CZU: 81`276.6`255.2:6 **https://doi.org/10.59295/drgl2026.18**

MODELLING SPECIALIZED LANGUAGE THROUGH AI: CHALLENGES IN TRAINING SYSTEMS FOR TECHNICAL DOMAIN TRANSLATION

Parascovia COZMA
MA, Lecturer
Moldova State University, Republic of Moldova
ORCID: https://orcid.org/0000-0002-7291-3138

Abstract: ***Modelling specialized language with artificial intelligence is a significant challenge for specialized translation, especially in fields where controlled terminology is essential, such as medicine, engineering, law, or the exact sciences. The performance of neural machine translation systems depends directly on the quality of the corpora used for training. However, obtaining clean, standardized, and representative corpora is difficult because access to technical documents is often limited, the structure of specialized texts varies, and terminology may be inconsistent across existing sources. As a result, unedited data can lead to semantic errors, incorrect generalizations, and conceptual confusion—issues that are particularly serious in fields where accuracy is critical.***

Another risk is the tendency of AI models to fill information gaps using statistical assumptions, which can lead to the use of inappropriate terms or to distorted conceptual relations. In this context, it becomes necessary to develop effective strategies for selecting, annotating, and cleaning corpora to ensure the terminological consistency required for training the models. This article proposes a hybrid AI–translator approach, in which technology and human expertise work together: AI increases the speed and volume of processing, while the translator validates terminology, checks conceptual coherence, and corrects semantic errors.

By analysing current challenges and presenting practical solutions, this contribution aims to show how artificial intelligence can be integrated responsibly into specialized translation, without compromising the precision and accuracy required in technical fields.

Keywords: *specialized language, AI, challenge, terminology, translation, training system*

Over the past decade, Machine Translation (MT) has undergone an impressive evolutionary journey, transitioning from statistical algorithms (SMT) to deep neural networks (NMT) and, most recently, to sophisticated Large Language Models (LLMs). In this era of global digital transformation, Artificial Intelligence (AI) has become an essential tool within the language services industry. Yet, despite the remarkable breakthroughs of systems such as GPT-4, Microsoft Translator or DeepL, the “last mile” of technical translation remains notoriously difficult to bridge.

The primary challenge emerges when moving from general-purpose language to Language for Specific Purposes (LSP). Specialized translation is far more than the mere substitution of words from one language to another; it is the precise conveyance of complex engineering, medical, or legal concepts. It requires deep modelling of the cognitive associations unique to a specific industry, being a challenge that cannot be met by simply increasing raw processing capacity.

At the heart of the issue lies a fundamental contradiction: AI, by its very nature, operates on probabilistic models and statistical inference, whereas high-stakes, knowledge-intensive domains demand absolute precision. In fields such as medicine, aerospace, or law, an error in a single term can result in catastrophic outcomes, ranging from the failure of multi-million-dollar equipment to direct threats to human life. Consequently, developing training strategies for systems that can move beyond mere probabilistic word prediction and ensure true conceptual fidelity has become a critically urgent task today.

Characteristics and Cognitive Modelling of Specialized Language

Language for Specific Purposes possesses a set of characteristics that fundamentally distinguish it from literary or colloquial language. Primarily, it is defined by a high terminological density and a rigid syntactic structure characterized by the frequency of passive constructions and impersonal sentences. In ideal conditions, LSP strives for monosemy, a state in which a single term strictly corresponds to a single concept, while its stylistic neutrality deliberately excludes metaphors and emotional shades.

In practice, however, technical language proves to be more fluid: even within a single industry, terms evolve and concepts may be replaced by synonyms over time. For artificial intelligence systems, modelling such a language entails more than mere word memorization; it requires the construction of a complex ontology and an understanding of conceptual hierarchies. The challenge lies in the fact that most contemporary models are trained on massive, general-purpose corpora (such as Common Crawl or Wikipedia). Accordingly, when deprived of a deep understanding of niche contexts, AI is prone to make inaccurate generalizations.

A striking example of this complexity within the Romanian technical lexicon is the polysemy of the term ***tensiune***. Depending on the specific domain, this single lexeme can refer to blood pressure in medicine (*tensiune arterială*), voltage in electrical engineering (*tensiune electrică*), or stress in mechanics (*tensiune mecanică*). Without clean, specialized datasets to provide context, AI systems tend to prioritize statistical probability over technical precision. This fails the model to achieve the necessary contextual depth by substituting a common term for the correct specialized one. For instance, a model might incorrectly translate *tensiune* as 'tension' or 'pressure' in a technical report where ***voltage*** is the only conceptually accurate equivalent. Thus, the lack of specialized training datasets leads to semantic errors that are unacceptable in fields with rigorous precision requirements.

As L. Bowker suggests, specialized translation is a knowledge-based activity. AI systems often fail because they lack the "world knowledge" or situational context that a human expert possesses. Current models operate on the probabilistic prediction of tokens, whereas technical domains demand deterministic accuracy. This gap between probability and certainty is the central challenge in modelling specialized language [1, 14 pp].

The Challenges of Data Imbalance and Quality

The performance of any AI translation system is fundamentally dictated by the quality and volume of its training data. In specialized domains, researchers face the "Long Tail" problem, where data for specific niches is rare or non-existent [6]. Most high-quality technical documentation: internal specifications, blueprints, or medical records, is protected by intellectual property laws or privacy protocols such as General Data Protection Regulation (GDPR) - the European Union's comprehensive data privacy law, effective since May 25, 2018, that sets strict standards for handling personal data of EU residents and Health Insurance Portability and Accountability Act of 1996 (HIPAA) - a US federal law designed to protect sensitive patient health information from disclosure without consent.

The corpus imbalance compels AI systems to operate through statistical approximation rather than semantic precision. The result is the emergence of semantic errors: instead of selecting a technically accurate term, the model defaults to the most probable lexeme found in general-purpose datasets. In scientific environments where conceptual accuracy is mandatory, such errors can lead to the misinterpretation of research data or the corruption of technical protocols.

To further illustrate the challenges of modelling specialized language, we propose the following terms in bioinformatics for the Romanian context as examples:

1. the term ***assembly (genome assembly)*** refers to the process of taking a large number of short DNA sequences and reconstructing them to form a representation of the original chromosomes, a concept translated into Romanian as ***asamblare genomică***. However, this term presents a potential risk when handled by general-purpose AI models, which are often trained extensively on industrial and IT-related data. In Romanian, *asamblare* is commonly associated with contexts such as *assembly lines* (*linie de asamblare*) or *assembly language (limbaj de asamblare*) in programming, rather than biological processes. As a result, an AI system may incorrectly translate specialized expressions like "*de novo assembly*" using terminology drawn from mechanical engineering or computer architecture, thereby obscuring the intended meaning of biological reconstruction.

2. the term ***coverage (sequencing depth)*** refers to the average number of times each nucleotide in a genome is read during sequencing, a concept translated into Romanian as ***acoperire***. However, this term poses a potential risk when used in AI-assisted translation or writing, because *acoperire* is also a common word in Romanian associated with *insurance coverage* (*acoperire de asigurare*) or *network signal coverage* (*acoperire de rețea*). As a result, in a technical context such as describing "*low coverage areas in the genome*", an AI system may misinterpret the meaning and involuntarily introduce associations with telecommunications or financial domains, producing misleading phrasing that implies a *lack of signal* or *insurance* rather than *insufficient sequencing depth*.

3. the term ***wild-type (WT)*** refers to the typical form of a species, gene, or characteristic as it naturally occurs, serving as a standard reference point distinct from mutant variants. In Romanian, this is translated as ***tip sălbatic***. However, this translation can introduce ambiguity, as the word *sălbatic* carries strong connotations of being undomesticated, aggressive, or uncultivated in everyday language. As a result, AI systems may misinterpret the scientific neutrality of the term, failing to recognize that *tip sălbatic* simply denotes the baseline, non-mutated form. This can lead to inaccurate translations, for example, rendering *wild-type population* in a way that suggests organisms are feral or violent, rather than correctly conveying that they represent the standard reference group in a biological context.

The analysis of specialized language modelling indicates that the complexities of technical translation cannot be resolved through increased processing capacity alone. As demonstrated by terminological discrepancies in Romanian bioinformatics, a fundamental challenge exists in the divergence between the statistical probabilities of general-purpose models and the rigorous requirements of specialized domains.

Morphosyntactic Ambiguity and Semantic Shift

The rich inflectional system of Romanian, marked by extensive noun declensions and verb conjugations, requires computational models to align a wide variety of inflected forms with the more analytic and relatively stable structures of English. This structural asymmetry frequently intensifies data sparsity issues, as specialized technical datasets often neglect rare morphological comprehensions. As noted by L. Sasu, these bidirectional

analyses are essential for translators and AI systems to attain message equivalence while respecting the natural constraints of the target language structure [2].

A significant obstacle for AI is the ambigeneric nature of Romanian neuter nouns, which function as masculine in the singular and feminine in the plural. This morphological transition often disrupts neural agreement mechanisms, particularly when the system has been trained on datasets that lack sufficient representation of these gender-shifting paradigms. For example, the term ***capăt*** (*end/head*) requires masculine agreement in the singular, as seen in *capăt filetat* (*threaded end*). However, in the plural, the noun shifts to feminine agreement, requiring the form *capete filetate*. AI failure typically manifests in the inability to execute this gender transition, resulting in grammatically incoherent outputs such as *capete filetați*, applying masculine plural agreement to a feminine plural noun.

English typically expresses possession and relational meaning through word order or the clitic 's, whereas Romanian relies on a system of Genitive and Dative case marking that requires consistent morphological agreement between a noun and its modifiers. This system involves the use of specific suffixes, which are often insufficiently represented in general-purpose corpora used for training AI models. For example, in the legal phrase ***drepturile persoanei autorizate*** (*The rights of the authorized person*), the adjective *autorizat* must take the feminine genitive suffix *-ei* to agree with the case-marked noun *persoanei*. AI systems may fail this requirement of inflectional agreement and instead produce the unmarked nominative form *persoanei autorizat*.

In English it can be said *the company's policy* or *the policy of the company*, where possession is indicated either by the clitic 's or by-word order. In Romanian, the same relation is expressed through case marking, as in *politica companiei*, where the noun *companie* takes the genitive form *companiei* to signal possession.

Semantic shift occurs when a term moves from a general context to a specialized one, or between two different specialized domains. A clear example is the Romanian term ***acțiune***, which represents a high-risk case of homonymy due to its multiple meanings across contexts. In everyday language, *acțiune* simply means *action*, while in the legal field it refers to a *lawsuit*, and in the financial domain it denotes a *share* or *stock*. This variation often leads to errors in AI-based translation systems, which tend to default to the general meaning. For instance, while the translation of *portfolio of shares* as *portofoliu de acțiuni* is correct, the reverse translation is frequently rendered by AI as *portfolio of actions*, a result that is clearly inappropriate and misleading in a financial context, such as an audit.

A semantic mismatch arises between the English term evidence and its Romanian equivalent ***evidență***, which are often mixed in AI-generated translations. In English technical and legal contexts, evidence denotes empirical proof supporting a claim, whereas Romanian *evidență* refers to *records* or *administrative tracking systems* (*evidența contabilă*). AI systems frequently fail to distinguish these meanings, producing translations that substitute *evidence* for *records* or vice versa. In technical and legal domains, such errors can alter the intended meaning of a document and affect its conceptual accuracy.

Differences in Romanian morphology and English structure, together with semantic ambiguity, create significant challenges for AI translation. These factors often lead to grammatical errors and misinterpretations, highlighting the need for more specialized data and linguistically informed models to ensure accuracy in technical contexts.

Modern AI Approaches to Specialized Language in Technical Translation

To overcome challenges such as lexical ambiguity and semantic errors, the field of computational linguistics is increasingly adopting hybrid approaches that combine neural and knowledge-based methods. While conventional neural machine translation (NMT)

systems perform well in general contexts, they often lack the precision required in specialized domains, where even minor terminological errors can lead to significant misinterpretations. In response, recent developments integrate the generative capacity of neural models with structured, externally sourced knowledge.

One prominent example of such an approach is *Retrieval-Augmented Generation* (RAG). RAG enables an AI system to consult a reliable, domain-specific database or glossary in real time before producing a translation [3]. In the context of translating legal terms such as *consideration* or *tort* into Romanian, RAG is particularly valuable because it ensures that the system uses standardized and authoritative terminology, such as that found in the database of the European Court of Justice. This reduces inconsistency and approximation by grounding output in officially recognized legal language rather than general linguistic patterns. Consequently, RAG not only improves accuracy but also enhances the reliability and credibility of AI-assisted legal translation.

Fine-tuning with *Low-Rank Adaptation* (LoRA), as another example of modern AI approaches, provides an efficient method for adapting pre-trained large language models to specialized domains without retraining them entirely [5]. This approach uses a relatively small, high-quality corpus to guide the model toward domain-specific usage. In banking, for example, LoRA can be applied by training on curated Romanian research texts, enabling the model to correctly interpret and generate specialized terminology such as *capital* in the sense of *financial assets* or *equity*, rather than its general meaning as a city or uppercase lettering.

Last but not the least, the Hybrid *Human-in-the-Loop* (HITL) model represents a highly effective strategy that integrates human expertise into AI-driven workflows [2]. In this approach, the AI system performs initial processing tasks, such as syntactic structuring and draft generation, thereby reducing the translator's cognitive load. Human experts then act as post-editors and data curators, refining outputs and feeding corrections back into the system through iterative learning processes. This combination of automated efficiency and human oversight enhances accuracy, ensures terminological consistency, and maintains the conceptual and technical integrity of the translated text.

Conclusion

The modelling of specialized language through artificial intelligence remains a significant challenge due to the innate tension between the probabilistic foundations of AI systems and the requirement for deterministic precision in technical domains. As demonstrated throughout this study, issues such as data scarcity, terminological inconsistency, morphosyntactic complexity, and semantic ambiguity significantly affect the performance of AI-driven translation systems. These limitations are particularly evident in languages like Romanian, where rich inflectional structures and domain-specific polysemy demand a level of contextual and conceptual understanding that current general-purpose models cannot consistently achieve.

At the same time, the analysis highlights that these challenges are not impossible but rather require a shift in methodological perspective. The integration of hybrid approaches, including Retrieval-Augmented Generation, fine-tuning techniques such as Low-Rank Adaptation, and Human-in-the-Loop frameworks, represents a promising direction for improving the accuracy and reliability of AI in specialized translation. By combining the computational efficiency of neural models with curated knowledge sources and human expertise, these approaches address both the linguistic and conceptual dimensions of Language for Specific Purposes, ensuring greater terminological precision and contextual adequacy.

Ultimately, the future of AI in specialized translation lies not in replacing human expertise but in augmenting it. A collaborative paradigm, in which artificial intelligence supports and enhances the work of domain specialists, offers the most effective path forward. Such an approach enables the responsible integration of AI technologies into professional translation practices, ensuring that innovation does not come at the expense of accuracy, reliability, or conceptual integrity in technical fields.

References:

1. BOWKER, L., & BUITRAGO CIRO, J. Machine translation and global research: Towards improved machine translation literacy in the scholarly community. Bingley, UK: Emerald Publishing, 2019. 128 pp. ISBN 9781787567221
2. SASU, L. Challenges in Romanian/English Translation of Technical Terminology. Research and Science Today, Vol. 2 (16), 2018, 166-175pp. ISSN-P: 2247-4455 / ISSN-E: 2285-9632 https://www.rstjournal.com/article/3253aba3-a0ad-4a53-ad04-c6aad67073db (accesat la 30.03.2026)
3. https://www.coursera.org/articles/what-is-rag-in-ai (accesat la 01.04.2026)
4. https://www.ibm.com/think/topics/human-in-the-loop (accesat la 01.04.2026)
5. https://www.ibm.com/think/topics/lora (accesat la 01.04.2026)
6. https://portable.io/learn/long-tail-data (accesat la 28.03.2026)