performance for the HAL-all corpus. LR indicates the length ratio.

Segment	#shots	CometKiwi	GEMBA-MQM	LR
		en–fr		
Sent	0	84.48	-6.82	1.427
3-sents	0	84.34	-5.09	1.385
Doc	0	84.10	-5.09	1.380
Doc	1	83.87	-5.01	1.374
Doc	2	83.32	-6.07	1.366
		fr–en		
Sent	0	85.04	-6.93	0.797
3-sents	0	85.00	-5.68	0.798
Doc	0	84.73	-3.40	0.794
Doc	1	84.80	-3.84	0.794
Doc	2	84.85	-4.71	0.792

Table 3: Comparison of MT performance for the HAL-Inria corpus. LR indicates the length ratio.

CometKiwi and GEMBA-MQM CometKiwi scores are similar for all segment sizes. For both language pairs, the sentence-level and 3-sentence-level scores are almost identical and higher than document-level scores.

These results are not mirrored exactly by GEMBA-MQM scores, as the translation where

	Sent.	3-sents	Doc 0-shot	Doc 1-shot	Doc 2-shot
Acc./Mistranslation	4 (5)	8	4 (5)	4 (6)	2
Acc./Omission	0	2	0	1	0
Flu./Grammar	5 (6)	5 (4)	4	4	2 (3)
Flu./Punctuation	3 (4)	4	0	3	4
Term./Inappropriate for context	2	2	1	1	2
Term./Inconsistent use	2	4 (5)	2	1	0
Style/Awkward	2	3 (4)	3	2	1

Table 4: Number of occurrences of each GEMBA-MQM error type across different translation segment sizes. Macro error types are acc(uracy), flu(ency) and term(inology).

the text was translated sentence by sentence is scored lower than the other two granularities. This is maybe unsurprising, as we applied GEMBA-MQM to whole abstracts, while CometKiwi was applied to sentence-like segments, therefore possibly missing important errors concerning consistency and coherence. In our pipeline, we cannot use GEMBA-MQM for cost and reproducibility reasons (Kocmi and Federmann, 2023a); we therefore use CometKiwi, which is nevertheless sensitive to the most severe problems.

We also find that few-shot examples tend to decrease CometKiwi scores. One-shot prompting can have a light positive or negative impact on GEMBA-MQM scores, but two-shot prompting always deteriorates scores except for the fr–en direction for the HAL-Inria corpus.

Analysis based on GEMBA-MQM Gemba-MQM allows us to perform finer-grained evaluation, offering insights into translation errors in addition to providing QE scores. The detected errors are classified into three categories: critical, major and minor, and each category is weighted differently in the calculation of the final quality score.

We perform qualitative analysis by reviewing the translations and the Gemba-MQM error annotations for the first 10 en–fr abstracts of the HAL-all corpus. Table 4 shows the number of occurrences of the most important error types. We find that some errors were misclassified or hallucinated by Gemba-MQM and therefore indicate the true count of errors in parentheses after our re-annotation. We present some misclassified and false errors in Appendices A and B respectively.

The most prevalent error type is mistranslation, reflecting the difficulty for LLMs to translate scientific terms accurately. Inconsistent translation of terms becomes more common for smaller translation segments (i.e. worse at the 3-sents level), as consistent use of terminology often requires context beyond the sentence or chunk level (Xiao et al., 2011). We also find some basic grammatical errors. They mainly concern subject-verb agreement and gender agreement. We provide below two examples of errors annotated by GEMBA-MQM. In the first example, the English source term *bio-oil*, is inconsistently translated into French as *bio-huile* and *huile biologique*. In the second example, the French word *dimensionnement* 'sizing' is preceded by a feminine determiner, even though it is a masculine noun.

(1) terminology/inconsistent use: "bio-huile" vs "huile biologique" (both acceptable but inconsistent within the text)

(2) fluency/grammar: la dimensionnement optimal" should be "le dimensionnement optimal" (gender agreement error: "dimensionnement" is masculine)

Based on the above findings, we have decided to implement document-level MT in our pipeline. We will continue experimenting with few-shot prompting and different strategies for example selection.

5 Conclusion and future work

We have presented the pipeline of the MaTOS project for the translation of English and French publication abstracts submitted to the HAL publishing platform. Our aim is to break down language barriers in research (where English dominates) by encouraging publication in other languages, in our case French. The design of the pipeline takes into account the need to filter submitted abstracts and the resulting translations in order not to discourage users from using the technology. The idea is for users to validate or post-edit the translations, which will also be publicly available as a useful resource for the MT community.

In the near future, we intend to experiment with more sophisticated prompting techniques, integrating translation of terms using bilingual glossaries and other domain-specific information.

Acknowledgements

This work was supported by the French national research agency (ANR) as part of the MaTOS project (grant number: ANR-22-CE23-0033).[11] Rachel Bawden was also partly funded by her chair position in the PRAIRIE institute funded by ANR as part of the "Investissements d'avenir" programme under reference ANR19-P3IA-0001. The authors are grateful to the anonymous reviewers for their insightful comments and suggestions.

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the first International Conference on Learning Representations*, San Diego, CA.

Bawden, Rachel and François Yvon. 2023. Investigating the translation performance of a large multilingual language model: the case of BLOOM. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 157–170, Tampere, Finland, June. European Association for Machine Translation.

Bawden, Rachel, Maud Bénard, Éric de la Clergerie, José Cornejo Cárcamo, Nicolas Dahan, Manon Delorme, Mathilde Huguin, Natalie Kübler, Paul Lerner, Alexandra Mestivier, Joachim Minder, Jean-François Nominé, Ziqian Peng, Laurent Romary, Panagiotis Tsolakis, Lichao Zhu, and François Yvon. 2025. MaTOS: Machine Translation for Open Science. In Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Samuel Läubli, Martin Volk, Miquel Esplà-Gomis, Vincent Vandeghinste, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 2*, pages 103–104, Geneva, Switzerland, June. European Association for Machine Translation.

Bénard, Maud, Alexandra Mestivier, Natalie Kubler, Lichao Zhu, Rachel Bawden, Eric De La Clergerie, Laurent Romary, Mathilde Huguin, Jean-François Nominé, Ziqian Peng, and François Yvon. 2023. MaTOS: Traduction automatique pour la science ouverte. In Boudin, Florian, Béatrice Daille, Richard Dufour, Oumaima El, Maël Houbre, Léane Jourdan, and Nihel Kooli, editors, *Actes de CORIA-TALN 2023. Actes de l'atelier "Analyse et Recherche de Textes Scientifiques" (ARTS)@TALN 2023*, pages 8–15, Paris, France, 6. ATALA.

Dahan, Nicolas, Rachel Bawden, and François Yvon. 2026. MetaDocEval: A Contrastive Framework for Evaluating Machine Translation Metrics at the Document-Level. In *Proceedings of the 26th Annual Conference of the European Association for Machine Translation*, Tilburg, the Netherlands, June. European Association for Machine Translation (EAMT).

Fernandes, Patrick, Kayo Yin, Emmy Liu, André Martins, and Graham Neubig. 2023. When Does Translation Require Context? A Data-driven, Multilingual Exploration. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, Toronto, Canada, July. Association for Computational Linguistics.

Freitag, Markus, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Gordin, Michael D. 2015. *Scientific Babel*. University of Chicago Press.

Guo, Jiaxin, Yuanchang Luo, Daimeng Wei, Ling Zhang, Zongyao Li, Hengchao Shang, Zhiqiang Rao, Shaojun Li, Jinlong Yang, Zhanglin Wu, and Hao Yang. 2025. Doc-Guided Sent2Sent++: A Sent2Sent++ Agent with Doc-Guided memory for document-level Machine Translation. arXiv Preprint:2501.08523 [cs], January.

Hardmeier, Christian. 2012. Discourse in Statistical Machine Translation: A Survey and a Case Study. *Discours - Revue de linguistique, psycholinguistique et informatique*, (11).

Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *CoRR*, abs/2302.09210.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Hu, Tianxiang, Pei Zhang, Baosong Yang, Jun Xie, Derek F. Wong, and Rui Wang. 2024. Large language model for multi-domain translation: Benchmarking and domain CoT fine-tuning. In Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5726–5746, Miami, Florida, USA, November. Association for Computational Linguistics.

[11] http://anr-matos.github.io/

Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomás Mikolov. 2016. Fasttext.zip: Compressing text classification models. *CoRR*, abs/1612.03651.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In Lapata, Mirella, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April. Association for Computational Linguistics.

Karpinska, Marzena and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore, December. Association for Computational Linguistics.

Kocmi, Tom and Christian Federmann. 2023a. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In Koehn, Philipp, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore, December. Association for Computational Linguistics.

Kocmi, Tom and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland, June. European Association for Machine Translation.

Kwon, Woosuk, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Lacunza, Iñaki, Javier Garcia Gilabert, Francesca De Luca Fornaciari, Javier Aula-Blasco, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2026. ACAData: Parallel dataset of academic data for machine translation. In Piperidis, Stelios, Núria Bel, Henk van den Heuvel, Nancy Ide, Simon Krek, and Antonio Toral, editors, *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 8498–8519, Palma, Mallorca, Spain, May. European Language Resources Association (ELRA).

Lambert, Patrik, Jean Senellart, Laurent Romary, Holger Schwenk, Florian Zipser, Patrice Lopez, and Frédéric Blain. 2012. Collaborative machine translation service for scientific texts. In Segond, Frédérique, editor, *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 11–15, Avignon, France, April. Association for Computational Linguistics.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Liu, Lei and Min Zhu. 2022. Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634, 12.

Lommel, Arle Richard, Aljoscha Burchardt, and Hans Uszkoreit. 2013. Multidimensional quality metrics: a flexible system for assessing translation quality. In *Proceedings of Translating and the Computer 35*, London, UK, November 28-29. Aslib.

Lopes, António, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. 2020. Document-level neural MT: A systematic comparison. In Martins, André, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal, November. European Association for Machine Translation.

Lu, Hongyuan, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024a. Chain-of-dictionary prompting elicits translation in large language models. In Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA, November. Association for Computational Linguistics.

Lu, Qingyu, Baopu Qiu, Liang Ding, Kanjian Zhang, Tom Kocmi, and Dacheng Tao. 2024b. Error analysis prompting enables human-like translation evaluation in large language models. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8801–8816, Bangkok, Thailand, August. Association for Computational Linguistics.

Lù, Xing Han. 2024. BM25S: orders of magnitude faster lexical search via eager sparse scoring.

Martins, Pedro Henrique, João Alves, Patrick Fernandes, Nuno Miguel Guerreiro, Ricardo Rei, M. Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report. *CoRR*, abs/2506.04079.

Miculicich, Lesly, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In Riloff, Ellen, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2947–2954, Brussels, Belgium, October-November. Association for Computational Linguistics.

Moslem, Yasmin, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In Nurminen, Mary, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland, June. European Association for Machine Translation.

Nguyen, Minh Van, Viet Dac Lai, Amir Pouran Ben Veyseh, and Thien Huu Nguyen. 2021. Trankit: A light-weight transformer-based toolkit for multilingual natural language processing. In Gkatzia, Dimitra and Djamé Seddah, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online, April. Association for Computational Linguistics.

O'Brien, Dayyán, Bhavitvya Malik, Ona de Gibert, Pinzhen Chen, Barry Haddow, and Jörg Tiedemann. 2025. DocHPLT: A massively multilingual document-level translation dataset. In *Proceedings of the Tenth Conference on Machine Translation*, pages 286–300, Stroudsburg, PA, USA, September. Association for Computational Linguistics.

Oncevay, Arturo, Charese Smiley, and Xiaomo Liu. 2025. The impact of domain-specific terminology on machine translation for finance in European languages. 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 2758–2775, Stroudsburg, PA, USA. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Peng, Ziqian, Rachel Bawden, and François Yvon. 2025a. Investigating length issues in document-level machine translation. In Bouillon, Pierrette, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 4–23, Geneva, Switzerland, June. European Association for Machine Translation.

Peng, Ziqian, Rachel Bawden, and François Yvon. 2025b. Self-retrieval from distant contexts for document-level machine translation. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 220–240, Suzhou, China, November. Association for Computational Linguistics.

Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Rei, Ricardo, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are References Really Needed? Unbabel-IST 2021 Submission for the Metrics Shared Task. In Barrault, Loic, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online, November. Association for Computational Linguistics.

Rei, Ricardo, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022a. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Rei, Ricardo, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Koehn, Philipp, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Tan, Weiting, Haoran Xu, Lingfeng Shen, Shuyue Stella Li, Kenton Murray, Philipp Koehn, Benjamin Van Durme, and Yunmo Chen. 2024. Narrowing the gap between zero- and few-shot machine translation by matching styles. In Duh, Kevin, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 490–502, Mexico City, Mexico, June. Association for Computational Linguistics.

Wang, Longyue, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In Bouamor, Houda, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore, December. Association for Computational Linguistics.

Xiao, Tong, Jingbo Zhu, Shujie Yao, and Hao Zhang. 2011. Document-level consistency verification in machine translation. In *Proceedings of Machine Translation Summit XIII: Papers*, Xiamen, China, September 19-23.

Zhu, Wenhao, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In Duh, Kevin, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico, June. Association for Computational Linguistics.

A Misclassified errors

In this section we show some examples of errors that were detected by GEMBA-MQM in our evaluation sample but were classified in the wrong category.

(3) fluency/grammar: some minor punctuation issues in the abbreviation list (missing commas or semicolons in a few places)

This was annotated as a *fluency/grammar* error, but we re-annotated is as a *fluency/punctuation* error, since it concerns missing punctuation marks.

(4) accuracy/mistranslation: "la hybridation" should be "l'hybridation" (missing elision of the article "la" before a vowel)

This was annotated by GEMBA-MQM as an *accuracy/mistranslation* error, even though it is a grammar error.

B False errors

In this section we show some examples of errors that were detected by GEMBA-MQM in our evaluation sample but are actually correct translations.

(5) terminology: the terms "sphère biogéographique" for "biogeographic realm," "dissemblance" for "dissimilarity" and other scientific terms are consistent and appropriate.

This was counted as a terminology error, even though the model itself explains that scientific terms are translated in a consistent and appropriate manner.

(6) style: the translation is formal and appropriate for the scientific context.

This was counted as a style error, even though the model itself explains that the translation is formal and appropriate.