LANGUAGE, DATA and KNOWLEDGE 2025

Proceedings of the 5th Conference on **Language, Data and Knowledge: Workshops**

UniorPress
Naples
2025

PROCEEDINGS OF THE 5TH CONFERENCE ON LANGUAGE, DATA AND KNOWLEDGE: WORKSHOPS

EDITORS:
Katerina Gkirtzou, ILSP “Athena” Research Center, Greece
Slavko Žitnik, University of Ljubljana, Slovenia
Jorge Gracia, University of Zaragoza, Spain
Dagmar Gromann, University of Vienna, Austria
Maria Pia di Buono, University of Naples “L’Orientale”, Italy
Johanna Monti, University of Naples “L’Orientale”, Italy
Maxim Ionov, University of Zaragoza, Spain

ISBN 978-88-6719-334-9

UniorPress — University of Naples “L’Orientale”
Via Nuova Marina 59 — 80133 Napoli (Italy)

UniorPress

Table of Contents

TermTrends 2025: Bridging the Gap between Terminological Resources and Large Language Models

TermTrends 2025

TermTrends 2025: Bridging the Gap between Terminological Resources and Large Language Models

The Trends in Terminology Generation and Modelling (TermTrends) workshop series started 3 years ago as a tutorial co-located with the 23rd International Conference on Knowledge Engineering and Knowledge Management (EKAW) 2022. This tutorial covered standardisation approaches for representing terminology, explored methods to accelerate terminology generation, demonstrated domain-specific use cases, and included a hands-on session with tools for terminology extraction, enrichment, and representation.

The concept of the tutorial quickly evolved into a workshop format, featuring a different topic each year based on the emerging trends and priorities within the field. Therefore, TermTrends 2023 embraced the Terminology in the Era of Linguistic Data Science, and it was co-located with the 4th Conference on Language, Data and Knowledge (LDK). In that edition, the papers presented addressed diverse challenges and innovations in multilingual and domain-specific terminology work, including the complexity of representing multilingual specialised knowledge, the development of terminology resources using OntoLex-Lemon, the theoretical and practical considerations of converting TBX to OntoLex-Lemon to ensure interoperability, and the formalisation of translation equivalence and lexicalsemantic relations in bilingual terminological resources, amongst other topics.

Last year, the workshop focused on Models and Best Practices for Terminology Representation in the Semantic Web, and it was co-located with the 3rd International Conference on Multilingual Digital Terminology Today. Design, representation formats and management systems (MDTT 2024). The contributions explored the intersection of lexical and domain knowledge, with a special focus on terminology representation and interoperability, including the linking of resources in the cybersecurity domain, the development of a three-layer multilingual terminology for oncological diseases, the representation of the TERMCAT glossaries in OntoLex-Lemon and the integration of the Lexical Markup Framework with the Terminological Markup Framework.

This year's edition of the workshop focused on a topic that has been rapidly adopted by nearly every sector of society and holds significant implications for language processing, knowledge representation, and artificial intelligence: Large Language Models. Therefore, the workshop is focused on Bridging the Gap between Terminological Resources and Large Language Models. In particular, this event intends to explore how terminological resources can be leveraged to enhance the performance of LLMs at different stages of their lifecycle, from pre-training and fine-tuning to evaluation and application. By identifying points of convergence between Linked Data-based terminological data and LLM processes, we seek to identify strategies to make these resources more interoperable, reusable, and impactful.

Specifically, TermTrends 2025 presents contributions on the performance of instruction-tuned Large Language Models in handling terminology-constrained translation for low-resource language varieties, the ISO/TC 37/WG 6 initiative that aims to integrate AI into standardised terminology management practices, and the capabilities of LLMs in inferring semantic relations between terms to enhance terminological resources.

Additionally, this year's edition of the workshop introduced a notable novelty: the inclusion of two keynote speakers, Giorgio Maria di Nunzio, Associate Professor at the University of Padova, and José Manuel Gómez Pérez, Director of Language Technology Research at expert.ai. Having such distinguished representatives from academia and industry, respectively, will provide comprehensive and contrasting insights into the application of terminological resources and LLMs, highlighting both theoretical advancements and practical implementations across different sectors.

With the papers presented and the insightful contributions from both academic and industry representatives, we have high hopes for this year's edition. TermTrends 2025 aims to bring together researchers and practitioners to explore practical ways of connecting terminological resources with Large Language Models, encouraging collaboration and adoption of these technologies. We look forward to productive discussions and new opportunities to advance the field together.

Patricia Martín Chozas, Elena Montiel Ponsoda, Sara Carvalho and Federica Vezzani
TermTrends 2025 Organising Committee

Organizing Committee

Patricia Martín Chozas, Universidad Politécnica de Madrid, Spain
Elena Montiel Ponsoda, Universidad Politécnica de Madrid, Spain
Sara Carvalho, University of Aveiro, Portugal
Federica Vezzani, University of Padova, Italy

Program Committee

Ana Ostroški Anić, Institute for Croatian Language and Linguistics, Croatia
Anas Fahad Khan, Istituto di Linguistica Computazionale, Italy
Antonio San Martín, University of Quebec, Canada
Beatriz Guerrero, Universidad de Salamanca, Spain
Bruno Almeida, NOVA FCSH / NOVA CLUNL, Portugal
David Lindemann, University of the Basque Country, Spain
Elena Montiel-Ponsoda, Universidad Politécnica de Madrid, Spain
Federica Vezzani, Università di Padova, Italy
Francesca Frontini, Istituto di Linguistica Computazionale, Italy
Giorgio Di Nunzio, Università di Padova, Italy
Giulia Speranza, Università degli Studi di Napoli L'Orientale, Italy
John McCrae, National University of Ireland, Ireland
Jorge Gracia, Universidad de Zaragoza, Spain
Laurent Romary, INRIA, France
Manuel Fiorelli, Tor Vergata University of Rome, Italy
Maria Pia di Buono, Università degli Studi di Napoli L'Orientale, Italy
Miguel Sánchez Ibáñez, Universidad de Valladolid
Natascia Ralli, Eurac Research, Italy
Nava Maroto, Universidad Politécnica de Madrid, Spain
Pamela Faber, University of Granada, Spain
Patricia Martín Chozas, Universidad Politécnica de Madrid, Spain
Patrick Drouin, Montreal University, Canada
Paul Buitelaar, National University of Ireland, Ireland
Penny Labropoulou, Institute of Language and Speech, Greece
Pilar León Arauz, Universidad de Granada, Spain
Sara Carvalho, Universidade de Aveiro / NOVA CLUNL, Portugal
Sigita Rackevičienė, Mykolas Romeris University, Lithuania
Silvia Piccini, Istituto di Linguistica Computazionale, Italy
Špela Vintar, University of Ljubljana, Slovenia

The *LegISTyr* Test Set: Investigating Off-the-Shelf Instruction-Tuned LLMs for Terminology-Constrained Translation in a Low-Resource Language Variety

Paolo Di Natale[1] **Egon W. Stemle**[1,2] **Elena Chiocchetti**[1] **Marlies Alber**[1]
Natascia Ralli[1] **Isabella Stanizzi**[1] **Elena Benini**[1,3]

[1]Eurac Research, Bolzano/Bozen, Italy [2]Masaryk University, Brno, Czech Republic
[3]University of Bologna, Bologna, Italy

{paolo.dinatale, egon.stemle, elena.chiocchetti, marlies.alber,
natascia.ralli, isabella.stanizzi}@eurac.edu
elena.benini99@gmail.com

Abstract

We investigate the effect of terminology injection for terminology-constrained translation in a low-resource language variety, with a particular focus on off-the-shelf instruction-tuned Large Language Models (LLMs). We compare a total of 9 models: 4 instruction-tuned LLMs from the Tower and EuroLLM suites, which have been specifically trained for translation-related tasks; 2 generic open-weight LLMs (LLaMA-8B and Mistral-7B); 3 Neural Machine Translation (NMT) systems (an adapted version of MarianMT and ModernMT with and without the glossary function). To this end, we release *LegISTyr*, a manually curated test set of 2,000 Italian sentences from the legal domain, paired with source Italian terms and target terms in the South Tyrolean standard variety of German. We select only real-world sources and design constraints on length, syntactic clarity, and referential coherence to ensure high quality. *LegISTyr* includes a homonym subset, which challenges systems on the selection of the correct homonym where sense disambiguation is deducible from the context. Results show that while generic LLMs achieve the highest raw term insertion rates (approximately 64%), translation-specialized LLMs deliver superior fluency (Δ COMET up to 0.04), reduce incorrect homonym selection by half, and generate more controllable output. We posit that models trained on translation-related data are better able to focus on source-side information, producing more coherent translations.

1 Introduction

While Neural Machine Translation (NMT) adaptation has demonstrated benefits from incorporating domain-specific terms (Farajian et al. 2018), it has yet to ensure consistent and unambiguous terminology enforcement. The delicate trade-off between the continuous expansion of parallel training corpora and the enforcement of lexical choices without hampering fluency further complicates cross-lingual alignment of terminologically relevant tokens (Alkhouli et al. 2018; Ferrando et al. 2022; Štefánik et al. 2023). This failure has even raised questions on the cost-effectiveness of maintaining termbases for MT purposes (Knowles et al. 2023). Terminology compliance has attracted notable interest and lead to the organization of shared MT tasks (Bawden et al. 2019; Alam et al. 2021b; Semenov et al. 2023).

Terminology compliance is a crucial quality aspect in high-stakes domains, such as the legal domain. Terminology mistakes in legal translation can have serious consequences (Mattila 2018), including legal disputes and infringement of basic linguistic and human rights (e.g., through incorrect use of critical terminology during interpretation in criminal trials). Furthermore, every legal system has its own specific set of rules and conceptual structures. Legal terminology expresses such specificities and is therefore always bound to a specific legal system (Gambaro and Sacco 2024). This system-boundness of legal terminology often results in conceptual incongruency between legal systems, even between legal systems sharing the same language. Consequently, correct terminology usage in languages with more than one recognized legal variety like Arabic, English, Spanish etc., pose notable translation challenges. In addition, the quality of MT in the legal domain hinges not so much on the language combination as on the legal subdomain (Quinci and Pontrandolfo 2023). Our experiments focus on South Tyrolean German, a minor standard variety of German spoken in Northern Italy that is used by the local public administration bodies and on legal terms from a range of different legal subdomains.

Researchers have addressed terminology enforcement using both NMT systems and Large Language Models (LLMs). In 2024, two new LLM suites specifically trained on translation-related tasks (TT LLMs) were released to the public (Tower and EuroLLM suites); yet to our knowledge their capabilities have not been adequately investigated thus far. To bridge the gap, we evaluate the instruction-following performances of these open models in terminology injection, comparing their term accuracy rate and overall translation quality against general-purpose LLMs as well as to adapted NMT models and their baseline. We also assess which models produce the most structured and clean outputs for easier downstream processing.

To this end, we curate *LegISTyr*[1], a test set comprising over 2,000 sentences from the legal domain in Italian. Each instance is annotated with both Italian source and South Tyrolean target terms, plus variants used in other German-speaking legal systems whenever available. We select the sentences from termbase contexts and other real-world sources while enforcing constraints on length, syntactic clarity, and referential coherence. The test set covers the usage of each term in multiple contexts. It also features a subset on the challenge of disambiguation between homonyms where the correct sense is deducible from the domain or context of use.

2 Background

2.1 Challenges of legal translation in South Tyrol

While the task of terminology injection presents an already serious challenge for highly resourced languages, difficulties grow when considering minor varieties of pluricentric languages (Clyne 1991). Such varieties often lack labeled datasets and resources in internationally standardized format (Lakew et al. 2020), and are under-represented in generic training corpora due to the sheer size of text produced by the respective speakers (Zampieri et al. 2020). This leads to dominant forms obfuscating diatopic variants (Koehn and Knowles 2017). The implication in statistical MT is that the stronger signal from numerically dominant varieties will tend to override terms from minority ones.

The situation of a minor variety used in the legal domain can be exemplified by South Tyrolean German. This is a standard variety of German (Ammon et al. 2016) used by about 300,000 members of the German language minority in Italy. In South Tyrol, German is a co-official language next to Italian. South Tyrolean German differs from other German varieties in various minor grammatical and lexical aspects, but most notably concerning food and legal vocabulary. Some of these differences are unlikely to affect terminology (e.g. the use of *sein* vs *haben* as an auxiliary for some verbs of position), others might (e.g. differences in grammatical gender like *Kataster*, cadastre, generally being masculine in South Tyrol and other Southern German areas but neuter in Germany).

In South Tyrol, there is a process of terminology standardization in place whereby the mandatory terminology for the local legal and administrative texts is being validated by a Terminology Commission (Chiocchetti 2021). The infringement of local terminology requirements may result in hindered clarity of government-citizen communications and in the denial of linguistic minority rights related to clear and consistent communication in the minority language (South Tyrolean German).

An analysis of machine-translated normative texts from Italian into South Tyrolean German has highlighted that terminology is the second most common type of mistake after mistranslations (De Camillis and Chiocchetti 2024). This can happen by effect of interference from most represented legal systems.

We also observe that Italian multiword terms (e.g., *decreto legislativo*, legislative decree; *decreto ingiuntivo*, payment order; *decreto di condanna*, penalty order, etc.) tend to be shortened to their headword within texts (i.e., to *decreto*) creating homonymous short forms. Context is needed for correct disambiguation when translating into German (i.e., choosing between *gesetzesvertretendes Dekret*, legislative decree; *Mahndekret*, payment order and *Strafbefehl*, penalty order), especially since there is a less pronounced tendency to reduce multiword terms or compounds to their headword in German. The notable presence of homonyms deriving from ellipsis of part of the term in Italian legal texts poses an issue of (domain) disambiguation in the correct selec-

[1] http://hdl.handle.net/20.500.12124/104

tion of the target term in German. Disambiguation is relevant even in cases where there is no shortening of a longer term (e.g. *procedura concorsuale* means 'bankruptcy proceedings' in insolvency law, in German *Insolvenzverfahren*, but 'open competitive employment procedure' in administrative law and should be translated with *Wettbewerbsverfahren*).

To evaluate translation quality into South Tyrolean German, we release *LegISTyr*, a highly curated test set (see 4.1). We use the test set to explore the success of different techniques aimed at terminology injection when translating from Italian into a minor standard variety of German, viz. South Tyrolean German.

2.2 Aligning LLMs to translation tasks

While LLMs have shown the capability to perform targeted translation tasks without task-specific training (Vilar et al. 2023; Hendy et al. 2023), top-notch performances are still bottlenecked to proprietary, large-scale models, which makes their adoption in low-resource environments excessively expensive. Another pitfall of generic LLMs is their tendency to exhibit undesired behavior, such as verbosity — the generation of explanatory text in addition to the proper translation (Briakou et al. 2024).

One line of research has sought to imbue text with the core features of translation tasks through in-context learning (ICL), prompting a generic model with exemplars in a predefined format (Brown et al. 2020). Despite being largely resource-efficient and effective in cross-lingual transfer for under-resourced languages (Zhu et al. 2024b; Zhu et al. 2024a), drawbacks still persist. In actuality, prompt design seems to be a hardly manageable strategy, as minimal permutations result in heavy performance volatility (Leidinger et al. 2023; Sclar et al. 2024; Weber et al. 2023), unreasonable templates give rise to acceptable outputs (Zhu et al. 2024b) and targeted exemplar selection for the prompt only has a limited impact (Zhang et al. 2023). In addition, ICL proves ever more effective as the scale of the model in use grows (Min et al. 2022). Overcoming these limitations might be possible by leveraging pre-trained, lightweight models specialized in translation-related tasks in order to attain comparable performances to larger models (Zeng et al. 2024).

Instruction tuning (IT) (Wei et al. 2022) consists of fine-tuning an existing model on instruction-output pairs, aligning next-token prediction with task-specific objectives to enhance adherence to user instructions (Zhang et al. 2024). As a matter of fact, (Zhou et al., 2023) (2023) highlight that a model's core knowledge is acquired during pre-training, while fine-tuning mainly influences interaction modalities and output format. This effect is further amplified when LLMs are pre-trained on multilingual corpora (Briakou et al. 2023), where alignment with downstream translation objectives begins at the initial training stage (Sia et al. 2024).

2.3 Terminology injection approaches

Researchers have explored diverse techniques to inject terminology and ensure exact term rendering.

In NMT models, relevant terminology can be marked at inference time by inserting inline term labels in the source, target, or both. This can be achieved either by replacing all terms in source and target with a placeholder (Dinu et al. 2019; Bergmanis and Pinnis 2021; Michon et al. 2020), which sacrifices semantic information, or by adding the target term to the source sentence (code-switching) (Song et al. 2019; Ailem et al. 2021). Variations of this approach include additional lemmatization or grammatical editing for increased fluency (Bergmanis and Pinnis 2021; Pham et al. 2021) as well as changes in the location of the label (Jon et al. 2021; Turcan et al. 2022). However, consistently predicting the term equivalent during training has proven challenging, particularly under more complex circumstances than those in experiment conditions (e.g., with more than one target term per sentence). This has resulted in the introduction of hard constraints – i.e. forced insertion of the term in the target sentence – that are in turn affected by fluency and grammatical issues (Post et al. 2019; Chen et al. 2020).

Other attempts directly act on the decoder behaviour. The decoder is the component that generates the target translation by sequentially predicting the next token, given the encoded representation of the source text and the incrementally generated output. One focus lies on the development of decoding algorithms which enforce the desired term (Molchanov et al. 2021; Hauhio and Friberg 2024) or exclude unsuitable terms from the search space (Bogoychev and Chen 2023). Despite their potential, these methods increase computational time and resource demands. Additionally, they fail to address issues such as incorrect morphological inflections and unintentional repetition of terminology (Dinu et al. 2019).

With the surge in the use of Large Language Models, prompt formulation has made it possible to exert greater control over features of the output, including "specific dialect" (Garcia and Firat 2022) and terminology injection. Initial experiments provided prompts augmented with glossary (Moslem et al. 2023a; Moslem et al. 2023b) and dictionary (Ghazvininejad et al. 2023) entries, also following an instruction tuning (Kim et al. 2024), where the retrieved target terms are injected in the prompt. Another strategy relies on post-editing existing translations (Bogoychev and Chen 2023; Chen et al. 2024; Liu et al. 2025, Sabo et al. 2024), which can be included within the wider concept of translation refinement (Feng et al. 2024; Koneru et al. 2024; Xu et al. 2024). This technique employs iterative prompting to adjust the generated translation until the terminological constraint is complied with.

3 Experimental target and limitations

In light of these insights, we evaluate off-the-shelf models that have undergone both pre-training (EuroLLM) and supervised fine-tuning (Tower) on translation tasks without additional adaptation to custom data. However, by deliberately excluding any modification to the model's hidden states representation, we restrict our investigation to the effect of terminology injection — particularly for terms likely under-represented during pre-training — on output fluency and decoding behavior under soft constraint conditions. This design choice may limit the performance upper bound for producing fully coherent South Tyrolean text, as some of the most effective approaches for handling language varieties often involve large-scale pre-training or continued training (Tejaswi et al. 2024; Nag et al. 2024), for machine translation (Kumar et al. 2021; Sousa et al. 2025) and evaluation tasks (Sun et al. 2023; Aepli et al. 2023) alike.

4 Methodology

4.1 Dataset curation — The *LegISTyr* Dataset

General principles

To make a fully comprehensive assessment of the models' term recognition, we impose stylistic and textual criteria in the collection of the test set sentences. We choose exemplars with a minimum of 8 and a maximum of 50 words, ignoring titles, truncated excerpts, captions, contents of tables and indexes, and bullet lists. All sentences are copied or adapted (e.g., shortened) from existing sources, including the contexts from *bistro*. The examples showcase the term of interest in different positions of the sentence with the possible variations in number but not in gender, as this would require a gender-specific equivalent in German and add a further layer of complexity. Subject and object are well defined, added manually when they are implicit, which is a common feature in Italian. Unresolvable co-reference relationships or ambiguous anaphoric references may appear in some exemplars but never affect the term. Parenthetical statements within brackets or dashes have been removed while maintaining the typical style of Italian legal language, which tends to use appositions and parenthetical material between commas.

Terms can be simple terms (e.g., *cittadinanza*, citizenship; *frode*, fraud) or complex terms (e.g., *decreto ministeriale*, Ministerial Decree; *capacità di intendere e di volere*, full possession of mental faculties). Most are nouns or noun phrases, with the exception of one collocation (*d'ufficio*, ex officio), which has a standardized German translation in South Tyrol. Almost all selected terms are available in *bistro* with their South Tyrolean variants and any terms used in other German-speaking legal systems.

The content of *LegISTyr* is largely based on the terminological data contained in the Information System for Legal Terminology *bistro*[2] (Ralli and Andreatta 2018). The latter collects the main legal concepts of the Italian legal system with their designations in Italian and in South Tyrolean German for use at the regional level, together with existing German language designations for any equivalent concepts from other legal systems that use German as an official language (i.e., Austria, Switzerland and Germany). *bistro* publishes over 13,000 fully-fledged term entries pertaining to a wide range of legal subdomains, for several legal systems (Italy, Austria, Germany, Switzerland, EU law, international law) and three languages (Italian, German, Ladin). However, many contexts in *bistro* are defining contexts, which are extremely useful to complement conceptual information given in the

[2] https://bistro.eurac.edu/

definitions but not always ideal to showcase a term in its most frequent usages. In addition, we wanted more than just one example (i.e., context) for each selected term. We therefore complemented *bistro* data with examples from legal texts and websites. The test set also contains data from subdomains that haven't been fully published in *bistro* to date (e.g. subsidised housing).

Dataset structure The dataset is divided into different sections:

- Standardized terminology: 250 sentences covering different legal subdomains (partly including but not limited to the subdomains in the other subsets), with 5 instances for each term with a mandatory, standardized equivalent in South Tyrolean German.
- Main terminology: 1,000 sentences from 4 different legal subdomains (criminal and criminal procedure law, family law, subsidized housing, occupational health and safety) with 5 instances for each term.
- Homonyms: 250 sentences with 5 term instances for each of the two or three sub-domains where the term is used.

Each entry comprises a source sentence, a source term and a target term. Alongside the main South Tyrolean term, other variants expressing the same concept in South Tyrol or in other German-speaking legal systems are specified — when existent. This lets us measure to what extent the interference from more represented legal systems impacts term translation. In Appendix A, the reader can find an illustrative example of a test set entry. Each entry is designed to evaluate the insertion of its corresponding designated target term alone; occurrences of other terms from the test set within the same sentence are disregarded for evaluation.

Additionally, the dataset contains a subset for abbreviated forms (i.e., acronyms, initialisms, and abbrevations), a frequent source of mistakes in legal translation. There is also a subset with strategies for gender-inclusive writing (e.g. gender-neutral agentives, split forms, terms with symbols and neomorphemes) because local legislation demands that legal and administrative texts be possibly inclusive. Since these two sections have not been used for this paper, we do not give further details.

Consistency We recognize that a key requirement for efficient terminology management is that term rendering be consistent throughout the entire document (Semenov and Bojar 2022). While our evaluation is conducted at the segment level, we approximate terminology consistency by designing a test set that includes five same-domain usage contexts for each term. This choice allows to evaluate over longer stretches of running text and simulate (insofar as possible) multiple occurrences in a text. It also helps assess the behavior of the model in the presence of domain-relevant context information systematically.

Homonyms We analyze the homonyms subset to monitor the circumstances when the notion of unambiguity (i.e., avoiding one-to-many translations) proposed in similar works (Bogoychev and Chen 2023) has to be dismissed. While the statement that source terms should be associated with one correspondent only is generally valid, we put the lens on homonymy to measure performances on a linguistic phenomenon that has proven extremely challenging to address. In these cases, the variability on the target side of a single surface-form source term is crucial to refer to the correct concept.

4.2 Model selection

4.2.1 Generic LLMs

Llama we use the **Llama-8B**-Instruct model, part of Meta AI's suite of open-weight transformer-based language models (Touvron et al. 2023). The 8B variant offers a balance between model capacity and computational efficiency. It has also been chosen because it is one of the base models onto which Tower suite models have been fine-tuned.

Mistral An open-weight model, **Mistral-7B**-Instruct is another decoder-only transformer architecture comparable to Llama in size. It achieves efficient inference thanks to sliding window attention and grouped-query attention (Jiang et al. 2023). It also serves as one of the base models for one of the Tower-suite fine-tuned models.

4.2.2 Translation-tuned LLMs

Tower (Alves et al. 2024) is a suite of multilingual LLMs fine-tuned to translation-related tasks. We test the two available configurations: **Tower-7B**-Instruct and **Tower-13B**-Instruct. The development of TOWER involves a two-stage process. Initially, the base model, TOWERBASE, undergoes continued pretraining upon the LLaMA-2 architecture (Touvron et al. 2023) on a 20-billion-token dataset comprising both monolingual and parallel data. Subsequently, the model is fine-tuned using a curated dataset, TOWERBLOCKS, which specializes the LLM for translation-related tasks. We also test an updated version of the model named **Tower-Mistral-7B**-Instruct (Rei et al. 2024) and based on Mistral-7b (Jiang et al. 2023).

EuroLLM (Martins et al. 2024) is a suite of open-weight multilingual language models trained from scratch and featuring all official EU languages. The models are trained on a filtered corpus of assorted web data, code, parallel corpora, and domain-specific texts. A byte-pair encoding (BPE) tokenizer is created to handle linguistic diversity and efficient subword segmentation. Pre-training is conducted with mixed multilingual objectives, followed by instruction tuning to enhance zero-shot and few-shot task performance. We test the configuration **EuroLLM-9B**-Instruct.

4.2.3 Neural systems

ModernMT is accessed via its adaptive API for enterprises with a licensed account. We use the unidirectional glossary function uploading all term pairs gathered in the test set (indicated as **ModernMT-glossary** in Table 1) and compare it with its baseline performances (**ModernMT-baseline**).

MarianMT We fine-tune the Italian to German version of the Opus-MT model[3], using the MarianMT architecture (Junczys-Dowmunt et al. 2018) through Hugging Face's Transformers library. The model was trained on an in-house parallel corpus including LEXB (Contarino 2021), MT@BZ (De Camillis et al. 2023), CATEX (Gamper 1999) and other internal translation memories — for a total of 223,716 training instances. As parameters, we use a batch size of 64, 15 epochs, a learning rate of 3e-4, and 5,000 warm-up steps. Mixed-precision is used. In Table 1, it is indicated as **MARIANMT-adapted**. We do not report the results of the baseline model as its inadequacy for the South Tyrolean variety has already been exposed in Oliver et al. 2024.

4.3 Experimental setup

We interface with the models via the vLLM inference framework[4]. Working with instruction-tuned models, we utilize the *chat* method for text generation[5] to make the best of their instruction-following capabilities. The vLLM framework automatically retrieves and applies the model's predefined chat template when processing chat-formatted inputs.

Appendix B contains the prompt structure. In the system message, we explain the task together with the expected features of the output (language variety and terminology awareness). The user message consists solely of the source sentence, enclosed within <> symbols, following the approach of Zhang et al. 2024 and Cettolo et al. 2024. This implicitly signals that the translation should also be enclosed within these delimiters, facilitating the exclusion of extraneous commentary. Because Tower Suite models do not rely on system messages for instructions due to their structured text generation pipeline, we provide the instruction and the appended source sentence in the user message for this class of models only. For the homonym set, all possible homonym translations are provided as options, without suggesting the correct one.

As custom inference parameters, we set top-p to 0.95 and temperature to 0.2. Low temperature is meant to limit output variability and ensure high-confidence text coherence in a domain that is rich in formulaic expressions and sensitive to meaning corruption, while a relatively high top-p setting allows potentially under-represented terms to remain into the sampling space.

[3] https://huggingface.co/Helsinki-NLP/opus-mt-de-it

[4] https://docs.vllm.ai/en/latest/

[5] https://docs.vllm.ai/en/latest/models/generative_models.html

4.4 Evaluation

Existing methods for measuring the enforcement of terminology into the output vary depending on the structure of available termbase resources, though always relying on some form of exact-match algorithm. While we acknowledge the limitations of a naive matching approach, the lack of a reference translation prevents us from implementing the solutions suggested by Alam et al. 2021a.

Hence, we define our evaluation criteria as follows:

- **Accuracy**: The frequency with which the target term appears in the output. This metric assesses whether the system has incorporated the instructed term in any form.

- **Fluency**: The estimated extent to which the term is used in a coherent, well-formed sentence. This criterion evaluates the overall linguistic fluency of the machine translation output.

To measure **accuracy**, we design a custom pipeline with two pre-processing steps. Given a target term (TT) and a target sentence (TS), we first apply lemmatization to all tokens in both groups. We then use the SpaCy PhraseMatcher[6] — which allows to match terminology lists from provided text — to determine whether TT appears in TS, returning a positive match if found. However, due to the high density and complexity of inflected forms in German, we observe the lemmatizer may occasionally struggle to identify the uniform lemma of inflected variants of complex terms across TT and TS. Therefore, we apply the *CharSplit* tool from Tuggener, 2016 to decompose all tokens into their most probable subcomponents. Lemmatization is then re-applied to these decomposed units, and the matching procedure from the first step is repeated. We find this additional step particularly beneficial to detect the cases of internal inflection within complex nouns. Subsequently, we adopt the same procedure for: alternative acceptable terms belonging to the South-Tyrolean variety, variants used in other German-speaking legal systems, and incorrect homonyms.

To assess **fluency**, we compute the COMET-Kiwi-XL score (Rei et al., 2023), a reference-less automatic machine translation quality estimation metric. This allows us to evaluate the overall quality of the translated segment beyond mere terminology injection. Following the holistic approach proposed by Alam et al. 2021a, this evaluation ensures that the imposed terminological constraints do not compromise the meaning of the generated sentence.

5 Results

As it can be appreciated in Table 1, generic LLMs (LLama and Mistral) achieve the highest term success rate, outperforming other model paradigms by at least 10 percentage points. However, this should not be naively accepted at face value as superior overall suitability. The well-documented tension between terminological accuracy and output fluency remains tangible. As the Comet-Kiwi-XL scores highlight, TT LLMs achieve higher levels of fluency at the system level. This divergence may suggest that generic LLMs fall into the patterns of hard-constrained decoding techniques, where the insertion of low-probability tokens causes a cascading degradation of output fluency by redistributing a lower probability mass across the whole output. In contrast, TT LLMs tend to insert the desired term only when its inclusion aligns with a high-probability token prediction, as determined primarily by the model's pre-trained hidden state representations, rather than external prompt conditioning.

In addition, the magnitude of the COMET score differential is non-trivial. According to Kocmi et al. (2024)'s tool[7], a 2-point delta in COMET-Kiwi corresponds heuristically to a 95% agreement rate with expert human judgments. This result reinforces the hypothesis that TT LLMs prioritize fluent output more effectively than other model paradigms.

Homonym set results point to this hypothesis too. The difficulty of detecting the correct domain lowers overall term enforcement rates, with the most pronounced decline observed in generic LLMs, hence narrowing their performance lead. Notably, generic LLMs exhibit a strong proneness to inserting any of the provided terms, even when contextual cues suggest otherwise. This results in more than twice as many incorrect homonym selections compared to TT LLMs — with critical consequences on meaning comprehension.

It could be argued that this divergence may be

[6] https://spacy.io/api/phrasematcher

[7] https://kocmitom.github.io/MT-Thresholds/

	Accuracy					Fluency
	Main + Standardized Terminology			Homonyms		
Models	Term Success Rate	Other South Tyrol	Other Legal System	Correct Homonym	Wrong Homonym	Comet Score
LLAMA 8B	**64.5**	1.68	3.10	**44.8**	21.2	0.6317
MISTRAL 7B	64.0	1.36	**2.55**	41.6	25.2	0.5983
TOWER 7B	39.2	1.84	10.0	37.2	10.4	0.6264
TOWER-MISTRAL 7B	43.6	2.40	7.64	34.0	11.6	0.6395
TOWER 13B	52.0	1.84	6.0	36.4	12.4	0.6534
EUROLLM 9B	46.5	3.28	8.72	36.8	**7.6**	**0.6717**
MMT-baseline	24.1	**5.84**	12.7	28.8	15.2	0.6693
MMT-glossary	51.3	1.70	4.80	32.4	26.4	0.6343
MARIANMT-adapted	49.5	4.53	3.7	34.8	14.0	0.5757

Table 1: Evaluation results according to accuracy and fluency criteria. The **Main + Standardized Terminology** section reports performance on the Standardized Terminology and Main Terminology subsets of the test set, as described in Section 4.1. *Term Success Rate* indicates the percentage of cases in which the exact instructed term was correctly injected into the translation output. *Other South Tyrol* denotes the proportion of translations — out of the sheer total — using an acceptable South Tyrolean variant instead of the specified term. *Other Legal System* reflects the relative frequency — computed only on applicable cases — of a term from another major legal variety of German being used in place of the target South Tyrolean form. The **Homonyms** section reports results on the homonym subset. The *Correct Homonym* column shows the percentage of cases the correct homonym has been inserted, while the *Wrong Homonym* column highlights the percentage of cases the erroneous homonym option has been used in place of the correct one. Finally, under **Fluency**, the *Comet Score* column reports the system-level evaluation scores performed with Comet-Kiwi-XL.

attributable to the different objectives of the respective tuning procedures. Specific training of TT LLMs on translation-specific tasks enables the models to learn source-target text alignment patterns, allowing its attention mechanisms to more effectively focus on source-side information. The higher learned attention to highly domain-relevant context in the source sentence reduces the likelihood of incorrect term selection. In contrast, instruction-tuning in generic LLMs often lacks explicit exposure to the parallel sentence structure typical of translation tasks. Therefore, contextual attention weights may end up excessively biased towards less important parts of the prompt.

Additional insights emerge from the generation of terms belonging to other major varieties of legal German (OLS) or other valid alternatives for South Tyrol (OST), in the cases where these were measurable. The higher selection rate of OLS variants by TT LLMs may signal that prompting embeddings cannot alone redress the token probability imbalance in the output distribution, which is presumably outmatched by major variant occurrences. However, we note that EuroLLM-9B achieves the highest OST selection rate among LLMs, despite these acceptable terms not being explicitly elicited in the prompt. Given its pre-training from scratch on European language corpora, this result suggests that including regionally-focused pre-training data may allow to better encode minor variant lexical forms.

Format-wise, Tower and EuroLLM have proven to be most stable models, with all translations enclosed in the desired delimiters and no signs of verbosity. While LLama 8B struggles to follow the requested format, spurious text remains a rare occurrence. Eventually, Mistral 7B shows a considerable amount of noise in the output, most notably in verbosity proneness, output in English language and repetition of parts of the prompt.

6 Discussion

The process of integrating LLMs into translation workflows is still at an early stage, with a major divide opening between the use of general-purpose models — leveraging scale-driven emergent capabilities — as opposed to models fine-tuned through task-specific, specialized training. While we register a higher overall success rate for generic LLMs, we also suggest that models trained specifically on translation data show greater promise to attend to the peculiar challenges of context-sensitive translation. On the technical side, future research should continue to explore strategies for conditioning more reliable terminology generation at decoding time.

Another promising direction involves exploiting the contextual metadata available in Linguistic Linked Data (LLD) resources (especially multilingual examples of use of term entries) for continued pre-training or task-adaptive fine-tuning. In this respect, while not having used terminology data as LLD for terminology injection in this first stage, we consider it a necessary follow-up step. Terminological data in machine-readable formats are likely to become crucial resources for terminology injection in future. We have shown that terminological data can be used to partly make up for the lack of training data in minor language varieties. Where training data from major varieties risks overriding language use in minor varieties and even leading to critical mistakes in high-stakes domains like the legal domain, available terminological resources could help to partially fill the gap.

The training data used for NMT systems and LLM models are skewed towards high-resource world languages and language combinations containing English. Further research is needed into non-English language combinations, which are a routine part of the work of many bodies with legislative, administrative and judicial powers that affect citizens daily lives. To name some examples in Europe, there is translation from German into a minor legal variety of Slovene in Austria for the Slovene-speaking minority in Carinthia, translation from Croatian into a minor legal variety of Italian in Croatia, translation from Finnish into a minor variety of Swedish in Finland. All these and other minority communities would profit from efficient strategies for adapting machine translation to their specific varieties and/or from terminology injection to fine-tune translation results.

7 Conclusion and limitations

We are aware of the limitations of our first experiments. Although proprietary LLMs set the current upper bound for performance, the opacity surrounding their training data precludes any assessment of whether results stem from explicit exposure to translation tasks. This constraint limits our ability to isolate and attribute observed advantages to translation-specific (pre-)training, thereby confounding the validation of the research hypothesis.

From a linguistic perspective, our data illustrates phenomena that apply to the South Tyrolean standard variety of German and to a non-English language combination. To assess whether similar results would apply, for example, to the standard variety used by the German-speaking community in Belgium and to translation from Dutch or French to another variety of German, we would need targeted studies. The same holds true for other minor (legal) varieties of major languages (e.g., Swiss French, Chilean Spanish etc.) and the language combinations that might be predominant in other minor or minority variety contexts. A further limitation consists in the lack of qualitative analyses that could shed better light on the results, which we are planning for the future.

8 Acknowledgements

Dataset creation was partly financed by the Autonomous Province of Bozen/Bolzano–South Tyrol within the project SSL-Laien "Optimising expert-lay communication. Case study: e-learning modules of the Autonomous Province of Bolzano – South Tyrol" awarded via the Programme Agreement for Eurac Research for 2022-2024. This financing only concerns the subset on occupational health and safety.

References

Noëmi Aepli, Chantal Amrhein, Florian Schottmann, and Rico Sennrich. 2023. A benchmark for evaluating machine translation metrics on dialects without standard orthography. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1045–1065, Singapore. Association for Computational Linguistics.

Melissa Ailem, Jingshu Liu, and Raheel Qader. 2021. Encouraging neural machine translation to satisfy terminology constraints. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1450–1455, Online. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Laurent Besacier, James Cross, Matthias Gallé, Philipp Koehn, and Vassilina Nikoulina. 2021a. On the evaluation of machine translation for terminology consistency. *CoRR*, abs/2106.11891.

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021b. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Transla-*

tion, pages 652–663, Online. Association for Computational Linguistics.

Tamer Alkhouli, Gabriel Bretschner, and Hermann Ney. 2018. On the alignment problem in multi-head attention-based neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 177–185, Brussels, Belgium. Association for Computational Linguistics.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Ulrich Ammon, Hans Bickel, and Alexandra N. Lenz. 2016. *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*, 2 edition. De Gruyter, Berlin.

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 biomedical translation shared task: Evaluation for MEDLINE abstracts and biomedical terminologies. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy. Association for Computational Linguistics.

Toms Bergmanis and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. In *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore. Association for Computational Linguistics.

Eleftheria Briakou, Colin Cherry, and George Foster. 2023. Searching for needles in a haystack: On the role of incidental bilingualism in PaLM's translation capability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9432–9452, Toronto, Canada. Association for Computational Linguistics.

Eleftheria Briakou, Zhongtao Liu, Colin Cherry, and Markus Freitag. 2024. On the implications of verbose llm outputs: A case study in translation evaluation. *Preprint*, arXiv:2410.00863.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mauro Cettolo, Andrea Piergentili, Sara Papi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2024. MAGNET - MAchines GeNErating translations: A CALAMITA challenge. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 1089–1093, Pisa, Italy. CEUR Workshop Proceedings.

Pinzhen Chen, Nikolay Bogoychev, Kenneth Heafield, and Faheem Kirefu. 2020. Parallel sentence mining by constrained decoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1672–1678, Online. Association for Computational Linguistics.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 181–190, Sheffield, UK. European Association for Machine Translation (EAMT).

Elena Chiocchetti. 2021. Effects of social evolution on terminology policy in south tyrol. *Terminology*, 27(1):110–139.

Michael Clyne, editor. 1991. *Pluricentric Languages. Differing Norms in Different Nations. Different Norms in Different Nations*. De Gruyter Mouton, Berlin, Boston.

Antonio Contarino. 2021. Neural machine translation adaptation and automatic terminology evaluation: A case study on italian and south tyrolean german legal texts. Master's thesis, Università di Bologna, Bologna, Italy.

Flavia De Camillis and Elena Chiocchetti. 2024. Machine-translating legal language: error analysis on an italian-german corpus of decrees. *Terminology science & research*, 27(1):1–27.

Flavia De Camillis, Egon W. Stemle, Elena Chiocchetti, and Francesco Fernicola. 2023. The MT@BZ corpus: machine translation & legal language. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 171–180, Tampere, Finland. European Association for Machine Translation.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

M. Amin Farajian, Nicola Bertoldi, Matteo Negri, Marco Turchi, and Marcello Federico. 2018. Evaluation of terminology translation in instance-based neural MT adaptation. In *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*, pages 169–178, Alicante, Spain.

Zhaopeng Feng, Ruizhe Chen, Yan Zhang, Zijie Meng, and Zuozhu Liu. 2024. Ladder: A model-agnostic framework boosting LLM-based machine translation to the next level. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15377–15393, Miami, Florida, USA. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8756–8769, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Antonio Gambaro and Rodolfo Sacco. 2024. *Sistemi giuridici comparati*, 4 edition. UTET, Milano.

Johann Gamper. 1999. Encoding a parallel corpus for automatic terminology extraction. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 275–276, Bergen, Norway. Association for Computational Linguistics.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *Preprint*, arXiv:2202.11822.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *Preprint*, arXiv:2302.07856.

Iikka Hauhio and Théo Friberg. 2024. Mitra: Improving terminologically constrained translation quality with backtranslations and flag diacritics. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 100–115, Sheffield, UK. European Association for Machine Translation (EAMT).

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *Preprint*, arXiv:2302.09210.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Josef Jon, Michal Novák, João Paulo Aires, Dusan Varis, and Ondřej Bojar. 2021. CUNI systems for WMT21: Terminology translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 828–834, Online. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in c++. *Preprint*, arXiv:1804.00344.

Sejoon Kim, Mingi Sung, Jeonghwan Lee, Hyunkuk Lim, and Jorge Gimenez Perez. 2024. Efficient terminology integration for LLM-based translation in specialized domains. In *Proceedings of the Ninth Conference on Machine Translation*, pages 636–642, Miami, Florida, USA. Association for Computational Linguistics.

Rebecca Knowles, Samuel Larkin, Marc Tessier, and Michel Simard. 2023. Terminology in neural machine translation: A case study of the Canadian Hansard. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 481–488, Tampere, Finland. European Association for Machine Translation.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2024. Contextual refinement of translations: Large language models for sentence and document-level post-editing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2711–2725, Mexico City, Mexico. Association for Computational Linguistics.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. 2021. Machine translation into low-resource language varieties. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 110–121, Online. Association for Computational Linguistics.

Surafel M. Lakew, Matteo Negri, and Marco Turchi. 2020. Low resource neural machine translation: A benchmark for five african languages. *Preprint*, arXiv:2003.14402.

Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232, Singapore. Association for Computational Linguistics.

Jiarui Liu, Iman Ouzzani, Wenkai Li, Lechen Zhang, Tianyue Ou, Houda Bouamor, Zhijing Jin, and Mona Diab. 2025. Towards global ai inclusivity: A large-scale multilingual terminology dataset (gist). *Preprint*, arXiv:2412.18367.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Heikki E. S. Mattila. 2018. Legal language. In John Humbley, Gerhard Budin, and Christer Laurén, editors, *Languages for Special Purposes: An International Handbook*, pages 113–150. De Gruyter Mouton.

Elise Michon, Josep Crego, and Jean Senellart. 2020. Integrating domain terminology into neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexander Molchanov, Vladislav Kovalenko, and Fedor Bykov. 2021. PROMT systems for WMT21 terminology translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 835–841, Online. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023a. 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.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023b. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics.

Arijit Nag, Soumen Chakrabarti, Animesh Mukherjee, and Niloy Ganguly. 2024. Efficient continual pre-training of llms for low-resource languages. *Preprint*, arXiv:2412.10244.

Antoni Oliver, Sergi Alvarez-Vidal, Egon Stemle, and Elena Chiocchetti. 2024. Training an NMT system for legal texts of a low-resource language variety south tyrolean German - Italian. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 573–579, Sheffield, UK. European Association for Machine Translation (EAMT).

Minh Quang Pham, Josep Crego, Antoine Senellart, Dan Berrebbi, and Jean Senellart. 2021. SYSTRAN @ WMT 2021: Terminology task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 842–850, Online. Association for Computational Linguistics.

Matt Post, Shuoyang Ding, Marianna Martindale, and Winston Wu. 2019. An exploration of placeholding in neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 182–192, Dublin, Ireland. European Association for Machine Translation.

Carla Quinci and Gianluca Pontrandolfo. 2023. Testing neural machine translation against different levels of specialisation: An exploratory investigation across legal genres and languages. *Trans-Kom. Journal of Translation and Technical Communication Research*, 16(1):174–209. Special Issue on Communicative Efficiency.

Natascia Ralli and Norbert Andreatta. 2018. Bistro – ein tool für mehrsprachige rechtsterminologie. *Trans-Kom. Journal of Translation and Technical Communication Research*, 11(1):7–44.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

Marek Sabo, Judith Klein, and Giorgio Bernardinello. 2024. Boosting machine translation with AI-powered terminology features. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 2)*, pages 25–26, Sheffield, UK. European Association for Machine Translation (EAMT).

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. *Preprint*, arXiv:2310.11324.

Kirill Semenov and Ondřej Bojar. 2022. Automated evaluation metric for terminology consistency in MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 450–457, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

Suzanna Sia, David Mueller, and Kevin Duh. 2024. Where does in-context translation happen in large language models. *Preprint*, arXiv:2403.04510.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 449–459, Minneapolis, Minnesota. Association for Computational Linguistics.

H. Sousa, S. Almasian, R. Campos, and A. Jorge. 2025. Tradutor: Building a variety specific translation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25183–25191.

Michal Štefánik, Marek Kadlcik, and Petr Sojka. 2023. Soft alignment objectives for robust adaptation of language generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8837–8853, Toronto, Canada. Association for Computational Linguistics.

Jiao Sun, Thibault Sellam, Elizabeth Clark, Tu Vu, Timothy Dozat, Dan Garrette, Aditya Siddhant, Jacob Eisenstein, and Sebastian Gehrmann. 2023. Dialect-robust evaluation of generated text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6010–6028, Toronto, Canada. Association for Computational Linguistics.

Atula Tejaswi, Nilesh Gupta, and Eunsol Choi. 2024. Exploring design choices for building language-specific LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10485–10500, Miami, Florida, USA. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Don Tuggener. 2016. Incremental coreference resolution for german. Master's thesis, University of Zurich, Faculty of Arts.

Elsbeth Turcan, David Wan, Faisal Ladhak, Petra Galuscakova, Sukanta Sen, Svetlana Tchistiakova, Weijia Xu, Marine Carpuat, Kenneth Heafield, Douglas Oard, and Kathleen McKeown. 2022. Constrained regeneration for cross-lingual query-focused extractive summarization. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2668–2680, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada. Association for Computational Linguistics.

Lucas Weber, Elia Bruni, and Dieuwke Hupkes. 2023. The icl consistency test. *Preprint*, arXiv:2312.04945.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. 2024. LLMRefine: Pinpointing and refining large language models via fine-grained actionable feedback. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1429–1445, Mexico City, Mexico. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2024. Teaching large language models to translate with comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19488–19496.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. 2024. Instruction tuning for large language models: A survey. *Preprint*, arXiv:2308.10792.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: less is more for alignment. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. 2024a. Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 388–409, Miami, Florida, USA. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024b. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Example of a test set entry

Source sentence	Per i videoterminalisti che non sono esposti a rischi aggiuntivi, è prevista una formazione di base di 4 ore più 4 ore di formazione per il rischio specifico.
Source term	videoterminalista
Target term	Bildschirmarbeiter
Other terms from South Tyrol	Bildschirmverwender
Terms from other legal systems	Bildschirmarbeitnehmer

Table 2: Example of an entry structure taken from the test set.

B Prompt Structure

```
[
    {
        "role": "system",
        "content": "This is a translation task. Translate the
           ↪following legal text from Italian into South-Tyrolean
           ↪German. There are terminological constraints you must
           ↪adhere to: {term_it} corresponds to {term_de}. You must
           ↪output only the translated text without any explanation.
           ↪ This is the text to be translated into South-Tyrolean
           ↪German:"
    },
    {
        "role": "user",
        "content": "<{source_sentence}>"
    }
]
```

Listing 1: Prompt structure for the Standard Terminology set. For the homonym set, both possible homonyms are provided as options, without suggesting the correct one.

Terminology Management Meets AI: The ISO/TC 37/SC 3/WG 6 Initiative

**Mohamed Khemakhem[1], Cristina Valentini[2], Natascia Ralli[3], Sérgio Barros[2],
Georg Löckinger[5], Federica Vezzani[6], Ana Salgado[9], Zhenling Zhang [7],
Sabine Mahr[4], Sara Carvalho[10], Klaus Fleischmann[8], Rute Costa[9]**

[1]MandaNetwork, [2]World Intellectual Property Organization,
[3]Eurac Research, Institute for Applied Linguistics, [4]Word b sign Sabine Mahr,
[5]University of Applied Sciences Upper Austria, [6]Università degli Studi di Padova,
[7]Computer Science College, Liaocheng University, Shandong Province, [8]Kaleidoscope,
[9]Centro de Linguística da Universidade NOVA de Lisboa, [10]Universidade de Aveiro

Abstract

The integration of artificial intelligence (AI) with terminology management (TM) has opened new avenues for enhancing efficiency and precision in both fields, necessitating standardized approaches to ensure interoperability and ethical application. The newly formed ISO/TC 37/SC 3/WG 6 represents the first dedicated initiative to study the standardization of the mutual improvements of AI and TM. This group aims to develop standardized frameworks and guidelines that optimize the interaction between AI technologies and terminology resources, benefiting professionals, systems, and practices in both domains. This article presents the state-of-the-art in the mutual relationship between AI and TM, highlighting opportunities for bidirectional advancements. It also addresses limitations and challenges from a standardization perspective. By tackling these issues, ISO/TC 37/SC 3/WG 6 seeks to establish principles that ensure scalability, precision, and ethical considerations, shaping future standards to support global communication and knowledge exchange.

1 Introduction

TM is critical to global communication by supporting the common understanding and exchange of specialized knowledge across languages and cultures. Within ISO's Technical Committee 37 "Language and terminology" (ISO/TC 37) [1], Subcommittee 3 (SC 3) [2] "Management of terminology resources" has played a pivotal role in formulating general principles, recommendations, and best practices for implementing effective, interoperable, and machine-readable TM systems. These systems support various computer language related processes, such as computer-assisted translation (CAT), controlled authoring, search engine optimization, and machine learning.

Advanced natural language processing (NLP) technologies – especially large language models (LLM) leveraged by generative artificial intelligence (GenAI) since 2022 – today offer new opportunities to enhance and further automate TM tasks which have the potential to make these processes more efficient and terminology resources more interoperable. In addition, the latest AI developments have highlighted how curated terminology can support AI-powered tools and enhance their quality, trustworthiness, and safety.

This bidirectional relationship, where AI functions both as a support technology for TM and as a domain requiring terminological data for improvement, is the focus of ISO/TC 37/SC 3/WG 6 "Terminology Management and Artificial Intelligence". This Working Group (WG) aims to develop standards that facilitate both the successful and ethical integration of AI into TM systems and the effective use of terminology resources to improve AI solutions. This paper discusses the challenges of standardizing this relationship within an unprecedented, rapidly evolving technological landscape.

2 State of the Art

2.1 AI for Terminology Management

The impact of AI on terminology builds upon technological shifts dating back to the early 1990s—when TM transitioned from paper-based to digital systems, as results of the integration of terminology with computational linguistics. Advances in computational power and data availability have subsequently enabled AI systems to perform large-scale data analysis, pattern extraction, and complex decision-making—capacities aligning with the knowledge-driven nature of terminology work.

AI promises to transform TM by enhancing and automating laborious and time-consuming processes, such as term extraction, or Automatic Term

[1] https://www.iso.org/committee/48104.html

[2] https://www.iso.org/committee/48136.html

Extraction (ATE), a TM task in which neural networks have been applied in various ways, including the generation of word embeddings as a contextual representation of terms and the development of classifiers for ATE (Lefever and Terryn, 2024; Tran et al., 2023).

Word embeddings are vector representations of words in a high-dimensional space, where words used in similar contexts have similar embeddings and tend to cluster together in that space. This capacity helps in revealing hidden relationships between terms, as seen in specialized domains such as maritime law (Mouratidis et al., 2022) and the broader evaluation of terminology extraction methods (Di Nunzio et al., 2023).

ATE classifiers operate by classifying tokens in sequence, using linguistic and statistical features to determine their status as lexical units. This approach supports a more efficient and context-sensitive TM workflow (Terryn et al., 2022). In enterprise settings [3], AI has been implemented to support terminology harmonization. For example, the Busch Group uses AI to extract terms, detect synonyms, and compare terminology across corporate databases (Beck and Fahlbusch, 2025). In Xu et al. (2025), LLMs are highlighted as a key to the future of terminology extraction tasks.

Recent advances in corpus alignment and domain-specific language modeling (Ye et al., 2024; Gururangan et al., 2020) indicate potential for LLMs trained on terminology-rich corpora to improve machine translation and domain understanding. Multilingual term alignment has also advanced, with embeddings facilitating the matching of equivalents across languages. Recent work has shown their success in aligning Arabic–French terminology (Setha and Aliane, 2023).

AI-backed chatbot tools have further enabled semi-automatic validation through document retrieval and concept checking. While these tools are not yet fully autonomous and require human oversight, they enhance TM (Bezobrazova et al., 2024). AI systems can be trained to assist conceptual validation by recognizing criteria such as term frequency, source authority, and definitional quality. AI also facilitates the automatic discovery of less intuitive concept relations, allowing for the scalable construction of domain-specific knowledge graphs. LLMs have recently gained attention for their ability to carry out concept system extraction (ISO 5394:2024; Gromann et al., 2022), hypernym detection (Cai et al., 2025), and term extraction from knowledge graphs (Pan et al., 2023; Cao et al., 2021).

AI can also help identify authoritative documents, which enhances corpus quality for terminology work (Nagendra and Chandra, 2022). Moreover, AI expedites terminology mining at scale—minimizing manual tasks for terminologists and improving efficiency in large organizations (Hamm, 2025).

2.2 Terminology Management for AI

While ontologies have historically guided the structuring of domain knowledge for AI systems, curated terminology resources in many cases remain underutilized. Terminology data—complete with designations, definitions, contexts and other data categories—provides valuable added structure for language models operating in specialized communication.

When LLMs are guided by curated terminology resources, either through training, fine-tuning, prompting, or retrieval, they exhibit improved factual grounding, contextual relevance, and reduced hallucination. Hallucination—wherein LLMs produce plausible but incorrect outputs—is especially problematic in critical fields. Terminologies act as semantic anchors, reinforcing verified content and supporting better error mitigation (Warburton, 2025).

Structured terminological data also expand reasoning in LLMs and support neuro-symbolic methods by integrating inferential pathways such as hierarchical (e.g., generic-specific) and non-hierarchical (e.g., associative) concept relationships. This aligns with the principles outlined in ISO 704:2022, which standardizes the representation of concepts, their relations, and definitions in terminology work, providing a foundational framework for knowledge organization in AI systems (Iantosca, 2022).

Conceptual Model-Augmented Generative AI (CMAG) leverages conceptual schemas to impose structure on Human-LLM interactions (Fill et al., 2024). CMAG encourages prompt generation within conceptual templates, facilitating cross-domain application in software engineering, heritage studies, and knowledge organization. Retrieval augmented generation (RAG), a prominent paradigm for augmenting LLMs with external knowledge - usually unstructured in text docu-

[3] https://kaleidoscope.at/en/blog/ai-and-terminology/

ments - to improve contextual relevance. Ontology-grounded RAG (OG-RAG) expands this idea: documents are modeled as hypergraphs using ontology-based clustering to enhance contextual understanding and decrease inference costs (Sharma et al.). Similarly, CLEAR (Clinical Entity Augmented Retrieval) utilizes entity recognition and ontology links to maximize retrieval precision in clinical contexts (Lopez et al., 2025). Such ontology-based approaches can inspire the leverage of conceptual knowledge in terminology resources to augment, improve, or evaluate LLM capabilities. Terminology Augmented Generation (TAG) (Fleischmann and Lang, 2025) is a recently proposed paradigm that integrates domain-specific terminologies into text retrieval processes, aiming at achieving high precision and computational efficiency.

Graph-structured retrieval (GraphRAG) offers further optimization by using Knowledge Graphs for query expansion, retrieval, and generative consistency (Han et al., 2025).

Terminology also plays a central role in adapting general-purpose LLMs to specialized domains. The VEGAD approach (Liu et al., 2024) automatically identifies and integrates optimal domain vocabularies into LLMs while mitigating catastrophic forgetting.

Terminology-based integration has been pivotal in neural machine translation tasks as well. Legacy efforts include structured incorporation of glossaries into machine translation (MT) models (Michon et al., 2020), while newer methods like trie-based extraction have led to efficient LLM training using limited specialized data (Kim et al., 2024). Instruction-based fine-tuning further enhances cross-linguistic term consistency, as evidenced by MT+G-Align's success in tasks like WMT 2023 (Zhao et al., 2024).

Collectively, these methods underscore terminology's central value in improving contextual relevance, factuality, and domain alignment of AI systems.

2.3 Limits

Despite the synergy between AI and TM, challenges persist. One concern is bias in training data, which can lead to culturally insensitive or non-inclusive terminological usage. Studies on the alignment of LLMs highlight its ethical criticality, inherent complexity, and the numerous challenges involved (Ji et al., 2024; Wang et al., 2024; Shen et al., 2023). Studies in gender-inclusive translation point toward unresolved ethical issues in AI integration (Gromann et al., 2023; Měchura, 2022; Corral and Saralegi, 2022; Touileb et al., 2022; Costa-jussà and de Jorge, 2020).

Another concern is the dependence of machine learning models on training data whose features and characteristics are unknown to its end-users, which makes it impossible to understand clearly how and why a given AI system makes certain decisions. This is a lack of transparency that could be addressed in the future, e.g. by building specialized LLMs (Ling et al., 2024), which are already being tested across several domains (e.g. Calamo et al. (2023) in e-justice, Chen et al. (2025) in biomedicine), as well as by investing on prompt engineering that could teach an AI system to make better decisions (Xu et al., 2025; Heinisch, 2024).

Another area of concern is the automatic extraction of terms and definitions. Bezobrazova et al. (2024) observed that while ChatGPT is not specifically designed for terminology extraction, it can perform the task effectively when properly prompted. However, compiling a thorough set of terms requires considerable time. Di Nunzio et al. (2024) corroborate this finding, noting that leveraging LLMs for term extraction, assessing term difficulty, and generating definitions for complex concepts is still in its early stages but holds potential for the future of ATE.

AI tools also face limitations in the automatic recognition of terminological variation, language varieties (e.g. American English, British English, Austrian German, Swiss German) and low-resource languages (e.g.. Ladin, South Tyrolean German) (Heinisch, 2024; Frontull and Moser, 2024; De Camillis et al., 2023).

Automated deep learning techniques for bias detection and inclusive language identification remain an emerging area (Savoldi et al., 2023; Piergentili et al., 2023). Ethical consideration must thus accompany technical innovation in deploying AI-enhanced terminology systems (Kumar et al., 2024).

3 Standardization Challenges in the Bidirectional Integration of Terminology Management and AI

3.1 Standardization Challenges in AI-Enhanced Terminology Workflows

AI is increasingly used to automate TM tasks. These capabilities promise greater efficiency and scalability—but also reveal substantial gaps in traceability, quality assurance, and semantic precision. These weaknesses are particularly problematic in domains where terminological accuracy is crucial for regulation, safety, or legal compliance.

Several challenges must be addressed through standardization efforts:

- **Validation Protocols:** There is a pressing need for standardized validation mechanisms to ensure that AI-generated terminological units are accurate, contextually appropriate, and domain-compliant. Human-in-the-loop validation workflows, quality thresholds, and expert review loops should be formally defined.

- **Metadata Schemas:** Current metadata standards do not accommodate the unique characteristics of AI-generated content. New schemas are needed to record confidence scores, sources of training data, contextual prompts, versioning information, and post-generation modifications.

- **Tool Interoperability:** AI-assisted TM tools must integrate smoothly with existing terminological ecosystems, such as termbases, content management systems (CMS), and translation tools. Standardized APIs and data exchange formats are essential to maintain consistency and usability across platforms.

- **Licensing and Attribution:** AI outputs frequently draw on heterogeneous training data, much of which may include third-party resources. Without appropriate licensing metadata, the reuse of such outputs in professional contexts raises intellectual property and compliance concerns. Machine-readable, standardized descriptors for licensing and use rights must be incorporated into the data pipeline.

- **Explainability and Traceability:** The opaque nature of neural networks and generative models hinders the auditing and validation of terminology-related outputs. Standards are needed to enable traceability of model behavior—potentially through techniques such as RAG, hybrid symbolic methods, or prompt-driven audit trails.

Collectively, these challenges call for the evolution of terminology standards to address emerging AI-specific requirements. Without rigorous validation frameworks, interoperable infrastructures, and transparent metadata, the integration of AI into TM risks undermining the precision and credibility traditionally associated with terminology work.

3.2 Standardization Challenges for the Integration of Terminologies into AI Systems

Integrating standardized terminologies into AI systems presents a complex set of challenges that span linguistic, technical, and ethical domains. AI systems—ranging from knowledge graphs to LLMs—often rely on representational formats (such as vector embeddings or ontology languages) that differ significantly from those used in traditional TM environments such as ISO 12620-1:2022, ISO 12620-2:2022, ISO 30042:2019, and ISO16642:2017. As a result, bridging the gap between terminology resources and AI's inference or generation processes requires addressing several foundational standardization issues.

Key challenges include:

- **Fragmentation Across Industries and Languages:** Terminology resources are often created in isolated institutional or national settings. In combination with a lack of adoption of standards, this results in a rather low level of interoperability. Cross-domain, multilingual frameworks are underdeveloped.

- **Under-Resourced Languages:** Most terminology resources exist only in high-resource languages. This skews AI inclusivity and limits access in culturally diverse contexts.

- **Access Limitations:** High-quality terminologies in medicine, law, and engineering are often proprietary. Licensing issues restrict open usage, especially in academic/non-commercial AI applications.

- **Complex and Costly Terminology Development:** Building terminology resources requires linguistic and domain expertise and is not easily scalable. Formats are often not compatible with contemporary AI pipelines.

These challenges point to the need for a new generation of interoperability and governance standards capable of:

- Mapping between traditional terminological formats and AI-ready knowledge structures (e.g., ontologies, embeddings) (Roche et al., 2009),

- Supporting modular, scalable, and semi-automated methods for terminology validation and enrichment,

- Embedding ethical and provenance metadata within terminology resources,

- Encouraging the creation and standardization of multilingual terminologies, especially in under-resourced language settings,

- Defining lifecycle-aware structures that enable alignment between terminological data and AI workflows.

3.3 Cross-Cutting Standardization and Governance Challenges

Efforts to integrate TM and AI systems face systemic challenges that go beyond technical interoperability.

A fundamental issue stems from the contrasting development dynamics of the fields: AI is characterized by rapid, iterative innovation cycles, often driven by data-centric experimentation and agile tooling. Conversely, TM evolves through slower, institutionalized processes grounded in expert validation, domain specificity, and long-term conceptual stability. These contrasting tempos result in mismatches between the needs of AI systems and the availability or granularity of terminology resources, complicating real-time integration, update synchronization, and cross-domain scalability.

Additionally, several cross-cutting challenges persist:

- **Fragmented Standardization Ecosystem:** ISO/TC 37 (language & terminology), W3C [4] (semantic web), and ISO/IEC JTC 1/SC 42 [5] (AI) are separate organizational frameworks. The lack of cross-committee collaboration creates isolated representations and challenges interoperability.

- **Legal and Licensing Barriers:** Terminology resources frequently lack standardized, machine-readable licensing information. This creates uncertainty around reuse, especially in AI workflows that involve automatic ingestion, transformation, and content generation. Initiatives to integrate licensing metadata—such as RDF-based descriptors or Creative Commons tags—in standardized terminology formats are still nascent and require broader institutional adoption.

- **Discipline-Specific Governance Models:** Governance structures in TM, AI, data ethics, and digital infrastructure are organized under different frameworks. As a result, collaboration between terminologists, AI developers, data stewards, legal experts, public-sector institutions, and domain knowledge authorities remains limited. Effective integration of TM into AI ecosystems demands multi-stakeholder governance models that support cross-disciplinary decision-making, accountability mechanisms, and shared compliance frameworks.

Addressing these challenges will require proactive coordination between standardization bodies, the development of bridging standards across communities, and the rethinking of governance models to ensure that integration strategies reflect both the conceptual rigor of terminology and the dynamic realities of AI innovation.

4 ISO/TC 37/SC 3/WG 6: A Strategic Player in Bridging Terminology Management and AI

4.1 Positioning the New WG in the Global Standardization and Regulatory Landscape

ISO/TC 37 develops international standards that support the language industry, emphasizing interoperability, consistency, and quality in multilingual content management. Within this structure, SC 3 focuses on standards for computer-assisted processes for managing terminological data. It is responsible for key infrastructural standards such as ISO 5078:2025, ISO 12620-1:2022, ISO 12620-2:2022, ISO 30042:2019, and ISO16642:2017.

In early 2024, ISO/TC 37/SC 3 initiated Ad Hoc Group 1 (AHG 1) to explore the growing synergies of TM and AI, ultimately leading to the formal establishment of WG 6 “Terminology Management and Artificial Intelligence” in 2025.

WG 6 occupies a unique niche within the broader standards ecosystem. While important AI and NLP

[4] https://www.w3.org/

[5] https://www.iso.org/committee/6794475.html

developments are underway in ISO/IEC JTC 1/SC 42—especially through its Joint Working Group 5 (JWG 5) on NLP—these initiatives often treat linguistic resources at a high level of abstraction and do not sufficiently address the specificities of controlled vocabularies and terminological data. In contrast, WG 6 focuses specifically on integrating TM with AI systems, with attention to multilingual and domain-specific precision.

WG 6's relevance extends beyond ISO's internal architecture. The adoption of the European Union's *Artificial Intelligence Act (AI Act)* (Parliamant and Council, 2024) provides an external regulatory catalyst for WG 6's mission. As the AI Act introduces legally binding requirements for transparency, traceability, and documented safety in AI systems—especially in high-risk domains such as medicine, law, and finance—terminological precision emerges as a crucial safeguard. Misuse or absence of validated domain-specific terminology in such contexts can result in legal liability, safety hazards, and public mistrust.

Against this backdrop, WG 6 not only contributes to ISO's standardization efforts but also functions as a strategic actor helping stakeholders operationalize the AI Act's compliance requirements. By aligning semantic resources with ethical, legal, and technical expectations, WG 6 enables AI systems to meet the regulatory demand for explainability, contextual accuracy, and trustworthy outputs. For example, the output created by AI systems needs to reflect best practices of data modelling and conform to relevant data categories (ISO 12620-1:2022 and ISO 12620-2:2022).

4.2 Strategic Objectives and Operational Approach

WG 6 was established with a dual mandate:

- To develop standards and technical guidance for the integration of AI technologies into terminology workflows—covering tasks such as term extraction, clustering, multilingual alignment, and automated definition generation.
- To support the integration of standardized terminologies into AI systems, enabling them to benefit from linguistic precision, multilingual equivalence, and concept-level clarity in interpretation and generation tasks.

Addressing this dual mission requires WG 6 to operate in new, adaptable ways that diverge from legacy standardization procedures. Traditional standardization timelines are outpaced by AI's rapid, weekly or monthly cycles. To address this, WG 6 introduced methods adapted to AI's fast and volatile developments.

One notable methodological adjustment occurred during the exploratory phase. AHG 1 was organized around four agile task forces, each assigned to focus on key dimensions stemming from the different stages of terminology work. This distributed structure greatly accelerated production timelines and was key to transforming AHG 1 into a fully operational WG, culminating in its first Technical Report (TR) project (ISO/AWI TR 25896) [6]. TR had been chosen as the appropriate type of deliverable for its considerably shortened production time.

This approach has proven useful: instead of relying solely on static deliverables designed for long-term generality, WG 6 embraces a portfolio of standardization formats—including TRs, Technical Specifications, and eventually International Standards—that are scalable, iterative, and adaptable to fast-moving technological ecosystems. More broadly, this flexibility reflects a growing consensus across standardization bodies that AI not only uses standards, but is itself a target for standardization—thus requiring a rethinking of the pace, process, and governance of standards development.

5 Conclusion and Outlook

In conclusion, the integration of AI with TM marks a transformative step toward enhancing efficiency, precision, and global communication in both fields. The establishment of ISO/TC 37/SC 3/WG 6 represents a pioneering effort to standardize the symbiotic advancements of AI and TM, fostering interoperable, scalable, and ethically sound frameworks. By addressing the challenges of aligning traditional standardization processes with the rapid evolution of AI technologies, WG 6 introduces innovative methodologies that balance stability with adaptability. This strategic approach ensures that emerging standards remain rigorous yet agile, enabling public and private stakeholders to responsibly leverage AI in complex, multilingual knowledge environments. Through these efforts, WG 6 is poised to shape future standards that support seamless knowledge exchange and drive ethical, effective applications of AI and TM worldwide.

[6] https://www.iso.org/standard/91875.html

Acknowledgments

We would like to express our sincere thanks to the International Organization for Standardization (ISO) for providing a collaborative framework to advance terminology and language resource standards. In particular, we are grateful to the members of ISO/TC 37/SC 3/WG 6 for the valuable discussions, insights, and collective expertise that have informed and enriched this work.

We also gratefully acknowledge the support of the authors' respective institutions and national standardization bodies, which have provided the academic and organizational environment necessary for the completion of this work.

This work has been supported, in part, by the *StandICT.eu* initiative through financial assistance provided to some of the authors. This support has facilitated their participation in standardization-related activities and contributed to the development of certain aspects of the research. We appreciate the initiative's continued efforts to promote European engagement in international ICT standardization. The views expressed in this paper are those of the authors and do not necessarily reflect those of the *StandICT.eu* initiative.

References

Katharina Beck and Fabian Fahlbusch. 2025. Termino-KI. Terminologiearbeit 4.0 mit BuschGPT. In *Terminologie in der KI - KI in der Terminologie. Akten des Symposions, Worms, 27-29 März 2025*, pages 39–50, München, Deutschland. Deutscher Terminologie-Tag e.V.

Anastasiia Bezobrazova, Miriam Seghiri, and Constantin Orasan. 2024. Benchmarking terminology building capabilities of ChatGPT on an English-Russian Fashion Corpus. *Computing Research Repository*, arXiv:2412.03242.

Guohui Cai, Jiangchuan Gong, Junliang Du, Hao Liu, and Anda Kai. 2025. Investigating Hierarchical Term Relationships in Large Language Models. *Journal of Computer Science and Software Applications*, 5(4).

Marco Calamo, Francesca De Luzi, Mattia Macrì, Tommaso Mencattini, and Massimo Mecella. 2023. CICERO: a GPT2-based writing assistant to investigate the effectiveness of specialized LLMs' applications in e-justice. In *ECAI 2023*, pages 3196–3203. IOS Press.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or Educated Guess? Revisiting Language Models as Knowledge Bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874, Online. Association for Computational Linguistics.

Linqing Chen, Weilei Wang, Yubin Xia, Wentao Wu, Peng Xu, Zilong Bai, Jie Fang, Chaobo Xu, Ran Hu, Licong Xu, Haoran Hua, Jing Sun, Hanmeng Zhong, Jin Liu, Tian Qiu, Haowen Liu, Meng Hu, Xiuwen Li, Fei Gao, and 14 others. 2025. Streamlining Biomedical Research with Specialized LLMs. *Computing Research Repository*, arXiv:2504.12341.

Ander Corral and Xabier Saralegi. 2022. Gender bias mitigation for NMT involving genderless languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 165–176.

Marta R Costa-jussà and Adrià de Jorge. 2020. Fine-tuning neural machine translation on gender-balanced datasets.

Flavia De Camillis, Egon Waldemar Stemle, Elena Chiocchetti, and Francesco Fernicola. 2023. The MT@ BZ corpus: machine translation & legal language. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation 12–15 June 2023, Tampere, Finland*, pages 171–180.

Giorgio Maria Di Nunzio, Stefano Marchesin, and Gianmaria Silvello. 2023. A systematic review of Automatic Term Extraction: What happened in 2022? *Digital Scholarship in the Humanities*, 38(Supplement_1):i41–i47.

Giorgio Maria Di Nunzio, Federica Vezzani, Vanessa Bonato, Hosein Azarbonyad, Jaap Kamps, Liana Ermakova, and 1 others. 2024. Overview of the CLEF 2024 SimpleText task 2: identify and explain difficult concepts. *Working Notes of CLEF*.

HG Fill, F Härer, I Vasic, D Borcard, B Reitemeyer, F Muff, S Curty, and M Bühlmann. 2024. CMAG: A Framework for Conceptual Model Augmented Generative Artificial Intelligence, ER2024: Companion. In *Proceedings of the 43rd International Conference on Conceptual Modeling: ER Forum, Special Topics, Posters and Demos, Pittsburgh, PA, USA*, pages 28–31.

Klaus Fleischmann and Christian Lang. 2025. Terminologie in der KI: Wie mit Terminologie der Output von LLMs und GenAI optimiert werden kann. In *Terminologie in der KI - KI in der Terminologie. Akten des Symposions, Worms, 27-29 März 2025*, pages 83–96, München, Deutschland. Deutscher Terminologie-Tag e.V.

Samuele Frontull and Georg Moser. 2024. Traduzione automatica per il ladino della Val Badia. *Ladinia XLVIII/2024*.

Dagmar Gromann, Manuel Lardelli, Katta Spiel, Sabrina Burtscher, Lukas Daniel Klausner, Arthur Mettinger, Igor Miladinovic, Sigrid Schefer-Wenzl, Daniela Duh, and Katharina Bühn. 2023. Participatory research as a path to community-informed, gender-fair machine translation. *Computing Research Repository*, arXiv:2306.08906.

Dagmar Gromann, Lennart Wachowiak, Christian Lang, and Barbara Heinisch. 2022. Extracting Terminological Concept Systems from Natural Language Text. In *European Language Grid: A Language Technology Platform for Multilingual Europe*, pages 289–294. Springer International Publishing Cham.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Julian Hamm. 2025. Terminologische Konsistenz und generative KI – ein Perfect Match? Produktiver Einsatz von Sprachmodellen im Terminologiemanagement und beim Post-Editing. In *Terminologie in der KI - KI in der Terminologie. Akten des Symposions, Wörms, 27-29 März 2025.*, pages 151–164, München, Deutschland. Deutscher Terminologie-Tag e.V.

Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A Rossi, Subhabrata Mukherjee, Xianfeng Tang, and 1 others. 2025. Retrieval-augmented generation with graphs (graphrag). *Computing Research Repository*, arXiv:2501.00309.

Barbara Heinisch. 2024. Large language model prompting in a university course on terminology. *Terminology Science & Research/Terminologie: Science et Recherche*, (27):28–51.

Michael J. Iantosca. 2022. From microcontent to neurons A practical guide for building a cognitive AI content supply chain for highly personalized user assistance. Online. Online article, accessed on 2025-05-29.

ISO 12620-1:2022. Management of terminology resources — Data categories. Part 1: Specifications. International Organization for Standardization. https://www.iso.org/standard/79078.html.

ISO 12620-2:2022. Management of terminology resources — Data categories. Part 1: Repositories. International Organization for Standardization. https://www.iso.org/standard/79018.html.

ISO 30042:2019. Management of terminology resources – TermBase eXchange (TBX). International Organization for Standardization. https://www.iso.org/standard/62510.html.

ISO 5078:2025. Management of terminology resources — Terminology extraction. International Organization for Standardization. https://www.iso.org/standard/81917.html.

ISO 5394:2024. Information technology — Criteria for concept systems. International Organization for Standardization. https://www.iso.org/standard/81656.html.

ISO 704:2022. Terminology work — Principles and methods. International Organization for Standardization. https://www.iso.org/standard/79077.html.

ISO16642:2017. Computer applications in terminology — Terminological markup framework. International Organization for Standardization. https://www.iso.org/standard/56063.html.

ISO/AWI TR 25896. Management of terminology resources - Artificial intelligence for terminology management. International Organization for Standardization. https://www.iso.org/standard/91875.html.

Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Tianyi Alex Qiu, Juntao Dai, and Yaodong Yang. 2024. Aligner: Efficient alignment by learning to correct. *Advances in Neural Information Processing Systems*, 37:90853–90890.

Sejoon Kim, Mingi Sung, Jeonghwan Lee, Hyunkuk Lim, and Jorge Froilan Gimenez Perez. 2024. Efficient Terminology Integration for LLM-based Translation in Specialized Domains. *Computing Research Repository*, arXiv:2410.15690.

Naveen Kumar, Xiahua Wei, and Han Zhang. 2024. Addressing Bias in Generative Ai: Challenges and Research Opportunities in Information Management. *Available at SSRN 4976889*.

Els Lefever and Ayla Rigouts Terryn. 2024. Computational Terminology. In *New Advances in Translation Technology: Applications and Pedagogy*, pages 141–159. Springer.

Chen Ling, Xujiang Zhao, Jiaying Lu, Chengyuan Deng, Can Zheng, Junxiang Wang, Tanmoy Chowdhury, Yun Li, Hejie Cui, Xuchao Zhang, Tianjiao Zhao, Amit Panalkar, Dhagash Mehta, Stefano Pasquali, Wei Cheng, Haoyu Wang, Yanchi Liu, Zhengzhang Chen, Haifeng Chen, and 5 others. 2024. Domain Specialization as the Key to Make Large Language Models Disruptive: A Comprehensive Survey. *Computing Research Repository*, arXiv:2305.18703.

Chengyuan Liu, Shihang Wang, Lizhi Qing, Kun Kuang, Yangyang Kang, Changlong Sun, and Fei Wu. 2024. Gold Panning in Vocabulary: An Adaptive Method for Vocabulary Expansion of Domain-Specific LLMs. *Computing Research Repository*, arXiv:2410.01188.

Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P Ma, April S Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, and 1 others. 2025. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine*, 8(1):45.

Michal Měchura. 2022. A taxonomy of bias-causing ambiguities in machine translation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 168–173.

Elise Michon, Josep M Crego, and Jean Senellart. 2020. Integrating domain terminology into neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937.

Despoina Mouratidis, Eirini Mathe, Yorghos Voutos, Klio Stamou, Katia Lida Kermanidis, Phivos Mylonas, and Andreas Kanavos. 2022. Domain-specific term extraction: a case study on Greek Maritime legal texts. In *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*, pages 1–6.

N Nagendra and J Chandra. 2022. A Systematic Review on Features Extraction Techniques for Aspect Based Text Classification Using Artificial Intelligence. *ECS Transactions*, 107(1):2503.

Jeff Z Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, and 1 others. 2023. Large language models and knowledge graphs: Opportunities and challenges. *Computing Research Repository*, arXiv:2308.06374.

European Parliamant and Council. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). Official Journal of the European Union.

Andrea Piergentili, Dennis Fucci, Beatrice Savoldi, Luisa Bentivogli, and Matteo Negri. 2023. Gender neutralization for an inclusive machine translation: from theoretical foundations to open challenges. *Computing Research Repository*, arXiv:2301.10075.

Christophe Roche, Marie Calberg-Challot, Luc Damas, and Philippe Rouard. 2009. Ontoterminology: A new paradigm for terminology. In *International Conference on Knowledge Engineering and Ontology Development*, pages 321–326, Madeira, Portugal.

Beatrice Savoldi, Marco Gaido, Matteo Negri, and Luisa Bentivogli. 2023. Test suites task: Evaluation of gender fairness in MT with MuST-SHE and INES. *Computing Research Repository*, arXiv:2310.19345.

Imene Setha and Hassina Aliane. 2023. Bilingual Terminology Alignment Using Contextualized Embeddings. In *Proceedings of the Workshop on Computational Terminology in NLP and Translation Studies (ConTeNTS) Incorporating the 16th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 1–8, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Kartik Sharma, Peeyush Kumar, and Yunqing Li. OG-RAG: Ontology-Grounded Retrieval-Augmented Generation For Large Language Models. *Computing Research Repository*, arXiv:2412.15235.

Tianhao Shen, Renren Jin, Yufei Huang, Chuang Liu, Weilong Dong, Zishan Guo, Xinwei Wu, Yan Liu, and Deyi Xiong. 2023. Large Language Model Alignment: A Survey. *Computing Research Repository*, arXiv:2309.15025.

Ayla Rigouts Terryn, Veronique Hoste, and Els Lefever. 2022. Tagging terms in text: A supervised sequential labelling approach to automatic term extraction. *Terminology: international journal of theoretical and applied issues in specialized communication*, 28(1):157–189.

Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2022. Occupational biases in Norwegian and multilingual language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 200–211.

Hanh Thi Hong Tran, Matej Martinc, Jaya Caporusso, Antoine Doucet, and Senja Pollak. 2023. The recent advances in automatic term extraction: A survey. *Computing Research Repository*, arXiv:2301.06767.

Han Wang, An Zhang, Nguyen Duy Tai, Jun Sun, Tat-Seng Chua, and 1 others. 2024. ALI-Agent: Assessing LLMs' Alignment with Human Values via Agent-based Evaluation. *Advances in Neural Information Processing Systems*, 37:99040–99088.

Kara Warburton. 2025. Terminology in the Age of AI. `https://multilingual.com/magazine/march-2025/terminology-in-the-age-of-ai/`. Online article, accessed on 2025-05-29.

Kang Xu, Yifan Feng, Qiandi Li, Zhenjiang Dong, and Jianxiang Wei. 2025. Survey on terminology extraction from texts. *Journal of Big Data*, 12(1):1–40.

Qi Ye, Zicheng Yao, Peihong Hu, Xiang Ji, Tong Ruan, and Ruihui Hou. 2024. Alignment of chinese-english medical terminology in small-sample scenarios: A two-stage approach. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3908–3911. IEEE.

Xuan Zhao, Chong Feng, Shuanghong Huang, Jiangyu Wang, and Haojie Xu. 2024. Incorporating Terminology Knowledge into Large Language Model for Domain-Specific Machine Translation. In *China Conference on Machine Translation*, pages 82–96. Springer.

Inferring Semantic Relations Between Terms with Large Language Models

Giulia Speranza
University of Naples "L'Orientale" - Via Chiatamone, 60/61 - 80121 Naples (Italy)
gsperanza@unior.it

Abstract

The purpose of this paper is to investigate the ability of Large Language Models (LLMs) to identify relations among terms, with the goal of facilitating and accelerating the construction of thesauri and terminological resources. We investigate whether the use of LLMs in this context can provide a valuable initial set of relations, serving as a basis upon which professional terminologists can build, validate, and enrich domain-specific knowledge representations.

1 Introduction

The identification and formalization of semantic relations among terms constitute a fundamental task in the development and refinement of thesauri and terminological resources.

In any specialized knowledge domain, terms do not exist in isolation but form a complex net of semantic relations. These relationships enable the structuring of domain knowledge, facilitate precise communication, and support tasks such as indexing, searching, and reasoning.

Early foundational work in terminology science, such as by Wüster (1975), emphasized the importance of defining clear hierarchical relations (broader-narrower) and associative relations to support knowledge organization and retrieval. Recent theoretical developments, such as Frame-Based Terminology (Faber, 2022), further highlight the central role of semantic relations in the organization of specialized knowledge.

Systematic terminological standards, including ISO 25964-1:2011 and ISO 704:2022 for thesauri, codify relations among terms into three main types:

- Hierarchical Relations: represent relations of superordination and subordination, where the superordinate concept represents a class or whole, and subordinate concepts refer to its members or parts. These relations form taxonomies or classification hierarchies that structure knowledge from general to specific. The most common forms are Broader Term (BT) and Narrower Term (NT);
- Equivalence Relations: establish links between terms that are near-synonyms within the domain, enabling consolidation of lexical variants and preventing ambiguity.
- Associative Relations: denote semantic connections that are neither hierarchical nor equivalence-based but are meaningful in context. Examples include Related Terms (RT) according to relationships of cause-effect, part-whole, or functionally related terms.

Formal models such as SKOS (Simple Knowledge Organization System) have standardized the representation of semantic relations in thesauri, providing an interoperable framework for semantic web applications (Miles and Bechhofer, 2009).

This work aims to analyze and compare the out-of-the-box performance of two LLMs (ChatGPT and Gemini) in identifying terminological semantic relations, with the aim of supporting the development of thesauri and other terminological resources. The case study and the gold standard dataset is in the domain of building materials and constructions.

2 Related Works

Historically, thesauri have been constructed manually by domain experts, who defined semantic relations based on domain knowledge. Hierarchical structures were typically conceptual while associative relations were more complex, often reflecting functional, causal, or contextual connections.

One of the first attempts to automatically identify hierarchical relations was proposed by Hearst (1992), who introduced lexico-syntactic patterns (e.g., "X such as Y", "Y is a type of X") as indicators of hyponymy.

One influential approach in this area is the use of knowledge patterns, as introduced by Meyer (2001), which have been instrumental in the extraction and identification of semantic relations. Automatic systems based on this approach have been developed to support the extraction of semantic relations from corpora. Notable examples include the EcoLexicon Semantic Sketch Grammar (ESSG) (León-Araúz and Martín, 2018), which applies knowledge patterns within a frame-based approach to identify and organize conceptual relations in domain-specific texts, and Corpógrafo (Maia and Matos, 2008), a tool designed to facilitate corpus-based terminological analysis by detecting and visualizing term relationships and co-occurrence patterns.

Studies by Cimiano et al. (2005); Velardi et al. (2013); Spiteri (2002) extended this by combining pattern-based extraction with distributional semantics, enabling more nuanced identification of various semantic relations beyond simple hierarchies.

In parallel, graph-based methods have gained traction for modeling and identifying term relations (Velardi et al., 2013). Tools like WordNet (Miller, 1992) exemplify a rich network of lexical-semantic relations structured as a graph, which has influenced the design of domain-specific thesauri.

Identifying these relations often requires domain expertise combined with semi-automated methods, such as supervised learning approaches trained on annotated corpora.

Studies such as Petroni et al. (2019) have shown that LLMs can retrieve encyclopedic facts through cloze tasks (e.g., “Paris is the capital of ___”), but the ability to infer more abstract conceptual relations (e.g., hypernymy, meronymy) remains underexplored.

Recently, some works have analyzed the implicit presence of structured semantic knowledge in LLMs. For example, Davison et al. (2019) investigated the extent to which LLMs can correctly answer questions about relations between entities. However, these studies almost always rely on rich contextual input (sentences, definitions, knowledge triples), whereas our work focuses on a minimal setting in which the model receives only a list of terms and must infer possible relations.

Trained on massive text corpora, LLMs have demonstrated a strong capacity to understand and generate human-like text. They are also capable of handling a wide range of linguistic tasks with little need for explicit guidance. As a result, recent studies have increasingly explored their effectiveness across a variety of NLP tasks. These capabilities raise a compelling question: can LLMs effectively support or even partially automate the process of finding relations among terms for terminological resource construction purposes?

In this regard, despite a growing interest in the use of LLMs in Automatic Terminology Extraction tasks for the creation of resources, glossaries, and thesaurus (Banerjee et al., 2024; Tran et al., 2024), there is still no systematic evaluation of their ability to infer semantic relations from simple term lists, without context, examples, or explicitly provided background knowledge.

3 Case study

Our study explores the zero-shot capabilities of LLMs in inferring relations between terms, without access to external texts or the use of fine-tuning techniques. We analyze the LLMs ability in this task by comparing their output against a thesaurus about construction materials of monuments, buildings, and structures taken as Gold Standard reference. No examples are provided in the prompt; the model relies only on prior knowledge.

3.1 LLM Selection

We selected two state-of-the-art large language models: ChatGPT and Gemini in order to conduct a comparative evaluation of different models’ designs in the context of terminology work. These models were chosen based on their widespread adoption, usage, and testing in various Natural Language Processing (NLP) tasks, their architectures, and training paradigms such as Reinforcement Learning from Human Feedback.

ChatGPT is developed by OpenAI and is based on the GPT architecture. For this study, we used the GPT-4 version, which represents a significant advance over previous iterations in terms of coherence and domain generalization capabilities. GPT-4 was trained on a massive corpus of publicly available text and licensed data using a transformer-based decoder architecture. It is capable of few-shot and zero-shot learning, meaning it can perform tasks with little to no task-specific fine-tuning, relying instead on prompt engineering.

Gemini is developed by Google DeepMind and represents the evolution of the earlier PaLM (Pathways Language Model) family. It is a multimodal LLM, trained not only on text but also on images and code, although our experiment utilizes the text-only capabilities of Gemini. The model is designed for factual grounding, task adaptability, native multimodality and long context window.

This approach allows us to compare the different LLMs strengths and weaknesses in handling the task and provides a broader perspective on how general-purpose language models can be employed in terminological tasks.

To simulate and closely reproduce a real-world use case scenario, we chose to interact with the selected LLMs in the same way a professional terminographer would in a practical work environment. This means that we did not apply any fine-tuning, domain-specific retraining, or technical modifications to the models. Instead, we relied on and evaluated exclusively their out-of-the-box capabilities, using the standard user interfaces provided by the respective platforms. This approach reflects the typical conditions under which language professionals, such as terminologists, translators, or linguists, without strong technical and engineering skills, would engage with LLM to perform terminology tasks. Our objective was to evaluate the actual usability and effectiveness of the models in supporting terminology work without requiring advanced technical expertise or custom integration efforts.

3.2 Prompting strategy

To guide the language models in generating domain-relevant and semantically coherent output, we adopted a persona prompting strategy. This approach involves framing the prompt in such a way that the model is instructed to assume the role of a specific expert or domain specialist—referred to as a persona. By embedding this expert persona into the prompt, we aimed to steer the models' responses toward more precise, terminologically consistent, and semantically appropriate outputs. The persona was specified at the beginning of the prompt and further contextualized with task-specific instructions, such as identifying hierarchical, associative, and equivalence relations among domain-specific terms. This technique leverages the models' ability to adapt to match the expectations associated with a particular professional role. To force the model to act like a specific person, adopting a certain perspective on the task to be performed, one effective strategy is to provide the persona's job title, which should elicit a set of associated attributes and competencies. The 'persona-pattern' or 'role-play' prompting techniques have been widely used in several studies for different tasks (Kong et al., 2023; Olea et al., 2024; Mzwri and Turcsányi-Szabo, 2025) as it is much more accessible, compared to fine-tuning the model from an engineering point of view. This prompting approach belongs to the so-called 'Output Customization category' according to White et al. (2023) or to the 'LLM Role-Playing' category according to Tseng et al. (2024).

For our experiment, we queried the language models using the prompt detailed in Example 1. It's important to highlight that all experimental sessions were carried out in May 2025.

LLM Prompt

```
You are an expert terminologist specialized
in the domain of building materials. You
are given a simple list of domain-specific
terms ordered alphabetically. Your task
is to identify all relevant semantic
relations (Broader Term, Narrower Term,
Related Term) among the terms in the list.

Instructions:

  1. Only consider relations valid within
     the context of building materials and
     construction.

  2. Use only the terms provided in the
     list: don't add or omit any term.

  3. Each relation should be directional
     where applicable (e.g., Term A is a
     broader term for Term B).
```

Example 1. Prompt with instructions

3.3 Dataset

As a case study, we used the FISH Building Materials Thesaurus[1], a controlled vocabulary employed for documenting primary construction materials of monuments, buildings, and structures associated with the built and buried heritage of the British Isles. This thesaurus is maintained by Historic England on behalf of the FISH (Forum on Information Standards in Heritage) Terminology Working Group and is openly available for reuse. It can be downloaded in multiple formats, including PDF (both alphabetical and hierarchical listings), CSV, and semantic web-compatible linked-data formats. A sample of the dataset can be found in Example 2.

[1] https://heritagedata.org/live/schemes/eh_tbm.html

GRANITE

SN: A hard, coarse grained, durable igneous rock. Can be used decoratively or as a main material. Found through-out the country.

BT: STONE

NT: ABERDEEN GRANITE, MOORSTONE, MOUNTSORREL GRANITE, PETERHEAD GRANITE, SHAP GRANITE

RT: ELVAN, GNEISS, GREENSTONE, PORPHYRY.

Example 2. Example of the FISH Building Materials Thesaurus

We fed the LLMs only with the list of terms (prefLabel) ordered alphabetically, without explicitly expressing the relations. This list of 232 terms (including TopConcepts) was obtained by querying the SPARQL endpoint. The complete dataset, including all term relations (broader, narrower, related), was retained as a Gold Standard reference for the subsequent evaluation phase.

4 Results and Evaluation

The following examples showcase representative outputs produced by each model, offering qualitative insight into their ability to identify semantic relations among the listed terms.

GRANITE

BT: Stone

NT: Aberdeen Granite, Granite, Peterhead Granite, Shap Granite

RT: Gneiss, Diorite, Feldspar

Example 3. ChatGPT output

GRANITE

BT: Stone

NT: Aberdeen Granite, Mountsorrel Granite, Peterhead Granite, Shap Granite

RT: Elvan, Gneiss, Greenstone

Example 4. Gemini output

To assess the ability of LLMs to correctly identify the semantic relation between terms, we adopted standard classification metrics such as Precision, Recall and F-Measure (the harmonic mean of Precision and Recall). The results of our evaluation are presented in Table 1, where we compare the performance of ChatGPT and Gemini.

Model	**Prec. (%)**	**Rec. (%)**	**F1 (%)**
ChatGPT	55.0	65.0	59.6
Gemini	75.4	84.2	79.6

Table 1: Evaluation results of LLM outputs

The comparative evaluation of ChatGPT and Gemini highlighted a significant difference in their ability to capture hierarchical depth in conceptual structures.

The recall is slightly lower than precision across models, indicating a tendency to be conservative in identifying relations, often abstaining when unsure.

In addition to the quantitative metrics, a qualitative evaluation was carried out to understand the nature of the errors and the strengths of each model.

While both LLMs performed almost adequately in detecting first-level (direct) conceptual relations, such as hypernymy (e.g., ‘granite’ or ‘limestone’ is a type of ‘stone’), none of the models showed great performance in capturing deeper taxonomic structures and inferring multi-step hierarchical relations, which can in some cases be very complex and reach deep levels, as in Figure1.

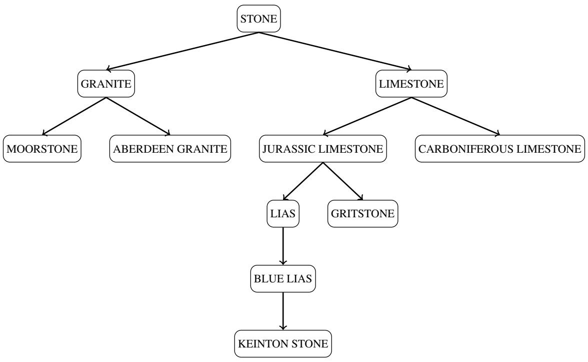

Figure 1: Hierarchical relations of stone types

Generally speaking, Gemini retains and operationalizes more granular domain knowledge, possibly due to a more extensive training dataset or improved alignment mechanisms. Conversely, ChatGPT tended to limit more often its inferences to surface-level relations and classifications (e.g., stone → limestone) and often failed to recognize nested or domain-specific sub-classifications, particularly when terms were more technical or less frequent. These findings suggest that Gemini may have stronger capabilities in ontology-oriented processing, making it better suited for tasks involving the structuring or expansion of specialized terminologies.

Both LLMs, however, may still provide valuable support to the terminographer, especially in the identification of first-level semantic relations—such as broader term (BT), narrower term (NT), and related term (RT) links—across well-represented conceptual domains. While Gemini may offer more depth in modeling nested hierarchies and domain-specific nuances, ChatGPT can nonetheless assist in generating baseline classifications or verifying established semantic patterns, particularly when supplemented with domain-specific prompts or curated input examples.

5 Conclusion and Future Works

This study investigated the ability of LLMs to infer semantic relations between terms in a minimal-input, zero-shot setting, without access to external context, training data, or fine-tuning.

Our results suggest that, while LLMs demonstrate promising capabilities in identifying first-level types of conceptual relations, they still struggle with a deeper level (second and third level) of analysis.

One of the key contributions of this work is to highlight that LLMs encode a certain degree of structured semantic knowledge that can be activated even in the absence of linguistic context or definitional cues. At the same time, our findings underline the limits of this implicit competence, especially when models are required to infer second-level hierarchies.

From a terminological practice perspective, these insights suggest that LLMs could serve as lightweight tools for semi-automatic relation extraction, particularly in the early stages of resource construction or when dealing with low-resource domains. However, their use should be complemented with expert validation or human-in-the-loop, given the high degree of variability and wrong hierarchy structuring.

As future works, several directions emerge from this initial study: further analysis could take into account different prompting strategies (e.g., few-shot, chain-of-thought) to improve the quality and explainability of relational inferences. Moreover, extending the range of relation types (e.g., made-of, causes, used-for) would allow for a broader assessment of the models' semantic competence. Future experiments may also focus on reproducing this experiment on different domain-specific term lists (e.g., in medicine, law, or engineering) as well as on other languages in order to generalize the output results and evaluation.

References

Shubhanker Banerjee, Bharathi Raja Chakravarthi, and John Philip McCrae. 2024. Large language models for few-shot automatic term extraction. In *International Conference on Applications of Natural Language to Information Systems*, pages 137–150. Springer.

Philipp Cimiano, Andreas Hotho, and Steffen Staab. 2005. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of artificial intelligence research*, 24:305–339.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, Hong Kong, China. Association for Computational Linguistics.

Pamela Faber. 2022. Frame-based terminology. In *Theoretical Perspectives on Terminology*, pages 353–376. John Benjamins Publishing Company.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Pilar León-Araúz and A San Martín. 2018. The ecolexicon semantic sketch grammar: from knowledge patterns to word sketches. *arXiv preprint arXiv:1804.05294*.

Belinda Maia and Sérgio Matos. 2008. Corpografo v4-tools for researchers and teachers using comparable corpora. In *quot; In Pierre Zweigenbaum; Éric Gaussier; Pascale Fung (ed) Proceedings of the 6 th International Conference on Language Resources and Evaluation; LREC 2008 Workshop on Comparable Corpora (LREC 2008)(Marrakech 28-30 May 2008; 31 May 2008) European Language Resources Association (ELRA); 79-82*. European Language Resources Association (ELRA).

Ingrid Meyer. 2001. Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. In *Recent advances in computational terminology*, pages 279–302. John Benjamins Publishing Company.

Alistair Miles and Sean Bechhofer. 2009. *SKOS Simple Knowledge Organization System Reference*. W3C Recommendation. World Wide Web Consortium, United States.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Kovan Mzwri and Márta Turcsányi-Szabo. 2025. The impact of prompt engineering and a generative ai-driven tool on autonomous learning: A case study. *Education Sciences*, 15(2):199.

Carlos Olea, Holly Tucker, Jessica Phelan, Cameron Pattison, Shen Zhang, Maxwell Lieb, Doug Schmidt, and Jules White. 2024. Evaluating persona prompting for question answering tasks. In *Proceedings of th e 10th international conference on artificial intelligence and soft computing, Sydney, Australia*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473., Hong Kong, China. Association for Computational Linguistics.

L Spiteri. 2002. Word association testing and thesaurus construction: Defining inter-term relationships. In *Proceedings of the 30th Annual Conference of the Canadian Association for Information Science*, pages 24–33.

Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Julien Delaunay, Antoine Doucet, and Senja Pollak. 2024. Is prompting what term extraction needs? In *International Conference on Text, Speech, and Dialogue*, pages 17–29. Springer.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3):665–707.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Eugen Wüster. 1975. Die ausbildung in terminologie und terminologischer lexikographie.

Author Index