TaMTAS: Terminology-Aware Machine Translation for Accessible Science

Corpus compilation, terminology extraction and data augmentation

Antoni Oliver
Universitat Oberta de Catalunya
aoliverg@uoc.edu

Sergi Alvarez-Vidal
Universitat Autònoma de Barcelona
sergi.alvarez@uab.cat

Abstract

This paper presents the TaMTAS project (Terminology-Aware Machine Translation for Accessible Science), a research project coordinated by the Universitat Oberta de Catalunya (UOC) to develop an open-source translation ecosystem for the Life Sciences. While we provide a general overview of the project's organization into seven Work Packages (WPs) and its collaborative consortium, this article focuses specifically on the work of WP2. Led by the UOC, this package is responsible for the parallel corpus compilation for five languages (English, Spanish, Catalan, Estonian, and Irish), the enhancement of TBX-Tools for terminology extraction, and the development of synthetic data augmentation strategies. These linguistic assets are essential to power the downstream Large Reasoning Models (LRMs) and Automatic Post-Editing (APE) modules, ensuring terminological consistency in highly specialized scientific domains.

1 Introduction

The dominance of English in scientific dissemination creates a significant barrier to knowledge access, affecting the ability of researchers, students, and the general public to access global research, while also limiting the capacity of researchers to disseminate their own findings in their native languages. To address this challenge, the TaMTAS project emerges as a 36-month initiative (15 December 2025 – 14 December 2028). Currently in its initial phase, the work focuses on the foundational data architecture and the functional setup of WP2 tasks. This collaborative effort is led by the Universitat Oberta de Catalunya (UOC), acting as project coordinator, in partnership with the Barcelona Supercomputing Center (BSC), the University of Surrey (UoS), Dublin City University (DCU), and the University of Tartu (UT). Focusing on the highly specialized Life Sciences domain, the project supports five languages with varying resource availability: English, Spanish, Catalan, Estonian, and Irish.

Unlike traditional neural machine translation (NMT) approaches, which often struggle with terminological consistency in long documents, TaMTAS introduces a paradigm shift by utilizing Large Reasoning Models (LRMs). Within this framework, translation is treated as a multi-step reasoning task. By leveraging chain-of-thought prompting, glossary-guided constraints, and self-correction mechanisms, the system ensures that specialized terminology is rendered accurately and consistently throughout entire texts.

2 Architecture and Work Packages

TaMTAS is structured as a comprehensive software and data pipeline, transitioning from raw data acquisition to end-user application. The ecosystem is built through collaborative Work Packages (WPs):

- **WP1 (UOC):** *Project Management and Coordination.* Administrative governance and inter-partner synergy.
- **WP2 (UOC):** *Corpus compilation, terminology extraction and data augmentation with terminology.* Compilation and curation of Life Sciences corpora and functional enhancement of TBX-Tools (Oliver and Vàzquez, 2015).
- **WP3 (BSC):** *Terminology-aware MT.* Development of Large Reasoning Models (LRMs) for domain-specific translation.
- **WP4 (UoS):** *Terminology-aware Quality Estimation and Automatic Post-Editing.* Implementation of QE and APE modules to ensure terminological integrity.
- **WP5 (UT):** *Post-Translation Text Augmentation.* Content adaptation through simplification

and explanatory enrichment for diverse audiences.

- **WP6 (DCU):** *Evaluation.* End-to-end validation with real-world stakeholders and scientific partners.
- **WP7 (DCU):** *Dissemination, Outreach and Impact.* Management of FAIR-compliant data release and strategic project communication.

3 WP2: Corpus Compilation, Terminology Extraction and Data Augmentation

Work Package 2 (WP2), led by the Universitat Oberta de Catalunya (UOC), plays a crucial role in the success of the TaMTAS project. Its main objective is to create and curate high-quality linguistic resources that are essential for training and validating terminology-aware machine translation models. These resources are vital for downstream components, such as the Large Reasoning Models (LRMs) and Automatic Post-Editing (APE) modules, ensuring terminological consistency and overall quality in highly specialized scientific domains. This work package focuses on three strategic areas:

- **Large-scale Corpus Curation:** The project involves compiling parallel and comparable corpora within the Life Sciences domain, focusing on high-quality alignment at the sentence and paragraph levels. Texts will be gathered from open-access academic journals, public health guidelines, and regulatory documents. The corpus will cover various text types, including research papers and medical guidelines, to create robust training material for the Large Reasoning Models (LRMs). Regular updates to the corpus will maintain its relevance and support the evolving needs of the field.
- **Functional Optimization of TBXTools:** WP2 will enhance TBXTools (Oliver and Vàzquez, 2015), UOC's open-source terminology software, to significantly improve the efficiency and accuracy of terminology extraction. The optimized TBXTools will be capable of automatically generating TBX-formatted termbases that are enriched with both morphological and syntactic metadata, allowing for a deeper understanding of the terms within their specific contexts. This enhancement will ensure that terminology remains consistent across translations, particularly in the specialized domain of Life Sciences. Moreover, the improved TBXTools will be able to handle complex multi-word expressions and highly specialized domain-specific terms, which are common in scientific texts. By enabling precise extraction and management of these terms, the system will contribute to higher accuracy and reliability in translations, ensuring that the final outputs are both technically correct and contextually appropriate.
- **Innovative Data Augmentation:** WP2 introduces a synthetic data augmentation strategy specifically designed to address the challenge of rare or out-of-vocabulary (OOV) terms. It leverages intelligent term substitution techniques within existing parallel segments of the corpus. By identifying similar, contextually appropriate terms from the corpus, the system can generate new training instances that expand the vocabulary without the need for manual data collection. This approach will prove particularly useful in low-resource languages such as Estonian and Irish, where obtaining extensive linguistic data can be challenging. As a result, the system's ability to handle specialized scientific terminology will be significantly enhanced.

4 Conclusion

The TaMTAS project represents a significant step forward in specialized machine translation. By framing translation as a reasoning task and backing it with the rigorous, high-quality data infrastructure developed by the UOC, the project will deliver a suite of practical products ready for integration into modern translation workflows. Consistent with our commitment to open science, all linguistic resources and software developed will be released under open-source licenses through the project website and FAIR-compliant repositories.

Acknowledgements

This work has been supported by project PCI2025-167063-2, funded by MICIU/AEI/10.13039/501100011033 and by the European Union (CHIST-ERA).

References

Oliver, Antoni and Mercè Vàzquez. 2015. TBXTools: a free, fast and flexible tool for automatic terminology extraction. In *Proceedings of the international conference recent advances in natural language processing*, pages 473–479.