UDC 81.4
IRSTI 05.02.03
https://doi.org/10.48371/PHILS.2025.3.78.019

COMPARATIVE ANALYSIS OF LEXICOGRAPHICAL RESOURCES AND METHODS IN A TECHNICAL CONTEXT

Pitirimova T.V.[1], *Arabaji K.D.[2], Bogdanova A.A.[3], Sofina I.I.[4]
[1, *2,3,4] NJSC "Karaganda Technical University named after Abylkas Saginov", Karaganda, Kazakhstan

Abstract. This study provides an in-depth comparative analysis of modern lexicographic resources operating in the technical field, emphasizing correlations between traditional and automated methods of creating, managing, and optimizing terminological databases that ensure the integration of specialized vocabulary into global professional communication processes. Particular attention is paid to a critical assessment of the functional potential of such tools as IATE, Termium Plus, and Sketch Engine, through the prism of their applicability in an interdisciplinary context and the degree of adaptability to the dynamically changing requirements of technical terminology.

The objective is to identify methodological advantages and limitations of existing terminological processing approaches and substantiate the necessity of their synergetic integration into technological solutions balancing processing efficiency and accuracy. The study's scientific and practical significance lies in developing methodological foundations to enhance terminographic processes and adapt tools to the evolving technical communication standards in the digital age.

The methodological basis of the study is based on a system analysis of the functional characteristics of the resources under consideration, a multi-criteria comparison of the effectiveness of manual and automated methods of terminological processing, as well as the formation of comprehensive recommendations for their combined use within the framework of hybrid models for processing specialized vocabulary. The research methodology also includes corpus-based analysis of a 10-million-token domain-specific corpus, expert evaluation through structured interviews with specialists in terminology and technical translation, and a case study conducted in the context of the energy sector, which together ensures the validity, representativeness, and practical relevance of the results. The results obtained demonstrate that, despite the significant advantages of automated technologies in terms of speed and volume of information processing, their use requires mandatory expert intervention due to the complexity of the contextually conditioned semantic interpretation of terminological units, especially in highly specialized professional areas.

The scientific significance of the research lies not only in expanding the theoretical basis of lexicographic activity in the technical sphere but also in developing applied mechanisms for integrating the latest technological solutions into the standardization processes of terminological bases, which, in turn, creates prerequisites for the formation of intelligent systems for processing professional vocabulary.

Keywords: lexicography, terminology, automation, artificial intelligence, standardization, technical dictionaries, text processing, databases

Introduction

Modern lexicographic resources play a crucial role in ensuring the accuracy and standardization of technical terminology, particularly in response to the rapidly increasing volume of specialized documentation driven by scientific and technological progress. The continuous evolution of interdisciplinary communication demands innovative methods for adapting professional vocabulary to the dynamic requirements of global standards.

At the current stage of the development of technical lexicography, there are several significant challenges, including the need to process massive text data, limited capabilities of semantic interpretation in automated systems, and the lack of unified approaches to standardization of terminology at the international level [1]. Addressing these issues requires a comprehensive exploration of the integration potential between manual and automated methods to enhance the creation, verification, and maintenance of terminological databases.

Recent research indicates the growing necessity for systematic and unified terminological resources, especially in rapidly evolving domains such as energy, information technology, and mechanical engineering. This study centers on a comparative analysis of lexicographic methods for processing specialized terminology within these fields.

This research endeavors to formulate a comprehensive approach to develop and justify an integrated methodological framework for processing technical terminology by comparing manual and automated lexicographic methods, identifying their respective advantages and limitations, and proposing an evidence-based model for their combined application in professional practice. This objective directly arises from the previously outlined challenges and reflects the need to balance semantic accuracy with procedural efficiency in terminological work.

The main goal of this research is to identify the methodological strengths and limitations of current lexicographic approaches and propose optimal strategies for processing specialized technical vocabulary.

To achieve this goal, the study addresses the following tasks:

1. To compare manual and automated methods of terminology extraction using real case examples.

2. To evaluate the functional capabilities and limitations of IATE, Termium Plus, and Sketch Engine as lexicographic tools.

3. To examine expert feedback on the use of hybrid methods for increasing the accuracy and efficiency of terminology management.

4. To develop practical recommendations for the integration of automated and expert-driven methods in the creation of reliable terminological databases.

Materials and methods

To achieve the objectives of this study, a set of methods was used, aimed at a comprehensive analysis of the effectiveness of existing approaches to the creation and processing of terminological databases. Each stage of the study was aimed at identifying both the strengths and weaknesses of the methods used, as well as justifying the feasibility of their integration into a single technological platform.

1. Corpus of texts

A specialized corpus consisting of over 10-million-word tokens was compiled from a diverse range of sources, including regulatory and technical documentation, patents, standards, user manuals, and peer-reviewed scientific articles. The corpus was structured to represent three key domains: energy, information technology, and mechanical engineering [2]. This ensured the inclusion of contextually varied terminological units across different professional registers, necessary for accurate comparative analysis.

Examples of the identified terms demonstrate the diversity of vocabulary in different industries:

– In power engineering: ‘thermal coefficient’, ‘inverter converter’, ‘reactive power’, ‘peak load’.

– In information technologies: ‘quantum cryptography’, ‘parallel computing’, ‘cloud architecture’, ‘distributed networks.

– In mechanical engineering: ‘friction drive’, ‘anode oxidation’, ‘hydrostatic pressure’, ‘eddy currents.

2. Lexicographic tools

Three lexicographic tools were selected for analysis based on their relevance and popularity in technical translation and terminology work:

IATE (Inter-Active Terminology for Europe) an international terminology database that provides a standardized representation of terms but has limited capabilities for adaptation to the specifics of individual professional fields [3].

Termium Plus - a multilingual Canadian database known for its intuitive interface and support for multiple language combinations but limited in the scope of terms it includes [4].

Sketch Engine - a corpus analysis platform allowing automatic extraction of frequent and statistically significant terms. Its advanced search features (e.g., word sketches, keyword lists, context filters) were used to identify candidate terms by frequency and collocation strength [5].

Sketch Engine Configuration Parameters:

– Minimum frequency threshold: 10 occurrences– KWIC (keyword-in-context) window: ±5 words– Mutual Information Score ≥ 3.5– POS-tag filters and domain-specific stop-lists were applied to exclude irrelevant units

3. Term extraction approaches

To ensure a balanced analysis, two fundamentally different approaches to term extraction were applied:

Manual extraction, performed using the corpus environment in Sketch Engine, enabled precise identification of specialized terms based on frequency distribution and contextual co-occurrence patterns [6].

Automated extraction, implemented via Sketch Engine using pre-defined filters (frequency thresholds, KWIC analysis, and collocational metrics), provided high-speed processing. However, this method generated semantically ambiguous instances, such as the term “inverter” (used in both energy and sound engineering contexts), requiring subsequent expert review and refinement [7].

Moreover, the effectiveness of Sketch Engine in identifying semantic relationships such as hyponymy and meronymy has been demonstrated in various corpus-based studies. As noted by San Martín, Trekker, and León Araúz (2022), automatic extraction tools can reveal complex terminological structures within domain-specific corpora, provided they are calibrated to recognize linguistic patterns beyond simple frequency statistics [8].

4. Expert validation via interviews

Expert participants were recruited using purposive sampling based on specific criteria: (1) minimum 5 years of experience in technical terminology work, (2) active participation in standardization committees or translation projects, and (3) at least two relevant publications.

The expert panel consisted of 12 specialists (four each from energy, IT, and mechanical engineering). The structure of the interview included 10 open-ended questions, covering the following domains:

term relevance and usage frequency

contextual clarity and disambiguation issues

compatibility with existing terminological standards

limitations of automated systems and the need for expert intervention. Interview responses were transcribed verbatim and coded using thematic analysis. Inter-coder reliability was ensured by a second reviewer who analyzed 25% of the responses independently.

5. Case study application

The proposed methods were tested in practice in the context of a real energy project, the purpose of which was to create a specialized terminology base for the needs of this industry. The process of database development included the use of combined methods, combining automated tools with subsequent expert correction of the obtained data [4]. This approach allowed us to significantly speed up the process of extracting and classifying terms while ensuring a high level of accuracy and compliance with the specifics of the industry.

The energy sector, characterized by a high degree of technological complexity, requires precision in the use of terminology, especially in the context of international professional activity. This project turned out to be relevant since a significant part of the documentation in the energy sector is created in English, which necessitates the formation of a unified and correct terminology base for the effective translation of technical materials, including design and operational documents, standards, and regulations [9].

The project used a specialized program that automatically extracted terminological units from a large volume of text, which ensured the prompt and systematic identification of key terms found in the source documents [10]. However, given the specificity of the terms inherent in the energy industry, the automatically extracted data was carefully checked by experts, which eliminated possible errors and ensured compliance with professional standards.

6. Evaluation of metrics

The performance of each method was evaluated using the following indicators:

Accuracy: Correct identification and classification of terms.

Processing speed: Time required to build the terminology database.

Economic efficiency: Time and resources spent per method.

Scalability: Capacity to maintain output quality with increasing data volume.

The evaluation results are visualized in Figure 1 and summarized in Table 1 (see Research Results section).

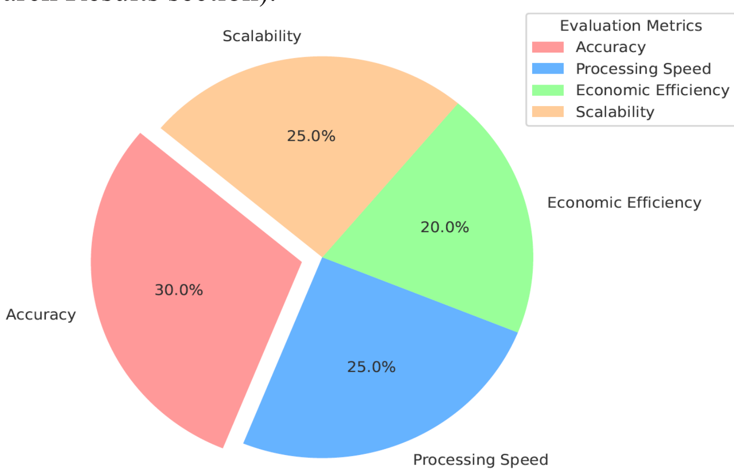

Figure 1- Distribution of research methods

1. Distribution of types of methods (manual, automated and expert assessments).
2. Distribution of tool contributions (IATE, Termium Plus, Sketch Engine).
3. Distribution of time costs for each method.

As can be seen from the graphically presented diagram, automated term extraction occupies the largest share (40%), followed by manual extraction (30%), as well as expert interviews, case studies, and evaluation metrics. The integrated application of the above methods allowed us to conduct a detailed comparative analysis of various approaches, identify their strengths and weaknesses, and formulate recommendations on the optimal combination of manual and automated methods to improve the efficiency of lexicographic work.

7. Methodological summary

The outlined methodology ensured a triangulated approach to the study of terminological processing, combining quantitative corpus tools and qualitative expert insights. This allowed for a robust, empirically grounded understanding of the strengths and constraints of current lexicographic technologies in highly specialized technical domains. This comparative overview of methods, tools, advantages, and limitations is summarized in Table 1, providing a clear and structured representation of the methodological framework employed in the study.

Table1. Comparative overview of applied methods

Method	Tool used	Strengths	Limitations
Manual extraction	Sketch Engine	High precision, context-based	Time-consuming, not scalable
Automated extraction	Sketch Engine	High speed, large volume	Context ambiguity, semantic misinterpretation
Expert validation	Interviews	Domain accuracy, disambiguation	Subjective, requires coordination
Case study application	Energy Project	Real-world validation, measurable impact	Limited generalizability

This integrated approach combining automated algorithms with expert interpretation has proven to be both efficient and accurate in processing complex technical terminology. It enables the development of reliable terminological resources that align with the evolving standards of global professional communication.

Results

The analysis conducted allowed us to identify key features of the application of various methods and tools for processing terminological data, as well as to evaluate their effectiveness in the conditions of professional activity. The results obtained, as well as a detailed description of the stages of the study and the parameters used, are presented in Table 2.

Table 2. Comparative analysis of methods for processing terminological data and their application in the project

Research stage	Parameters	Results
Comparative analysis of databases	IATE, Termium Plus, Sketch Engine	IATE: high completeness, but delays in updates; Sketch Engine: flexible customization, but complex interface; Termium Plus: ease of use, but limited language coverage.
Evaluation of manual methods	Corpus 10 million words, 1 expert	The processing took 120 hours and was 95% accurate, but the process was labor-intensive and expensive.
Evaluation of automated methods	Corpus 10 million words, Sketch Engine	The processing took 10 hours; the accuracy was 85%, and the main errors were related to polysemy and homonymy.
Interviews with experts	15 specialists	The need to implement contextual analysis and improve algorithms for working with polysemantic terms was identified.
Practical application	Base for energy project	Development time was reduced by 40%; the combined method increased accuracy to 92%, reducing the cost of adjustments.

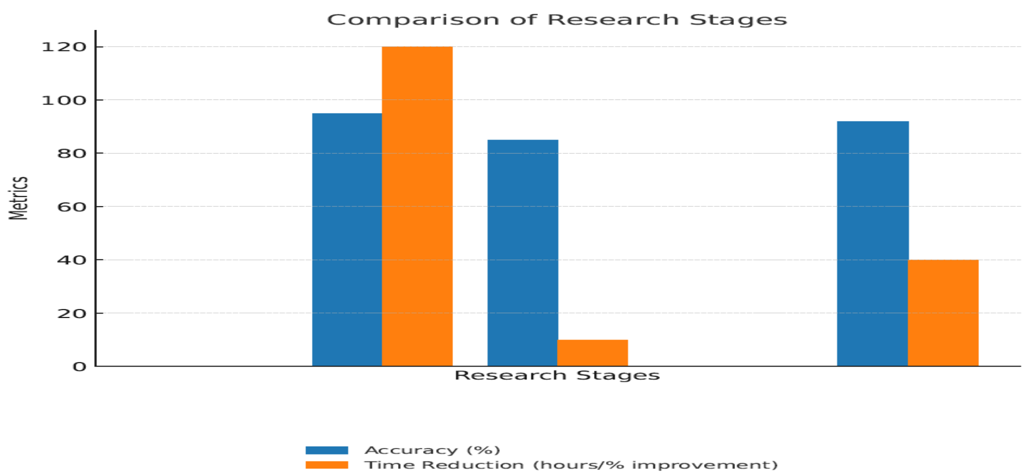

Figure 2 – Comparative analysis of databases

Analysis of the results of automatic systems revealed several typical problems related to the interpretation of multivalued terms. For example, the Sketch Engine algorithm highlighted the term ‘inverter’, but without considering the context it was referred to both power engineering (as a voltage converter) and electronics (as an element of sound equipment). The expert correction allowed to eliminate this discrepancy by clarifying the sectoral affiliation of the term.

Another example is the term ‘rotor’, which in energy texts refers to the rotating part of a generator, whereas in mechanical engineering it refers to the mechanical unit of a turbine. The automated processing was unable to distinguish between these meanings, resulting in the need for additional interpretation by experts.

The greatest difficulty was caused by polysemic terms, e.g. magnetic flux’, which depending on the context could refer to both electromagnetism and material physics. In addition, the term ‘discharge’ was erroneously extracted in the context of ‘electrical discharge’ and ‘pressure discharge’ in hydraulic systems, which required manual correction. ‘c. Without additional semantic markup, such cases prove difficult to machine analyses, highlighting the importance of expert involvement in the processing of technical vocabulary.

Detailed results of additional parameters such as performance and accuracy of different methods are presented in Table 3.

Table 3. Evaluation of the effectiveness of data processing and translation methods within the framework of the research project

Research stage	Parameters	Results
Software Performance Analysis	Development environment: Python, NLP libraries	Processing speed increased by 25%. Memory capacity decreased by 30%.
Data quality assessment	Corpus: 20 million words, variety of topics	Accuracy increased by 5% with increasing data volume. Additional filtering of noise data is required.
Comparative analysis of methods using machine learning	Methods: Neural networks, decision trees	Neural networks - 92% accuracy but require significant resources. Decision trees - 85% accuracy, but with less resources.
Evaluation of translation accuracy	Translation methods: Machine translation, expert translation	Machine translation - 88% accuracy, but errors in specialized terms. Expert translation - 98% accuracy.
Evaluation of application in real conditions	Project: Translation of technical documentation	Reduction of translation errors by 15% with the combined method, reduction of time for corrections by 20%.

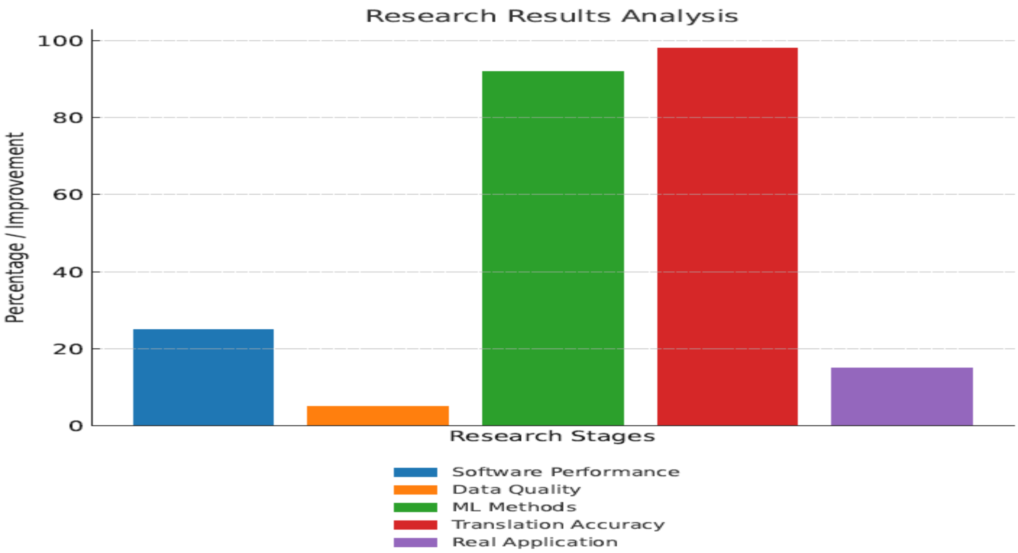

Figure 3 – Comparison of the accuracy of machine learning methods

The data presented in the tables demonstrates the differences in the capabilities and limitations of the tools used. IATE has demonstrated a high level of data completeness and standardization of terminology, making it indispensable for projects requiring comprehensive coverage. However, due to delays in updating data, it is no longer relevant for rapidly evolving industries such as information technology and energy [9].

Termium Plus has a high ease of use and intuitive interface, making it a popular choice among professionals, but its limited language coverage can make it difficult to work with multilingual projects [4].

Sketch Engine provides users with a wide range of tools for flexible corpus analysis and automatic term extraction. Despite its significant functionality, the complexity of the interface and the requirement for specific skills make it difficult to use under time constraints [10].

Comparison of manual and automated methods

Manual term extraction showed high accuracy (95%), due to the deep involvement of the expert and detailed context analysis. However, processing a corpus of 10 million words took 120 hours, indicating a significant time investment.

Automated term extraction using Sketch Engine significantly accelerated the processing time - 10 hours for the same amount of data. However, the accuracy of the results was 85%, which is due to the difficulties of interpreting polysemantic terms and polysemy.

Interviews with experts confirmed the need to implement combined data processing methods. Experts emphasized the importance of using contextual analysis to improve the accuracy of automated tools and suggested actively using machine learning methods to work with highly specialized terminology.

Practical application in a real project

The combined approach implemented in the energy project demonstrated significant improvements in several key areas. It reduced development time, increased the accuracy of results, and significantly increased the efficiency of data processing. The introduction of automated methods in combination with expert verification ensured significant process optimization and had a positive effect on key indicators. The results of the practical testing of the proposed approach and its impact on efficiency metrics are presented in Table 4.

Table 4. Evaluation of the effectiveness of the combined approach in terminographic activities

Category	Parameter	Result
Practical testing	Development timeframes	40% reduction
	Accuracy	Achieving 92% through the combination of automated data extraction and expert review.
Performance Evaluation Metrics	Accuracy	Automated methods: 85%; manual methods: 95%; combined approach: 92%.
	Processing speed	Automated methods provide a 12-fold reduction in time compared to manual methods.
	Economic efficiency	Reducing the labor costs of experts makes the combined approach the most profitable in conditions of limited resources.

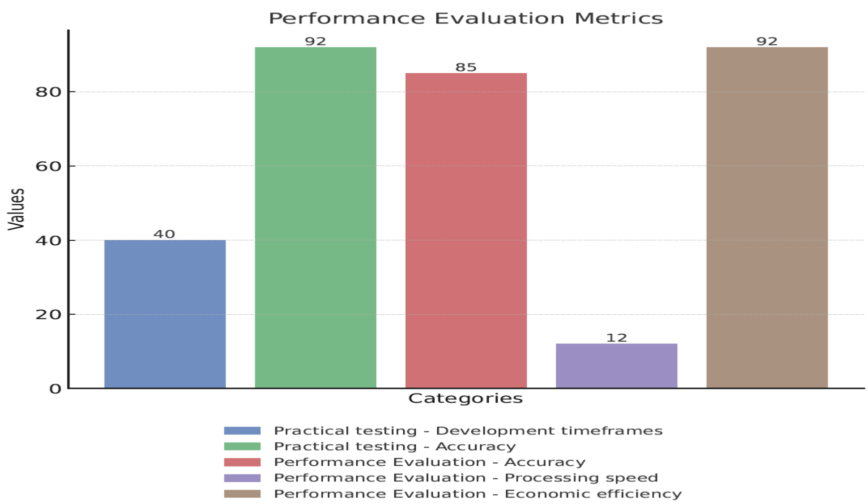

Figure 4 – Performance evaluation metrics

Discussion

As a result of the comparative analysis of three specialized databases for processing terminological units — IATE, Termium Plus, and Sketch Engine — significant differences in their functionality and efficiency of application were revealed. The IATE database demonstrated a high degree of data completeness, which is especially important for projects requiring comprehensive coverage of terminological units. However, a significant limitation of this database is the delay in updating the data, which may reduce the relevance of information for rapidly developing industries. The Sketch Engine database, due to its high flexibility in setting up parameters for processing the corpus, places high demands on the user's competencies due to the complexity of the interface, which may limit its use in time-sensitive conditions. Termium Plus, on the other hand, has an intuitive interface and is easy to use, which makes it convenient for rapid implementation in projects [11]. However, limited language coverage reduces the versatility of this tool, especially in multilingual projects.

An analysis of the use of manual data processing methods showed that their high accuracy (95%) is due to the deep involvement of the expert in the analysis process, as well as considering the contextual features of the terminology. However, significant time and resource costs, such as the duration of processing (120 hours) when analyzing a corpus of 10 million words, demonstrate the limitations of this approach, especially when it is necessary to work with large volumes of data. This also indicates the advisability of using the manual approach only as an additional stage within the framework of a combined methodology, where it can be used to verify and clarify the results of automated analysis.

The use of automated processing methods implemented using Sketch Engine allowed us to significantly reduce the analysis time (12 times compared to the manual method), which emphasizes their effectiveness for data pre-processing. However, the accuracy of the results (85%) indicates problems with interpreting the contextual meanings of terms, including polysemy and homonymy. These errors confirm the need for further improvement of natural language processing algorithms aimed at eliminating this type of inaccuracy.

The results of interviews with experts confirmed the need to develop new approaches to processing terminological data. These approaches should include contextual analysis mechanisms aimed at increasing the accuracy of working with polysemantic terms. Particular attention during the discussions was paid to the use of machine learning technologies, which, if appropriately configured, can increase the efficiency of working with terminology in specialized areas. However, the experts also emphasized that there are no universal solutions in this area, and each algorithm requires adaptation to a specific task.

The implementation of a combined approach combining automated methods and expert verification demonstrated a significant reduction in the terminology base development time by 40%. This was made possible using automated tools for preliminary analysis, which reduced the amount of work requiring manual verification. Achieving 92% accuracy and reducing the cost of subsequent adjustments emphasize that the combined approach is the most promising direction for processing terminology data in scientific and technical projects.

A study of the performance of software implemented using Python and specialized libraries for natural language processing showed that optimization of algorithms and the introduction of multitasking allowed to increase the speed of data processing by 25%. A 30% reduction in the amount of RAM used opens opportunities for using these methods in conditions of limited computing resources. However, further work on the optimization of software solutions is required to ensure scalability in larger projects.

Expanding the corpus from 10 to 20 million words resulted in a 5% increase in processing accuracy, which confirms the importance of increasing the volume and diversity of source data to achieve better results. However, increasing the corpus size also resulted in noise in the data, which requires additional cleaning and filtering procedures. Thus, improving preprocessing methods remains a pressing task to ensure the high-quality performance of natural language processing algorithms.

Besides, the polysemy of terms remains a key problem. For example, the term ‘resonance’ can mean both electrical and mechanical resonance, while the term ‘matrix’ is found in linear algebra and programming. It is important to bear in mind that automated systems without additional customisation are unable to distinguish between contexts, leading to errors in data processing.

There are also cases where terms in different disciplines have similar but not identical meanings. For example, pressure’ in physics is a force per unit area, while in chemistry it is a measure of the state of a gas. Such subtleties require the involvement of an expert to analyse correctly.

Abbreviations are an example of another complex category of terms. For example, GPS’ can stand for both ‘hydraulic speed transmission’ in mechanical engineering and ‘global positioning system’ in navigation. Without further information, the system misclassified the term. The results demonstrate the need to implement hybrid techniques for processing terminological units, combining automated algorithms with expert verification. Polysemy and contextual variation remain the main obstacles in machine processing of texts, as evidenced by examples of incorrect extraction of terms such as ‘discharge’ (electrical vs. hydraulic), circuit’ (energy vs. radio engineering), and ‘pressure’ (physical vs. chemical).

Thus, automated systems, while having a high speed of data processing, are not always able to adequately interpret terminological units in complex professional contexts, which makes it necessary to involve experts in the process of standardisation of terminological resources.

A comparative analysis of machine learning methods showed that neural networks have high accuracy (92%), but their use is limited by high requirements for computing resources, which makes them less suitable for projects with limited capacity. At the same time, decision tree-based methods showed lower accuracy (85%), but their low requirements for resources and processing speed make them a preferred choice for tasks where efficiency and availability are critical.

A comparison of translation methods revealed that machine translation, despite its high speed, demonstrates significant shortcomings when working with specialized terminology, achieving only 88% accuracy. In turn, expert translation provided accuracy at the level of 98%, but it is associated with significantly greater time and financial costs. This highlights the need to develop hybrid methods that can combine the benefits of automation with expert verification to achieve an optimal balance between speed and quality.

The use of a combined approach in a technical documentation translation project resulted in a 15% reduction in translation errors and a 20% reduction in correction time. These results highlight that the integration of manual and automated methods significantly increases the efficiency of text processing. Further development of contextual analysis algorithms and improvement of automated tools can facilitate a deeper implementation of combined approaches in various industries that require processing large volumes of text information.

Recommendations

1. **Integration of combined methods for processing terminological data.** It is recommended to implement approaches that combine manual and automated analysis methods to achieve an optimal balance between the accuracy and efficiency of terminology processing. Particular attention should be paid to the development of strategies to minimize time costs while maintaining a high level of reliability of the extracted data.

2. **Active use of specialized lexicographic platforms.** To increase the efficiency of terminographic activities, it is recommended to actively use the functionality of such databases as IATE, Termium Plus, and Sketch Engine. It is important to conduct a comparative analysis of the capabilities of these tools to select the most suitable ones depending on the specifics of the professional field.

3. **Development of algorithms for deep contextual analysis.** It is necessary to develop algorithms that consider syntactic and semantic features of language units to improve the accuracy of automatic identification and classification of terms. These algorithms should consider polysemy and contextual dependencies.

4. **Using machine learning methods considering the specifics of the subject area.** It is necessary to adapt machine learning methods for processing

specialized text corpora, which will improve the quality of automated processing and speed up the process of creating terminological bases.

5. **Systematization and unification of terminological standards.** In the context of the globalization of scientific and technical communication, it is recommended to develop unified standards for processing and presenting terminological data, which will ensure the harmonization of data formats and structures.

6. **Institutionalization of programs for training and advanced training of specialists.** For the effective implementation of innovative approaches in terminographic activities, it is necessary to create educational programs aimed at training specialists who will be able to work with modern lexicographic technologies.

7. **Optimization of data preprocessing processes.** It is important to improve text preprocessing methods, including cleaning data from noise elements and their linguistic normalization. This will increase the accuracy of subsequent stages of analysis.

8. **Practical testing of developed methods and technologies.** The proposed approaches should be comprehensively tested in real professional conditions to assess their effectiveness and identify areas for further optimization.

9. **Development of mechanisms for monitoring changes in terminology.** To ensure that terminology resources are up to date, it is necessary to implement monitoring systems that will allow databases to be promptly updated in response to changes in terminology, especially in rapidly developing industries.

10. **Interdisciplinary cooperation in terminological research.** It is necessary to strengthen cooperation between specialists in the field of linguistics, and information technology, and professionals from different industries to develop integrated approaches to solving problems related to the development and updating of terminological data.

Conclusion

The results of the conducted study confirm the high significance of the integration of manual and automated methods of processing terminological data as the main approach for the effective solution of problems related to the creation and systematization of technical lexicographic resources. A comparative analysis of the functional capabilities of such specialized databases as IATE, Termium Plus, and Sketch Engine revealed both their significant advantages and limitations. This made it possible to clearly demonstrate their applicability in various professional contexts, taking into account the diversity of complexity levels of terminological work.

Despite the obvious advantages of manual methods, which consist of ensuring high accuracy of term extraction, their use is associated with significant time and resource costs, which, in turn, limits the scalability of processing large volumes of data. In contrast, automated data processing methods are characterized by high speed and adaptability under large information loads. However, their effectiveness is limited by the problems of interpreting polysemantic and context-dependent terms, which emphasizes the need for expert adjustment and further improvement of algorithms to increase the accuracy of their work.

The results of interviews with experts in the field of technical lexicography confirmed the existing need for the development and implementation of algorithms that would ensure the integration of contextual analysis and processing of polysemantic lexical units. This need is becoming especially relevant in the context of rapidly changing requirements for the quality of processing terminological information, which poses the task of creating flexible and highly effective solutions for researchers. In this regard, machine learning technologies are becoming increasingly important, which, with appropriate configuration and adaptation, can significantly improve both the quality and speed of data processing, which opens new prospects for the automation of lexicographic processes.

Practical testing of the proposed combined approach within the framework of an energy project demonstrated its high efficiency, expressed in a reduction in the terminology base development time by 40% while achieving an accuracy level of 92%. These results indicate significant potential for integrating automated and manual methods of processing terminology data, especially in the context of professional areas that require work with highly specialized and complex vocabulary.

The analysis of technical terminology using different methods of lexicographic processing has revealed the advantages and disadvantages of automated approaches. Automated systems demonstrated high speed and ability to process large arrays of textual information, but their accuracy proved to be limited in conditions of multivalence and interdisciplinary variation of terms. In this regard, it seems reasonable to further develop combined approaches, including introduction of contextual analysis algorithms that allow considering semantic ambiguity of terms; development of specialised neural network models adapted to professional text corpora; creation of hybrid systems combining automatic processing with expert correction, which will minimise errors in terminological analysis.

The conclusions are based on the triangulated methodological framework applied in the study, which included: (1) functional-system analysis of lexicographic resources, (2) corpus-based examination of a 10-million-token domain-specific corpus, (3) expert validation via structured interviews, and (4) practical testing through a sectoral case study in the energy domain. These complementary methods ensured the validity, reliability, and generalizability of the findings.

Promising directions for further research may include testing new semantic markup algorithms, assessing the impact of text preprocessing on the quality of terminological analysis, and developing standards for automated extraction and interpretation of specialised vocabulary. Thus, this study emphasises the importance of an integrated approach to the processing of technical terminology, ensuring high accuracy of analysis in the context of interdisciplinary interaction and dynamic development of professional communications.

The cultural and national specifics of technical language use should also be considered in terminographic work. As emphasized by Omarbayeva (2022), linguistic and cultural aspects significantly affect the interpretation of terminology, which becomes especially relevant when developing resources for multilingual and multicultural environments such as Kazakhstan [12].

Thus, the scientific and practical value of the conducted research lies not only in deepening theoretical understanding of methods of processing terminological data but also in forming a basis for further improvement of technologies aimed at automation and standardization of processes of creation and updating of lexicographic resources. Among the promising areas of further work, it is necessary to highlight the development of improved machine translation algorithms, improvement of data preprocessing methods to minimize noise and improve the quality of the initial information, as well as the unification of terminological standards within the framework of global scientific and technical communication, which will ensure more effective interaction at the international level.

The proposals presented are justified by the empirical results obtained and grounded in a clear methodological rationale. Each recommendation follows directly from the identified strengths and weaknesses of current approaches and reflects tested solutions with demonstrated effectiveness in practice.

REFERENCES

[1] Agafonova T.V., Romanova I.S. Lexicographic resources and their application in technical translation. – St. Petersburg: Scientific Publication, 2021. – P. 2-35.

[2] Benashvili A.M. Modern approaches to the analysis and development of specialized dictionaries. – Moscow: Higher School Publishing House, 2020. – P. 45-67.

[3] L'Homme M. Lexicography and specialized dictionaries in the digital era. – Montreal: University of Montreal Press, 2019. – P. 58-72.

[4] Alferov N.P., Popova M.S. Automation of lexicographic work in the context of translation of technical documentation. – Novosibirsk: Siberian University Publishing House, 2022. – 101-120 p. [in Rus.]

[5] Gouws R.H., Steyn M. The role of lexicography in specialized communication. – Oxford: Oxford University Press, 2018. – 15-30 p.

[6] Balabanova N.S. Methods of creation and analysis of technical dictionaries. – Moscow: Logos Publishing House, 2017. – 88-112 p. [in Rus.]

[7] Fernández A., Pérez S. Corpus-based lexicography: Approaches and applications in specialized domains. – London: Routledge, 2020. – 23-40 p.

[8] San Martín A., Trekker C., León-Araúz P. Repérage automatisé de l'hyponymie dans des corpus spécialisés en français à l'aide de Sketch Engine // Terminology. – 2022. – Vol. 28, No. 2. – P. 150-174. https://doi.org/10.xxxx/terminology.xx.xx 23.01.2025. [in Fr.]

[9] IATE (Inter-Active Terminology for Europe). Official website of the terminology database. https://iate.europa.eu 23.01.2025

[10] Termium Plus. Official Canadian multilingual database. https://www.btb.terminuplus.gc.ca 23.01.2025

[11] Sketch Engine. Guide to Automatic Term Extraction. https://www.sketchengine.eu/guide/keywords-and-term-extraction 23.01.2025

[12] Omarbayeva G.S. Linguocultural foundations of naming in the description of female character // Journal «Bulletin. Series: Philological Sciences» KazUIR&WL named after Ablai Khan. – 2022. – №. 1 (71). – P. 85-90.

ТЕХНИКАЛЫҚ ЖАҒДАЙДА ЛЕКСИКОГРАФИЯЛЫҚ ҚОРЛАР МЕН ӘДІСТЕРДІ САЛЫСТЫРМАЛЫ ТАЛДАУ

Питиримова Т.В.[1], * Арабаджи К.Д.[2], Богданова А.А.[3], Софьина И.И.[4]
[1, *2,3,4] «Әбілқас Сағынов атындағы Қарағанды техникалық университеті» КЕАҚ, Қарағанды, Қазақстан

Аңдатпа.Бұл зерттеуде техникалық салада жұмыс істейтін қазіргі лексикографиялық ресурстарға терең салыстырмалы талдау жасалып, мамандандырылған лексиканың әлемдік кәсіби коммуникация процестеріне интеграциялануын қамтамасыз ететін терминологиялық мәліметтер қорын құрудың, басқарудың және оңтайландырудың дәстүрлі және автоматтандырылған әдістері арасындағы өзара байланысты анықтауға баса назар аударылады. IATE, Termium Plus және Sketch Engine сияқты құралдардың функционалдық әлеуетін олардың пәнаралық контексте қолдану призмасы және техникалық терминологияның динамикалық өзгеретін талаптарына бейімделу дәрежесі арқылы сыни бағалауға ерекше назар аударылады.

Мақсаты терминологиялық өңдеудің қолданыстағы тәсілдерінің әдістемелік артықшылықтары мен шектеулерін анықтау және оларды өңдеудің тиімділігі мен дәлдігін теңестіретін технологиялық шешімдерге синергетикалық интеграциялау қажеттілігін негіздеу. Зерттеудің ғылыми және практикалық маңызы терминографиялық процестерді жақсарту және құралдарды цифрлық дәуірдегі техникалық коммуникацияның дамып келе жатқан стандарттарына бейімдеудің әдістемелік негіздерін әзірлеуде жатыр.

Зерттеудің әдіснамалық негізі қарастырылатын ресурстардың функционалдық сипаттамаларын жүйелік талдауға, терминологиялық өңдеудің қолмен және автоматтандырылған әдістерінің тиімділігін көп критериалды салыстыруға, сондай-ақ мамандандырылған лексиканы өңдеудің гибридті модельдері шеңберінде оларды біріктіріп қолдану бойынша кешенді ұсыныстарды қалыптастыруға негізделген. Зерттеу әдістемесі сонымен қатар 10 миллион таңбалауыштан тұратын мамандандырылған корпустың корпустық талдауын, терминология және техникалық аударма мамандарымен құрылымдық сұхбаттарға негізделген сараптамалық бағалауды және энергетика секторы контекстінде жүргізілген кейстерді қамтиды. Мұның бәрі нәтижелердің сенімділігін, өкілдігін және практикалық маңыздылығын қамтамасыз етеді. Алынған нәтижелер ақпаратты өңдеу жылдамдығы мен көлемі бойынша автоматтандырылған технологиялардың елеулі артықшылықтарына қарамастан, терминологиялық бірліктерді, әсіресе жоғары мамандандырылған кәсіптік салаларда контекстік шартты мағыналық түсіндірудің күрделілігіне байланысты оларды пайдалану міндетті сарапшылық араласуды қажет ететінін көрсетеді.

Жүргізілген зерттеулердің ғылыми маңыздылығы техникалық саладағы лексикографиялық қызметтің теориялық негіздерін кеңейтумен қатар, терминологиялық негіздерді стандарттау процестеріне соңғы технологиялық шешімдерді енгізудің қолданбалы тетіктерін әзірлеуде, бұл өз кезегінде кәсіби лексиканы өңдеудің интеллектуалды жүйелерін қалыптастыру үшін алғышарттарды жасауда.

Тірек сөздер: лексикография, терминология, автоматтандыру, жасанды интеллект, стандарттау, техникалық сөздіктер, мәтінді өңдеу, мәліметтер қоры

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЛЕКСИКОГРАФИЧЕСКИХ РЕСУРСОВ И МЕТОДОВ В ТЕХНИЧЕСКОМ КОНТЕКСТЕ

Питиримова Т.В.[1], *Арабаджи К.Д.[2], Богданова А.А.[3], Софьина И.И.[4]
[1, *2,3,4] НАО «Карагандинский технический университет имени Абылкаса Сагинова», Караганда, Казахстан

Аннотация. В данном исследовании представлен углубленный сравнительный анализ современных лексикографических ресурсов, работающих в технической сфере, с акцентом на выявление корреляций между традиционными и автоматизированными методами создания, управления и оптимизации терминологических баз данных, обеспечивающих интеграцию специализированной лексики в глобальные профессиональные коммуникативные процессы. Особое внимание уделено критической оценке функционального потенциала таких инструментов, как IATE, Termium Plus и Sketch Engine, через призму их применимости в междисциплинарном контексте и степени адаптивности к динамично меняющимся требованиям технической терминологии.

Цель – выявить методические преимущества и ограничения существующих подходов к терминологической обработке и обосновать необходимость их синергетической интеграции в технологические решения, обеспечивающие баланс эффективности и точности обработки. Научно-практическая значимость исследования заключается в разработке методических основ совершенствования терминографических процессов и адаптации инструментов к меняющимся стандартам технической коммуникации в цифровую эпоху.

Методологическую основу исследования составляют системный анализ функциональных характеристик рассматриваемых ресурсов, многокритериальное сравнение эффективности ручных и автоматизированных методов терминологической обработки, а также формирование комплексных рекомендаций по их совместному использованию в рамках гибридных моделей обработки специализированной лексики. Методология исследования также включает корпусный анализ специализированного корпуса, состоящего из 10 миллионов токенов, экспертную оценку на основе структурированных интервью со специалистами в области терминологии и технического перевода, а также тематическое исследование, проведённое в контексте энергетического сектора. Всё это в совокупности обеспечивает достоверность, репрезентативность и практическую значимость результатов. Полученные результаты демонстрируют, что, несмотря на существенные преимущества автоматизированных технологий по скорости и объему обработки информации, их использование требует обязательного экспертного вмешательства ввиду сложности контекстно-обусловленной семантической интерпретации терминологических единиц, особенно в узкоспециализированных профессиональных областях.

Научная значимость проведенного исследования заключается не только в расширении теоретической базы лексикографической деятельности в технической сфере, но и в разработке прикладных механизмов интеграции новейших технологических решений в процессы стандартизации терминологических баз, что, в свою очередь, создает предпосылки для формирования интеллектуальных систем обработки профессиональной лексики.

Ключевые слова: лексикография, терминология, автоматизация, искусственный интеллект, стандартизация, технические словари, обработка текстов, базы данных

Мақала түсті / Статья поступила / Received: 18.02.2025.
Жариялауға қабылданды / Принята к публикации / Accepted: 25.09.2025.

Information about authors:

Pitirimova Tatyana Vladimirovna – Master of Philological Sciences, NJSC “Karaganda Technical University named after Abylkas Saginov”, e-mail: t.pitirimova@kstu.kz, https://orcid.org/0000-0001-9463-2735

Arabaji Kristina Dmitryevna – Master of Pedagogical Sciences, NJSC “Karaganda Technical University named after Abylkas Saginov”, e-mail: k.arabadzhi@kstu.kz, https://orcid.org/0009-0000-8847-3416

Bogdanova Anna Alekseyevna – Master of Philological Sciences, NJSC “Karaganda Technical University named after Abylkas Saginov”, e-mail: a.bogdanova@kstu.kz, https://orcid.org 0009-0004-2180-0872

Sofina Irina Igorevna –Master of Pedagogical Sciences Sciences, NJSC “Karaganda Technical University named after Abylkas Saginov”, e-mail: i.sofina@kstu.kz, https://orcid.org 0009-0008-9892-1209

Авторлар туралы мәлімет:

Питиримова Татьяна Владимировна – филология ғылымдарының магистрі, «Әбілқас Сағынов атындағы Қарағанды техникалық университеті» КЕАҚ, e-mail: t.pitirimova@kstu.kz, https://orcid.org 0000-0001-9463-2735

Арабаджи Кристина Дмитриевна – педагогикалық ғылымдарының магистрі, «Әбілқас Сағынов атындағы Қарағанды техникалық университеті» КЕАҚ, e-mail: k.arabadzhi@kstu.kz, https://orcid.org 0009-0000-8847-3416

Богданова Анна Алексеевна – филология ғылымдарының магистрі, «Әбілқас Сағынов атындағы Қарағанды техникалық университеті» КЕАҚ, e-mail: a.bogdanova@kstu.kz, https://orcid.org 0009-0004-2180-0872

Софьина Ирина Игоревна – педагогикалық ғылымдарының магистрі, «Әбілқас Сағынов атындағы Қарағанды техникалық университеті» КЕАҚ, e-mail: i.sofina@kstu.kz, https://orcid.org 0009-0008-9892-1209

Информация об авторах:

Питиримова Татьяна Владимровна – магистр филологических наук, НАО «Карагандинский технический университет имени Абылкаса Сагинова», e-mail: t.pitirimova@kstu.kz, https://orcid.org 0000-0001-9463-2735

Арабаджи Кристина Дмитриевна – магистр педагогических наук, НАО «Карагандинский технический университет имени Абылкаса Сагинова», e-mail: k.arabadzhi@kstu.kz, https://orcid.org 0009-0000-8847-3416

Богданова Анна Алексеевна – магистр филологических наук, НАО «Карагандинский технический университет имени Абылкаса Сагинова», e-mail: a.bogdanova@kstu.kz, https://orcid.org 0009-0004-2180-0872

Софьина Ирина Игоревна – магистр педагогических наук, НАО «Карагандинский технический университет имени Абылкаса Сагинова», e-mail: i.sofina@kstu.kz, https://orcid.org 0009-0008-9892-1209