sciendo
DOI 10.2478/jazcas-2025-0021

REDEFINING TERMINOLOGY WORK: THE ROLE OF GLOSSARIES AND TMS IN CUSTOMIZED AI-DRIVEN MACHINE TRANSLATION

JAKUB ABSOLON
Department of British and American Studies, Faculty of Arts, University of Ss. Cyril and Methodius, Trnava, Slovakia (ORCID: 0009-0008-2365-1048)

ABSOLON, Jakub: Redefining Terminology Work: The Role of Glossaries and TMs in Customized AI-driven Machine Translation. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 237 – 246.

Abstract: This study investigates whether traditional terminology work in the customization of machine translation (MT) systems can be effectively replaced by translation memories (TMs) alone. Given the growing reliance on AI-driven translation tools, we evaluated three MT configurations using English–Slovak technical documentation: a baseline (non-customized system), a system customized with both TMs and a glossary, and a system customized with TMs only. Since the text corpora for the given area were sufficient, we used the LLM model to generate additional training data. Results show that TM-only customization can achieve terminology translation accuracy nearly equivalent to setups that include glossaries—particularly when supported by high-quality, domain-specific bilingual data. Nonetheless, glossary-based customization further improves consistency, and terminology errors persist across all systems. This suggests that although automation of translation processes can reduce dependence on traditional terminology building, terminology databases remain essential for ensuring the quality (QA) of the output text. The study offers practical guidance for translators, terminologists, and developers of translation tools by emphasizing the importance of collaboration between automated and human-driven translation processes. It also underscores both the promise and limitations of LLM-generated data for domain adaptation in low-resource language settings.

Keywords: terminology, machine translation, customization, translation memory, glossary, AI, domain adaptation, low-resource languages, quality assurance, post-editing

1 INTRODUCTION

The rapid evolution of artificial intelligence (AI) has transformed machine translation (MT), reshaping traditional translation workflows. The translation process used to be linear and the privilege of translators, human beings. Nowadays, translation increasingly utilizes artificial intelligence, which speeds up and streamlines the process; however, it also introduces the risk of unpredictability in the quality of the final text and the possibility of critical errors in high-risk areas. Consequently, customized MT solutions, leveraging domain-specific adaptations, are emerging as a practical middle-ground to balance automation with quality.

Customizable MT engines allow users to incorporate domain-specific resources, such as translation memories (TMs) and glossaries, potentially challenging the traditional role of terminological work in ensuring translation accuracy and consistency—particularly in technical and scientific contexts.

1.1 Theoretical framework

The theoretical foundation for our research is grounded in two dominant schools of terminology theory:

1. Socioterminology, pioneered by (Gaudin 1993; as cited in Temmerman 2000), positions terminology as an inherently socially situated phenomenon. Terms emerge and evolve within expert communities, reflecting usage variation and social context.
2. Sociocognitive terminology, introduced by Temmerman (2000), emphasizes the cognitive and contextual dimensions of term usage. This approach emphasizes the role of contextual, cognitive, and discursive influences in the creation of meanings and variations of concepts.

Furthermore, terminology developed by Faber (2012), based on a cognitive framework, integrates cognitive semantics and terminology management.

1.2 Research focus

Building on this theoretical grounding, our study examines whether TMs alone can effectively customize MT systems, potentially substituting traditional glossary-based terminology practices. Since glossary creation requires considerable resources and the use of translation memory-based solutions is growing, research on this issue provides important practical and academic insights.

2 RELATED WORK

2.1 Traditional terminology work

Terminology work ensures consistency in technical domains through the collection and management of terms. The dynamics of language are shaped by technological, social, and cultural factors, and therefore require approaches to terminology work that are capable of responding to changing contexts and shifts in the meaning of terms. Translators must handle:

- **Conceptual deviation**: Term meanings may diverge across domains or cultures.
- **False friends**: Lexical homonyms with different meanings in different languages, which increases the risk of misinterpretation in translation.

- **Cultural/Contextual gaps**: MT systems often lack nuance, requiring human input.

This complexity drives a shift toward integrating AI with traditional term workflows.

2.2 AI-enhanced terminology strategies

Modern research suggests that combining AI technologies with traditional terminology workflows can mitigate many of the above limitations:

- **Terminology-aware MT**: Techniques like constrained decoding improve term accuracy (Bogoychev and Chen 2023).
- **WMT 2023 Shared Task**: Term injection during training/inference improved accuracy, though BLEU gains varied (Semenov et al. 2023).
- **Human-AI synergy**: Translators now focus on creativity and QA while machines handle routine tasks (Gao 2022).

The above-mentioned strategies underscore the importance of designing work processes that consider both cognitive mechanisms and technical efficiency.

2.3 LLM-based synthetic terminology training

Large language models (LLMs) such as GPT-4 are increasingly used to augment training data when bilingual corpora are scarce. A study by Moslem et al. (2023) during WMT 2023 demonstrated that synthetic parallel sentences generated by LLMs, followed by fine-tuning and human post-editing, led to notable improvements in terminology translation accuracy, from 37% to over 70% for domain-specific terms (ACL Anthology).

However, the synthetic data often requires rigorous human curation, as LLM outputs may introduce semantic simplifications or hallucinate context.

2.4 Domain-specific terminology: The volcanology case

A study by Harris et al. (2017) explored the translation of volcanological terms across multiple languages, emphasizing that terminological consistency and scientific accuracy are essential in high-risk fields. Their work confirmed that machine translation alone cannot guarantee conceptual clarity or cross-cultural appropriateness without human oversight.

3 MACHINE TRANSLATION (MT) CUSTOMIZATION

Machine translation (MT) customization is a critical area of both applied practice and ongoing research, particularly when aiming to improve translation

quality in specialized domains. The process of adapting machine translation typically involves the use of domain-specific resources, primarily translation memories and glossaries, to increase the accuracy and relevance of the output. This section reviews current customization techniques and outlines best practices based on recent empirical findings.

3.1 Customization techniques

3.1.1 Fine-tuning and data selection

Fine-tuning MT models using in-domain bilingual data has been shown to significantly improve both terminology translation accuracy and overall contextual fidelity. In our English–Slovak case study, fine-tuning yielded measurable performance gains. Selecting high-quality training segments—especially with the aid of document classification tools—enables more efficient domain adaptation, often outperforming generic MT systems trained on larger but less relevant datasets.

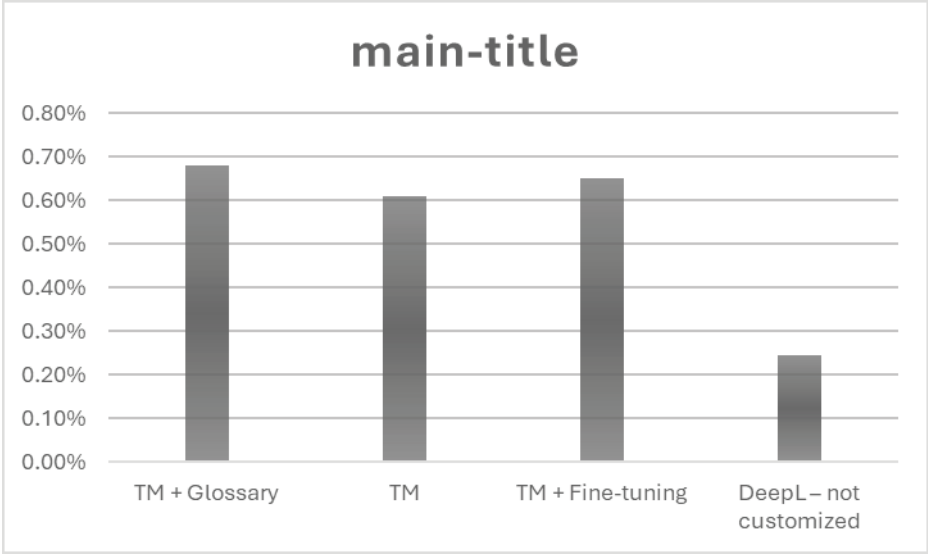

Fig. 1. Terminology translation accuracy

3.1.2 Terminology integration

Integrating user-defined glossaries into MT engines ensures that domain-critical terminology is translated consistently and accurately. The process of adapting machine translation typically involves utilizing specific domain resources, primarily translation memories and glossaries, to enhance the accuracy and relevance of the output text. By defining preferred translation equivalents, glossaries guide MT system output towards consistency and compliance with industry standards.

3.2 Implications and future directions

Integrating TMs and glossaries into MT customization is a proven strategy for enhancing translation quality, particularly in specialized domains. These approaches not only increase accuracy and consistency but also the ability of MT systems to respond to specific user needs. Future research should focus on developing scalable, cost-effective solutions such as AI-supported glossary generation and adaptive MT systems with real-time customization capabilities.

4 METHODOLOGY

4.1 Source text description

The source text used for evaluation is a specialized technical manual for industrial packaging equipment. Although originally authored in English, its lexical patterns and syntactic structures indicate influence from Italian, making it representative of multilingual industrial documentation. The manual is intended for use by technicians and maintenance personnel in manufacturing environments.

It includes detailed operational instructions, safety guidelines, component specifications, and references to EU regulations (e.g. Directive 2006/42/EC). The text is characterized by high terminological density, frequent use of compound noun phrases, imperative forms, and structurally consistent formatting. Key terms include film tensioning system, pre-stretch carriage, photocell sensor, vacuum chamber, and emergency stop function — all of which pose a challenge for accurate machine translation and make the material well-suited for evaluating terminology handling and consistency in customized MT systems.

4.2 Machine translation configurations

According to Akhulkova (2023), the Language Technology Atlas identifies 111 MT solutions that are currently available, with 33 offering customization capabilities. The Intento "State of Machine Translation 2024" report evaluated 52 MT engines and LLMs. Among 28 MT engines, 7 supported both TM and glossary customization, 2 supported glossary-only customization, and 1 supported TM-only customization. Among the LLMs assessed, 9 supported only glossary customization and 15 supported both glossary and TM customization via techniques such as fine-tuning, retrieval-augmented generation (RAG), or prompt engineering.

Despite the broader contextual understanding of LLMs, their tendency to hallucinate factual content made them less suitable for high-precision translation in this study. Therefore, we focused on customizable neural MT (NMT) systems, evaluating three widely used platforms: DeepL, Microsoft Translator, and Google Translate.

For the English–Slovak language pair, Microsoft Custom Translator was selected due to its technical feasibility, affordability, and robust language support. Notably, the platform supports both inference-time glossary integration and weakly

supervised fine-tuning using translation memory (TM) data—an approach also mirrored by several top-performing systems in the WMT 2023 Terminology Shared Task (Semenov et al. 2023). In collaboration with ASAP-translation.com, s.r.o., we prepared a domain-specific dataset, which includes a TMX file containing 29,334 bilingual sentence pairs and a glossary of 39 verified English–Slovak term pairs.

4.3 MT system configurations tested

To assess the impact of TMs and glossaries on translation quality, we tested three MT system configurations:

- **Non-Customized MT (Baseline)**: A generic MT system with no domain-specific adaptation.
  → DeepL was chosen for this configuration, based on its empirical performance in the EN–SK language pair.
- **Customized MT with TM + Glossary**: A system trained with both a domain-specific translation memory and a glossary containing the target terminology.
  → Implemented using Microsoft Custom Translator, model MT Custom 1.0.
- **Customized MT with TM Only**: A system trained solely on the translation memory, without an integrated glossary.
  → Implemented using Microsoft Custom Translator, models MT Custom 1.1 and 1.2.

For Model MT Custom 1.2, the dataset was extended to include additional translation units (TUs) containing five selected test terms, ensuring sufficient exposure during fine-tuning. Where authentic parallel data was insufficient, we generated synthetic training data using GPT-4.0 (OpenAI). To investigate the impact of term frequency on translation accuracy, we varied the number of training instances per term as follows:

- pulley – remenica: 25 TUs
- transpallet – paletový vozík: 50 TUs
- transit – posun: 100 TUs
- drawbar – ťahadlo: 100 TUs
- carriage – unášač: 200 TUs

This variation was designed to assess whether increased exposure to specific terms enhances their translation accuracy across various system configurations.

4.4 Evaluation and assessment criteria

The evaluation focused on both terminology-specific performance and overall translation quality. We employed three core evaluation dimensions:

1. **Terminology Handling**: Accuracy of term translation in context, fidelity to the intended technical meaning, and alignment with glossary entries (where applicable).
2. **Terminology Consistency**: Consistent use of terminology throughout the translated text, minimizing synonym variation or inconsistent renderings.
3. **Overall Translation Quality**: General fluency, adequacy, and faithfulness of the translations, assessed via both human and automated methods.

Human evaluation was conducted independently by two professional translators with expertise in technical translation. They assessed the accuracy and consistency of terminology on a subset of translated segments. In addition, the **BLEU (Bilingual Evaluation Understudy)** score was used as an automatic metric to supplement human judgments and facilitate comparison across MT configurations.

5 RESULTS AND DISCUSSIONS

5.1 Terminology handling and consistency

The effectiveness of different MT configurations in handling domain-specific terminology was evaluated based on the accurate rendering of five target terms across 69 segment occurrences. The results are as follows:

- **MT Custom 1.0 (TM + glossary)**: 68.1% accuracy (47/69)
- **MT Custom 1.1 (TM only)**: 60.9% accuracy (42/69)
- **MT Custom 1.2 (TM with fine-tuned synthetic data)**: 65.2% accuracy (45/69)
- **DeepL (non-customized baseline)**: 24.6% accuracy (17/69)

The highest accuracy was achieved using the configuration that incorporated both a domain-specific translation memory and glossary (MT Custom 1.0). Surprisingly, the fine-tuned model (MT Custom 1.2) achieved slightly lower accuracy, despite being supplemented with additional LLM-generated examples. This suggests that while synthetic data may help bridge gaps in terminology coverage, it cannot fully replace curated, human-validated content.

The TM-only configurations still performed reasonably well, confirming that a high-quality translation memory can support robust terminology handling even without a glossary. In contrast, the non-customized DeepL system struggled with specialized terms, reinforcing the need for domain adaptation.

All systems demonstrated vulnerabilities when contextual cues were insufficient, even those with glossary integration. This observation supports the notion that glossaries alone are insufficient for ensuring terminology accuracy and that leveraging full-sentence parallel data remains critical for robust customization.

Inconsistent term usage, even within customized models, further highlights the importance of post-editing and terminological quality assurance (QA). Moreover, the synthetic data generated by LLMs showed a tendency to simplify terminology, necessitating human review for effective integration into training workflows.

5.2 Overall translation quality

BLEU scores provide a supplementary measure of overall translation performance across the four system configurations:

- **MT Custom 1.2 (TM + synthetic fine-tuning)**: 75.18
- **MT Custom 1.0 (TM + glossary)**: 71.99
- **MT Custom 1.1 (TM only)**: 71.05
- **DeepL (baseline)**: 52.69

These results confirm that MT customization—particularly when augmented with fine-tuning on in-domain data—significantly enhances translation quality. While all customized configurations outperformed the baseline, the highest BLEU score was achieved by the system using LLM-generated supplemental training data, suggesting that targeted augmentation can improve general fluency and lexical adequacy, even if terminology fidelity remains a challenge.

Microsoft Custom Translator's fine-tuning mechanism proved effective, especially when trained with adequate domain-specific content. However, the marginal difference between the TM-only and TM+glossary configurations suggests that in some contexts, high-quality TMs alone can achieve near-equivalent performance.

5.3 Implications for terminology work

The results underscore the critical role of high-quality, human-validated data in effective MT customization. While glossaries enhance precision, their creation and maintenance remain resource-intensive. TM-only approaches, particularly when paired with synthetic data augmentation, offer a cost-effective alternative with reasonable performance.

This challenges the traditional view of terminologies as the core carriers of meaning in translation, as noted by Semenov et al. (2023), who argue that such assumptions may be overstated, especially given the comparable performance of systems relying solely on high-quality TMs.

Nevertheless, terminology errors persist, especially in complex technical texts. Consistent, expert-driven terminology work remains essential for both

training data quality and post-editing workflows. The increasing availability of AI-driven tools, such as automated term extraction, can help alleviate some of the manual burden. However, these tools require careful human curation to ensure that termbases remain accurate, contextually appropriate, and aligned with evolving domain standards.

Ultimately, scalable and high-quality localization will depend not on replacing traditional terminology work but on transforming it into a curation-centered, collaborative process, with translators, terminologists, and AI systems working in concert.

6 CONCLUSION AND FUTURE WORK

This study examined the interplay between translation memories (TMs), glossary integration, and traditional terminology work in the context of customized machine translation (MT) for the English–Slovak language pair. By evaluating multiple MT configurations, including TM-only customization, TM combined with a glossary, and TM fine-tuned with LLM-generated data, we explored the extent to which TMs can replace or complement conventional terminological resources in domain-specific translation workflows.

Our findings indicate that systems combining TMs with glossaries achieved the highest terminology translation accuracy. However, TM-only configurations delivered a comparable performance, particularly when enhanced with synthetic training data from large language models (LLMs). This suggests that well-constructed translation memories may, in some cases, reduce the need for exhaustive glossary compilation—especially in cost-sensitive or time-constrained settings.

Despite these gains, terminology inconsistencies persisted across all configurations. General-purpose MT systems like DeepL performed poorly with specialized terms, underscoring the importance of domain adaptation. Interestingly, the MT engine often prioritized TM-derived patterns over glossary entries, emphasizing the continued value of validated and well-curated termbases, particularly for post-editing and quality assurance (QA) processes.

Future research should further investigate how factors such as TM quality, term frequency, and domain variability influence terminology handling across language pairs. LLM-generated bilingual data for fine-tuning appears promising but requires rigorous human validation due to risks of semantic simplification and context loss. Additionally, the integration of AI-based term extraction and dynamic glossary adaptation during translation represents a key area for innovation.

As the scale and speed of localization increase, the field is gradually shifting from terminology creation to terminology curation. Supporting this shift will require deeper collaboration between human experts and machine learning systems, ensuring that automation enhances, rather than compromises, translation quality.

ACKNOWLEDGEMENTS

The author would like to thank ASAP-translation.com, s.r.o., for their collaboration and for providing domain-specific translation data and technical infrastructure essential to the successful execution of this research.

References

Akhulkova, Y. (2023). The 2023 Nimdzi Language Technology Atlas. Nimdzi Insights. Accessible at: https://www.nimdzi.com/language-technology-atlas/.

Buysschaert, J., and Kovács, L. (2017). Challenges encountered during the compilation of a multilingual termbase in the domain of communication. Terminology, 23(1), pp. 1–18. Accessible at: https://doi.org/10.18460/ANY.2017.1.001.

Faber, P. (ed.). (2012). A Cognitive Linguistics View of Terminology and Specialized Language (Applications of Cognitive Linguistics, Vol. 20). De Gruyter Mouton. Accessible at: https://doi.org/10.1515/9783110277203.

Forcada, M. L. (2017). Making sense of neural machine translation. Translation Spaces, 6(2), pp. 291–309. Accessible at: https://doi.org/10.1075/ts.6.2.04for.

Gao, J. (2022). The impact of digital technologies on the structure of translation activities. Litera, 10, pp. 72–86. Accessible at: https://doi.org/10.25136/2409-8698.2022.10.39067.

Harris, A. J. L., Belousov, A., Calvari, S., Delgado-Granados, H., Hort, M., Koga, K. T., Wulan Mei, E. T., Harijoko, A., Pacheco, J., Prival, J.-M., Solana, C., Þórðarson, Þ., Thouret, J.-C., and van Wyk de Vries, B. (2017). Translations of volcanological terms: Cross-cultural standards for teaching, communication, and reporting. Bulletin of Volcanology, 79(7), 57 p. Accessible at: https://doi.org/10.1007/S00445-017-1141-9.

Intento. (2024). The State of Machine Translation 2024: Independent Evaluation of MT Engines and LLMs. Accessible at: https://www.inten.to.

Li, B. (2023). Conceptual deviation in terminology translation. Terminology. Accessible at: https://doi.org/10.1075/term.00073.li.

Microsoft. (2024). Microsoft Custom Translator Documentation. Microsoft Learn. Accessible at: https://learn.microsoft.com/en-us/azure/ai-services/translator/custom-translator/overview [30/03/2025].

Moslem, Y., Romani, G., Molaei, M., Haque, R., Kelleher, J. D., and Way, A. (2023). Domain Terminology Integration into Machine Translation: Leveraging Large Language Models. In: P. Koehn – B. Haddow – T. Kocmi – C. Monz (eds.): Proceedings of the Eighth Conference on Machine Translation, pp. 902–911. Association for Computational Linguistics. Accessible at: https://doi.org/10.18653/v1/2023.wmt-1.82.

Semenov, K., Zouhar, V., Kocmi, T., Zhang, D., Zhou, W., and Jiang, Y. E. (2023). Findings of the WMT 2023 Shared Task on Machine Translation with Terminologies. In: P. Koehn – B. Haddow – T. Kocmi – C. Monz (eds.): Proceedings of the Eighth Conference on Machine Translation, pp. 663–671. Association for Computational Linguistics. Accessible at: https://doi.org/10.18653/v1/2023.wmt-1.54.

Temmerman, R. (2000). Towards New Ways of Terminology Description: The Sociocognitive Approach. Terminology and Lexicography Research and Practice, Vol. 3. Amsterdam/Philadelphia: John Benjamins Publishing Company.