Leveraging Large Language Models for maintaining scientific multilingual documents

Yehor Korotenko[1][0009−0002−4570−2391]

Laboratoire Interdisciplinaire des Sciences du Numérique, Gif-sur-Yvette, France
yehor.korotenko@universite-paris-saclay.fr

Abstract. We explore the specific challenges of authoring and maintaining multilingual computational scientific narratives, like course notes, textbooks, or reference manuals, and the design space for leveraging adaptive machine translation to assist authors.

Keywords: Machine Translation · Large Language Models · Markup Languages · NLP

1 Introduction

Over the years, machine translation has remained a prominent area of research in computer science, continually benefiting from advancements in neural network architectures.

Although current machine translation systems achieve high-quality results for individual text translations and numerous tools exist to support the translation process, they remain inadequate for the continuous maintenance of evolving multilingual projects. In particular, these tools face significant challenges when handling documents written in *markup* languages such as *Markdown*, *LaTeX*, or *MyST*.

Moreover, existing tools typically do not retain the author's stylistic intent—a factor that plays a critical role in the clarity, tone, and pedagogical effectiveness of written materials.

We introduce a methodical framework and supporting toolchain that facilitates the sustainable translation of evolving scientific documents, with explicit preservation of *structural layout*, *stylistic consistency*, and *domain-specific syntax*. We design a markup-aware pipeline that parses, tokenizes, and synchronizes document segments using placeholder-preserving translation logic and author-style adaptation heuristics using the context of the document.

2 Approach

We divide the translation challenge into two complementary aspects:

- **Syntax preservation**: protecting non-linguistic elements such as LaTeX macros, code, and references.
- **Style preservation and long-term maintainability**: ensuring consistent terminology and authorial style across document revisions.

2.1 Baseline Translation with Syntax Preservation

A key difficulty in translating scientific documents is preserving markup syntax, which standard MT systems often corrupt. Our approach is to analyze the document beforehand, split it into context-aware chunks, and represent each chunk as a single `<TEXT>` element containing natural

language, with inline placeholders `<PH original="..."/>` for non-linguistic spans (e.g., LaTeX math, commands, or code).

This design has two advantages. First, it enables correct word reordering between languages (e.g., adjective–noun inversion in French → English) without breaking the surrounding markup. Second, the `original` attribute of each `<PH>` tag provides the model with contextual information. For example:

```
Let <PH original="$f$"/> be an endomorphism.
```

produces more accurate results than the contextless:

```
Let <PH/> be an endomorphism.
```

Among the tested strategies, this method was most effective: it minimizes ambiguity by explicitly labeling translatable versus non-translatable spans, reducing the number of sub-tasks the model must implicitly solve.

2.2 Style Preservation and Long-Term Maintenance

As documents grow in size and evolve over time, maintaining translation quality and consistency becomes increasingly challenging. Retranslating unchanged text is both wasteful and undesirable, as it discards valuable human post-edits and terminology choices.

To address this, our system introduces a *translation memory and correspondence database*. Each chunk of source or translated text is stored in a language-specific directory, where the file name is the checksum of the chunk. A correspondence table links checksums across languages. This design allows the system to:

- detect and reuse unchanged chunks without invoking a model,
- identify slightly modified chunks (e.g., one sentence added, or a few words reordered),
- provide the model with the previous source–translation pair as contextual guidance.

This mechanism not only reduces cost and energy consumption by avoiding redundant translation, but also preserves the author's style and terminology over multiple revision cycles. When small edits occur, the system encourages minimal updates to existing translations rather than full re-generation, helping ensure stylistic continuity across languages and versions.

2.3 Domain Vocabulary Support

Many translation projects involve domain-specific terminology whose meaning differs from everyday language, leading to frequent ambiguities. For example, the English word *thread* may translate into French as *fil* in a textile context, but as *thread d'exécution* in a computing context. Similar issues arise with terms such as *ring* (mathematics vs. jewelry) or *cell* (biology vs. prisons).

To mitigate this, our system allows users to provide a custom *vocabulary dictionary* of source–target pairs (`original` → `translation`). During translation, these entries act as high-priority overrides: they are applied only to natural-language segments and never to code or markup placeholders.

This feature ensures terminological consistency across large projects, reduces post-editing effort, and gives authors control over critical word choices in specialized domains.

3 Evaluations and Preliminary Results

State-of-the-art large language models (LLMs) achieve excellent performance in plain-text translation, but their ability to handle scientific documents with mixed natural language and markup remains less clear. To assess this, we conducted preliminary evaluations of different model families, focusing on their trade-offs in translation quality, structural fidelity, and resource requirements.

We considered three representative models:

- `Gemini-2.0-flash`: proprietary, highly energy-efficient, offering the strongest overall quality;
- `Llama-3.3-70B`: a medium-size open-weight model hosted at Université Paris-Saclay;
- `Gemma3-27B`: a smaller open-weight model that can run locally on personal hardware.

Since `Gemini-2.0-flash` consistently produced the best results, our comparisons emphasize the two open-weight models, which are more attractive for privacy-preserving and institutionally hosted workflows. Tests were divided into two categories: (i) natural language translation without markup syntax, and (ii) translation of syntax-heavy documents (LaTeX, Markdown, Jupyter).

Plain text. In the absence of markup, *Gemma3-27B* often outperformed *Llama-3.3-70B* in lexical choice and fluency, especially in less-resourced languages such as Ukrainian. However, its smaller capacity sometimes led to grammatical inconsistencies, where LLaMA's larger context window provided more stable results.

Syntax-preserving translation. When markup elements were present, the trend reversed: *LLaMA-3.3-70B* was markedly more robust, while *Gemma3-27B* frequently omitted or altered placeholders, left elements untranslated, or dropped content entirely. This suggests that Gemma's smaller effective context window and parameter count limit its reliability for structure-aware translation tasks.

Overall, while *Gemma3-27B* is attractive for its ability to run locally and low resource usage, its reliability in syntax-heavy translation is limited. *LLaMA-3.3-70B* is more robust but requires institutional infrastructure and is less reliable then *Gemini-2.0-flash* that remains the strongest overall, but its proprietary nature limits its integration in privacy-sensitive workflows.

4 Conclusion

We presented a framework for adaptive translation of scientific documents that combines XML-based syntax preservation, translation memory for long-term style continuity, and domain vocabulary support. Preliminary evaluations across different model families confirm that while open-weight models can achieve competitive results on plain text, they struggle with syntax-heavy documents, underscoring the need for our model-agnostic strategies. Our open-source toolchain[1] offers a practical basis for privacy-preserving, reproducible multilingual authoring workflows. Future work will extend chunking to additional markup formats and integrate tighter human-in-the-loop feedback for style guidance.

[1] https://github.com/DobbiKov/sci-trans-git, DOI: 10.5281/zenodo.16983036

References

1. Moslem, Yasmin and Haque, Rejwanul and Kelleher, John D and Way, Andy.: Adaptive machine translation with large language models. In: arXiv preprint arXiv:2301.13294. https://arxiv.org/abs/2301.13294