Parallel Corpora of Scholarly Documents for English-French Machine Translation

Ziqian Peng[1,3], **Lichao Zhu**[2], **Rachel Bawden**[3], **Maud Bénard**[2], **Éric de la Clergerie**[3], **Mathilde Huguin**[4], **Natalie Kübler**[2], **Paul Lerner**[1], **Alexandra Mestivier**[2] and **François Yvon**[1]

[1] ISIR, CNRS & Sorbonne Université, Paris, France [2] ALTAE, Université Paris Cité, Paris, France
[3] Inria, Paris, France [4] INIST, CNRS, Nancy, France
contact@anr-matos.fr

Abstract

The growing ability of large language models (LLMs) to process long-range context opens new perspectives for document-level machine translation (MT), especially in scholarly communication. In fact, translating scholarly texts requires to integrate both local and long-range contextual information to ensure the consistency and coherence across the full document. However, document-level parallel corpora for such text types remain scarce, limiting both evaluation and domain adaptation of MT systems for this task. To address this gap, we introduce PARAEPS (Earth and Planetary Sciences Bilingual Corpus) and PARANLP (Natural Language Processing Bilingual Corpus), two new parallel corpora covering 14k abstracts and 105 full-length articles in two scientific domains to be used for fine-tuning and evaluation purposes. We compare the performance of eight MT systems on these test sets and find that fine-tuning on document-level data closes the gap between open systems based on Large Language Models (LLMs) and commercial systems. We also find that the performance of recent LLMs can worsen when translating full articles instead of translating them on a per paragraph basisfine-tuning. These experiments underscore the need for corpora such as PARAEPS and PARANLP.

Keywords: Machine Translation, Parallel Corpus, Scientific Documents, Long-context Modelling, Large language models

1. Introduction

The development of large language models (LLMs) creates new possibilities for document-level machine translation (MT), owing to their ability to process long-range dependencies (Karpinska and Iyyer, 2023; Peng et al., 2024a; Wang et al., 2024, 2025; Zhu et al., 2025). However, document-level parallel corpora remain scarce,[1] particularly for scholarly documents. For such texts, existing resources are mostly limited to sentence-aligned parallel texts (Roussis et al., 2022; Esalati et al., 2024) without document boundary information or restricted to the medical domain (Ive et al., 2016; Névéol et al., 2018). Although some resources preserve the document structure (Abdul Rauf and Yvon, 2024), they generally comprise only abstracts (Kleidermacher and Zou, 2025). Consequently, the training of MT systems for scientific texts mainly relies on short texts, thus failing to represent the actual complexity of scholarly articles, which often contain complex formulas, citations and long-range contextual dependencies. Moreover, given the lack of document-level test sets, the evaluation of recent LLMs to translate scholarly articles is often restricted to reference-free metrics (Zhu et al., 2025) or human judgments (Kleidermacher and Zou, 2025).

In this work, we present the curation of additional resources for document-level translation in two scientific fields: Earth and Planetary Sciences and PARANLP for Natural Language Processing. Our corpora, PARAEPS and PARANLP, consist of both abstracts and full-length articles; each contains training, validation and test sets of parallel abstracts, and a test set of complete parallel articles. These parallel articles are constructed using one of the three different approaches: 1) human translations opportunistically collected by the authors, 2) human post-edits of machine translated texts, 3) combination of human translations, automatic post-editions and machine translations derived from comparable articles published in both English and French.

Using our test sets of full articles, we compare the performance of six MT systems translating on a per paragraph basis, and of two LLMs translating chunks of varying sizes. Experimental results show that 1) fine-tuning on paragraph-level datasets closes the gap between the performance of medium-size LLMs and commercial MT system such as DeepLPro,[2] for the translation of abstracts and 2) there is no systematic performance gain when using recent LLMs to translate scientific texts at the article level as opposed to applying them at

[1] A situation that is changing, at least for Web pages (O'Brien et al., 2025).

[2] https://deepl.com

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 84–95
11 May 2026.

the paragraph or sentence level. These results illustrate the utility of our corpora and the need to improve recent models for MT tasks addressing the increasing needs of scholarly communication across languages.[3] Our main contributions are as follows:

- the collection and construction of parallel scholarly documents in two domains, including 14k abstracts and 105 full-length articles.
- an original pipeline that constructs parallel articles from comparable articles extracted from scholarly publications in the NLP domain; as these parallel texts are partly machine-generated, partly human generated, we dub this corpus a silver reference corpus;
- benchmark results of two commercial systems and recent LLMs, fine-tuned or not on these newly created corpora.

We openly release our corpora and the code of the corpus construction pipeline under a permissive license.[4]

2. Related Work

Parallel corpora are central for training and evaluating MT systems, but most existing resources for scholarly documents are only aligned at the sentence level (Roussis et al., 2022; Esalati et al., 2024; Roussis et al., 2024) disregarding document boundary information, or restricted to specific domains (e.g., for the medical sciences (Ive et al., 2016; Névéol et al., 2018; Abdul Rauf and Yvon, 2024), or both (for instance the Taus Corona Crisis corpus[5] and the Mlia Covid corpus[6])). Even when document structure is preserved, documents mostly correspond to abstracts (Kleidermacher and Zou, 2025), failing to represent (a) very long-range dependencies (e.g., between the Introduction and Conclusion sections) and (b) translation issues that are specific to scholarly texts such as the translation of captions, table cells, or the insertion of citations in the discursive flow.

The lack of availability of full-length parallel documents also means that the evaluation of MT systems in their ability to translate scholarly content mainly relies on reference-free metrics computed at the paragraph level (Zhu et al., 2025), human evaluation (Kleidermacher and Zou, 2025) or the analysis of translated abstracts (Sebo and de Lucia, 2024).

Recently, several parallel corpora comprising long documents have been introduced in literary domains (Jiang et al., 2022; Wang et al., 2024a,b). In addition, O'Brien et al. (2025) recently introduced DocHPLT, a massive collection of parallel documents extracted from the Internet Archive for multiple language pairs. However, they do not focus on scholarly texts, nor do they provide sufficiently informative domain tags. Another recent resource is ACADATA (Lacunza et al., 2025), a collection of parallel abstracts across 12 languages harvested from public academic web sites and archives. These documents are however relatively short (about 1000 characters), and lack gold domain tags. Finally, the ACL 60-60 initiative aimed to produce reference translations for NLP abstracts in multiple languages, as well as a large number of automatically generated translations of papers and talks.[7] One outcome of this effort was the release of development and test data for the IWSLT 2023 shared task (Salesky et al., 2023): each set contains a post-edited version of the translations (in 11 languages) of 10 presentations delivered during the ACL 2022 conference[8].

Our work complements these developments by introducing two document-level parallel corpora of scholarly documents, including full-length parallel-articles, aimed to mitigate the scarcity of resources required to study discourse-aware MT.

3. Dataset Creation

We create new parallel resources for English–French translation in two fields of study: Earth and Planetary Sciences (EPS) and natural language processing (NLP). Each of the two subsets, PARAEPS and PARANLP, comprises four data splits: train, dev and test sets of *abstracts* (TRAIN, DEV and TEST) and a second test set (TEST-LONG) of *full articles*. We collect texts from multiple sources and use various techniques to construct the datasets, including manual translation, post-editing and, in the case of the NLP domain, some partial automatic translation to complement manually translated examples. We describe the sources and dataset creation process for PARAEPS and PARANLP in Sections 3.1 and 3.2 respectively; statistics for both datasets can be found in Table 1.

3.1. EPS Dataset (PARAEPS)

For the EPS domain, we collected 11k abstracts and 29 articles in both English and French. This section presents data collection, text processing, alignment and the construction of the data splits.

[3] https://www.helsinki-initiative.org/.
[4] https://anr-matos.github.io/pages/resources.html
[5] https://md.taus.net/corona
[6] http://eval.covid19-mlia.eu/task3/
[7] https://acl6060.org/
[8] https://2022.aclweb.org/

Split	#docs	# sents	#toks/doc ($\mu \pm \sigma$) en	fr
PARAEPS				
TRAIN	10,577	83,036	347 ± 180	474 ± 233
DEV	400	3,273	344 ± 144	483 ± 192
TEST	391	3651	401 ± 203	544 ± 271
BSGF	*132*	*1311*	*472 ± 196*	*622 ± 263*
CRAS	*100*	*677*	*277 ± 118*	*388 ± 168*
CRG	*59*	*364*	*260 ± 101*	*360 ± 135*
THESES$_{EPS}$	*100*	*1299*	*512 ± 211*	*707 ± 281*
TEST-LONG	29	5,133	7,773 ± 2,755	10,539 ± 3,613
MERSENNE	*19*	*3,532*	*7,673 ± 2,828*	*10,652 ± 3,814*
STUDENT	*10*	*1,601*	*7,962 ± 2,600*	*10,322 ± 3,186*
PARANLP				
TRAIN	2,723	24,085	287 ± 159	429 ± 234
DEV	96	1,024	353 ± 135	523 ± 210
TEST	346	2,022	176 ± 124	264 ± 181
rTAL	*246*	*1,015*	*121 ± 46*	*184 ± 68*
THESES$_{NLP}$	*100*	*1,007*	*310 ± 150*	*463 ± 216*
TEST-LONG	76	14,467	6,064 ± 2,783	8,388 ± 3,820
NLPGOLD	4	533	*4,477 ± 2,999*	*6,679 ± 4,436*
NLPSILVER$_{EN-FR}$	36	7,025	*6,028 ± 2,399*	*8,433 ± 3,331*
NLPSILVER$_{FR-EN}$	36	6,909	*6,275 ± 3,045*	*8,532 ± 4,145*

Table 1: Statistics for PARAEPS and PARANLP and their data splits. TRAIN, DEV and TEST splits are composed of abstracts. TEST-LONG is composed of full articles. English (en) and French (fr) token counts are based on TOWERBASE tokens.

3.1.1. Data Collection and Processing

Abstracts We collected over 11k parallel scientific abstracts (89k sentences, and 2.2M and 2.6M words in English and French respectively) from seven sources (listed in Table 2 with the statistics after quality filtering), by extracting the plain texts from the HTML pages, and aligning them across the two languages. We use `langdetect`[9] to filter out noisy abstracts written in other languages, and we also disregard abstracts of source-to-target length ratio is smaller than 0.5.

We applied NFC normalization using unicodedata,[10] before segmenting abstracts in sentences using Trankit (Nguyen et al., 2021), which reliably disambiguates the multiple interpretations of the dot ('.') symbol in scholarly documents (e.g. in numbers or abbreviations, in addition to the sentence-final punctuation mark). We aligned the resulting bilingual segments using a slightly modified version of BertAlign (Liu and Zhu, 2022), which robustly supports many-to-many sentence alignments.[11] We use the value 0.001 for the `skip` parameter, which improves zero-to-one alignments. We also introduce a new parameter `len_slack` (with value 0.15), which prevents to apply a length penalty for parallel segments having length ratio close to 1; this tends to reduce the number of spurious many-to-any alignments.

Then we filtered the aligned sentences using quality estimation scores from TransQuest (Ranasinghe et al., 2020). Additional details concerning the filtering are provided in Section 3.1.2, as we also use the alignment scores when selecting data for the different data splits.

Full Articles We collected parallel articles from two sources. Firstly, ten English articles and their translations were obtained from a specialised translation course. The original articles were either sourced from the ISTEX database[12] (Maurel et al., 2019) or were Open Access. The translations, which were produced by master's students, were the result of either translation from scratch or post-edition of MT (both standard approaches used by translation specialists) followed by proofreading. We refer to these texts as the STUDENT collection.

Secondly, we also collected 19 articles (five in English and fourteen in French) published in the "Compte rendus Géosciences" journal,[13] which were then automatically translated and post-edited by a professional translator using the MateCat platform (Federico et al., 2014). We refer to these texts as the MERSENNE collection. Unlike the STUDENT collection, we had to extract parallel texts from MERSENNE. We used pandoc[14] to extract the plain text from the HTML documents, removing empty lines, the symbol `\xa0` and carrying out NFC normalization. We extract the blocks of text (hereafter referred to as paragraphs),[15] equations and tables.[16] Since the English and French versions contain the same number of paragraphs, we trivially align them, before segmenting into sentences and aligning the sentences using the same pipeline as for the abstracts. Non-aligned sentences were manually realigned. We validated the resulting alignment using TransQuest (Ranasinghe et al., 2020): all paired sentences had an alignment score of at least 0.75, indicated good alignment throughout.

[9] https://pypi.org/project/langdetect/

[10] https://docs.python.org/3/library/unicodedata.html

[11] https://github.com/ANR-MaTOS/bertalign

[12] https://www.istex.fr/

[13] https://comptes-rendus.academie-sciences.fr/geoscience

[14] https://pandoc.org/

[15] In practice, these also correspond to section titles and captions in addition to paragraphs.

[16] We store equations and tables as complementary information to be used in future work.

[17] We collected abstracts from the following scientific journals: *Hydrogeology Journal*, *Mineralogy and Petrology*, *Swiss Journal of Geosciences*, *Geodinamica Acta*, *Journal of South American Earth Sciences*, etc. in ISTEX, which is a data portal of multilingual scientific data.

	Source of abstracts	#segments	#abstracts (all)	#abstracts		
				TRAIN	DEV	TEST
PARAEPS	BSGF (Bulletins de la Société Géologique de France)	1,311	132	-	-	132
	CanMin (Canadian Mineralogist)	8,140	793	793	-	-
	CJES (Canadian Journal of Earth Sciences)	37,525	4,624	4,524	100	-
	CRAS (Comptes Rendus de l'Académie des Sciences - Earth and Planetary Sciences, 1995-2001)	9,620	2,026	1,826	100	100
	CRG (Comptes rendus Géoscience)	364	59	-	-	59
	ISTEX (Infrastructure de services pour la fouille de textes)[17]	15,190	2,117	2,017	100	-
	THESES (Database of PhD abstracts)	17,810	1,617	1,417	100	100
	TOTAL	89,960	11,368	10,577	400	391
PARANLP	ISTEX (Infrastructure de services pour la fouille de textes)	8,099	1,309	1,309	-	-
	rTAL (revue TAL)	1,015	246	-	-	246
	THESES (Database of PhD abstracts)	18,987	1,610	1,414	96	100
	TOTAL	30,543	3,165	2,723	96	346

Table 2: Data sources and statistics for the abstracts in the PARAEPS and PARANLP TRAIN, DEV and TEST splits, after quality filtering.

3.1.2. Dataset splits

PARAEPS consists of TEST-LONG, containing the 29 parallel articles, and TRAIN, DEV and TEST splits composed of parallel abstracts.

EPS-TEST-LONG is composed of the full articles from the MERSENNE and STUDENT collections aligned at the sentence level. For MERSENNE articles, we also preserve paragraph-level boundary information, comprising 915 paragraphs.

EPS-TEST is composed of abstracts from four of the collections listed in Table 2: BSGF, CRAS, CRG and THESES. To ensure the quality of the test set, we computed the average alignment score for sentence pairs within each abstract, and empirically excluded abstracts with an alignment score below 0.5. We then kept the remaining abstracts from BSGF and CRG due to their small size, and selected the 100 most recent abstracts from CRAS and THESES.

EPS-DEV contains a total of 400 abstracts. Only considering parallel abstracts with an alignment score above 0.5, we randomly sample 100 abstracts from the 200 most recent abstracts from each of the CJES, CRAS, ISTEX and THESES collections.

EPS-TRAIN contains the remaining 10,577 parallel abstracts after filtering the most unreliable alignments (i.e. those whose average alignment score is below 0.4 when all sentences are aligned, or below 0.5 if at least one sentence is unmatched).

The right side of Table 2 displays the distribution of abstracts in EPS-TEST, EPS-DEV, EPS-TRAIN, broken down by data source.

3.2. NLP Dataset (PARANLP)

For the NLP domain, we collected 3k parallel abstracts and 76 parallel articles. The complete articles are derived from four human translated articles and 36 comparable human-written articles, each turned into two parallel texts as described below. The corresponding statistics are reported in the bottom part of Table 1.

3.2.1. Data Collection and Processing

Abstracts We collected abstracts from NLP publications for which both English and French versions are available. These raw texts include 246 parallel abstracts extracted from the French NLP journal *revue TAL* (rTAL),[18] 1358 NLP abstract retrieved from various journal articles available in the ISTEX archive, and 1701 abstracts from PhD dissertations (THESES).[19] We processed the abstracts using the same pipeline as for PARAEPS, i.e. first segmenting them with Trankit, then performing sentence alignment using BertAlign. Furthermore, we filtered the resulting alignments using `TransQuest` scores.[20] These abstracts are then split into three parallel corpora NLP-TRAIN, NLP-DEV and NLP-TEST, as detailled in Section 3.2.2. NLP-TEST was aligned using hunalign[21] (Varga et al., 2005) in the early stage of our data preparation process. To ensure its quality, we manually reviewed all the sentence pairs having a TransQuest score lower than 0.3. The statistics of PARANLP after quality filtering are in the bottom of Table 2.

[18] https://www.atala.org/revuetal

[19] https://theses.fr/

[20] We use the same filtering heuristics as for PARAEPS.

[21] https://github.com/danielvarga/hunalign

	NLPSILVER$_{FR-EN}$				NLPSILVER$_{EN-FR}$						
	total	mean	min	max	total	mean	min	max	correct	total	TER
Copy	2205	61	17	132	1677	46	13	104	41	126	24.6
APE	3478	96	38	246	4031	111	43	256	51	199	27.5
MT	1226	34	3	176	1317	36	5	184	12	42	21.5
all	6909	191	103	408	7025	195	112	477	109	367	26.2

Table 3: An analysis of the composition of our silver NLP corpus in number of sentences, with translations being produced through (i) copying the original aligned human-written translations (Copy), (ii) postediting them (APE) or (iii) machine-translating from scratch (MT). The three columns on the right correspond to a quality assessment, based on 3 articles that were manually post-edited by one author. For these, we report the number of correct translations out of all sentence pairs (total), and the TER score based on a comparison between the silver and the post-edited versions.

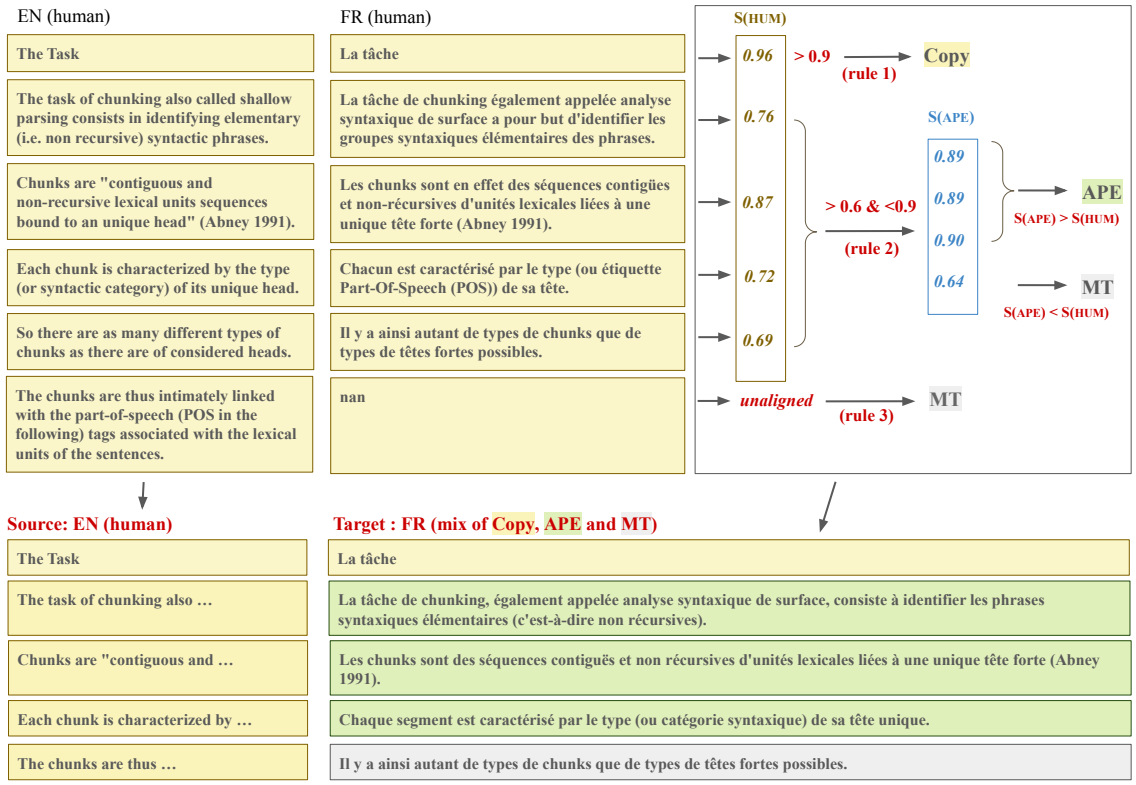

Figure 1: The pipeline to construct parallel articles from comparable articles, with examples from NLPSILVER$_{EN-FR}$. We denote S(HUM) and S(APE) the cosine similarity between an English (EN) sentence and its human-written translation (HUM) or automatic post-edition (APE), based on LaBSE embeddings.

Full articles We collect full parallel articles from two sources: (i) regular human translations (NLPGOLD), consisting of two English articles translated into French and two French articles translated into English, collected opportunistically by the authors, which we complement with (ii) a larger set of silver translations (NLPSILVER), which we assembled using a combination of automatic and manual operations, as described below.

In addition to full human translations (NLPGOLD), we also construct NLPSILVER, which is derived from articles from the ACL Anthology.[22] While the ACL Anthology mostly stores English articles, it also contains a small number of French articles, published in French journals and conferences. A fraction of these also have an English counterpart, as some French speaking authors wish to publish their research in both languages, possibly using MT and post-editing to speed up the process. Given that there is no guarantee that the English and French versions are strictly parallel, we apply a more complex alignment process than for the previously described datasets. We first extracted the text version of all English and French papers, segmented them into sentences,

[22] https://aclanthology.org/

filtering short segments or segments containing mostly non-alphabetic characters (likely equations or OCR errors, for older articles) and embedded them in a joint multilingual space using LaBSE (Feng et al., 2022).[23] We then indexed the embeddings using FAISS (Johnson et al., 2021), and, for each French sentence, searched for its closest English neighbours. We manually inspected the French articles with a large number of sentences whose neighbours were found in the same English documents, resulting in a final list of 36 pairs of articles.

We convert these articles from pdf to markdown using `pymupdf4llm`[24], remove noisy texts (page numbers, pdf headers, algorithms, etc.) using regular expression and manual verification, extract and keep aside tables, resulting in clean markdown files, which are finally converted into plain texts for sentence segmentation.

For each of pair of articles, we first performed sentence alignment using the same pipeline as described previously (using Trankit and Bertalign). We then derived a fully parallel *English-French* version, denoted as NLPSILVER$_{\text{EN-FR}}$, as follows. We process each English article, and for each source segment e and its aligned French counterpart f and apply the following rules, which are also illustrated in Figure 1:[25]

1. if the alignment score between e and f is above 0.9, we keep the pair (e, f), assuming that they are mutual translations (Copy).
2. if the alignment score between e and f is between 0.6 and 0.9, we use automatic post-editing (APE) of (e, f) with TowerInstruct-13B-v0.1[26] to generate f'. If f' has a better alignment score with e than f, we keep f'; otherwise we retranslate e from scratch (MT).
3. if the alignment score between e and f is below 0.6, or if e is not aligned with any French counterpart, we retranslate e from scratch as for case 2 (MT).

We used the same heuristics to create a fully parallel *French-English* version, which we refer to as NLPSILVER$_{\text{FR-EN}}$, comprising all the human-written French texts and their corresponding translations derived from the pipeline of Figure 1. We provide the corresponding number of segments produced by each rule for each version in Table 3.

[23] https://huggingface.co/sentence-transformers/LaBSE.

[24] https://github.com/pymupdf/pymupdf4llm

[25] e and f each correspond to one or more consecutive sentences resulting from BertAlign's many-to-many alignment.

[26] https://huggingface.co/Unbabel/TowerInstruct-13B-v0.1

In order to evaluate the confidence of each type of rule, a native French speaker with expertise in NLP post-edited a random sample of three articles from the English-French version.[27] TER scores (Snover et al., 2006) were also computed using SacreBLEU (Post, 2018). The results of this analysis are in the right part of Table 3. Sentences produced through the *Copy* action (i.e. the translation was taken from the original article) are slightly more likely to be fully correct than segments generated with the other rules, but the edit distance to acceptable references is still non-negligible.

3.2.2. Dataset Splits

As for PARAEPS, PARANLP consists of TEST-LONG, containing full parallel articles, and TRAIN, DEV and TEST splits composed of parallel abstracts.

NLP-TEST-LONG is composed of the parallel articles just described (NLPGOLD, NLPSILVER$_{\text{EN-FR}}$ and NLPSILVER$_{\text{FR-EN}}$).

NLP-TEST contains 346 parallel abstracts, corresponding to all RTAL abstracts and 100 randomly selected abstracts from THESES$_{\text{NLP}}$.

TRAIN and DEV NLP-DEV is composed of 96 abstracts randomly sampled from THESES$_{\text{NLP}}$ (non-overlapping with those selected for TEST). NLP-TRAIN contains the remaining abstracts extracted from ISTEX and THESES$_{\text{NLP}}$.

4. Experiment Settings

We provide benchmarking experiments to illustrate the usefulness of the two datasets for both fine-tuning and evaluation of document-level MT for scholarly documents.

MT Engines We test the translation performance of two multilingual LLMs: TowerBase-7B[28] (TOWER) (Alves et al., 2024) and EuroLLM-9B[29] (EUROLLM) (Martins et al., 2025), when translating abstract test sets at the paragraph level, before and after fine-tuning on the corresponding training set (EPS-TRAIN and NLP-TRAIN). For comparison, we also evaluate the translation quality of two com-

[27] Using the MateCat platform, without knowledge of each segment origin.

[28] https://huggingface.co/Unbabel/TowerBase-7B-v0.1

[29] https://huggingface.co/utter-project/EuroLLM-9B

mercial systems: DeepLPro (DeepL)[30] and SystranPro (Systran).[31][32]

For the translation of complete articles, we compare the performance of Llama3.1-8B-Instruct[33] (Llama3) (Grattafiori et al., 2024), which has a context length of 128k tokens, and Qwen3-8B[34] (Qwen3) (Yang et al., 2025) with a context length greater than 32k on the MERSENNE subset of EPS-test-long.

Fine-tuning and Inference We perform supervised fine-tuning using QLoRA (Dettmers et al., 2023), following the quantization and LoRA configurations proposed by Moslem et al. (2023, Section 2.3) for Tower and EuroLLM. The QLoRA learning rate is $2e{-}4$ adjusted by a cosine schedule, with neither warm-up steps nor packing. The batch size is set to 8.

For NLP we fine-tune Tower and EuroLLM for two epochs on NLP-train, with two gradient accumulation steps to produce FT-Tower-NLP and FT-Euro-NLP respectively. Similarly, we fine-tune both models on EPS-train to produce FT-Tower-EPS and FT-Euro-EPS, although due to the fact that the training set is larger, we do so for one epoch only and set the gradient accumulation steps as 4.

Inference is performed without additional in-context examples, with bfloat16 and greedy search, using the Huggingface implementation, except for the inference of Llama3 and Qwen, which is carried out with vLLM (Kwon et al., 2023). For Qwen3, we use the suggested configuration for the non-thinking mode using a hybrid decoding method that combines top-k and top-p, with temperature, top-p, top-k values of 0.7, 0.8, and 20 respectively.

Prompts We prompt base models following their HuggingFace model cards, using the following two prompts for Tower and EuroLLM respectively:

(1) `English: SRC\nFrench:`

(2) `English: SRC French:`

We use the following prompt for fine-tuning and fine-tuned models:

(3) `Translate the following text from English into French.\nEnglish: SRC\nFrench: TGT`

For the translation of full articles with Qwen3, we use the following prompt:

(4) `Translate the following text from English into French.\nEnglish: SRC\nFrench:`

For full article translation with Llama3, we use the following system prompt:

(5) `You are a good translator! Translate the following text from English into French. Reply only with the translated text.`

and the following user prompt template:

`English: SRC\nFrench:`

Metrics To evaluate translation quality, we use standard BLEU (Papineni et al., 2002) and its document-level variant, denoted ds-BLEU (Peng et al., 2024b).[35,36] We also report the document-level COMET (d-COMET) score (Vernikos et al., 2022) with `wmt22-comet-da` (Rei et al., 2022). While ds-BLEU can be applied as-is to documents without requiring sentence alignment, to evaluate BLEU and d-COMET, we first have to realign abstracts, paragraphs, and articles at the sentence level.[37]

5. Results and Analysis

5.1. Paragraph-level MT

Tables 4 reports the translation quality of the six MT systems in translating abstracts from test sets in EPS-test and in NLP-test.

For EPS, DeepLPro and FT-Euro-EPS are ranked as the top two MT systems. DeepLPro achieves the best d-COMET scores for all test sets, while FT-Euro-EPS results in higher BLEU scores for three out of four subsets of EPS-test. We obtain performance gain through fine-tuning both LLMs on EPS-train, except for CRG when using the fine-tuned EuroLLM.

[30] https://deepl.com

[31] https://www.systransoft.com/

[32] DeepL and Systran were accessed in March 2026 for abstracts and October 2025 for MERSENNE articles.

[33] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[34] https://huggingface.co/Qwen/Qwen3-8B

[35] For ds-BLEU, each document is considered as a single segment (all sentences concatenated), 'sentence-level' BLEU is applied to each document and the average score is calculated over documents.

[36] We use SacreBLEU (Post, 2018) with the signature `nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.4.0` to compute BLEU. For ds-BLEU, effective order is activated with `eff:yes`

[37] We do this by first aligning at the character level between the automatic translation and its reference using `edlib` (https://pypi.org/project/edlib/) then segmenting sentences with respect to the sentence boundaries of the reference.

	BSGF			CRAS			CRG			$\text{THESES}_{\text{EPS}}$		
	BLEU	ds-BLEU	d-COMET	BLEU	ds-BLEU	d-COMET	BLEU	ds-BLEU	d-COMET	BLEU	ds-BLEU	d-COMET
Systran	**42.8**	**44.3**	78.5	33.3	33.0	79.5	48.1	48.2	82.6	47.3	48.4	82.4
DeepL	41.2	43.2	**80.1**	33.5	33.2	**81.1**	47.5	47.9	**83.6**	45.8	47.2	**83.0**
Tower	36.1	38.2	75.6	33.2	33.1	78.6	48.5	48.3	81.7	43.9	45.5	80.6
FT-Tower-EPS	38.7	40.0	77.9	34.9	34.4	79.5	49.1	48.9	82.0	45.0	46.7	81.9
EuroLLM	40.7	42.1	78.6	35.8	35.5	79.9	**51.5**	**51.5**	82.8	47.5	49.2	82.7
FT-Euro-EPS	41.1	42.5	78.5	**36.6**	**35.8**	80.1	50.9	50.2	82.5	**48.0**	**49.5**	82.6

	rTAL			$\text{THESES}_{\text{NLP}}$		
	BLEU	ds-BLEU	d-COMET	BLEU	ds-BLEU	d-COMET
Systran	34.5	33.4	76.1	**43.0**	**43.1**	78.9
DeepL	34.2	33.6	**78.0**	41.7	42.4	**80.4**
Tower	32.1	31.1	75.5	40.0	40.1	77.7
FT-Tower-NLP	32.8	32.1	76.4	41.5	41.6	78.6
EuroLLM	34.6	33.5	76.4	**43.0**	42.7	79.2
FT-Euro-NLP	**35.0**	**33.8**	76.9	42.8	**43.1**	79.3

Table 4: BLEU, ds-BLEU and d-COMET scores for each subset of EPS-TEST (top) and NLP-TEST (bottom), corresponding to abstract translation. Best and second-best scores are bold and underlined, respectively

For NLP, DEEPLPRO and FT-EURO-NLP achieve the best and second-best performance for all metrics. As for EPS, we observe that fine-tuning is beneficial for both TOWER and EUROLLM.

5.2. MT of full articles

To investigate the capacity of LLMs with large context lengths to translate full scientific articles, we evaluate LLAMA3 and QWEN3 on the full-length articles from the MERSENNE subset of EPS-TEST-LONG. We also calculate the scores at the paragraph level in order to also compare the LLMs to the MT models from the previous experiments, whic have limited context window sizes.

Scores for paragraph-level translation are given in Table 5: EUROLLM and FT-EURO-EPS perform the best.

To study the effect of increasing the size of translation segments, we compare the translation quality (BLEU scores) of LLAMA3 and QWEN3 when translating MERSENNE at the sentence, paragraph, and article level. The results reported in Table 6 show that the translation quality of LLAMA3 slightly improves when translating paragraphs instead of sentences, although it degrades when translating full-length articles, despite the lengths of input articles fitting within the context window size. This is consistent with the findings of Wang et al. (2024) and Peng et al. (2025). The value of the brevity penalty observed suggests that under-translation is one of the reasons apart from the quality degradation. In contrast, the BLEU scores of QWEN3 increase when translating full articles with respect to paragraph-level translation, suggesting its robustness in long-context MT.

MT system	BLEU	ds-BLEU	d-COMET
Systran	57.3	58.0	87.3
DeepL	56.2	58.6	**88.0**
Tower	56.2	55.5	85.8
FT-Tower-EPS	57.1	58.4	87.0
EuroLLM	**61.7**	**62.7**	87.6
FT-Euro-EPS	59.7	60.7	87.6

Table 5: Scores for MT systems translating the MERSENNE subset of EPS-TEST-LONG (full articles) at the paragraph level. Best and second-best scores are bold and underlined, respectively.

Model	sent2sent	par2par	doc2doc
Llama3	51.2 (1.00)	51.8 (1.00)	49.2 (0.96)
Qwen3	48.4 (1.00)	48.3 (1.00)	50.8 (0.99)

Table 6: BLEU score (and brevity penalty) for LLMs translating MERSENNE articles at the sentence (sent2sent), paragraph (par2par), and full document (doc2doc) levels.

6. Conclusion

To address the scarcity of parallel documents in scientific fields for document-level MT, we constructed two English–French parallel corpora, PARAEPS and PARANLP, consisting of parallel abstracts and full-length parallel articles in Earth and Planetary Sciences and in Natural Language Processing respectively. Each corpus comprises TRAIN, DEV and TEST sets of abstracts and a second test of full articles (TEST-LONG). While most of the translations are produced through manual effort, the NLP-TEST-LONG is partly made up of silver translations constructed from comparable versions of the same article in English and French.

We present the original pipeline by which we derive silver parallel articles from those comparable human-written articles. Our experiments demonstrate the usefulness of our corpora, as fine-tuning LLMs on our dataset improves translation quality, and the parallel articles provide resources for the evaluation of article-level MT. Our future work will explore more precisely how document-level phenomena are handled by machine translation systems. This includes discourse phenomena that require extra-sentential context, but also terminological variation (in terms of consistency and logical use of term variants such as acronyms and reduced forms). Another short term goal will be to develop parallel scholarly corpora for a more diverse set of scientific domains, including other scientific domains that are characterised by domain-specific notation and formulae.

7. Ethical Considerations

BSGF, CRAS and CRG articles are released under a permissive CC BY 4.0 license. PARANLP abstracts are considered as part of the metadata of published documents and are therefore not copyrighted. The status of the CanMin and CJES abstracts is more restricted. Therefore, we open-source the scripts to collect and process the abstracts. This situation could change if specific permission of the respective publishers is received.

The parallel articles will also be released in accordance with the licenses of raw texts. *Comptes Rendus Géoscience* has been distributed since 2020 in partnership with the Mersenne Centre for Open Scientific Publishing based on a diamond open access policy, the journal articles and their translations distributed under a CC-BY 4.0 licence. The translated articles from the STUDENT collection are either Open Access or included in the ISTEX database.

The tools and LLMs used in our experiments are open-source or open-weights except for the commercial MT systems Systran and DeepL. We do see any ethical issues with this work.

8. Acknowledgements

This work was supported by the French national agency (ANR) as part of the MaTOS project under reference ANR-22-CE23-0033.[38] 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 wish to thank Célia Vaudaine and Caroline Rossi for giving access to the MERSENNE corpus, to Maxime Bouthors for early work on the abstract corpus, and to Nicolas Dahan for an ex-post analysis of the data quality. The authors are also grateful to the anonymous reviewers for their insightful comments and suggestions.

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

9. Bibliographical References

Sadaf Abdul Rauf and François Yvon. 2024. Translating scientific abstracts in the bio-medical domain with structure-aware models. *Computer Speech and Language*, 87:101623.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Marcello Federico, Nicola Bertoldi, Marco Trombetti, and Alessandro Cattelan. 2014. MateCat: an open source CAT tool for MT post-editing. In *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas: Tutorials*, Vancouver, Canada. Association for Machine Translation in the Americas.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, and Bobbie Chern et al. 2024. The Llama 3 Herd of Models. Preprint arXiv:2407.21783.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. 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*.

Lei Liu 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.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, 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. Preprint arXiv:2506.04079.

Yasmin Moslem, Rejwanul Haque, John D Kelleher, and Andy Way. 2023. Adaptive Machine Translation with Large Language Models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Minh Van Nguyen, 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 *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 80–90, Online. Association for Computational Linguistics.

Kishore Papineni, 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024a. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*.

Ziqian Peng, Rachel Bawden, and François Yvon. 2024b. À propos des difficultés de traduire automatiquement de longs documents. In *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, pages 2–21, Toulouse, France. ATALA and AFPC.

Ziqian Peng, Rachel Bawden, and François Yvon. 2025. Investigating length issues in document-level machine translation. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 4–23, Geneva, Switzerland. European Association for Machine Translation.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. TransQuest: Translation quality estimation with cross-lingual transformers. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5070–5081, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Dimitris Roussis, Sokratis Sofianopoulos, and Stelios Piperidis. 2024. Enhancing scientific discourse: Machine translation for the scientific domain. In *Proceedings of the 25th Annual*

Conference of the European Association for Machine Translation (Volume 1), pages 275–285, Sheffield, UK. European Association for Machine Translation (EAMT).

Paul Sebo and Sylvain de Lucia. 2024. Performance of machine translators in translating French medical research abstracts to English: A comparative study of DeepL, Google Translate, and CUBBITT. *PLOS ONE*, 19(2):1–13.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the seventh conference of the Association for Machine Translation in the America (AMTA)*, pages 223–231, Boston, Massachusetts, USA.

Dániel Varga, Péter Halaácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005 Conference*, pages 590–596.

Giorgos Vernikos, Brian Thompson, Prashant Mathur, and Marcello Federico. 2022. Embarrassingly easy document-level MT metrics: How to convert any pretrained metric into a document-level metric. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 118–128, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking and improving long-text translation with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7175–7187, Bangkok, Thailand. Association for Computational Linguistics.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025. DelTA: An online document-level translation agent based on multi-level memory. In *The Thirteenth International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, and Jiaxi Yang et al. 2025. Qwen3 technical report. Preprint arXiv:2505.09388.

Ziming Zhu, Chenglong Wang, Shunjie Xing, Yifu Huo, Fengning Tian, Quan Du, Di Yang, Chunliang Zhang, Tong Xiao, and Jingbo Zhu. 2025. LaTeXTrans: Structured LaTeX Translation with Multi-Agent Coordination. Preprint arXiv:2508.18791.

Sonia Zulfiqar, M. Farooq Wahab, Muhammad Ilyas Sarwar, and Ingo Lieberwirth. 2018. Is Machine Translation a Reliable Tool for Reading German Scientific Databases and Research Articles? *Journal of Chemical Information and Modeling*, 58(11):2214–2223.

10. Language Resource References

Esalati, Mersad and Dousti, Mohammad Javad and Faili, Heshaam. 2024. *Esposito: An English-Persian Scientific Parallel Corpus for Machine Translation*. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). ELRA and ICCL. PID https://huggingface.co/datasets/universitytehran.

Ive, Julia and Max, Aurélien and Yvon, François and Ravaud, Philippe. 2016. *Diagnosing High-Quality Statistical Machine Translation Using Traces of Post-Edition Operations*. International Conference on Language Resources and Evaluation - Workshop on Translation Evaluation: From Fragmented Tools and Data Sets to an Integrated Ecosystem (MT Eval 2016).

Hannah Calzi Kleidermacher and James Zou. 2025. *Science Across Languages: Assessing LLM Multilingual Translation of Scientific Papers*. PID https://arxiv.org/abs/2502.17882.

Iñaki Lacunza and Javier Garcia Gilabert and Francesca De Luca Fornaciari and Javier Aula-Blasco and Aitor Gonzalez-Agirre and Maite Melero and Marta Villegas. 2025. *ACADATA: Parallel Dataset of Academic Data for Machine Translation*. PID https://huggingface.co/datasets/BSC-LT/ACAData.

Denis Maurel, Enza Morale, Nicolas Thouvenin, Patrice Ringot, and Angel Turri. 2019. Istex: A database of twenty million scientific papers with a mining tool which uses named entities. *Information*, 10(5).

Névéol, Aurélie and Jimeno Yepes, Antonio and Neves, Mariana and Verspoor, Karin. 2018. *Parallel Corpora for the Biomedical*

Domain. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA). PID https://github.com/biomedical-translation-corpora/corpora.

Dayyán O'Brien and Bhavitvya Malik and Ona de Gibert and Pinzhen Chen and Barry Haddow and Jörg Tiedemann. 2025. *DocHPLT: A Massively Multilingual Document-Level Translation Dataset*. PID https://huggingface.co/datasets/HPLT/DocHPLT.

Roussis, Dimitrios and Papavassiliou, Vassilis and Prokopidis, Prokopis and Piperidis, Stelios and Katsouros, Vassilis. 2022. *SciPar: A Collection of Parallel Corpora from Scientific Abstracts*. Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association. PID https://live.european-language-grid.eu/catalogue/corpus/20067.

Salesky, Elizabeth and Darwish, Kareem and Al-Badrashiny, Mohamed and Diab, Mona and Niehues, Jan. 2023. *Evaluating Multilingual Speech Translation under Realistic Conditions with Resegmentation and Terminology*. Association for Computational Linguistics. PID https://iwslt.org/2023/multilingual.

Wang, Longyue and Du, Zefeng and Jiao, Wenxiang and Lyu, Chenyang and Pang, Jianhui and Cui, Leyang and Song, Kaiqiang and Wong, Derek and Shi, Shuming and Tu, Zhaopeng. 2024a. *Benchmarking and Improving Long-Text Translation with Large Language Models*. Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics. PID https://github.com/longyuewangdcu/Document-MT-LLM.

Wang, Longyue and Liu, Siyou and Lyu, Chenyang and Jiao, Wenxiang and Wang, Xing and Xu, Jiahao and Tu, Zhaopeng and Gu, Yan and Chen, Weiyu and Wu, Minghao and Zhou, Liting and Koehn, Philipp and Way, Andy and Yuan, Yulin. 2024b. *Findings of the WMT 2024 Shared Task on Discourse-Level Literary Translation*. Proceedings of the Ninth Conference on Machine Translation. Association for Computational Linguistics. PID https://www2.statmt.org/wmt24/literary-translation-task.html.