УДК 920.183.7. 563.70
https://doi.org/10.31548/philolog14(4).2023.05

# Linguocognitive Approach to Extracting Terms from a Corpus of Veterinary Texts

**Yurii ROZHKOV**
PhD in Philology,
Associate professor at the Foreign Philology and Translation Department,
National University of Life and Environmental Sciences of Ukraine
15 Heroyiv Oborony Street, Kyiv 03041, Ukraine
https://orcid.org/0000-0002-6830-9130

**Abstract.** This research delves into the intricate landscape of computational linguistics with a focused exploration of term identification challenges within the domain of veterinary medicine. A comprehensive analysis was conducted, balancing the difficulties associated with the automated extraction of single-word terms and the structured patterns observed in two-word terms within veterinary dictionaries and scientific literature.

The study commenced with a meticulous manual identification of 462 single-word terms, emphasizing the inherent challenges in automating the extraction of terms characterized by linguistic diversity and potential ambiguity. Simultaneously, the investigation of two-word terms unveiled structured patterns, particularly in dictionaries, offering contrasting simplicity for identification through conventional frequency-based methods.

The chosen text type, a veterinary dictionary, revealed its own intricacies with a standardized template governing entry construction. The revelation that only 59% of terms find placement in the title section underscored the need for adaptive extraction methods attuned to the varied distribution of terms within dictionary structures. Scientific texts further complicated the term identification landscape by showcasing varying term frequencies, prompting a critical evaluation of standard lexeme selection methods.

Building on these insights, the research proposes strategies for refining automated term identification processes. This includes leveraging advanced natural language processing techniques for single-word terms and advocating for adaptive extraction methods for dictionaries, while also proposing a hybrid approach for scientific texts.

The interdisciplinary nature of the research is underscored by the recognition of collaboration between linguists, computational scientists, and domain experts as crucial for developing sophisticated models and ontologies that accurately capture the unique linguistic nuances of veterinary medicine.

As the digital landscape evolves, this research not only contributes to the advancement of computational linguistic methodologies but also envisions the creation of terminological resources reflecting the dynamic nature of language within the veterinary domain. Through a comprehensive exploration of challenges and opportunities, this research aspires to pave the way for more accurate and adaptable automated systems, offering implications for the broader field of computational linguistics.

**Key words:** corpus, corpus analysis, frequency analysis, veterinary terminology, cognitive science.

**Introduction.** In computational linguistics, the accurate identification and extraction of domain-specific terminology pose intricate challenges that demand nuanced solutions. This research delves into the complexities associated with term identification, with a particular focus on the distinctive features of veterinary medicine texts. The preliminary analysis, undertaken through the lens of a comprehensive examination of a veterinary dictionary and scientific literature, has unearthed intriguing dynamics governing the distribution and frequency of terms.

The impetus for this investigation stems from the dual nature of term identification challenges. On one hand, the manual identification of 462 single-word terms serves as a quality control measure, emphasizing the difficulties inherent in automatically extracting terms of this nature. Single-word terms, owing to their linguistic diversity and potential for ambiguity, require sophisticated natural language processing techniques for accurate identification. On the other hand, the exploration of two-word terms in the context of dictionaries and scientific texts uncovers structured patterns that offer contrasting simplicity for identification.

The chosen text type, a veterinary dictionary, introduces its own set of peculiarities. A uniform template governs the construction of dictionary entries, comprising a title and a definition, often incorporating hyperonyms and additional information. However, the revelation that only 59% of terms find their place in the title section underscores the need for adaptive extraction methods that accommodate the varied distribution of terms within dictionary structures.

Scientific texts further compound the challenges, showcasing scenarios where certain terms exhibit high-frequency occurrences, while others remain sporadic. This dynamic nature prompts a critical evaluation of standard methods for selecting lexemes, such as TF-IDF and LDA, raising questions about their efficacy in capturing the nuanced frequencies inherent in scientific language.

Beyond the technical intricacies, the research recognizes the interdisciplinary nature of the endeavor. Collaboration between linguists, computational scientists, and domain experts is envisioned as crucial for creating sophisticated models and ontologies that capture the unique linguistic nuances of veterinary medicine.

As the digital landscape continues to evolve, this research is poised to contribute not only to the advancement of computational linguistic methodologies but also to the creation of terminological resources that reflect the dynamic nature of language within the veterinary domain. Through a comprehensive exploration of challenges and opportunities, this research aspires to pave the way for more accurate and adaptable automated systems in the field of veterinary medicine and beyond.

**Literature review.** Internationally renowned figures in the field of corpus linguistics, such as Baker M. (Baker, 2015, 2019) and Biber D., (Biber, 2018, 2012) have laid the groundwork for understanding language structures and the application of corpora in linguistic analysis. Notable Ukrainian linguists like Zhabotynska S. (Zhabotynska, 2014) and Perkhach R. (Perkhach, 2015, 2017) explored corpus linguistics within the context of Ukrainian languages.

Researchers in veterinary terminology, including Rozhkov Yu. And Syrotin O. have addressed challenges in defining and standardizing veterinary terms (Rozhkov, Syrotin, 2021). Pioneers in megacorpora research, such as Oostdijk N. (Oostdijk, 2010, 2021) and Bowker L., (Bowker, 2016) have significantly advanced our understanding of language use. By synthesizing insights from these internationally renowned linguists and Ukrainian scientists, this study **aims to** establish a comprehensive theoretical framework. It endeavors to contribute to a nuanced understanding of veterinary medical terminology in both the global and local contexts, integrating the perspectives of corpus linguistics, veterinary medicine, and language education.

**Results and discussion.** If you use the advantages of the knowledge base in terms of storing and presenting information about the terms of the subject area and take into account the linguistic characteristics of the terms, their interpretation and examples of use, then you can create a fundamental contextual dictionary of the subject area. The main purpose of such a dictionary is a comprehensive presentation of the semantics of the term in an indirect way, for example, by providing an answer to an information request about the term with a sufficient number of generalized examples – contexts of the use of this term. It is characterized by the high complexity of defining basic concepts and selecting equivalents for foreign language terms. In such cases, the only way to present a term in a dictionary is to define it by demonstrating it in different contexts (Gonzales, 2022: 65-66).

The advantage of the contextual method of defining a term is, on the one hand, a visual illustration of the functioning of the term in the text, and on the other hand, a reflection of its semantic connections with other terms of the given subject area. Thus, the property of systematic terminology is most productively used in the contextual dictionary. From a methodological point of view, the advantage of the contextual dictionary in comparison with other lexicographic sources focused on a systematic approach when displaying terminology is that, using the ideology of the knowledge base in terms of information storage and presentation, the contextual dictionary supplies the participant of the analytical-synthetic process (for example, the translator) with model texts that can be used both for understanding the input text (original) and for generating the output text (translation) (Wiebe, 2010: 60-61).

In the theoretical sense, the context dictionary, in terms of the ideology of its organizational structure and actual content, is fully consistent with the ideology of the processes of perception, information storage, thinking and representation of linguistic forms of communication, which are oriented to the theoretical principles of building knowledge bases, since most often the source of these theoretical principles is the fact that that a person, learning a new situation, chooses from his memory some known image (called a frame in the theory of thinking), and then, preserving and combining these situations by changing the recognizable details, creates a new image (frame) suitable for understanding the new situation and assimilation of specific new knowledge (Zhabotynska, 2015: 81-82).

Just as a frame in the structure of a knowledge base serves to represent a stereotypical situation, so a generalized example in a contextual dictionary can be used to present one and generate another stereotypical text. In such a case, if a generalized example is given in parallel in two or more languages in a contextual dictionary, it becomes a particularly useful model suitable for such an analytical-synthetic process as translation, since the version in one language can be used to understand the meaning (of the original). and the version in another language – for text generation (translation) (Geeraerts, 2011: 587-588).

A number of linguistic problems have been solved thanks to the ideas of machine translation. The view of language as a code became the reason for applying the method of statistical research of texts for the purpose of identifying certain linguistic regularities in them. According to Biber D. (Biber, 2012: 115), it was at this time that the main provisions of the distributive theory were formulated, the main principle of which was the study of the textual behavior of language elements for their further comprehensive characterization. The distributional technique, combined on the basis of statistical techniques, greatly stimulated the formation of theoretical linguistics, in which modeling implemented for the first time (Biber, 2012: 58).

Such a technique involves the study of fairly large arrays of source texts to obtain reliable data. According to Zhukovska V., the study of distribution allows: a) to determine the model of meaning, that is, the composition of the main components that together form the meaning of a given lexical unit; b) establish a model of connectivity of this lexical unit with other lexical units; c) define the formal structure of the lexical unit. At one time, the distributional-statistical approach played a major role not only in research on machine translation, but also in theoretical linguistics, allowing building distributional-statistical models of language styles. At the level of meaningful understanding, the distribution of a language element is its occurrence in certain contexts, a set of language elements and text units adjacent to it (Zhukvoska, 2013: 50-51).

At the same time, the depth of the context can be set based on the possibilities of the research. The development of corpus linguistics and the growing attention to statistical methods of processing language material in recent years have led to the development of a number of methods that are associated with the use of parallel or close texts in different languages. These techniques and methods, mainly based on statistics and mathematical linguistics (information theory, theory of algorithms, etc.), allow obtaining objective data on the composition, structure and functioning of language units (Oostdijk, 2021: 97).

It cannot be said that these techniques themselves are something completely new. Thus, statistical research in linguistics has a rather long history and is particularly characteristic of the 1950s of the XX century, when computers first appeared and the possibility of using them to solve linguistic problems. Research tools such as frequency dictionaries, concordances, etc. have also been known for quite some time (Bowker, 2016: 35).

In the dynamic intersection of linguistics and computational science, corpus linguistics has emerged as a transformative field, revolutionizing the analysis of language through the systematic examination of extensive text collections. This article navigates the intricate terrain where corpus linguistics converges with the nuanced process of term retrieval, unraveling the methodologies, challenges, and transformative implications for linguistic analysis. As we delve into the symbiotic relationship between these realms, we illuminate the profound impact on our understanding of language evolution, educational practices, and specialized research. The following exploration takes us through the key components, methodologies, and ethical considerations, painting a comprehensive portrait of the pivotal role that corpus linguistics and term retrieval play in advancing linguistic inquiry (Perkhach, 2017: 71-72).

Unveiling Insights in Linguistic Analysis Corpus linguistics, a multidisciplinary field at the intersection of linguistics and computer science, has revolutionized the study of language by providing researchers with powerful tools to analyze vast collections of texts systematically. At the heart of corpus linguistics lies the corpus – a structured and extensive assemblage of written or spoken language samples. This discussion delves into the symbiotic relationship between corpus linguistics and the intricate process of term retrieval, shedding light on the methodologies, challenges, and transformative implications for linguistic analysis (Gonzales, 2022: 65-66).

Corpus linguistics leverages the compilation of diverse text sources, ranging from literature and newspapers to academic journals and social media. This composition ensures a representative snapshot of language use in various contexts, enabling researchers to uncover patterns and trends. Before delving into analysis, corpus linguists employ preprocessing techniques like tokenization, stemming, and removal of stop words to refine raw text data. These steps are vital for transforming unstructured text into analyzable units, facilitating subsequent linguistic investigations.

Concordance analysis is a cornerstone of corpus linguistics, enabling the examination of word occurrences in context. By presenting words in their original contexts, concordances empower researchers to discern semantic nuances and linguistic patterns. Frequency-based methods: Term frequency (TF) and document frequency (DF) are fundamental measures used in corpus linguistics. TF-IDF, a widely adopted approach, combines these metrics to identify the significance of terms within a corpus. This method underpins various linguistic analyses, including term retrieval. Term Retrieval: Unraveling Language Specifics Contextual understanding through part-of-speech tagging: Part-of-speech tagging enhances the accuracy of term retrieval by providing contextual information about the roles words play in sentences. This is particularly crucial in languages with intricate grammatical structures and multiple word meanings (Oostdijk, 2010: 12-14).

In languages where word relationships influence meaning, dependency parsing becomes invaluable. This technique explores the syntactic dependencies between words, contributing to a more nuanced understanding of term relationships within a corpus. Machine learning techniques, such as supervised and unsupervised learning, have found applications in term retrieval. These algorithms, when trained on relevant data, can identify patterns and relationships, offering a complementary approach to traditional frequency-based methods. Challenges in domain-specific corpora: While generic corpora provide a broad overview of language use, domain-specific corpora, such as those in veterinary medicine or legal discourse, pose unique challenges (Bowker, 2016: 31-32).

Domain-specific terminology, abbreviations, and context-dependent meanings require tailored approaches for effective term retrieval. Ethical considerations: The use of corpora, especially when sourced from sensitive domains, demands ethical considerations. Respecting privacy, obtaining consent, and ensuring data anonymization are essential tenets in conducting ethical linguistic research. Transformative Implications: Advancing Linguistic Inquiry Insights into language evolution: Corpus linguistics, coupled with term retrieval, facilitates the tracking of language evolution over time (Baker, 2019: 165).

By analyzing diachronic corpora, researchers gain insights into shifting language patterns, semantic drift, and the emergence of new terms. Informing language teaching: Corpora aid language educators by providing authentic examples of language use. Term retrieval allows the identification of key vocabulary, enabling educators to tailor language instruction to real-world linguistic contexts. Facilitating specialized research: For researchers in specialized fields, such as medical or legal experts, corpus linguistics supports the identification and analysis of domain-specific terminology.

This, in turn, contributes to advancements in specialized knowledge and terminology standardization. In conclusion, the symbiosis between corpus linguistics and term retrieval unveils a rich landscape of linguistic analysis. From unraveling language intricacies to informing educational practices and advancing specialized research, the transformative implications underscore the indispensable role of these methodologies in our evolving understanding of language. As technology advances and linguistic inquiries become more sophisticated, the synergy between corpus linguistics and term retrieval promises to be at the forefront of linguistic research.

The methodological framework employed for term retrieval from the corpus of veterinary texts is a critical cornerstone, significantly shaping the accuracy and relevance of the identified terms. This extensive discussion encompasses the intricacies, challenges, and far-reaching implications of the chosen approach, dissecting key considerations involved in the process (Abraham-Barna, 2011: 87-88).

The bedrock of effective term retrieval lies in the meticulous selection and composition of the veterinary texts corpus. The corpus's pivotal role in ensuring a representative array of diverse subfields within veterinary medicine cannot be overstated. A comprehensive corpus, spanning clinical reports, research articles, and specialized veterinary literature, not only enriches the depth and breadth of the identified terms but also enhances the robustness of subsequent analyses. However, it is essential to acknowledge and address potential biases introduced by the corpus composition, such as the inadvertent overrepresentation of certain subfields, to ensure a nuanced interpretation of results. The preprocessing stage is pivotal, involving techniques like tokenization, stemming, and stop-word removal to refine raw text data for subsequent term extraction. Decisions made during preprocessing profoundly impact the granularity of identified terms and the subsequent analysis. Striking the right balance between preserving the specificity of terms and minimizing noise requires a judicious approach, as overly aggressive preprocessing may inadvertently result in the loss of crucial information (Oostdijk, 2010: 201).

Incorporating part-of-speech tagging and dependency parsing is instrumental for contextual understanding, particularly concerning single-word terms. These techniques significantly contribute to disambiguating term meanings within diverse contexts, thereby enhancing the precision of term identification. Nevertheless, the inherent intricacies of veterinary language, characterized by domain-specific jargon and abbreviations, may introduce challenges that need to be systematically addressed to prevent misinterpretations (Zhukovska, 2013: 116).

Frequency-based methods, exemplified by Term Frequency-Inverse Document Frequency (TF-IDF), stand as standard practices in term extraction processes. A key focal point for discussion revolves around the suitability of these methods for the veterinary domain. The varying frequencies of terms within different subfields, as illuminated in the preliminary analysis, necessitate a nuanced approach. Potential adjustments to weighting mechanisms or the incorporation of domain-specific metrics may be warranted to accurately capture the significance of terms in their respective contexts (Baker, 2015: 231-232).

The potential enhancement of the methodology through the incorporation of domain-specific lexicons and ontologies emerges as a promising avenue. Developing specialized resources tailored to veterinary terminology could significantly bolster the accuracy of term identification. However, challenges may arise in maintaining and updating such resources to reflect the dynamic nature of veterinary language and accommodate emerging terminologies.

The integration of machine learning and advanced natural language processing (NLP) techniques introduces a layer of both complexity and opportunity to the methodology. Supervised and unsupervised learning models, when appropriately trained on domain-specific data, have the potential to discern patterns and relationships not immediately apparent through traditional methods. However, the interpretability of machine-generated results and the need for annotated datasets present challenges in the widespread adoption of these advanced techniques. Discussions on the selection of appropriate evaluation metrics and validation methods are imperative. The quantification of the accuracy of identified terms against a gold standard or expert-supervised list is essential for robustly assessing the reliability of the methodology. The intricacies of domain-specific evaluation, with its potential for synonymous terms and context-dependent meanings, should be acknowledged to ensure a nuanced interpretation of results (Bowker, 2016: 50-51).

The scalability and efficiency of the chosen methodology are paramount considerations, especially in the face of the increasing volume of veterinary literature. Robust computational resources capable of handling large-scale term extraction become imperative. Discussions around optimization strategies, parallel processing, and the potential integration of cloud-based solutions become pertinent to ensure the feasibility and timeliness of the analysis.

The ethical implications of term retrieval should not be overlooked in the pursuit of scientific inquiry. The utilization of veterinary texts may involve sensitive information, necessitating clear ethical guidelines regarding data privacy and consent. Ensuring that the methodology aligns with stringent ethical standards in data usage and publication is crucial for maintaining the integrity of the research and upholding ethical principles in the scientific community (Perkhach, 2017: 173-174).

In the realm of computational linguistics, ensuring the accuracy and effectiveness of automated systems for term extraction is a critical aspect of text processing. In this particular analysis, a meticulous quality control procedure was undertaken, involving the manual identification of 462 single-word terms. The focus on single-word terms is driven by the recognition of the inherent challenges associated with their automatic extraction. Single-word terms often pose difficulties due to their diverse linguistic characteristics and potential for ambiguity (Edward, 2015).

Conversely, the text under consideration presents an intriguing scenario regarding two-word terms. In this context, these terms exhibit a discernible and structured pattern, typically following an adjective + noun format. This structural clarity facilitates their identification through conventional frequency-based methods, offering a stark contrast to the complexities associated with single-word terms.

The choice of the text type, a dictionary, introduces its own set of unique characteristics and challenges. Dictionaries, as repositories of lexical knowledge, adhere to a standardized template. Each entry typically comprises a title and a definition. The definition, in turn, is often constructed using hyperonyms and supplementary information, contributing to a comprehensive understanding of the term. Interestingly, terms from the subject area may manifest in both the title and definition sections of dictionary entries.

A notable revelation from the analysis is that only 59% of the terms find their place in the title section of the dictionary. This structural regularity eases the computational burden associated with text processing. However, it also underscores the need for nuanced approaches when dealing with specific text types, as terminological nuances may be distributed across different sections of a given text.

Moving beyond the structural peculiarities of dictionaries, the text explores the challenges inherent in scientific texts. Within this domain, a common occurrence involves certain terms from the subject area having frequent appearances, while others have sporadic occurrences. For instance, the term "BANK" is cited 826 times, constituting 3% of all non-service words in the corpus, exemplifying a high-frequency term. Conversely, terms like "anaplasma" and "cytoplasm" make only isolated appearances.

The analysis delves into the consequences of employing standard methods for selecting lexemes, such as TF-IDF (Term Frequency-Inverse Document Frequency) and LDA (Latent Dirichlet Allocation). These methods, while effective in various contexts, may inadvertently discard both frequently occurring terms, crucial for a comprehensive understanding of a domain, and those with sparse occurrences, which might still be of considerable significance.

Moreover, among words with moderate frequency, a substantial portion consists of general vocabulary. This dynamic introduces a layer of complexity in the identification and prioritization of specialized terminology, a crucial consideration in computational linguistic endeavors.

A concrete example is presented to illustrate the intricacies of term frequency distribution. The term "SCALPEL," with an absolute frequency of 45 occurrences in the corpus, holds the 56th rank in terms of noun frequency out of 98. In a frequency dictionary constructed based on the corpus, 19 related nouns, spanning the range from the 50th rank (frequency 51) to the 62nd rank (frequency 39). This example showcases the diversity and distribution of related terms, each contributing to the nuanced understanding of the core term "SCALPEL" (McFarland, 2013: 115).

The intricacies uncovered in the analysis prompt a deeper examination of potential strategies to refine automated term identification processes. As the nuances of term distribution vary across different text types, a tailored approach is essential. For single-word terms, which pose a pronounced challenge, implementing advanced natural language processing techniques such as part-of-speech tagging and contextual analysis could enhance accuracy. The identification of single-word terms demands a nuanced understanding of their linguistic context, which can be achieved through the integration of contextual analysis algorithms.

In the context of dictionaries, where terms can manifest in both the title and definition sections, an adaptive extraction method becomes imperative. Leveraging semantic analysis and linguistic patterns specific to dictionary structures can contribute to a more precise identification of terms. Recognizing that terms may be distributed across various sections necessitates an approach that is attuned to the idiosyncrasies of dictionary formats.

Moreover, in scientific texts characterized by varying term frequencies, a hybrid approach is proposed. Instead of outrightly discarding infrequent terms, an intelligent weighting system, perhaps based on semantic relevance or contextual significance, could be employed. This would prevent the loss of potentially valuable terms while still prioritizing those with higher contextual importance.

The identified challenges also point towards the need for a comprehensive terminological resource that captures the dynamic nature of term usage in different contexts. A curated corpus specifically tailored to the intricacies of veterinary medicine, encompassing diverse sources including dictionaries, research articles, and clinical notes, could serve as a valuable reference. Such a resource would not only aid in the refinement of automated systems but also contribute to the continuous evolution of veterinary terminology.

The implications of this analysis extend beyond the immediate challenges of term extraction. They underscore the significance of interdisciplinary collaboration, bringing together linguists, computational scientists, and domain experts to create more sophisticated models. The collaboration can involve the development of specialized ontologies that encapsulate the intricate relationships between terms in the veterinary domain.

**Conclusion.** This exploration highlights the multifaceted nature of term identification, particularly in specialized texts like dictionaries and scientific literature. The dualities observed in term frequency distribution underscore the need for sophisticated computational methods that can discern and prioritize terms based on their linguistic characteristics and contextual relevance. As computational linguistics continues to evolve, refining the methodologies for term extraction becomes pivotal for accurately mapping and interpreting the intricacies of language in diverse textual landscapes.

The challenges posed by the extraction of veterinary terminology, as illuminated by this analysis, open avenues for innovative solutions and interdisciplinary collaborations. The refinement of computational linguistic methodologies is not only essential for accurate term identification but also crucial for advancing our understanding of domain-specific languages. As technology continues to shape the landscape of linguistic analysis, an adaptable and nuanced approach becomes paramount for harnessing the full potential of automated systems in the field of veterinary medicine and beyond.

**Acknowledgement.** No.

**Conflict of Interest.** No.

## Список використаної літератури

Жаботинська С. А. Когнітивна лінгвістика: принципи когнітивного моделювання. Лінгвістичні студії.– Черкаси: Сіяч, 2013. С. 3-11.

Жаботинська С. А. Концептуальний аналіз мови: фреймові мережі. Мова. Науково-теоретичний часопис з мовознавства, (9), 2015. С. 81-92.

Жаботинська С. А. Ім'я як текст: концептуальна мережа логічного значення (аналіз назви емоції). Когніція, комунікація, дискурс. (6), 2014. С. 47-76.

Жуковська, В. В. Вступ до корпусної лінгвістики: навчальний посібник. Житомир: Видавництво ЖДУ ім. Івана Франка, 2013. 146 с.

Перхач Р. Застосування комп'ютерних технологій при дослідженні медичної термінології. Львів: Видавництво львівської політехніки, 2015. С. 186-188.

Перхач Р. Ю. Корпус інструкцій до медичних препаратів як метод дослідження медичної термінології. Філологічні студії. Науковий вісник Криворізького державного педагогічного університету, (13), 2017. С. 171-175.

Abraham-Barna C. G. Creating a French-Romanian Bilingual Terminology of Veterinary Parasitology. Bulletin UASVM Horticulture, 68(2), 2011. P. 87-95.

Baker M. Corpora in translation studies: An overview and some suggestions for future research. Target. International Journal of Translation Studies, 7(2), 2015. P. 223-243.

Baker P. Glossary of corpus linguistics. Edinburgh: Edinburgh University Press, 2019. 187 p.

Biber D. Corpus linguistics: Investigating language structure and use. Cambridge University Press, 2018. 341p.

Biber D. Corpus-based and corpus-driven analyses of language variation and use. The Oxford handbook of linguistic analysis. Oxford University Press., 2012. 545 p.

Bowker L. Towards a corpus-based approach to terminography. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 3(1), 2016. P. 27-52.

Edward B. Black's Veterinary Dictionary, 2015. 514 p.

Geeraerts D. Introduction: Prospects and problems of prototype theory. Linguistics, 27(4), 2011. P. 587-612.

Gonzales M. C. Variation in Spanish Accounting Terminology Implications for Translators. Terminology, 28(1), 2022. P. 65-102.

McFarland C. Veterinary Dictionary and Horseman's Guide. Phoenix: Western Horseman, 2013. 192 p.

Oostdijk N. A corpus linguistic approach to linguistic variation. Amsterdam: Literary and Linguistic Computing, 3(1), 2010. P. 12-25.

Oostdijk N. Corpus linguistics and the automatic analysis of English. Amsterdam-Atlanta(GA): Rodopi, 2021. 269 p.

Rozhkov Yu., Syrotin, O. Verbalization of the concepts of disease and animal disease in English. Cogito multidisciplinary research journal. Bucharest. Vol. XIII (4), 2022. P. 224-233.

Wiebe J. Learning subjective adjectives from corpora. New-Mexico: New-Mexico State University, 20(4), 2010. P. 54-70.

## References

Zhabotynska, S. A. (2013). Kohnityvna linhvistyka: pryntsypy kohnityvnoho modelyuvannya [Cognitive linguistics: principles of cognitive modeling]. Linguistic studies.–Cherkasy: Siyach, P. 3-11.

Zhabotynska, S. A. (2015). Kontseptual'nyy analiz movy: freymovi merezhi [Conceptual analysis of language: frame networks]. Language. Scientific and theoretical journal of linguistics, (9), P. 81-92.

Zhabotynska, S. A. (2014). Imya yak tekst: kontseptual'na merezha lohichnoho znachennya (analiz nazvy emotsiyi) [The name as a text: a conceptual network of logical meaning (analyzing the name of an emotion)]. Cognition, communication, discourse. (6), P. 47-76.

Zhukovska, V. V. (2013). Vstup do korpusnoyi linhvistyky: navchal'nyy posibnyk [Introduction to corpus linguistics: a study guide]. Zhytomyr: ZhDU Publishing House named after Ivan Franko, 146 p.

Perkhach, R. (2015). Zastosuvannya komp"yuternykh tekhnolohiy pry doslidzhenni medychnoyi terminolohiyi [Application of computer technologies in the study of medical terminology]. Lviv: Lviv Polytechnic Publishing House, P. 186-188.

Perkhach, R. (2017). Korpus instruktsiy

do medychnykh preparativ yak metod doslidzhennya medychnoyi terminolohiyi [Corpus of instructions for medical preparations as a method of researching medical terminology]. Philological studies. Scientific Bulletin of Kryvyi Rih State Pedagogical University, (13), P. 171-175.

Abraham-Barna, C. G. (2011). Creating a French-Romanian Bilingual Terminology of Veterinary Parasitology. Bulletin UASVM Horticulture, 68(2), P. 87-95.

Baker, M. (2015). Corpora in translation studies: An overview and some suggestions for future research. Target. International Journal of Translation Studies, 7(2), P. 223-243.

Baker, P. (2019). Glossary of corpus linguistics. Edinburgh: Edinburgh University Press, 187 p.

Biber, D. (2018). Corpus linguistics: Investigating language structure and use. Cambridge University Press, 341p.

Biber, D. (2012). Corpus-based and corpus-driven analyses of language variation and use. The Oxford handbook of linguistic analysis. Oxford University Press., 545 p.

Bowker, L. (2016). Towards a corpus-based approach to terminography. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, 3(1), P. 27-52.

Edward, B. (2015). Black's Veterinary Dictionary, 514 p.

Geeraerts, D. (2011). Introduction: Prospects and problems of prototype theory. Linguistics, 27(4), P. 587-612.

Gonzales, M. C. (2022).Variation in Spanish Accounting Terminology Implications for Translators. Terminology, 28(1), P.65-102.

McFarland, C. (2013). Veterinary Dictionary and Horseman's Guide. Phoenix: Western Horseman, 192 p.

Oostdijk, N. (2010). A corpus linguistic approach to linguistic variation. Amsterdam: Literary and Linguistic Computing, 3(1), P. 12-25.

Oostdijk, N. (2021). Corpus linguistics and the automatic analysis of English. Amsterdam-Atlanta(GA): Rodopi, 269 p.

Rozhkov, Yu., Syrotin, O. (2022). Verbalization of the concepts of disease and animal disease in English. Cogito multidisciplinary research journal. Bucharest. Vol. XIII (4), P. 224-233.

Wiebe J. (2010). Learning subjective adjectives from corpora. New-Mexico: New-Mexico State University, 20(4), P. 54-70.

## Лінгвокогнітивний підхід до вилучення термінів з корпусу ветеринарних текстів

**Юрій РОЖКОВ**
доктор філософії з філології,
доцент кафедри іноземної філології і перекладу,
Національний університет біоресурсів і природокористування України,
03041, Героїв Оборони, 15, Київ, Україна
E-mail: yuriev694@gmail.com
https://orcid.org/0000-0002-6830-9130

**Анотація.** Це дослідження заглиблюється в складний ландшафт комп'ютерної лінгвістики з цілеспрямованим дослідженням проблем ідентифікації термінів у сфері ветеринарної медицини. Був проведений всебічний аналіз, включаючи автоматизоване виділення однослівних термінів та структуровані моделі, що спостерігаються в двослівних термінах у ветеринарних словниках та науковій літературі.

Дослідження розпочалося з докладної ідентифікації 462 однослівних термінів. Особливий акцент був зроблений на проблемах, притаманних автоматизації виділення термінів, що характеризуються мовною різноманітністю та потенційною двозначністю.

Виявлено, що лише 59% термінів розміщуються в розділі заголовка, це підкреслює потребу в адаптивних методах вилучення, налаштованих на різноманітний розподіл термінів у структурі словника. Наукові тексти ще більше ускладнили ландшафт ідентифікації термінів, демонструючи різну частотність термінів, спонукаючи до критичної оцінки стандартних методів відбору лексем.

Спираючись на ці ідеї, дослідження пропонує стратегії вдосконалення процесів автоматизованої ідентифікації термінів. Це включає в себе використання передових методів обробки природної мови для однослівних термінів і підтримку адаптивних методів вилучення для словників, а також пропозицію гібридного підходу для наукових текстів.

Міждисциплінарний характер дослідження підкреслюється визнанням співпраці між лінгвістами, науковцями з обчислювальної техніки та експертами в галузі як вирішальної для розробки складних моделей і онтологій, які точно відображають унікальні лінгвістичні нюанси ветеринарної медицини.

Оскільки цифровий ландшафт продовжує свій розвиток, це дослідження сприяє не лише розвитку комп'ютерних лінгвістичних методологій, але й передбачає створення термінологічних ресурсів, які відображають динамічну природу мови у сфері ветеринарії. Завдяки всебічному аналізу проблем і можливостей, це дослідження прагне вибудувати шлях для більш точних і адаптованих автоматизованих систем, відкриваючи перспективи для розширення області комп'ютерної лінгвістики.

**Ключові слова:** корпус, корпусний аналіз, частотний аналіз, ветеринарна термінологія, когнітивістика.