The Role of Artificial Intelligence in Modern-Term Formation

Nigar Sadigova [1]

[1] Azerbaijan National Academy of Science, Institute of Linguistics named after Nasimi
115 H. Cavid Avenu, Baku, Azerbaijan

DOI: 10.22178/pos.115-26

LCC Subject Category: PE1001-1693

Received 27.02.2025
Accepted 25.03.2025
Published online 31.03.2025

Corresponding Author:
nigarismaylova@gmail.com

Abstract. The rapid expansion of scientific knowledge and digital content in the 21st century has significantly increased the demand for accurate, efficient, and scalable terminology development processes. Term formation – the creation, extraction, validation, and standardisation of specialised vocabulary – has traditionally been a manual endeavour by linguists, terminologists, and subject matter experts. However, with the growing complexity and volume of information, traditional methods are no longer sufficient to keep pace. Recent advancements in Artificial Intelligence (AI), particularly in Natural Language Processing (NLP), machine learning, and knowledge representation, are transforming terminology science. AI-powered systems can process vast multilingual and domain-specific corpora to automatically identify candidate terms, evaluate their relevance using semantic and statistical criteria, and suggest standardised forms and multilingual equivalents. These technologies support the automation of term extraction, filtering, clustering of semantic variants, and alignment across languages and disciplines. This article explores the core applications of AI in term formation, including automatic term extraction, term validation and filtering, semantic clustering, and standardisation. It also addresses integrating AI tools in multilingual environments and constructing terminological resources compatible with ontologies and knowledge graphs. While AI introduces speed, scalability, and contextual awareness to terminology management, it also raises challenges related to accuracy, cultural and linguistic sensitivity, and algorithmic bias. The study concludes that a hybrid model – combining AI's computational capabilities with expert human judgment – offers the most promising path toward creating dynamic, inclusive, and globally coherent terminological systems.

Keywords: artificial intelligence (AI); term formation; terminology management; machine learning.

INTRODUCTION

Terminology – the system of terms used within a specific knowledge domain – forms the backbone of effective communication, knowledge dissemination, and data management in every scientific, technical, and professional field. It serves not only as a linguistic tool but also as a conceptual framework that enables individuals and institutions to share ideas, define concepts clearly, and advance research and innovation in a structured manner. From scientific publications and policy documents to technical manuals and academic curricula, precise and standardised terminology ensures mutual understanding and avoids ambiguity.

Traditionally, the process of term formation has relied heavily on human expertise. Linguists, terminologists, and domain specialists have collaborated to coin new terms, validate existing ones, and maintain terminological databases [1-2]. This work involves meticulous research, including analysing primary texts, etymological studies, comparison of term usage across languages, and alignment with existing classification systems. While this manual approach ensures linguistic and conceptual accuracy, it is inherently time-consuming, resource-intensive, and often unsustainable in the face of rapid scientific and technological change.

In the digital era, the pace at which new knowledge is produced has grown exponentially. Fields such as artificial intelligence, biotechnology, medicine, nuclear energy, environmental science, and quantum computing generate new concepts almost daily, requiring timely and accurate term formation. The global nature of knowledge production and dissemination demands that terminologies be developed and maintained across multiple languages and cultural contexts [3]. These trends pose significant challenges for traditional terminology work and highlight the urgent need for scalable, efficient, and adaptive solutions.

Artificial Intelligence (AI) has emerged as a transformative force in addressing these challenges. Through its various subfields – particularly *Natural Language Processing, deep learning, machine learning, and knowledge representation* – AI offers tools and techniques to automate, accelerate, and enhance many terminology development and management aspects. AI systems can process vast corpora of domain-specific texts, identify patterns and concepts, suggest candidate terms, and even generate multilingual equivalents with remarkable speed and consistency.

These capabilities not only support the work of terminologists but also democratise access to terminological tools for a broader community of users, including educators, translators, policymakers, and researchers. Furthermore, AI-driven approaches allow for real-time updating of terminological resources, ensuring they remain relevant and reflective of current usage.

In this context, applying Artificial Intelligence in term formation is not merely a convenience but a necessity. It represents a paradigm shift in how we conceptualise and manage language in the knowledge society – one that blends computational power with linguistic insight to meet the demands of our increasingly complex and interconnected world.

RESULTS AND DISCUSSION

Automatic term extraction (ATE) is one of the most foundational applications of Artificial Intelligence in terminology science. It involves the identification of domain-specific terms from large volumes of unstructured or semi-structured texts. These terms often represent core concepts, processes, or entities within a specialised field and are essential for knowledge organisation, information retrieval, and communication.

AI systems can process vast corpora using advanced Natural Language Processing techniques – from academic publications and technical manuals to patent documents and web content. These techniques analyse the syntactic structure of sentences (such as part-of-speech tagging, noun phrase chunking), semantic relations (such as concept co-occurrence or contextual embeddings), and statistical measures (such as term frequency-inverse document frequency, or TF-IDF) [4].

Unlike traditional manual methods, which are time-consuming and prone to inconsistency, AI-powered approaches enable rapid and scalable identification of candidate terms. These methods can recognise not only single-word terms (e.g., "radiation", "algorithm") but also complex multi-word expressions (e.g., "photoelectric absorption coefficient", "machine learning model") that are more descriptive and often more informative in technical discourse.

Some of the widely used tools in this domain include:

*TermoStat***:** A term extraction tool that analyses the frequency and specificity of terms within a target corpus compared to a general reference corpus.

*AntConc***:** A corpus analysis toolkit that, while not AI-based in the modern sense, is a powerful reference for contrastive and frequency-based term studies.

*spaCy***:** A modern NLP library in Python that can be used for linguistic preprocessing, entity recognition, and pattern matching in text, forming the backbone for many AI-based extraction pipelines.

BERT-based Models: Pre-trained language models like BERT [5] (Bidirectional Encoder Representations from Transformers) offer powerful context-aware embeddings. These models can be fine-tuned to detect technical terminology by learning from labelled domain-specific corpora.

AI models can be trained to adapt to different subject areas, such as medicine, engineering, or law, allowing domain-sensitive term extraction. Deep learning techniques also allow these models to learn implicit patterns and recognise emerging terms that might not yet be included in formal glossaries or dictionaries.

AI-based systems can continuously improve through iterative feedback loops, incorporating expert validation and corrections into the model, making them more robust and accurate. This continuous learning capability makes AI an indispensable ally for modern terminologists, researchers, and policymakers who need timely access to evolving terminological landscapes.

Term Validation and Filtering. Extracting candidate terms from large corpora is the first step in terminology development. Not every extracted phrase or word qualifies as a valid term in the context of a specific domain. Therefore, *term validation and filtering* is a critical stage that involves evaluating the extracted candidates' linguistic, statistical, and semantic relevance. This step ensures that only meaningful, contextually appropriate, and domain-specific terms are selected for inclusion in terminological databases or glossaries.

AI significantly enhances this phase by automating the evaluation process by applying *frequency analysis, contextual embeddings, semantic similarity measures*, and *co-occurrence networks* [6]. These methods collectively allow AI systems to analyse how often a term appears, where, how, and in what semantic context it is used.

Frequency Analysis: AI models compute how frequently a candidate term appears in a target domain corpus compared to a general-language reference corpus. Terms with disproportionately high frequency in the target domain (e.g., "neutron flux" in nuclear physics) are more likely to be domain-specific. However, frequency alone is insufficient to determine termhood, especially for common words that may appear frequently but lack domain relevance.

Contextual Embeddings: Deep learning models such as BERT or RoBERTa generate vector representations of words and phrases based on their usage in context. These embeddings allow the AI system to evaluate whether a candidate term consistently appears in similar semantic environments – a key indicator of terminological stability and domain relevance. For instance, the term "dose-response relationship" would consistently appear in pharmacological or toxicological discussions, reinforcing its validity.

Co-occurrence Networks: These networks analyse how terms appear together within texts. Valid domain-specific terms often have strong and stable associations with other known terms. For example, in a biomedical corpus, the candidate term "cytokine storm" would frequently co-occur with "immune response," "inflammation," and "infection," which supports its validation as a recognised concept.

In addition to these techniques, *machine learning classifiers* – such as support vector machines (SVM), random forests, or neural networks – can automatically be trained on labelled datasets to distinguish between valid and invalid terms. These models learn from various features, including part-of-speech patterns, word length, morphological structures, contextual consistency, and semantic specificity. By analysing these attributes across thousands of examples, the models become increasingly accurate at filtering out irrelevant or ambiguous candidates, such as generic adjectives or overly broad nouns.

AI in validation drastically reduces the manual workload typically required by terminologists, enabling them to focus their expertise on refining and interpreting validated terms rather than sifting through raw data. Moreover, AI systems can continuously improve through feedback loops: as human experts confirm or reject AI-generated suggestions, the models learn from these corrections and enhance their future performance.

Term Standardisation is a crucial phase in terminology management that ensures consistency, clarity, and interoperability in scientific, technical, and professional communication. It involves selecting a preferred form for each concept and establishing uniform usage across documents, databases, institutions, and languages. Inconsistent terminology – such as using different words or acronyms for the same concept – can lead to confusion, reduce information retrieval efficiency, and hinder collaboration across disciplines and borders.

AI plays a transformative role in modernising and automating this standardisation process. It does so by leveraging *semantic analysis, context-aware embeddings*, and *clustering algorithms* to identify and group related terms, suggest standard equivalents, and even highlight outdated or ambiguous usage.

One of the key challenges in term standardisation is managing *term variants* – different expressions that refer to the same concept. These include:

Synonyms: e.g., "magnetic resonance imaging" vs. "MRI"

Spelling Variants: e.g., "tumour" vs. "tumour."

Morphological Variants: e.g., "radiological exposure" vs. "radiation exposure."

Abbreviations and Acronyms: e.g., "polymerase chain reaction" vs. "PCR".

AI models, particularly *embedding-based models* like BERT, word2vec, and fastText [7], can capture the semantic meaning of terms based on the context in which they appear. These models represent words and phrases as vectors in a high-dimensional space, where similar meanings cluster closely together. By analysing these semantic proximities, AI systems can group different lexical forms that refer to the same or closely related concepts. For instance, if "COVID-19," "coronavirus," and "SARS-CoV-2" frequently appear in similar contexts, the system can infer their conceptual relatedness and recommend standardised usage based on authoritative sources.

Furthermore, *clustering algorithms* such as k-means or DBSCAN can automatically be applied to these embeddings to group conceptually similar terms. This allows terminologists or domain experts to review term clusters and select preferred nomenclature based on domain standards, frequency of use, and clarity.

In multilingual settings, AI also aids in *cross-lingual standardisation* by aligning terms from different languages that describe the same concept. Machine translation models and multilingual embeddings (e.g., mBERT, XLM-R) facilitate the mapping of terms across languages, ensuring terminological consistency in international documentation, treaties, technical manuals, and scientific literature.

AI can also assist in *detecting deprecated or obsolete terms*, flagging them for review or replacement. This is especially useful in the fast-evolving medical, information technology, and climate science fields, where discoveries or policy shifts can quickly render older terminology outdated.

Beyond just suggesting term alignments, advanced AI systems can learn *domain-specific preferences* by analysing existing authoritative corpora (e.g., ISO standards, scientific journals, regulatory documents). Based on these insights, they can recommend standardised terms that align with institutional practices and international norms.

Multilingual Term Generation and Translation (MT) has long been a tool for bridging linguistic gaps. Still, recent advances in *neural machine translation (NMT)* and large language models (LLMs) have significantly enhanced the quality, fluency, and contextual accuracy of translated content – particularly in *technical terminology*. Modern AI-powered models, such as *Google's NMT system* and *OpenAI's GPT*, can translate general-purpose language and generate *accurate equivalents of domain-specific terms* across various languages and disciplines [8].

These AI models are trained on vast multilingual corpora and can capture *linguistic patterns and conceptual relationships* between terms. As a result, they can infer the correct equivalents of technical terms—even in cases where those terms have not been formally standardised or documented. This capability is particularly beneficial in *emerging scientific and technological fields* such as quantum computing, synthetic biology, climate modelling, and artificial intelligence. New concepts are constantly being introduced and often lack standardised terminology in many languages. For instance, when translating a newly coined term like "neural radiance field" into languages where the term hasn't been officially adopted, GPT-based models can analyse its components and generate a plausible, semantically accurate equivalent based on how similar terms have been translated. This process may involve breaking the term into interpretable parts, translating each component accurately, and then reconstructing a linguistically coherent and contextually meaningful term in the target language.

These models are highly adaptable. Their performance improves dramatically when fine-tuned or guided by domain-specific data (e.g., bilingual glossaries, multilingual scientific texts). This allows researchers, terminologists, translators, and policymakers to leverage AI in rapidly prototyping *terms* across languages – facilitating early communication, peer collaboration, and international dissemination of scientific findings.

Another significant benefit is AI's support for *multilingual documentation and localisation*. In international organisations, industries, and collaborative research projects, consistent terminology across user manuals, regulatory documents, safety guidelines, and educational materials is often necessary. AI-driven machine translation helps ensure that the *same concepts are described consistently and culturally appropriately*, even in regions with limited access to professional terminologists.

AI tools can also be integrated into *computer-assisted translation (CAT) environments*, where human translators review and refine AI-suggested equivalents. This human-in-the-loop approach enhances productivity and accuracy, translating technical content faster and more reliably than ever.

Ontology and Knowledge Graph Integration. In modern terminology science and knowledge management, *ontologies* and *knowledge graphs* are critical in organising and representing the relationships between terms, concepts, and entities within a specific domain [9]. Ontologies define structured frameworks that capture the hierarchical, associative, and semantic relationships between concepts (e.g., "is-a," "part-of," and "related-to"). At the same time, knowledge graphs provide a visual and computational representation of this interconnected information.

AI significantly enhances these semantic structures' construction, enrichment, and dynamic updating. By aligning newly formed or extracted terms with existing ontological frameworks or knowledge graphs, AI ensures that new knowledge integrates seamlessly into established conceptual ecosystems – thereby supporting consistency, interoperability, and machine-readable understanding.

Traditionally, ontology development required extensive manual input from domain experts and knowledge engineers. However, AI, particularly *Natural Language Processing* and *machine learning***,** now enables the *semi-automated or fully automated creation* of ontologies. These AI systems analyse large corpora of domain-specific text to:

Extract candidate terms and concepts using syntactic and semantic pattern recognition.

Identify conceptual relationships between terms based on linguistic cues (e.g., "X is a type of Y") and statistical co-occurrence.

Propose hierarchical structures (e.g., subclass relationships) and infer associations (e.g., causality, functionality, temporal order). For instance, if AI identifies that the term "alpha particle" frequently occurs in similar contexts with "ionising radiation" and "helium nucleus," it can propose that "alpha particle" is a subtype of "ionising radiation" and is structurally related to "helium."

Knowledge graphs (KGs), such as *Wikidata***,** *ConceptNet***,** or *DBpedia*, are large-scale, structured semantic networks that describe relationships between entities and concepts. AI systems can enrich these graphs by:

Inserting new nodes and edges representing novel terms and their semantic links.

Updating outdated or deprecated information with current terminology.

Detecting and resolving semantic ambiguities using context-aware models like BERT can distinguish between homonyms and polysemous terms based on usage.

AI tools like *Open Information Extraction (OpenIE)* and *ConceptNet* use deep NLP techniques to extract triples (subject–predicate–object relationships) from natural language texts and insert them into knowledge graphs [10]. For example, from the sentence "Neutrons are uncharged particles found in atomic nuclei," an AI system can extract:

(Neutron) is a Particle
(Neutron) has charge **-** *None*
(Neutron) located in the *Atomic nucleus*

These triples can then be mapped to existing concepts in a domain ontology or used to create new entries when no match is found, thereby expanding the knowledge base intelligently.

CONCLUSIONS

Artificial Intelligence (AI) is fundamentally transforming the field of terminology science, marking a paradigm shift in how terms are identified, validated, standardised, and maintained across languages and disciplines. As the global flow of information accelerates, the demand for precise, consistent, and accessible terminology has never been greater. AI technologies are uniquely positioned to meet this challenge by providing scalable, data-driven solutions that complement and enhance human expertise. From the automatic extraction of domain-specific candidate terms to their semantic validation and eventual standardisation, AI tools offer speed, accuracy, and adaptability far beyond traditional manual methods. These tools can process vast multilingual corpora, detect nuanced contextual relationships, align synonyms and variants, and suggest culturally appropriate and linguistically accurate equivalents across languages. In doing so, AI supports terminologists and linguists in their work and expands the reach and relevance of terminology science to fields such as information retrieval, data science, policy-making, healthcare, and education.

AI has the potential to democratise term formation by making advanced linguistic tools accessible to non-specialists, including subject matter experts, educators, and multilingual content creators. Users can interact interactively with terminology resources through intuitive interfaces and human-in-the-loop systems, promoting broader participation in developing and refining specialised vocabulary.

However, integrating AI in term formation is not without its challenges. Ensuring *semantic accuracy, domain relevance*, and *cultural sensitivity* remains critical. AI systems can sometimes produce contextually inappropriate or biased suggestions due to limitations in training data or algorithmic design. There is also the risk of over-standardisation, where linguistic diversity and nuance may be lost in favour of efficiency. These concerns highlight the continued importance of human oversight and collaboration between computational and linguistic communities.

The synergy between AI and human expertise is promising for building dynamic, inclusive, and forward-looking terminological ecosystems. As AI systems become more context-aware and explainable, their role in supporting interdisciplinary communication, preserving linguistic diversity, and fostering innovation in knowledge representation will only grow. The future of terminology science lies not in replacing human judgment with machines but in *harnessing the strengths of both* to navigate the complexity of language in a rapidly changing world.

Acknowledgement

This project has received funding from the Azerbaijan Science Foundation AEF-MGC-2024-2(50)-16/04/1-M-04

REFERENCES

1. IAEA. (2022). *Nuclear Safety and Security Glossary*. doi: 10.61092/iaea.rrxi-t56z

2. Sadigova, N. (2024). Terminology of Nuclear Energy Field based on the English Language. *Path of Science, 10*(9), 5007–5013. doi: 10.22178/pos.108-14

3. Sadigova, N., & Aliyeva, S. (2025). Methods of terms formation in nuclear medicine. *Open Research Europe, 5*, 46. doi: 10.12688/openreseurope.18941.1

4. Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology, 9*(1), 99–115. doi: 10.1075/term.9.1.06dro

5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT* (pp. 4171–4186). Retrieved from https://arxiv.org/abs/1810.04805

6. Cambria, E., & White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research. *IEEE Computational Intelligence Magazine, 9*(2), 48–57. doi: 10.1109/mci.2014.2307227

7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, 30. Retrieved from https://arxiv.org/abs/1706.03762

8. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). *Google's Neural Machine Translation System: Bridging the human and machine translation gap*. Retrieved from https://arxiv.org/abs/1609.08144

9. Navigli, R., & Velardi, P. (2004). Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics, 30*(2), 151–179. doi: 10.1162/089120104323093276

10. Haciyeva, N., & Sadiqova, N. (2023). Terminology in nuclear energy in the Azerbaijani language based on English. *Terminology Issues, 01*, 18. doi: 10.59849/2663-8967.2023.1.18