Thesauri in the modern world: Research and prospects for application

Arap Yespenbetov, Kuralay Tulebayeva, Assem Kassymova, Baurzhan Yerdembekov and Akmaral Smagulova

Abstract: In today's information society, where the amount of available information is constantly growing, the issues of semantic classification and data organization are becoming more and more relevant. Efficient information retrieval and analysis play a key role in scientific and applied fields, requiring innovative tools for semantic processing of texts and words. The study aims to analyse the structure, role, and potential of thesauri by means of statistical and structural analysis methods, as well as analytical-synthetic and comparative methods. The results emphasized the importance of thesauri in providing accurate and structured access to information in various fields. Statistical results showed that the broadest thesaurus categories in Library of Congress Subject Headings (LCSH) were *art*, *library systems*, *medicine*, *culture*, and *media*, followed by *scientific research*, *linguistics*, and *semantics*. The study presented a hierarchy between the subject area of research, thesaurus categories, narrowly focused terms, and ways to improve the classification and presentation of information. For example, the subject area *art and culture* included such thesaurus categories as *sculpture*, *literature*, *painting*, at the same time, the category *sculpture* can include such terms as *sculpture group*, *statue*, *bust.* Among the prospects of thesaurus development, we suggest improvement of information classification quality, efficiency of data analysis, optimization of catalogue search, development of new thesaurus structures, identification of interrelations between terms by means of semantic analysis, improvement of information accessibility of materials in libraries. The practical significance of the research lies in providing a basis for the development of effective strategies for thesaurus tools application in information technology, medicine, education, and art.

Keywords: semantic connections, information technology, text mining, lexical resources, semantic information retrieval

1. Introduction

A thesaurus is a lexical resource that organizes terms by their semantic relationships, such as synonymy and antonymy, to facilitate search and classification (Aitchison et al., 2000). Unlike a standard dictionary, which provides definitions and explanations of word meanings, a thesaurus focuses on the relationships between words, helping you choose the most appropriate term depending on the context (Gilchrist, 2003). A thesaurus does not provide detailed descriptions, but only groups words according to their meaning and function (Bushman et al., 2015). Popular thesauri include WordNet (a lexical database with established relationships between synonyms, hyponyms, and meronyms), Roget's Thesaurus (a thesaurus of English words and phrases), LCSH (Miller et al., 1990).

The study of thesauri, as specialized dictionaries reflecting the comprehensive array of concepts within a particular field of knowledge, is crucial for the systematization of terminology. Competently classifying terms not only enhances the ability to navigate scientific literature but also facilitates the selection of pertinent sources (Shu et al., 2017). In this context, the development of thesauri, based on the analysis of existing systems' quantitative data, is highly relevant. By leveraging pre-existing models, improved thesauri can be constructed, offering larger datasets, more efficient searches, and better categorization of materials. The practical construction of thesauri includes planning and design, vocabulary research, and retrieval systems (Aitchison et al., 2000). Specific tasks include search optimization, improving semantic and terminological relationships between scientific subdisciplines, automatic extraction of terms from text corpora, transforming subject headings and arranging them in priority order, as well as developing relationships between thesauri (Hearst, 1992).

In the current information society, the expanding volume of available data underscores the need for more efficient semantic classification and data organization (Balabekova and Khan, 2022). Effective information retrieval and analysis are pivotal for both scientific inquiry and practical applications. Understanding the structure and interrelationships of words plays a crucial role in optimizing these processes (Martinez and Mammola, 2021). This present research focuses on the role and applications of thesauri, their contribution to the systematization of terminology, and the identification of semantic relationships between terms. Studies in this area have already been carried out, especially on issues of the creation and use of thesauri, as well as methods of their integration into information systems. For example, the work of Kairaktyzy (2022) focuses on the relevance of systematization of terms and automatic processing of texts by means of building thesauri. This study investigated methods and technologies aimed at effective creation of thesaurus structures, which may be important in the context of processing Kazakh textual content. However, no statistical analysis of search queries by subject areas was presented.

Sayakhmet and Tumanova (2022) consider the influence of thesaurus structures on the language training of students, which may be important for pedagogical practices, but their work did not analyse thesauri by subject categories by quantitative methods. The research by Balabekova and Khan (2022) is based on the study of the basics of convergence in media texts, which may be relevant in the context of the modern media environment and its impact on language structures. These findings highlight the importance of thesauri for semantic classification and organization of information. What was lacking, however, is to consider promising directions in the development of thesauri.

At the present stage, there is a need for supervised multilingual thesauri on narrow industry terminology, used to describe specific subject areas. In the work of Bayekeyeva et al. (2021), the authors focused on the systematization of industry terminology in Kazakh, as well as the study of equivalent terminology in Russian and English. Yet, the principles of thesaurus construction were not studied. Since document classification is an important problem, in particular, for state institutions, special attention is paid to text databases. A multilingual EuroVoc thesaurus was investigated by Avram et al. (2021), but the study emphasized the technical capabilities of thesauri at the expense of their structural features.

The presented study aims to help fill these research gaps and to provide a better understanding of the structure and role of thesauri in today's world, as well as to consider the prospects for developing new methods and applications to help optimize information retrieval and analysis. The remainder of the paper is structured as follows. In Section 2 we briefly describe our methodology, and motivate why we chose to base the study on the Library of Congress Subject Headings (LCSH) (2023). Section 3 provides the quantitative and qualitative findings, discussed further in Section 4. We conclude in Section 5 with a summary, and suggestions for future research.

2. Methods

To achieve the goal of the study, which is to analyse in depth the structure, role, and potential of thesauri, we applied a form of methodological triangulation. The main methods – statistical and structural analysis – played a key role in data collection and interpretation. Statistical analysis allowed not only to obtain quantitative indicators but also to display the structure of terms used in different subject areas. Finding the most common terms and their topical organization facilitated the optimization of search engines and the classification of information. Structural analysis revealed hierarchical relationships between terms of different levels, in particular between categories and highly specialized terms. This allowed us to create a clear picture of how terms interact and are distributed across levels, which contributes to a more efficient use of thesauri in information retrieval. The applied comparative analysis allowed for comparisons between different subject areas and terms within them. This allowed us to identify patterns, differences, and possible inaccuracies in the classification of terms, which is important for refining and improving existing thesauri.

The Library of Congress (USA) uses the LCSH system and LCSH thesaurus (2023) to classify and organise collections on various topics, as well as to standardise and facilitate information retrieval in library catalogues. We used it to study thesaurus construction in information systems and their effectiveness in real-life conditions. We chose this system as an example of thesaurus construction due to its high level of detail and accuracy, while also considering the diversity of subject areas such as medicine, linguistics, librarianship, media resources, and information technology.

The analytical and synthetic approaches allowed us to assess the prospects for the development of thesauri, taking into account modern technological advances, such as automated text analysis and semantic technologies. This method helped to develop recommendations for the implementation of new approaches to building thesauri, in particular by integrating the latest technologies, which allows for increasing their efficiency and adaptability to the needs of the modern information environment.

In this paper, we provide a detailed assessment of the construction of thesauri through the study of the LCSH (Library of Congress Subject Headings) system, which is used to classify and organize information resources in US libraries. As mentioned above, we applied various methods of analysis to evaluate the effectiveness of this system and on this basis suggest ways to improve the structure of thesauri through a comparative analysis of different subject areas and terms.

3. Results

The number of results produced by subject headings in the LCSH system (2023) of the U.S. Library is presented in Table 1.

Table 1. Correlation between the number of results issued for different subject areas in the LCSH

Subject area	Number of output results
Art	3,385,841
Library systems	2,585,406
Medicine	2,321,377
Culture	1,229,786
Media	386,622
Scientific research	373,041
Linguistics	35,159
Semantics	13,799

Source: compiled by the authors based on LCSH (2023).

The analysis showed a high degree of structuring of information in various fields. For example, in the subject area of art, it is possible to categorize works by world artists into collections. For example, the query *painting, French* facilitates the search for works by French artists. It yielded the following results: materials available online – 1,102,336, all materials – 1,110,238. The main genres included: newspapers (1,055,940), printed materials (42,770), web pages (1,566), manuscripts (1,097), periodicals (711), photos and drawings (564), films and videos (386), legislative materials (207). The online format included images (1,100,352), PDFs (1,100,288), online texts (1,099,707), web pages (767), videos (382). The materials were categorized by date: from 1900 to 1999, 578,930; from 1800 to 1899, 519,690. The largest number of submissions was for the United States (1,059,248), specifically Washington, D.C. (96,809), District of Columbia (78,235), New York (48,587), and Ohio (44,322).

The number of searches in the medical field was about half that of the subject area of art. For example, the query *cancer treatment* retrieved only 422,726 items, among which 417,796 were in online access. In terms of genre, the following materials were retrieved: newspapers –

403,852, books – 10,259, legislation – 2,517, web pages – 535, periodicals – 506, manuscripts – 414, films and videos – 93, and audio recordings – 23. The output of the results for this query showed that the number of PDFs was 416,550, online texts were 416,026, images were 415,194, web pages were 2.627, videos were 93, and audios were 22. By date, the most materials were from 1900 to 1999, 225,197, and from 1800 to 1899, 189,032. By location, the materials were concentrated in the United States – 404,684, particularly in Washington, D.C., Columbia, Ohio, Virginia, California, and New York. The largest proportion of resources were in English (416,465), with resources in Spanish (2,839), German (1,824), French (1,139), Hawaiian (116), Czech (108), Danish (71) also represented.

The following results were obtained for the query *quantum physics* in LCSH (2023): a total of 8,057 materials are available and 2,755 are available online. In terms of genre, the following categories were distinguished: books and printed materials 1,861, newspaper materials 742, web pages 352, periodicals 57, and film and video 28. Manuscripts (19), legislative materials (17), audio files (2), and web archives (1) were also represented. Online formats included the following: PDF (2,641), images (2,320), online texts (2,197), web pages (46), videos (28), and EPUB (5). The time ranges from 1800 to 1899 (1,176), and 1900 to 1999 (959) had the most materials. By location, the materials were from the United States (825), specifically the District of Columbia (349), Washington, D.C. (333), New York (76), Virginia (61), and Indiana (52). Materials were provided in English (2,374), German (28), Latin (22), English (15), Italian (10), French (4).

We noted the high accuracy in providing information not only on subject areas, but also on complex topics. In a test of the LCSH (2023) system, the query *environmental consequences of plastic waste in the ocean* yielded 1,263 available items, which were distributed in the following genre categories: newspapers – 671, books and printed materials – 454, web pages – 106, manuscripts – 49, periodicals – 44, legislative materials – 29, films and videos – 5, maps – 1. The following online formats were represented: PDF (1,249), pictures (1,091), online texts (1,048), web pages (34), videos (5). By date, most of the materials were from 2000 to 2024, and by location, the most numerous groups were as follows: USA (750), Washington DC (617), DC (593), Europe (34), Eurasia (29). The studies for this request were submitted in English, German, French, and Italian.

Entering the query *impact of climate change on the global economy* in the Library of Congress Subject Headings system (2,023) identified newspapers (2,536), books (2,536), legislation (447), web pages (332), manuscripts (269), periodical literature (135), films and audio recordings (59). A total of 4.664 materials were found, of which 4.993 were freely available. For the query *research into cancer treatment using immunotherapy* LCSH (2023), recent scientific papers, clinical trials, and publications from 2000-2024 with accurate and up-to-date results were identified (157 in total, 183 in the free access).

When searching for literature on the *renaissance in art* period, the LCSH (2023) system provided high accuracy in providing metadata, taking into account the historical context, authorship, and art features of the era. A total of 84,478 materials were identified, among which 77,652 are online. In terms of genre characteristics, newspaper articles (67,188), books (9,007), web pages (1,056), manuscripts (244), periodicals (235), videos and films (224), photos and drawings (120), legislation (115), audio (7), maps (3), 3D objects (1) were presented. By date, the categories with the most material were 1900 to 1999 (57.082) and 1800 to 1899 (18,255). By location, Washington, D.C. (11,844), Columbia (10,973), New York (9,846), and Florida (5,597) were key. Results were also differentiated by language: English (75,489), French (888), German (884), Spanish (340), Italian (59).

For example, the LCSH system (2023) showed a high relevance of the metadata when searching for materials on the author Gabriel Garcia Márquez, reflecting the work and the time frame, which facilitates the understanding of the context of the works. A total of 4,800 materials were found, of which 3,839 are freely available. The genre specificity of the results produced was as follows: periodicals – 1,955, newspapers – 1,256, books – 532, web pages – 172, films, and videos – 20. By location, the majority of materials were related to the United States (1,344) and

Mexico (991), and by language criterion, materials were in Spanish (2,183), English (1,657), and Portuguese (34).

When using LCSH (2023), a high level of usability was observed, as reflected in accessible subject headings, logical classification structure, and intuitive interface. For example, stable expressions such as *sustainable development* and *renewable energy* were perceived by the system as coherent terminological units, which contributes to a satisfactory search experience.

When analysing search queries, it was also important to examine the geographical or national coverage of a particular topic. For example, the study of queries on *cybersecurity* showed that for Italian society the term is primarily associated with legislative aspects: *regulation of artificial intelligence in selected jurisdictions*, *law for improved digital network systems*. For American society, the query *cybersecurity* was associated with information security: *cybersecurity and infrastructure security agency*, *cybersecurity and privacy in business transactions*, *cybersecurity risk in election infrastructure*. The results of the LCSH (2023) analysis highlighted the importance of optimizing searches in library catalogues, in particular the use of thesaurus structures significantly reduces the time required to find relevant resources. For example, a query for *ecology* revealed related topics such as *protection of nature* or *sustainable development*.

The use of thesauri in semantic analysis of texts allowed us to reveal connections and relations between terms. For example, when analysing scientific articles on the query *artificial intelligence in medicine* the hierarchy between related terms was determined: from terminology with broader semantics *healthcare industry – healthcare innovation*, *deep medicine* to terminology with narrower semantics: *clinical trial efficiency*, *personalized disease treatment*. The analysis of thesaurus construction principles, such as the analysis of semantic relations of words, allowed the development of approaches to automatic creation of thesaurus structures. The practical significance of thesaurus development consisted in the gradual development of scientific fields. For example, the development of a thesaurus for *blockchain technologies in medicine* has helped advance the study of this promising field, particularly in areas such as *public health hazards* and *resilient data-sharing infrastructure*.

Statistical analysis of subject area data has highlighted the importance of optimizing thesaurus links to improve search accuracy and relevance. An examination of the structure of thesaurus terms in LCSH (2023) revealed a hierarchy reflecting relationships between the field of study, thesaurus categories and highly specialized concepts, as shown in Table 2. This contributed to effective navigation and understanding of the meaning of each term in the context of the overall knowledge system.

Thus, the results of the study emphasized not only the effectiveness of thesauri in search, but also their practical applicability in educational institutions. On this basis, it can be recommended that developers of information systems should pay attention to the integration of thesauri into search engines and educational platforms.

Table 2. Representation of the hierarchy between subject areas, thesaurus categories and specialized terms, as well as their practical application

Field of study	**Examples of thesaurus categories**	**Narrow terminology**	**Practical use**
Arts and culture	Sculpture	*Sculptural group, statue, bust*	Organization of art collections and cultural archives
	Literature	*Historical novel, stylistic figure*	
	Painting	*Landscape, battle scene, portrait*	
Medicine and healthcare	Oncology	*Cancer treatment, adenocarcinoma, oncological disease*	Classification of medical resources, search optimization
	Medical technology	*Pharmaceutical technology, telemedicine, robotic technology*	
	Diseases	*Hepatitis, coronavirus, bronchitis*	
Linguistics and semantics	Semantic analyses	*Semantic core, output of results, request*	Development of automated methods for creating thesauri
	Linguistic methods	*Linguistic experiment, linguistic analysis*	
Scientific research	Artificial intelligence	*Intelligent computer system, training using artificial intelligence*	Analysis of topics of scientific articles, identification of key directions
	Biotechnology	*Microbial biotechnology, genetic engineering, genetic manipulation*	
Media and public debate	Climate change	*Cyclicity in climate change, climate memory*	Comparative analysis of topics in news sources
	Technologies	*High technology, chemical technology, crystallization*	
Library systems	Information Technology	*Communication means, telephone networks, cellular communications*	Optimizing searches in library catalogues, improving accessibility
	Education	*Educational system, accreditation, family education*	

Source: compiled by the authors based on LCSH (2023).

4. Discussion

The results of the study emphasized the importance of thesauri in today’s information environment. It was found that these tools not only improve the processes of data analysis and classification, but may also contribute to the formation of more accurate queries, which is of great importance in the context of increasing amount of information on the web. Thesauri may enhance developing students’ vocabulary by creating links between terms, and thus providing a ground for the understanding of semantic relationships (Zaitseva et al., 2023). Further, their role in improving the efficiency of information retrieval is undeniable, making them an important component for modern search engines and analytical tools.

The study found that incorporating thesaurus technologies into educational programmes may lead to a deeper understanding of the material, leading to improved analytical skills and the development of a critical view of information. This aspect also emphasizes the need to integrate thesaurus technologies into modern information technology. Users' interaction with such tools builds skills for effective information retrieval and analysis, which is important in the long run (Abdrakhmanov et al., 2024).

The results of this study confirmed that thesauri are valuable tools in modern information technology. The impact of thesaurus technologies ranges from providing search efficiency to improving learning processes and developing innovative applications (Kolbayev et al., 2024). Despite similar issues in current research, such as improving automatic processing of search queries, improving the categorization of materials, different researchers have focused on different aspects of this technology. For example, the studies of Palekha and Alieksieienko (2022) and Braichevskyi (2022) focused not so much on general educational and technological aspects, but on studying the possibility of systematizing data using thesauri in the field of jurisprudence. The results of our research showed that there is a need to improve search systems for both subject areas and highly specialized terminology.

Organization of thesauri with the help of subject headings is necessary because they help to screen out irrelevant search results (Tkachenko et al., 2024). A subject heading, in particular, represents standardized lexemes that are assigned to a certain concept. As argued by Panda (2022), when creating thesauri, it is important to take into account the principles of cataloguing materials by subject area. The author came to this conclusion by analyzing the following thesauri built on the catalogue principle: Sears List of Subject Headings (SLSH) and Medical Subject Heading (MeSH). Based on the experience of the LCSH study in our own research, it should be noted that the catalogues are organized according to several important criteria such as location, language, genre principle, which greatly facilitates the search for necessary information. But it should also be noted that there is a need to improve the categorization of materials by thematic (content) principle or by subject areas.

Modern thesauri, such as WordNet, Roget's Thesaurus, are not just linear lists of synonyms, but complex systems of interrelated words and concepts (Skliarenko et al., 2019). These tools are becoming important for linguists, researchers, and engineers involved in natural language processing because they reflect the structure and semantics of the language as a whole. For example, in the work of Mandal (2021), the author designed a thesaurus using TemaTres and Visual Vocabulary, claiming advantages such as providing easy access to the library, displaying neighbouring terms in the search engine, and using controlled vocabulary. The study of thesauri in our own research has demonstrated the need to develop new mechanisms and improve the filtering of search queries. Thus, the more term sorting functions are presented in the catalogue, the more efficient and faster the information retrieval will be.

The evolution of thesauri is accelerating against the background of information technology development. These tools play an important role in natural language processing and the formation of new linguistic paradigms. Expansion of their functionality includes adaptation to terminology of new branches and striving to incorporate cultural and social contexts into the content (Lewinski, 2015). For example, in the study by Mandal and Malik (2022), the authors presented a variant of thesaurus creation using different data published using Apache Jena and Apache Jena Fuseki. Thus, modern thesauri have the ability to access metadata located on multiple interfaces simultaneously.

It is important to acknowledge that in modern society the need for accuracy and richness of language is increasing due to the emergence of new technologies including artificial intelligence, machine learning and automated data processing systems (Iklassova et al., 2024; Nurbatyrova et al., 2024). Thesauri are becoming a tool for enriching language processing and improving the accuracy of text comprehension by computers. Yi and Chan (2010) and Bashmanivskiy (2016) focused on problems that are related to syntactic and hierarchical structures. For example, they emphasized problematic headings that include offensive lexemes, combine several ideas, or use infrequent terminology. Based on the results of our study, it can be noted that there is a need to differentiate terminological units that belong to different subject areas. This is necessary to improve the efficiency of outputting results for a specific query.

The role of thesauri in the formation of intellectual capital of society is vividly manifested. These linguistic resources not only reflect linguistic dynamics, but also influence the understanding of the world. Thesauri, becoming a key element of the linguistic infrastructure, contribute to the development of knowledge, technology, and culture, emphasizing their continuous relevance (Symonenko et al., 2019). The issues of improving the linguistic construction of thesauri were addressed by Usmonov and Usmonova (2020). This work indicated that the main problem on the way to accurate translation is the ambiguity of data in dictionary entries. So, standardization of terminology is necessary not only to improve awareness and accessibility of information, but also to improve communicative processes (Mizin et al., 2023).

However, despite all the advantages, there is a need to constantly update and expand thesauri to reflect modern linguistic changes and new terms used in science and technology. It should be noted that adjectives, which exist as stand-alone terms, are not sufficiently considered in the development of thesauri because they are primarily oriented towards the terminological practice of using nouns. A study by Bonadonna (2020) focused on building syntagmatic and paradigmatic relations between term-adjectives and term-nouns and fixing these relations in dictionaries. Based on the results of our research, it is worth identifying the need to improve the automatic processing of established terminological units combining different parts of speech.

5. Conclusions

The study of thesauri in the modern world is of profound importance for understanding the evolution of language, its influence on technological progress and the need for constant updating to meet the rapidly changing language trends and challenges of modernity. According Hazem et al. (2020), the quality of thesauri depends on content construction, that is, the ability to extract terminology efficiently and solve the structural design problem. Due to the development of modern technologies, there is a goal to create such a thesaurus that would use automatic methods to identify relationships between terms. Our research emphasized the need for enriching semantic relationships and thematic enrichment of thesaurus categories, which can be achieved using automatic text processing methods. Thus, further research on thesauri becomes strategically important in providing a deep understanding of language in our current information society.

The study found that thesauri can serve as an important tool for structuring knowledge and their use provides systematization of information, which could have a positive impact on the effectiveness of teaching, research and management decisions. The results highlighted the need to optimize searching in scientific and library catalogues. This can be done by leveraging the principles of natural language processing and improving the efficiency of machine learning algorithms. Thus, thesaurus systems can be successfully integrated into educational processes, contributing to the creation of interactive educational tools and improving teaching methods, which allows students to more effectively assimilate knowledge and develop creative thinking.

The importance of thesauri in today's information landscape is also due to the stimulation of new research directions through the development and improvement of thesaurus structures. The study of terminological relations between subject areas, thesaurus categories and highly specialized concepts is necessary from the point of view of the development of automated systems for semantic text processing. In the course of research of the thesaurus database Library of Congress Subject Headings it was revealed that the largest number of issued results referred to such subject areas: *art*, *library systems*, *medicine*, *culture*, *media*, *scientific research*, *linguistics*, and *semantics*. The hierarchy of the thesaurus structure was represented as follows: *subject area – thesaurus category – narrowly focused terminology*.

In general, research in the field of thesauri provides a wide range of opportunities for improving the organization of knowledge in the modern information society. Among the promising directions for the development of this area are the following: optimization of thesaurus links to improve search relevance, automation of the process of results delivery, ensuring adaptation of thesauri to changing models of their construction. These factors could significantly improve the efficiency of information retrieval in library and information systems. The development of more detailed and consistent thesauri can contribute to the improvement of automated text processing, semantic analysis and the expansion of terminology control systems. Further prospects include the integration of such developments into various fields of knowledge, which could facilitate more accurate and faster results in scientific and practical research.

Acknowledgments

Research was implemented as part of the scientific grant financing project AP14869894 for 2022-2024 on the topic “Thesaurus of Abai's words of edification”. We wish to thank an anonymous reviewer and the editors for many helpful suggestions.

References

Abdrakhmanov, R., Kenesbayev, S.M., Berkimbayev, K., Toikenov, G., Abdrashova, E., Alchinbayeva, O., & Ydyrys, A. (2024). Offensive Language Detection on Social Media using Machine Learning. *International Journal of Advanced Computer Science and Applications, 15*(5), 575-582. https://doi.org/10.14569/IJACSA.2024.0150557

Aitchison, J., Bawden, D., & Gilchrist, A. (2000). *Thesaurus construction and use: A practical manual.* London: Routledge. https://doi.org/10.4324/9780203402986

Avram, A.M., Pais, V., & Tufis, D.I. (2021). PyEuroVoc: A tool for multilingual legal document classification with EuroVoc descriptors. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing* (pp. 92-101). Shumen: INCOMA Ltd. https://aclanthology.org/2021.ranlp-1.12

Balabekova, M.A., & Khan, N.K. (2022). Linguistic convergence in mass media texts. *Bulletin of Karaganda University, Philology Series, 106*(2), 13-19. https://doi.org/10.31489/2022ph2/13-19

Bashmanivskiy, O. (2016). The problems of automated translation of business correspondence using free software. S*ociety. Document. Communication*, *1*(2), 79-90.

Bayekeyeva, A., Tazhibayeva, S., Beisenova, Zh., Shaheen, A., & Bayekeyeva, A. (2021). Controlled multilingual thesauri for Kazakh industry-specific terms. *Social Inclusion and Multilingualism: The Impact of Linguistic Justice, Economy of Language and Language Policy, 9*(1), 35-44. https://doi.org/10.17645/si.v9i1.3527

Bonadonna, M.F. (2020). Using lexical functions to describe adjectives in terminography. Terminology. *International Journal of Theoretical and Applied Issues in Specialized Communication, 26*(1), 7-32. https://doi.org/10.1075/term.00040.bon

Braichevskyi, S.M. (2022). Creation of thesauruses of regulatory and legal information in conditions of digitization. *Information and Law, 40*(1), 61-71. https://doi.org/10.37750/2616-6798.2022.1(40).254343

Bushman, B., Anderson, D., & Fu, G. (2015). Transforming the medical subject headings into linked data: Creating the authorized version of MeSH in RDF. *Journal of Library Metadata, 15*(3-4), 157-176. https://doi.org/10.1080/19386389.2015.1099967

Gilchrist, A. (2003). Thesauri, taxonomies, and ontologies – An etymological note. *Journal of Documentation, 59*(1), 7-18. https://doi.org/10.1108/00220410310457984

Hazem, A., Daille, B., & Lanza, C. (2020). Towards automatic thesaurus construction and enrichment. In *Proceedings of the 6th International Workshop on Computational Terminology* (pp. 62-71). Marseille: European Language Resources Association.

Hearst, M.A. (1992). Automatic acquisition of hyponyms from large text corpora. In *The 14th International Conference on Computational Linguistics* (pp. 539-545). Nantes: Association for Computational Linguistics.

Iklassova, K., Aitymova, A., Kopnova, O., Shaporeva, A., Abildinova, G., Nurbekova, Z., Almagambetova, L., Gorokhov, A., & Aitymov, Z. (2024). Ontology modeling for automation of questionnaire data processing. *Eastern-European Journal of Enterprise Technologies, 5*(2-131), 36-52. https://doi.org/10.15587/1729-4061.2024.314129

Kairaktyzy, A.A. (2022). Construction of thesauruses of the Kazakh language in an automatic text processing system. In *Proceedings of the XXV International Scientific and Practical Conference "Innovative Trends of Science and Practice, Tasks and Ways to Solve Them"* (pp. 487-492). Athens: International Science Group. *Library of Congress Subject Headings.* (2023). https://www.loc.gov/search/?in=&q=Library+systems&new=true

Kolbayev, N., Tuyenbayeva, K., Seitimbetova, D., & Apakhayev, N. (2024). Methods of Modelling Electronic Academic Libraries: Technological Concept of Electronic Libraries. *Preservation, Digital Technology and Culture, 53*(2), 81-90. https://doi.org/10.1515/pdtc-2024-0001

Lewinski, P. (2015). Commentary: Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Frontiers in Psychology, 6*(NOV), article number 1832. https://doi.org/10.3389/fpsyg.2015.01832

Lykke Nielsen, M. (2004). Thesaurus construction: Key issues and selected readings. *Cataloging & Classification Quarterly, 37*(3-4), 57-74. https://doi.org/10.1300/J104v37n03_05

Mandal, S. (2021). Controlled vocabulary system for knowledge organization: Designing a framework. *World Digital Libraries, 14*(2), 111-124. https://doi.org/10.18329/09757597/2021/14206

Mandal, S., & Malik, S. (2022). Transforming thesaurus records into MARC 21 and MADS: Designing a framework for libraries. *Library Philosophy and Practice, 7477*, 1-15.

Martinez, A., & Mammola, S. (2021). Specialized terminology reduces the number of citations of scientific papers. *Proceedings of the Royal Society: Biological Sciences, 288*(1948), 1-5. https://doi.org/10.1098/rspb.2020.2581

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K.J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography, 3*(4), 235-244. https://doi.org/10.1093/ijl/3.4.235

Mizin, K., Slavova, L., Letiucha, L., & Petrov, O. (2023). Emotion concept disgust and its German counterparts: Equivalence determination based on language corpora data. *Forum for Linguistic Studies, 5*(1), 72-90. https://doi.org/10.18063/FLS.V5I1.1552

Nurbatyrova, R., Japarov, B., Apakhayev, N., Abdulaziz, B., & Khushkeldiyeva, S. (2024). Digital Transformation of Archives in the Context of the Introduction of an Electronic Document Management System in Kazakhstan. *Preservation, Digital Technology and Culture, 53*(3), 147-155. https://doi.org/10.1515/pdtc-2024-0017

Palekha, Y., & Alieksieienko, K. (2022). Narrow and wide definition of the term *document*. *Society. Document. Communication*, *7*(1), 100-115. https://doi.org/10.31470/2518-7600-2022-14-100-115

Panda, S. (2022). *The work architecture of the Sears List of Subject Headings (SLSH), Library of Congress Subject Headings (LCSH), and Medical Subject Heading (MeSH): An introduction. Authorea.* https://doi.org/10.22541/au.166212050.03129709/v1

Sayakhmet, S.S., & Tumanova, A.B. (2022). Formation of thesaurus of linguistic personality of the future specialist-philologist from the perspective of lingo-didactics. *Bulletin of Karaganda University, Philology Series, 106*(2), 151-160.

Shu, F., Dinneen, J.D., Asadi, B., & Julien, Ch.A. (2017). Mapping science using Library of Congress Subject Headings. *Journal of Informetrics, 11*(4), 1080-1094. https://doi.org/10.1016/j.joi.2017.08.008

Skliarenko, O., Akimova, A., & Svyrydenko, O. (2019). Psycholinguistic peculiarities of contextual realisation of concept «macht» in linguistic and cultural space of German's.

Psycholinguistics, 26(2), 321-340. https://doi.org/10.31470/2309-1797-2019-26-2-321-340

Symonenko, S., Zaitseva, N., Titova, O., & Vynogradova, M. (2019). Development of communicative competence as a precondition of competitive software engineer formation. In *Modern Development Paths of Agricultural Production: Trends and Innovations* (pp. 307-315). Cham: Springer. https://doi.org/10.1007/978-3-030-14918-5_32

Tkachenko, O., Goncharov, V., & Jatkiewicz, P. (2024). Enhancing Front-End Security: Protecting User Data and Privacy in Web Applications. *Computer Animation and Virtual Worlds, 35*(6), article number e70003. https://doi.org/10.1002/cav.70003

Usmonov, Y.M.U., & Usmonova, Z.B.Q. (2020). The role and importance of terminology in the study of specialized language. *International Journal on Economics, Finance and Sustainable Development, 1*(3), 19-23. https://doi.org/10.31149/ijefsd.v1i3.323

Yi, K., & Chan, L.M. (2010). Revisiting the syntactical and structural analysis of Library of Congress Subject Headings for the digital environment. *Journal of the American Society for Information Science and Technology, 61*(4), 677-687. https://doi.org/10.1002/asi.21295

Zaitseva, N.V., Symonenko, S.V., Titova, O.A., Osadchyi, V.V., & Osadcha, K.P. (2023). Fostering communication and collaboration skills for computer science students by means of ICT tools. *CEUR Workshop Proceedings, 3553*, 43-56.

About the authors

Arap Yespenbetov is a Full Doctor, Professor at the Department of Kazakh Philology and Journalism, Shakarim University, Semey, Republic of Kazakhstan. He focuses on the semantic processing of textual data and the development of thesauri to optimize data organization and retrieval in library systems.
yespen-betov@outlook.com
https://orcid.org/0009-0006-5984-682X

Kuralay Tulebayeva (corresponding author) is a PhD, Head of the Department of Philology, Alikhan Bokeikhan University, Semey, Republic of Kazakhstan. She investigates the role of thesauri in improving access to medical information, exploring semantic analysis and the hierarchical structuring of terms for electronic health record systems.
tulebayevakuralay@gmail.com
https://orcid.org/0000-0003-1636-8822

Assem Kassymova is a PhD, Senior Researcher at the Research Center “Abai and National Spirituality”, Shakarim University, Semey, Republic of Kazakhstan. Her research interest includes integration of thesauri in digital art and cultural heritage archives.
askassymova@hotmail.com
https://orcid.org/0000-0002-8783-888X

Baurzhan Yerdembekov is a Full Doctor, Professor at the Department of Kazakh Philology and Journalism, Shakarim University, Semey, Republic of Kazakhstan. He concentrates on the application of thesauri in linguistic and semantic research, as well as comparative analyses across domains like linguistics, media, etc.
yerdembe.kov@outlook.com
https://orcid.org/0000-0003-0503-4640

Akmaral Smagulova is a PhD, Head of the Research Center “Abai and National Spirituality”, Shakarim University, Semey, Republic of Kazakhstan. She focuses on the role of thesauri in education and develops innovative structures to enhance access to learning materials.
a.smagulova3@hotmail.com
https://orcid.org/0000-0002-7509-7025