

Assessment of terminological density in scientific publications on physical culture

Sergii Iermakov^{1ABCDE}, Georgiy Korobeynikov^{2,3,4ABCDE}, David Curby^{5ABCDE}

¹ Department of Methodologies of Cross-Cultural Practices, Kharkiv State Academy of Design and Arts, Ukraine

² Uzbek State University of Physical Education and Sports, Uzbekistan

³ Institute of Psychology, German Sport University Cologne, Germany

⁴ Department of Combat Sports and Power Sports, National University of Physical Education and Sport, Ukraine

⁵ International Network of Wrestling Researchers, USA

Authors' Contribution: A – Study design; B – Data collection; C – Statistical analysis; D – Manuscript Preparation; E – Funds Collection

Abstract

Background and Study Aim Scientific publications in the field of physical culture demonstrate considerable diversity in terminological usage and structural organization. With increasing standards for the quality of academic writing, the need for an objective and quantitative evaluation of terminological density has become more pressing. The aim of this study was to develop and apply a method for automated assessment of terminological density in scientific articles on physical culture using adapted thematic dictionaries.

Material and Methods The study was based on articles retrieved from the Web of Science (WoS) database. A total of 16 593 bibliographic records related to physical culture were extracted over the past five years. Two dictionaries were employed for analysis: the official Medical Subject Headings (MeSH) in XML format and a thematic dictionary constructed from the WoS document corpus. The analysis included full-text PDF articles from 12 scientific journals, of which 6 were categorized as Q3, 1 as Q4, 3 were indexed in DOAJ, and 2 were not indexed. Terminological density was calculated in Python using the pandas library and evaluated on a scale ranging from very low to high.

Results The assessment covered 12 journals in the field of physical culture. An optimal density level (0.010–0.019) was identified in 2 journals (16.7%), corresponding to a “balanced use of scientific terminology.” Three journals (25.0%) demonstrated low density (<0.01), characterized as “insufficient elaboration of the topic in scientific language”. In 7 journals (58.3%), a higher density (0.020–0.039) was observed, interpreted as either an “attempt to enhance scientific rigor” or an “excessive terminological load”.

Conclusions The evaluation of terminological density provides an objective measure of the scientific style of publications in the field of physical culture. The differences identified across journals highlight variability in approaches to presenting scientific material. The integration of specialized dictionaries and the application of relative indicators offer a robust basis for ongoing monitoring and optimization of scientific discourse.

Keywords: terminological density, sport and exercise sciences, scientific publications, thematic dictionaries, academic discourse, Medical Subject Headings (MeSH), Web of Science, bibliographic analysis, quantitative evaluation, scientific style

Introduction

Contemporary scientific communication requires adherence to established standards of structure and content in scholarly publications. Among these standards, the use of professional terminology plays a central role, ensuring the accuracy of scientific expression, thematic identification of research, and effective indexing in digital databases. The application of precise scientific vocabulary influences not only the perception and interpretation of information but also the indexing of articles in major bibliographic systems such as Web of Science and Scopus. This, in turn, determines the visibility and citation impact of publications, as well as

the overall influence of academic journals. Such practices provide a foundation for both quantitative and qualitative analyses of the linguistic dimension of scientific texts, reveal the depth of thematic elaboration, and contribute to addressing tasks related to the optimization of scientific discourse in the field of physical culture and sport.

In recent years, automated analysis of scientific texts has become an important tool in bibliometrics and the evaluation of research output. One of the key methods is TF-IDF (term frequency–inverse document frequency), which is applied to determine the significance of terms within a text [1, 2]. This approach is widely used for thematic indexing, keyword analysis, and the assessment of scientific publication quality [3, 4].

In medicine and related disciplines, the MeSH (Medical Subject Headings) thesaurus, developed by the U.S. National Library of Medicine, is frequently applied [5]. It provides a standardized terminology and is extensively used in systematic reviews and in the classification of scientific articles [6, 7]. The Web of Science database also employs its own keywords, which are listed separately from the authors' keywords.

At the intersection of these approaches (TF-IDF and MeSH), effective methods have emerged for the quantitative and content-based evaluation of scientific publications. Their applicability has been confirmed in several studies, including those in the field of physical culture and sport sciences. Kiss et al. [8] conducted a bibliometric analysis of publications on sports nutrition using MeSH keywords and science mapping techniques to identify major research themes. Venâncio et al. [9] performed a bibliometric and scoping review to trace the evolution of research on strength training among competitive swimmers. Jagiello and Lochbaum [10, 11] applied Python-based algorithms to analyze publications in the pedagogy of physical activity. Yermakova [12, 13] employed topic modeling methods to investigate literature on sports injuries and rehabilitation.

The use of terminological dictionaries, both controlled (MeSH) [14] and empirical (constructed from WoS data), provides a more objective approach to evaluating terminological density. This enables a quantitative characterization of thematic rigor, scientific maturity, and the relevance of publications, as well as the comparison of journals and articles according to unified criteria.

The use of thematic lexicons, both manually constructed and automatically extracted from scientific databases, has long been applied in bibliometrics, thematic analysis, and the evaluation of scientific content. The effectiveness of controlled vocabularies (e.g., MeSH [5]) as well as empirically derived dictionaries (e.g., Keywords Plus from Web of Science [15]) has been well documented in the literature.

Scientometric analysis tools such as VOSviewer and SciMAT rely on frequency-based term dictionaries from WoS and Scopus for constructing science maps and thematic clusters [16, 17]. At the same time, the MeSH thesaurus [5] has traditionally been employed in biomedical analyses, including automatic classification of scientific texts [18], profiling of academic journals [Kim2016], assessment of thematic relevance [19], and standardization of density intervals [20]. Several studies suggest combining free keywords (WoS) with controlled descriptors [14] to improve the accuracy of thematic classification [21]. A similar approach was applied by Wang et al. in their analysis of literature on physical activity and health, where both WoS and PubMed data were

used simultaneously [22].

Comparative studies of the terminological structures of Scopus and Web of Science have demonstrated that the combined volume of author keywords and Index Keywords in Scopus is more than twice that of the corresponding field in WoS [23]. Furthermore, van Eck and Waltman [24] showed that Scopus keywords are well suited for constructing thematic maps and publication clusters. Comparable approaches are implemented in software tools such as Bibliometrix, where author and indexed keywords from Scopus serve as the basis for semantic and co-word analysis [25].

Threshold values for interpreting terminological density are based on empirical observations and indexing standards. For example, Leblanc et al. demonstrated that incorporating a sufficient number of MeSH terms into the search strategies of systematic reviews significantly improves retrieval completeness, indirectly highlighting the importance of high terminological density [19]. Breuer et al. emphasized that concentrating key terms and identifying a "core" set of documents increases representativeness when analyzing reduced samples [26]. At the same time, values exceeding 4% may impair text readability even for specialists [27].

Within the framework of information standards, the U.S. National Library of Medicine recommends indexing publications using 5–15 MeSH terms per 1, 000 words, corresponding to a density of approximately 1–1.5% [5]. Similarly, Scopus employs keywords to enhance thematic search functions and provides the basis for automated topic matching [28].

Thus, despite the absence of a strict hierarchy, the terminological data in Scopus represent a well-grounded empirical resource for dictionary construction and the evaluation of scientific texts. Moreover, the practice of creating adapted dictionaries based on MeSH in combination with frequency-derived empirical terms from WoS and Scopus terminological data is supported in the literature and is methodologically justified.

Terminological density represents a quantitative indicator that reflects the proportion of specialized scientific terms within the total volume of significant words. This measure is widely applied in bibliometric studies, automatic indexing, and the evaluation of the scientific orientation of texts. The rationale for employing this approach lies in the attempt to objectively capture the thematic focus and academic style of scholarly publications.

In a study by Ding et al. [29], co-word analysis was used to construct maps of research directions in the field of information retrieval, demonstrating the effectiveness of frequency-based term analysis for identifying themes and subdisciplines. A similar approach was applied by Haunschild et al. [30], who utilized automatically generated Keywords Plus

terms from the Web of Science database to analyze publications on climate change. These findings confirmed the value of automatically extracted terminology for developing thematic profiles of scientific domains.

From the perspective of computational text processing, the TF-IDF method remains one of the most widely applied approaches for identifying significant terms. Wang [3] demonstrated that combining TF-IDF with semantic analysis enables the effective extraction of keywords from texts, including scientific publications. This approach can be used to evaluate both the density and orientation of the academic style of an article. In databases with well-developed controlled vocabularies, such as those in medicine, the MeSH thesaurus is extensively employed. Bekhuis et al. [31] examined the coverage of concepts related to comparative studies in MeSH and Emtree, underscoring the importance of standardized terminology in the systematization of scientific content. Similarly, Koloski et al. [32] demonstrated how neural network-based keyword extraction methods can be complemented by TF-IDF and aligned with pre-defined terminological lists, thereby enhancing the accuracy of thematic analysis.

To determine the appropriate number of documents for analyzing terminological density, the Pareto principle (80/20 rule) is often recommended. Valkanas and Diamandis argued that approximately 20% of publications may account for up to 80% of citations, reflecting the uneven distribution of scientific attention [33]. In bibliometric research, the Pareto rule is frequently applied to identify core journals, where about 20% of titles generate around 80% of citations or usage, thereby ensuring representativeness even with a reduced dataset [34].

Thus, terminological density can serve as a reliable indicator of scientific relevance, thematic precision, and stylistic rigor of texts. Its application is justified both in tasks of automated classification of publications and in expert evaluation of scholarly materials.

The assessment of terminological density in scientific articles makes it possible to identify their thematic orientation and to conduct a structural analysis of scientific rigor. However, excessive concentration of terms may indicate stylistic oversaturation and the use of “pseudo-scientific” vocabulary without sufficient semantic depth. This underscores the need to define an optimal range of terminological density.

Solnyshkina et al. [35] emphasized the importance of lexical density as an indicator of complexity and stylistic level in educational texts, which can also be applied to scientific publications. Recommendations of the U.S. National Library of Medicine [5] suggest that the number of MeSH terms should range from

5 to 15 per 1000 words, corresponding to a density of approximately 1–1.5%. According to Halliday’s concept of lexical density [36], term overload (>4%) reduces accessibility and makes comprehension less effective, even for specialists.

Other studies provide indirect evidence supporting the proposed range of terminological density through textual descriptions or methodological approaches. Some works focus on linguistic analysis and the evaluation of the “scientificness” of texts, including terminological richness and distinctions between specialized and popular materials across subject areas [37, 38]. Other studies examine computational methods of text processing, including TF-IDF, topic modeling, automatic term extraction, and lexical density analysis, which may be employed to determine the relative proportion of terms in a text [39, 40].

One of the key factors influencing the accuracy of assessing terminological density in scientific texts is the consideration of synonymic and morphological variations of terms. Fu et al. proposed the SynGen method, which employs regularizers for synonym generalization in biomedical NER, thereby improving the completeness of concept extraction even beyond predefined dictionaries [41]. This is particularly relevant in the analysis of precise terms, where word forms and synonyms may differ substantially. Slater et al. [42] demonstrated that extending terms through inter-ontology synonymy significantly increases coverage in medical text analysis. Moreover, methods that combine contextual analysis with structural relationships enable the extraction of synonyms, hypernyms, and hyponyms, enhancing both terminological completeness and consistency [43]. In the field of NLP, Thießen et al. [44] showed that large language models (BERT, RoBERTa, GPT-3) are capable of detecting scientific synonyms through clustering of hidden representations, which directly contributes to more accurate term normalization.

Studies of terminological density in the context of physical culture illustrate approaches to identifying and quantitatively assessing terms in applied domains. For example, Pans et al. proposed a methodology for evaluating the terminological contribution of key concepts across different ranges of research [45]. Another approach demonstrated that collocations and derivative forms play a dominant role in shaping terminological density within sports-related vocabulary [46, 47]. The historical and cultural grounding of terminology, along with its standardization, enhances the relevance of professional communication and improves the efficiency of thematic searches [48, 49]. In digital sports discourse, the systematic identification and classification of terms help assess their frequency, context, and semantic variability, thereby improving sample representativeness

and fostering interdisciplinary interaction [50]. Corpus-based observations of synonymy and variability confirm that high terminological density is characteristic of specialized texts and reflects the specificity of individual sports [51]. In this regard, the thematic classification of sports terminology serves as a foundation for the quantitative evaluation of terminological composition and for analyzing the informational richness of materials [52].

Thus, the inclusion of synonyms and word forms in terminological dictionaries: increases the sensitivity of analysis without compromising specificity; accounts for linguistic variability in the expression of the same conceptual entity; standardizes terminological analysis in the evaluation of scientific texts; and enhances the robustness of assessments when analyzing large document collections.

A review of numerous studies has shown that assessing terminological density through the use of thematic dictionaries and structured metadata is an effective tool for exploring scientific discourse. Scholars emphasize that such approaches contribute to more objective comparisons of publications, the identification of research trends, and the standardization of scientific vocabulary across disciplines. These methods are equally applicable to the domain of physical culture and sport. However, although some publications have addressed linguistic or content-based analyses in this field, they have rarely focused on term-centered evaluations of full-text documents. Such studies do not fully reveal important aspects of scientific communication, including lexical rigor, thematic relevance, and terminological density. This highlights the need for a more comprehensive and systematic analysis of scientific publications in this area.

The aim of this study was to develop and implement a method for the automated assessment of terminological density in scientific articles in the field of physical culture using adapted thematic dictionaries.

Materials and Methods

Sources of Information

The Web of Science (WoS) platform was used as the primary database. A total of 16 593 bibliographic records of articles published over the past five years were retrieved based on the search query. From the selected corpus, five thematic frequency-based dictionaries of terms were compiled, containing the relevant keywords. In addition, the official Medical Subject Headings (MeSH) thesaurus [14], obtained in XML format (desc2025.xml), was employed.

Full-text articles in PDF format were collected from 12 Ukrainian journals indexed in WoS and/or Scopus: six journals categorized as Q3, one journal

categorized as Q4, three journals indexed in DOAJ, and two journals without indexing. For each journal, articles were obtained from the first available issues published in 2025.

Study Design

The study employed a descriptive-analytical design incorporating bibliometric and linguistic methods. To construct the document corpus, a search query was formulated in the Web of Science (WoS) database using the following keywords: “physical activity” OR “physical education” OR “physical culture” OR “physical fitness” OR “aerobic exercise” OR “resistance training” OR “exercise physiology” OR “motor skills” OR “sports science” OR “athletic performance” OR “training load” OR “endurance” OR “strength training” OR “health-related fitness” OR “sports training” OR “youth sports” OR “exercise intervention” OR “exercise behavior.” The search was restricted to a five-year time frame and to articles published in English.

In addition to keyword filtering, six Web of Science subject categories were applied: Sport Sciences, Public Environmental Occupational Health, Education Educational Research, Hospitality Leisure Sport Tourism, Rehabilitation, and Physiology. These categories were selected as the most relevant to the field of physical culture and related disciplines. They encompass key dimensions ranging from sports science, pedagogy, health promotion, and rehabilitation to the biophysiological foundations of physical activity.

Beyond thematic relevance, the selection was also guided by the informational richness of these six subject categories (Table 1, Figure 1), which accounted for approximately 77.7% of all documents retrieved on the topic of physical culture over the past five years. By contrast, the next 14 categories contributed only 14.7% of publications, while an additional 37 categories represented about 6.7%. The remaining 50 categories contained fewer than 10 publications each, collectively amounting to less than 1% of the total dataset. This distribution provided the rationale for restricting the corpus to the most representative and thematically significant domains.

Table 1. Distribution of categories by number of publications

Group of Categories	Range of Publications per Category	Number of Categories
Six core categories	1400–6000	6
Next 14 categories	100–500	14
Next 37 categories	10–94	37
Remaining 50 categories	<10	50

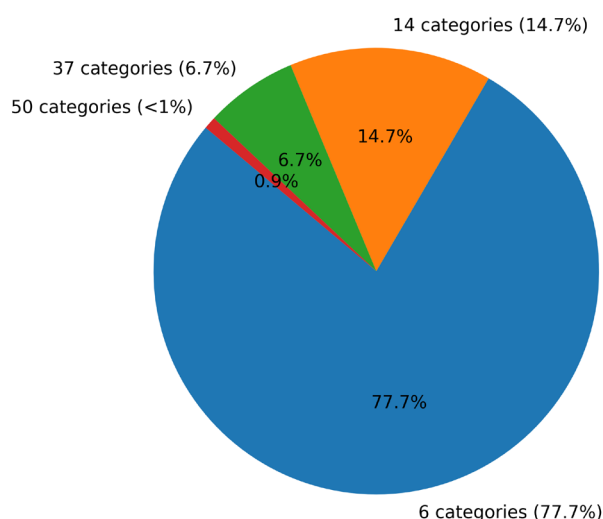


Figure 1. Distribution of publications across subject categories.

As a result, 16593 unique records that met the specified criteria were selected. Using the “Export Records to Plain Text File” function, the records were downloaded in batches of 1000 documents. A total of 17 batches were exported and subsequently merged into a CSV file for analysis.

For the unit of analysis, full-text PDF articles from 12 scientific journals were considered. The inclusion criteria comprised the availability of full texts, relevance to the subject area, and publication in English. Data processing involved extracting the main body of the articles (up to the “References” section), removing noise and technical symbols, eliminating short and insignificant words, and counting terms based on five thematic dictionaries, with attention to synonyms and word forms. All analyzed data are publicly available and do not contain any personal information.

Five different lexical dictionaries were created, each reflecting a specific aspect of the terminological composition of the texts. Each dictionary contained keywords and their frequency of occurrence in the corpus. In addition, the sixth dictionary, MeSH [14], was employed.

The dictionaries were constructed on the basis of 16593 bibliographic records. Using the CountVectorizer method [53], the 1000 most frequent n-grams (1–3 words) were extracted. From this set, the top 300 terms were selected for further analysis, in line with Zipf’s law and recommendations for building controlled vocabularies, where the core semantic field is captured by a limited number of frequent units [54, 55].

Subsequently, semantic filtering was applied: constructions containing numbers, abbreviations, and phrases consisting solely of short words (≤ 3 characters) were removed. From the cleaned dataset, the top 300 terms were retained as the most representative and thematically relevant.

For all dictionaries, a procedure was conducted to identify lexically and morphologically related forms. New terms were generated based on the stem or plural form of the original words. These derivative forms were grouped with their corresponding base terms and recorded in a separate column of the table.

For the first four dictionaries (Dictionary_1, Dictionary_2, Dictionary_3, and Dictionary_4), the most frequent expressions were extracted from the full dataset. Filtering was then applied to remove duplicates, non-specific and overly general phrases, as well as elements containing numerical characters. Newly generated derivative forms—synonymic and morphological variants—were added to the dictionaries. From the cleaned lists, the top 300 terms were selected, representing stable thematic categories used for the formal classification of publications. As a result, the following dictionaries were constructed:

- Dictionary_1 (AU). Created from the corpus of author texts, including TI (article title), AB (abstract), and DE (author keywords).
- Dictionary_2 (DE). Created from the corpus of author keywords (DE). A total of 16, 512 unique DE terms were identified.
- Dictionary_3 (ID). Created from the corpus of keywords added by WoS (ID). A total of 9, 018 unique ID terms were identified.
- Dictionary_4. Created from the combined corpus of author keywords (DE) and WoS-assigned keywords (ID).

In addition, Dictionary_5 (Adapted Dictionary) was developed, containing author terms that also appear in the MeSH thesaurus.

Finally, Dictionary_6 was compiled from the official version of MeSH, downloaded in XML format from the U.S. National Library of Medicine (desc2025.xml) [14]. From the hierarchical tree structure of MeSH, descriptors relevant to the core categories—Physical Activity, Sports, Rehabilitation, Health, and related domains—were extracted.

To evaluate the terminological structure of scientific publications, an algorithm for the automated processing of PDF files was developed in Python using the PyMuPDF (fitz) library and a set of predefined dictionaries.

After constructing five dictionaries and converting the sixth MeSH dictionary into an analyzable format, a term search was conducted across all six dictionaries. Unique matches with each dictionary were identified within the text of each article, and terminological density was calculated accordingly. A composite indicator (Overall Score) was then computed as the mean value of densities across all dictionaries, which was subsequently used to generate a ranking of articles within each journal. For each journal, the final dataset also included a summary row with the average values of the metrics.

The results were stored as CSV files for each

journal and were further used for cross-journal comparisons and the construction of summary analytical tables.

For the assessment of terminological density, the following indicator was applied:

Terminological density (or the relative weight of terms) represents the proportion of unique matched terms from a dictionary within the text of a scientific article relative to the total number of meaningful words. Density was calculated according to the formula:

$$D = M / N,$$

where D is the terminological density (relative weight), M is the number of unique terms identified in the text, and N is the total number of meaningful words (excluding stop words, numbers, short forms ≤ 2 characters, and punctuation).

Density was computed separately for each of the dictionaries. Terms were considered matched if they appeared in the text in any of the accounted forms (synonyms or morphological variants). For each document, the following additional indicators were also calculated:

- the total number of meaningful words;
- the number of matched terms from each dictionary;
- the final combined score, defined as the average of the dictionary-based density values.

For the objective interpretation of terminological density (the relative weight of matched terms), an empirical interpretation scale was developed. The scale was grounded in evidence from contemporary studies on terminological analysis, bibliometrics, and information retrieval practices [19, 26, 27].

To interpret the results, a scale of terminological density was applied, based on empirically validated thresholds (Table 2). This approach made it possible to account for the relative position of an article within the overall distribution and to mitigate the influence of differences in text length and total number of terms. The scale ensured an objective comparison of publications regardless of their structure or length.

Table 2. Classification of terminological density levels in scientific texts

Terminological Density (Relative Weight)	Level	Interpretation
< 0.005	Low	Weak thematic saturation
0.005–0.01	Below average	Low saturation
0.01–0.02	Medium	Optimal saturation
0.02–0.04	Above average	Increased saturation
> 0.04	High	Terminological overload

To summarize the characteristics of terminological density across scientific journals, an algorithm for the automated compilation of an aggregated table was implemented based on previously calculated metrics for individual publications. The input data consisted of the term analysis results for each journal, stored in CSV format.

The algorithm included the following steps:

1. Loading the reference table (list_journals.csv) containing the correspondence between journal codes and their full titles.
2. Iterating through all final result files (results_*.csv) containing term analysis data for individual journals.
3. Extracting from each file the row marked as “Average for journal,” which contained the mean metric values across the journal’s articles.
4. Assigning the full journal title based on the reference table and incorporating it into the aggregated list.
5. Constructing the final summary table, which included the following fields:
 - total number of words,
 - number and density of matched terms for each of the five dictionaries (WoS, MeSH, DE, ID, AU),
 - the integrated indicator of terminological density (Overall Score),
 - the journal’s rank position in the comparative list.

The summary table (journals_summary_form_de_id_au.csv) was saved in a separate directory for further analysis, visualization, and interpretation. This step ensured the transition from article-level analysis to journal-level generalization, enabling the identification and quantitative evaluation of differences in thematic density among journals.

For the quantitative evaluation of terminological density in scientific publications, all individual articles published in the selected journals were analyzed. In total, 103 articles were processed, and key terminological metrics were calculated for each based on five dictionaries.

Stage 1: Construction of a unified article database.

At the first stage, results previously obtained for each journal were aggregated. From all tables containing metrics for individual articles, only the rows with data for specific publications (excluding the “Average for journal” row) were selected. For each article, the identifier of the corresponding journal was additionally recorded. All entries were merged into a single table containing:

- the number of words,
- the number and density of matched terms from the five dictionaries (WoS, MeSH, DE, ID, AU),
- the overall terminological density indicator (Overall Score).

The resulting table (all_articles_combined_de_id_

au.csv) included data for all publications and served as the basis for subsequent quantitative analysis.

Stage 2: Correlation analysis of terminological metrics. At the second stage, a correlation analysis of term densities across the five dictionaries was conducted. Spearman's correlation was applied to assess the degree of concordance among the rankings of terminological densities between articles, regardless of the scale of the values.

A matrix of pairwise Spearman coefficients was calculated for the following indicators:

- WoS Density
- MeSH Density
- DE Density
- ID Density
- AU Density.

Statistical Analysis

Data processing was carried out using the Python programming language (version 3.11). The pandas library was employed for data aggregation and analysis. Terminological density was calculated as a relative indicator, expressed as the proportion of matched terms from the five lexical sources (WoS, MeSH, DE, ID, AU) in relation to the total number of meaningful words in the text. At the stage of summary analysis, mean values were computed for each journal, ranking was performed according to the overall integrated score, and the percentage distribution of terminological density levels was determined based on the established classification. To normalize the results, a linear transformation scale ranging from 0 to 10 was applied, ensuring comparability of outcomes regardless of text length or number of terms. In addition, group-level analysis of density was carried out across

the following categories: *very low*, *low*, *moderate*, *increased*, and *high* terminological density, which provided a quantitative characterization of the publication corpus. For the purpose of identifying relationships between lexical sources, correlation analysis was performed using Spearman's rank correlation coefficient at the level of individual articles. The resulting correlation matrix included all five terminological density indicators.

Results

Table 3 presents the values of terminological density for individual publications from one of the journals across the five dictionaries: WoS, MeSH, DE, ID, and AU. According to the established benchmark, the optimal range of terminological density is 0.01–0.02 terms per word. Comparison of the observed values with this reference range provides the basis for a qualitative assessment of the lexical structure of the publications.

The analysis of term density in the journal's articles (Table 3) revealed the following trends. A comparison of terminological density with the recommended range of 0.01–0.02 terms per word indicates a systematic exceedance across all five dictionaries. Hyper-saturation is most pronounced in the WoS and AU dictionaries, where values in individual publications reached 0.27 and 0.24, respectively—exceeding the upper threshold by a factor of 10–12. Even in the least saturated article (Overall Score = 0.039), density remained approximately twice above the recommended level. While the formalized dictionaries (MeSH, DE, ID) yielded comparatively lower values, they likewise exceeded the standard thresholds.

These findings indicate a high lexical density

Table 3. Assessment of terminological density in the journal

File	Word Count	WoS Hits	WoS Density	MeSH Hits	MeSH Density	DE Hits	DE Density	ID Hits	ID Density	AU Hits	AU Density	Overall Score	Rank
Article 1.pdf	615	169	0.2748	49	0.07967	68	0.11057	74	0.12033	149	0.24228	0.16553	1
Article 2.pdf	860	181	0.21047	52	0.06047	90	0.10465	96	0.11163	179	0.20814	0.13907	2
Article 3.pdf	480	89	0.18542	23	0.04792	32	0.06667	44	0.09167	87	0.18125	0.11459	3
Article 4.pdf	898	146	0.16258	38	0.04232	56	0.06236	66	0.07350	149	0.16592	0.10134	4
Article 5.pdf	2771	162	0.05846	57	0.02057	64	0.02310	77	0.02779	183	0.06604	0.03919	5
Article 6.pdf	3319	187	0.05634	50	0.01506	80	0.02410	100	0.03013	205	0.06177	0.03748	6
Article 7.pdf	2733	158	0.05781	44	0.01610	61	0.02232	76	0.02781	172	0.06293	0.03739	7
Article 8.pdf	3943	202	0.05123	63	0.01598	89	0.02257	100	0.02536	219	0.05554	0.03414	8
Average for journal (ppcs)	1952.38	161.75	0.13214	47.00	0.03726	67.50	0.05454	79.13	0.06353	167.88	0.13048	0.08359	4.5

Note. *WoS Density*— density of keywords derived from both author and WoS sources. *MeSH Density*— density of keywords corresponding to the MeSH standard. *DE Density*— density of author-provided keywords (DE). *ID Density*— density of WoS-assigned keywords (ID). *AU Density*— density of author terms derived from the title, abstract, and keywords of the publication. *Overall Score*— arithmetic mean of the five densities. *Rank*— position of the article in descending order of overall density score.

characteristic of scientific publications in applied domains. None of the articles fell within the normative range, underscoring the need to address standards of terminological load in the preparation and editing of scholarly texts.

As part of the summary analysis, an aggregated table was compiled to reflect the indicators of terminological density for 12 scientific journals in the field of physical culture. Table 4 presents the numerical values of terminological density for each of the five dictionaries (WoS, MeSH, DE, ID, AU), as well as the integrated indicator (Overall Score), which represents the mean value of the relative densities.

Using the normative interval of 0.01–0.02 terms per word as the reference range for terminological density, the following tendencies were identified:

1. The indicators of all journals substantially exceed the normative range across most dictionaries. This is particularly evident for:
 - *WoS Density* – with leading journals reaching values of up to 0.35;
 - *AU Density* – averaging above 0.20.
2. The densities of *MeSH*, *DE*, and *ID* are lower, but in many cases also exceed the normative range.
3. None of the journals fall entirely within the normative range across all dictionaries.
4. The top-ranked journal (Overall Score = 0.199) demonstrates the highest overall saturation, with marked exceedances across all densities, particularly *WoS* (0.355) and *AU* (0.247).

A more detailed interpretation of the obtained

results is presented in Figure 2, which shows the deviations of actual terminological density values from the normative range (0.01–0.02 terms per word). The calculation was performed for each of the five dictionaries: *WoS*, *MeSH*, *DE*, *ID*, and *AU*. Positive deviation values indicate an excess over the upper limit of the norm, reflecting potential oversaturation of the text with terms.

The analysis of the data in Figure 2 indicates the following tendencies:

1. All journals demonstrate values exceeding the normative range across each of the five dictionaries. None of the indicators fall within the interval 0.01–0.02, confirming the oversaturation of texts with specialized terminology.
2. The largest deviations are observed in:
 - *AU Density* – up to +0.226 (journal N1),
 - *WoS Density* – up to +0.335 (the same journal), making these sources the primary contributors to terminological load.
3. Deviations in *MeSH*, *DE*, and *ID* are considerably lower but consistently positive (on average 0.03–0.09), reflecting a more moderate but still formalized level of terminologization.
4. Journal N1 stands out as the leader in all indicators, demonstrating the maximum deviations across all metrics.
5. Even the lowest-ranked journals (those with the smallest Overall Score) show deviations not falling below +0.03, which highlights the general

Table 4. Summary indicators of terminological density for 12 scientific journals in the field of physical culture

Rank	Journal	Word Count	WoS Hits	WoS Density	MeSH Hits	MeSH Density	DE Hits	DE Density	ID Hits	ID Density	AU Hits	AU Density	Overall Score
1	N1	1385.0	91.8	0.35507	26.8	0.12951	27.0	0.15644	33.0	0.10979	91.2	0.24655	0.19947
2	N2	1011.0	91.4	0.25944	29.6	0.08645	31.8	0.07651	36.4	0.09449	85.8	0.20547	0.14447
3	N3	1131.6	105.8	0.21123	30.2	0.05058	39.2	0.06708	45.4	0.07384	110.0	0.24393	0.12934
4	S1	1463.1	120.3	0.20116	36.1	0.06309	41.1	0.07521	48.8	0.06122	116.7	0.14768	0.10967
5	S2	1869.1	115.4	0.15521	33.7	0.04295	41.1	0.04887	48.6	0.05124	122.0	0.15498	0.09065
6	S3	1747.4	149.6	0.13725	44.4	0.04171	55.6	0.05259	66.4	0.06090	158.6	0.14628	0.08775
7	S4	1952.4	161.8	0.13214	47.0	0.03726	67.5	0.05454	79.1	0.06353	167.9	0.13048	0.08359
8	S5	1902.4	147.6	0.13194	42.5	0.03946	54.5	0.04684	63.5	0.04480	156.0	0.12183	0.07697
9	S6	2495.5	143.8	0.08506	43.9	0.02878	52.8	0.02708	65.6	0.03481	150.0	0.08496	0.05213
10	S7	3217.0	179.4	0.08186	56.2	0.02516	64.9	0.02733	81.7	0.03565	188.1	0.08163	0.05032
11	N4	2138.3	117.2	0.07679	34.8	0.02283	37.5	0.02524	49.8	0.03237	137.5	0.09389	0.05022
12	N5	4629.8	171.2	0.05370	52.6	0.01599	66.6	0.01872	82.2	0.02402	189.4	0.06051	0.03459

Note. The list of journals with abbreviations is as follows (N = non-indexed journals; S = journals indexed in WoS/Scopus): N1 – *Health Technologies*; N2 – *Health-saving Technologies, Rehabilitation and Physical Therapy*; N3 – *Physical Culture, Recreation and Rehabilitation*; S1 – *Rehabilitation and Recreation*; S2 – *Slobozhanskyi Herald of Science and Sport*; S3 – *Physical Education of Students*; S4 – *Pedagogy of Physical Culture and Sports*; S5 – *Physical Education Theory and Methodology*; S6 – *Health, Sport, Rehabilitation*; S7 – *Physical Rehabilitation and Recreational Health Technologies*; N4 – *Journal of Learning Theory and Methodology*; N5 – *Pedagogy of Health*

tendency toward terminological redundancy in the text corpus.

Thus, the analysis demonstrates that the terminological density of scientific publications in the field of physical culture systematically exceeds the established normative range.

To analyze the relationships between different lexical sources in the structure of scientific publications, a Spearman correlation matrix was calculated at the level of individual articles (Figure 3).

Five indicators of terminological density were considered (Figure 3):

1. WoS Density (lexicon from Web of Science),
2. MeSH Density (medical subject headings),
3. DE Density (keywords from descriptions),
4. ID Density (identifiers),
5. AU Density (author keywords).

The data in Figure 3 indicate the following tendencies:

1. All correlations fall within the range of 0.964 to 0.991, reflecting a high level of consistency in term distribution across the dictionaries.
2. Particularly strong associations are observed between:
 - WoS and AU ($r = 0.991$), indicating the proximity of author-provided vocabulary to the general scientific terminology actively used in the articles.
 - DE and ID ($r = 0.981$), showing a strong relationship between formalized lexemes from descriptions and index tags.
3. AU Density integrates into the correlation structure as reliably as the other indicators, confirming its importance in modeling the terminological profile of the text.

The results demonstrate the complementarity of lexical dictionaries in the analysis of scientific publications. The high degree of correlation confirms that terminological density possesses a stable structure regardless of the type of term source. This enables the use of both combined and isolated dictionaries in semantic evaluation, thematic mapping, and editorial analysis.

Discussion

The aim of this study was to develop and test a method for the automated assessment of terminological density in scientific publications in the field of physical culture using adapted lexical dictionaries. The obtained results confirmed the effectiveness of the proposed approach: all 12 analyzed journals demonstrated moderate, increased, or high levels of terminological saturation according to the WoS and MeSH scales. Particularly high density was observed when using the WoS dictionary, which is likely associated with the inclusion of author keywords and the stylistic

features of the texts. Terminological density based on MeSH ranged from optimal to high levels, indicating the gradual integration of biomedical terminology into scientific publications on physical culture and sports.

Our results are consistent with previous studies that addressed issues of lexical saturation and its consequences. For example, Haunschild et al. demonstrated that the automatic use of Keywords Plus from WoS substantially increases term density and may overload the text [30]. Kiss et al. [8] and Venâncio et al. [9] emphasized the variability of terminological structures in sports-related publications, which is also confirmed by our observations.

Lu et al. [56] showed that the analysis of author keywords' frequency can effectively reveal thematic trends; however, such an approach carries the risk of terminological redundancy. Another study [57] employed automatic identification of data-related articles, which also relied on the detection of specific key expressions. These findings indirectly confirm that our method of quantifying terminological density highlights the same issue – concentrations of terms above the optimal level.

Moreover, Kim and Kim [58] demonstrated that even emerging scientific fields (e.g., the metaverse and sport) are characterized by the active introduction of specific terminology, which does not always align with established standards. This finding resonates with our conclusions: local publications in the field of physical culture often rely on their own lexical frameworks, thereby widening the gap with international descriptors.

The novelty of our study lies in the use of two types of dictionaries: controlled (MeSH) and empirical (WoS) for the analysis of the publication corpus. This approach made it possible to identify the distinction between standardized vocabulary and authorial practice, which has rarely been considered in previous research. Unlike studies with a narrow focus, our journal-level comparison expands the scope of bibliometric expertise and provides more generalized conclusions. A similar approach was applied by Hammerschmidt et al. [59], who conducted a bibliometric analysis of publication activity and thematic dynamics across five leading sport management journals over the period 2011 to 2020. Likewise, Shilbury [60] examined citation patterns in several sport management journals and emphasized the importance of journal-level analysis for understanding academic influence and citation structures.

In addition to integrating two dictionaries, a significant element of novelty was the creation of a domain-specific dictionary based on 16593 bibliographic records from the Web of Science database over the past five years. Unlike ready-made controlled resources, this dictionary reflects

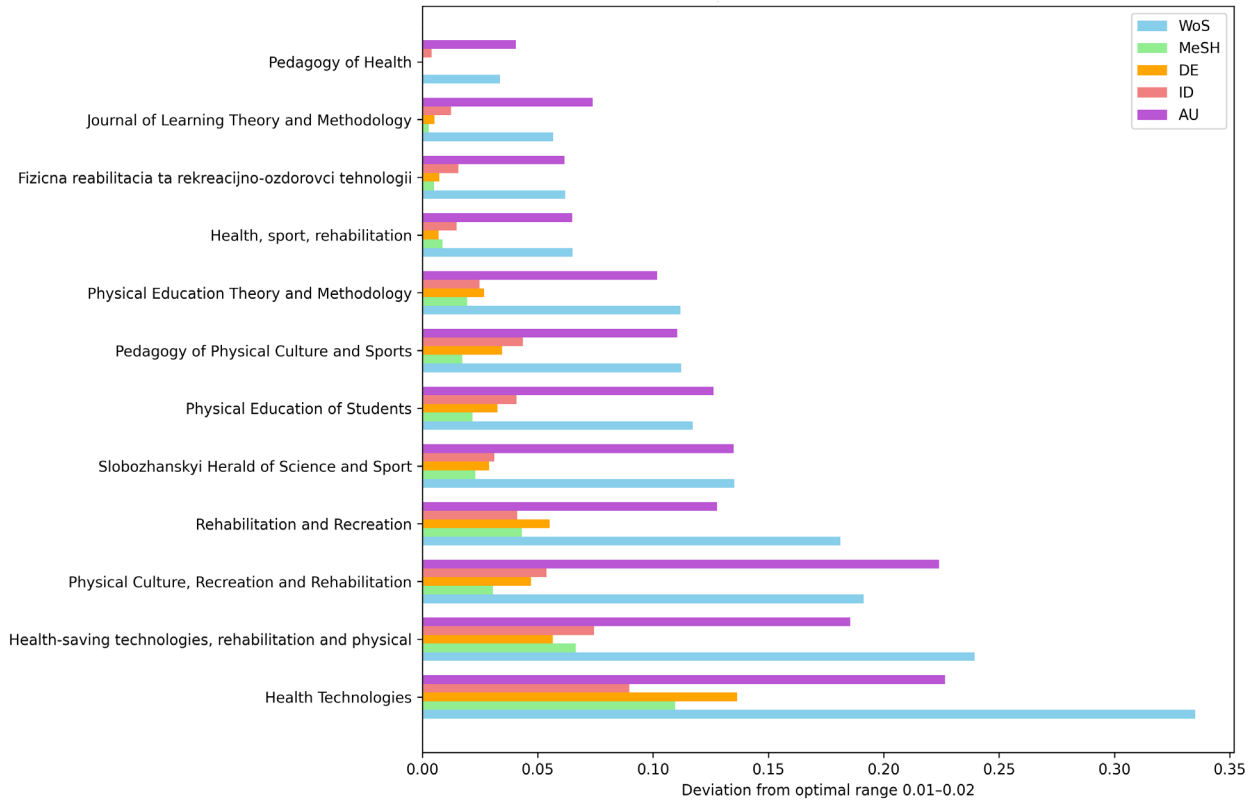


Figure 2. Deviations of terminological density from the normative range (0.01–0.02)

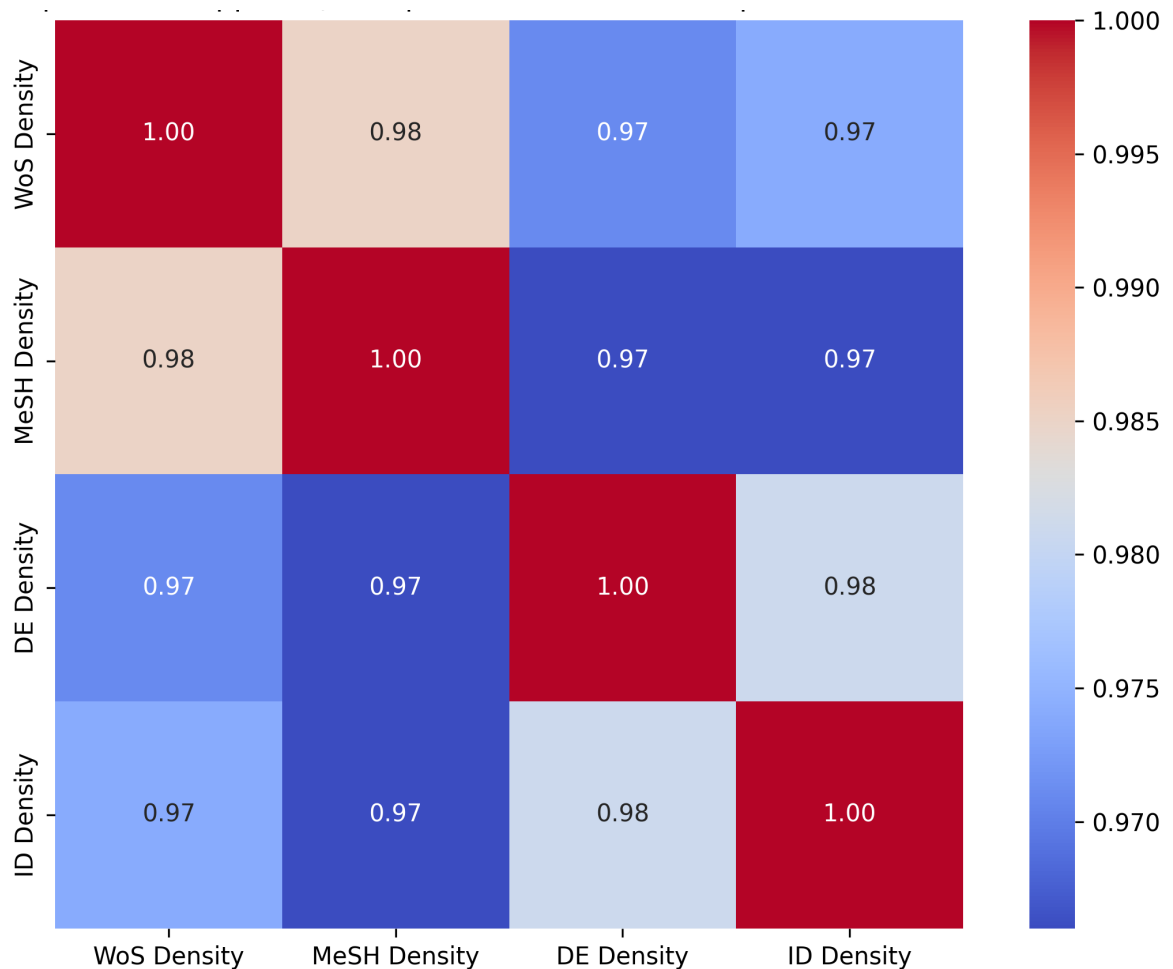


Figure 3. Spearman correlation matrix at the level of individual articles

the actual vocabulary of scientific publications in the field of physical culture. It represents stable expressions and professional terminology. Similar practices have a strong basis in the international research tradition. For example, Yan et al. [61] proposed a domain-independent method of term extraction for the identification of disciplinary dictionaries from full-text articles. Another study demonstrated the potential of a corpus-based approach for the automatic construction of semantic lexicons [62]. A new approach to the semi-automatic creation and expansion of a multilingual terminological thesaurus was presented by Horák et al. [63]. These examples show that the development of specialized dictionaries is a recognized direction in scientometrics and computational linguistics. However, the application of such a practice in the field of physical culture, based on such an extensive corpus of sources, has not been identified in the available data.

Another aspect of novelty is the high degree of automation in the assessment of terminological density. Unlike traditional content-analytical studies, we implemented algorithms for automated extraction of text from PDF files, corpus cleaning, consideration of word forms and synonyms, as well as the calculation of relative terminological density using Python (libraries `pandas`, `re`, `fitz`). This approach ensures reproducibility, eliminates subjective errors, and allows the processing of large datasets. Comparable methods of automation are reported in bibliometrics and computational linguistics. For example, Riloff and Shepherd [62] demonstrated the potential of corpus-based approaches for building semantic lexicons. Another study used automatic term extraction from full-text publications [61]. Horák et al. [63] showed the effectiveness of semi-automatic construction and expansion of a multilingual thesaurus. Such approaches increase the objectivity of evaluating publication practices [64]. In our case, these principles were applied for the first time to the evaluation of scientific journals in the field of physical culture. No similar approaches were identified in the available sources.

The approach applied in our study, which integrated two dictionaries supplemented with synonyms and word forms, provided more accurate terminological matching and increased the sensitivity of the analysis. This is consistent with the recommendations of Manning and Schütze [54] and Ahmad et al. [65], who emphasized the importance of reducing variability in academic writing. Our results also revealed a high correlation of metrics across the dictionaries, which confirms the stability of terminological density structures and demonstrates the applicability of automated methods for semantic evaluation of scientific publications.

The developed scale of terminological density based on the MeSH and WoS dictionaries

demonstrated the ability to objectively classify publications and compare journals with each other. This expands the toolkit of bibliometrics and editorial practice by identifying imbalances in scientific style. Similar to the findings of Ahmad et al. [65], who showed differences in academic writing across disciplines, our scale reflects lexical specificity and can be used to adjust publication practices. The use of scales to measure textual complexity and terminological saturation is well established in linguistics and bibliometrics: Halliday [36] introduced the concept of lexical density as an indicator of text complexity; Solnyshkina et al. [35] applied readability formulas for building scales of educational texts; Leblanc et al. [19] demonstrated that the number of MeSH terms in systematic reviews can serve as a basis for standardized density intervals; Breuer et al. [26] proposed a scale for identifying the “core documents” based on the concentration of key concepts; Nasser and Thompson [66] highlighted the significance of lexical density and diversity for distinguishing L1 and L2 texts; Bakuuro [67] linked lexical density with readability and perceived text complexity. However, in studies on physical culture and sport, similar scales have not been identified. Existing research has focused mainly on citation analysis and conceptual apparatus: Khatra et al. [68] conducted a bibliometric review of the most cited articles in sports medicine, and Staunton et al. [69] addressed the misuse of the term “load”. Yet these studies did not include measurement of terminological density, which underscores the methodological novelty of our proposed scale, adapted for analyzing publications in the field of physical culture.

Thus, the proposed scale goes beyond existing bibliometric approaches and demonstrates methodological innovation, as it is the first to focus on measuring terminological density in publications on physical culture. The development of this direction is closely related to the automation of text analysis, which allows for broader applications and greater objectivity in the assessment of publication practices.

The characteristics of publications in the field of physical culture are shaped by the interdisciplinary nature of this domain, where pedagogy, sports medicine, physiology, and psychology intersect. Terminological density in such texts arises from a combination of specialized concepts (“load,” “adaptation,” “motivation,” “injury prevention”) and a more general scientific vocabulary. This balance influences how texts are perceived by both professional and academic audiences. Unlike medicine or pedagogy, where terminology is more established, sports science demonstrates greater variability and frequent borrowing of concepts, which highlights the need for objective metrics in text evaluation.

In this context, bibliometric studies in sport have so far focused primarily on citation analysis and the identification of leading research areas. For instance, Khatra et al. [68] reviewed the most cited publications in sports medicine, showing the dominance of clinical and physiological research. Staunton et al. [69] drew attention to the misuse of the term “load,” underlining the importance of clarifying and standardizing terminology. However, none of these studies addressed the measurement of structural lexical saturation. The proposed scale of terminological density thus fills this gap by providing a quantitative tool for evaluating texts in the field of physical culture.

The application of the developed scale in this field has both theoretical and practical relevance. On the one hand, it provides editorial boards of national journals with an objective tool to assess manuscript quality in comparison with international standards. On the other hand, it offers authors a means to adjust their academic writing strategies in accordance with indexing requirements. This is particularly significant in the context of increasing competition among journals, where a balanced use of terminology becomes a marker of academic maturity and an important factor of international visibility.

In scientific literature, terminological density is regarded as an indicator of both the quality and complexity of academic texts. This interpretation originates from linguistic studies on translation and language [70], while a number of applied works have demonstrated its significance in fields such as healthcare and education, where terminology serves as a marker of professional quality and suitability of publications for indexing [56, 71, 72]. These studies confirm that the balance between terminological richness and clarity of expression directly affects the perception of research articles and their integration into the international academic space. In the field of physical culture, this balance is particularly critical, as publications are simultaneously addressed to the academic community and to practitioners in sport and education. Thus, the approach requires clarity of language while maintaining scientific rigor.

Our study extends this approach to the field of physical culture. The analysis of texts from leading Ukrainian journals demonstrated that terminological density can serve as an indicator of academic maturity and readiness of publications for international communication. The use of quantitative and bibliometric methods has already been established in related domains of physical literacy and physical activity. For instance, Mendoza-Muñoz et al. [73] conducted a global review of publications on *physical literacy* employing spatial and thematic visualization. Similarly, Memon et al. [74] analyzed the most cited studies on sedentary lifestyle, highlighting

the need for qualitative indicators to assess textual influence. Li et al. [75] examined the thematic scope of research on physical activity and health in the context of osteoporosis, underlining the relevance of quantitative assessment methods in sport and health sciences. In addition, Arnal Gómez et al. [76] applied bibliometric techniques to analyze a physical therapy journal on aging, identifying key journals and citation trends. Furthermore, Buhin Pandur et al. [64] demonstrated the potential of topic modeling in social sciences based on Web of Science data, confirming the versatility of such approaches in interdisciplinary research.

In the context of growing competition and the need for indexing in international databases, terminological density can serve as a tool for evaluating both authorial writing strategies and editorial policies in journals on physical culture. This opens the possibility of comparing different domains of sports science, from biomechanics and sports medicine to pedagogy of physical culture, thereby identifying their degree of terminological maturity.

Thus, the analysis demonstrated that terminological density is a reliable indicator of the academic quality of publications, and its systematic excess in texts on physical culture indicates the need to adjust editorial standards. The integration of controlled and empirical vocabularies, taking into account synonymy and word forms, provides a basis for objective analysis and can be extended to other disciplines. For editors of journals on physical culture and sport, such a scale may serve as a practical tool for quality control of manuscripts, helping to identify excessive terminology, justify editorial revisions, and increase the transparency of author requirements in preparation for indexing. Future prospects include the use of contextual language models and more advanced morphological processing to deepen the analysis. A further step in the field of sports science may involve automated monitoring of terminological density in specific areas (sports medicine, training loads, student physical activity), which will make it possible to track the dynamics of disciplinary development.

Limitations

Several limitations should be acknowledged. Lemmatization and morphological normalization of terms were not fully implemented, which may have resulted in a partial loss of accuracy when calculating terminological density. In addition, the study was limited to a sample of 12 journals, which reduces the generalizability of the findings. Differences in indexing policies of WoS, Scopus, and PubMed may also have influenced the interpretation of the results. It is important to note that the focus on journals in physical culture and related fields limits the applicability of the conclusions

to a broader range of scientific disciplines. Future research should expand the sample of journals in sports science to refine the identified trends and enhance the reliability of generalizations.

Conclusions

The study demonstrated that terminological density can be regarded as a reliable indicator of the academic quality of publications in the field of physical culture and sports. The proposed scale, based on a combination of controlled and empirical vocabularies, showed methodological novelty and created opportunities for the objective comparison of journals, as well as for identifying imbalances in style and academic writing standards.

The analysis revealed that a high level of terminological density does not always correspond to the academic maturity of a text and may indicate the need for editorial adjustments. At the same time, a balanced use of terminology supports the integration of publications into the international scientific community, enhances their citation potential, and increases the transparency of research

communication.

The findings confirm the potential of applying terminological density scales for evaluating publishing practices in physical culture and related disciplines. Further development should focus on the automation of text analysis, the expansion of the journal sample, and the use of contextual language models, which will enable deeper interpretation and improve the reproducibility of results.

Conflict of Interest

One of the authors (Sergii Iermakov) serves as the Editor-in-Chief and Publisher of this journal. To ensure an objective review process, the manuscript was handled by an independent editorial board member, and the peer review was conducted by external reviewers who had no affiliations with the authors. The Editor-in-Chief did not participate in the review or editorial decision-making process regarding this manuscript. The other co-authors (Georgiy Korobeynikov and David Curby) declare that they have no conflict of interest related to this publication.

References

- Han J, Kamber M, Pei J. *Data mining: concepts and techniques*. 3rd ed. Amsterdam Boston: Elsevier/Morgan Kaufmann; 2012.
- Ramos J. Using TF-IDF to determine word relevance in document queries. In: *Proceedings of the First Instructional Conference on Machine Learning, Vol. 242, Citeseer*; 2003. P. 29–48.
- Wang Y. Research on the TF-IDF algorithm combined with semantics for automatic extraction of keywords from network news texts. *Journal of Intelligent Systems*, 2024;33(1): 20230300. <https://doi.org/10.1515/jisys-2023-0300>
- Wang W, Zhang J, Zhou F, Chen P, Wang B. Paper acceptance prediction at the institutional level based on the combination of individual and network features. *Scientometrics*, 2021;126(2): 1581–1597. <https://doi.org/10.1007/s11192-020-03813-x>
- National Library of Medicine. MeSH Indexing Manual. Bethesda, MD: U.S. Department of Health & Human Services; 2022.
- Mao Y, Lu Z. MeSH Now: automatic MeSH indexing at PubMed scale via learning to rank. *Journal of Biomedical Semantics*, 2017;8(1): 15. <https://doi.org/10.1186/s13326-017-0123-3>
- Kim S, Yeganova L, Wilbur WJ. Meshable: searching PubMed abstracts by utilizing MeSH and MeSH-derived topical terms. *Bioinformatics*, 2016;32(19): 3044–3046. <https://doi.org/10.1093/bioinformatics/btw331>
- Kiss A, Temesi Á, Tompa O, Lakner Z, Soós S. Structure and trends of international sport nutrition research between 2000 and 2018: bibliometric mapping of sport nutrition science. *Journal of the International Society of Sports Nutrition*, 2021;18(1): 12. <https://doi.org/10.1186/s12970-021-00409-5>
- Venâncio TF, Costa MJ, Santos CC, Batalha N, Hernández-Beltrán V, Gamonales JM, et al. Evolution of documents related to strength training research on competitive swimmers: a bibliometric review. *Frontiers in Sports and Active Living*, 2025;7: 1603576. <https://doi.org/10.3389/fspor.2025.1603576>
- Jagiello M, Lochbaum M. Pedagogical strategies for enhancing physical activity: a systematic review of trends and approaches. *Pedagogy of Health*. 2024;2(2):37–43. <https://doi.org/10.15561/health.2024.0201>
- Jagiello M, Lochbaum M. Modern methods and means of physical culture in the rehabilitation of various population groups: a systematic review. *Physical Culture, Recreation and Rehabilitation*, 2024;3(2): 34–45. <https://doi.org/10.15561/phycult.2024.0201>
- Yermakova T. Risk factors and prevention of falls in children under 3 years: a systematic review. *Physical Culture, Recreation and Rehabilitation*, 2025;4(1): 17–34. <https://doi.org/10.15561/phycult.2025.0103>
- Yermakova T. Patterns and risk factors of falls among older adults: a systematic review. *Pedagogy of Health*. 2025;1(1):11–21. <https://doi.org/10.15561/health.2025.0102>
- National Library of Medicine. Download MeSH Data. XML Format [cited 2025 May 17]. Available from: <https://www.nlm.nih.gov/databases/download/mesh.html>
- Clarivate. *KeyWords Plus generation, creation, and changes* [Internet]. 2025 [updated 2025 Jul 21; cited 2025 Jul 22]. Available from: <https://support.clarivate.com/ScientificandAcademicResearch/s/article/KeyWords-Plus-generation-creation-and>

- changes?language=en_US
16. Van Eck NJ, Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 2010;84(2): 523–538. <https://doi.org/10.1007/s11192-009-0146-3>
 17. Cobo MJ, López-Herrera AG, Herrera-Viedma E, Herrera F. SciMAT : A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, 2012;63(8): 1609–1630. <https://doi.org/10.1002/asi.22688>
 18. Trieschnigg D, Pezik P, Lee V, De Jong F, Kraaij W, Rebholz-Schuhmann D. MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics*, 2009;25(11): 1412–1418. <https://doi.org/10.1093/bioinformatics/btp249>
 19. Leblanc V, Hamroun A, Bentegeac R, Le Guellec B, Lenain R, Chazard E. Added Value of Medical Subject Headings Terms in Search Strategies of Systematic Reviews: Comparative Study. *Journal of Medical Internet Research*, 2024;26: e53781. <https://doi.org/10.2196/53781>
 20. Lipscomb CE. Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association*, 2000;88(3), 265–266.
 21. Chen C, Hu Z, Liu S, Tseng H. Emerging trends in regenerative medicine: a scientometric analysis in CiteSpace. *Expert Opinion on Biological Therapy*, 2012;12(5): 593–608. <https://doi.org/10.1517/14712598.2012.674507>
 22. Sujarwo, Paramitha ST, Hasyim AH, Ramadhan MG, Setiawan I. A bibliometric analysis of research on physical activity and fitness among preschool children in Asia (2020–2024). *Edu Sportivo: Indonesian Journal of Physical Education*, 2024;5(3): 243–257. [https://doi.org/10.25299/esijope.2024.vol5\(3\).19085](https://doi.org/10.25299/esijope.2024.vol5(3).19085)
 23. Pradhan P, Zala LN. Bibliometrics analysis and comparison of global research literatures on research data management extracted from Scopus and Web of Science during 2000–2019. *Library Philosophy and Practice (e-journal)*, 2021;5519:1–17.
 24. Van Eck NJ, Waltman L. Citation-based clustering of publications using CitNetExplorer and VOSviewer. *Scientometrics*, 2017;111(2): 1053–1070. <https://doi.org/10.1007/s11192-017-2300-7>
 25. Aria M, Cuccurullo C. Bibliometrix: An R-tool for comprehensive science mapping analysis. *J Informetrics*. 2017;11(4):959–975. <https://doi.org/10.1016/j.joi.2017.08.007>
 26. Breuer T, Schaer P, Tunger D. Relevance assessments, bibliometrics, and altmetrics: a quantitative study on PubMed and arXiv. *Scientometrics*, 2022;127(5): 2455–2478. <https://doi.org/10.1007/s11192-022-04319-4>
 27. Han O, Demydenko O. Terminological richness of english-language scientific-popular and media texts in physics. *Advanced Linguistics*, 2023;(12). <https://doi.org/10.20535/2617-5339.2023.12.290971>
 28. Elsevier. *Scopus Author Guidelines*. Amsterdam: Elsevier; 2021.
 29. Ding Y, Chowdhury GG, Foo S. Bibliometric cartography of information retrieval research by using co-word analysis. *Inf Process Manag*. 2001;37(6):817–842. [https://doi.org/10.1016/S0306-4573\(00\)00051-0](https://doi.org/10.1016/S0306-4573(00)00051-0)
 30. Haunschild R, Bornmann L, Marx W. Climate change research in view of bibliometrics. *PLoS One*. 2016;11(7):e0160393. <https://doi.org/10.1371/journal.pone.0160393>
 31. Bekhuis T, Demner-Fushman D, Crowley R. Comparative effectiveness research designs: An analysis of terms and coverage in Medical Subject Headings (MeSH) and Emtree. *J Med Libr Assoc*. 2013;101(2):92–100. <https://doi.org/10.3163/1536-5050.101.2.004>
 32. Koloski B, Pollak S, Škrlj B, Martinc M. Extending Neural Keyword Extraction with TF-IDF tagset matching. In: *Proc EACL Hackashop on News Media Content Analysis and Automated Report Generation*; 2021. p. 22–29. <https://aclanthology.org/2021.hackashop-1.4.pdf>
 33. Valkanas K, Diamandis P. Pareto distribution in virtual education: challenges and opportunities. *Canadian Medical Education Journal*, 2021; <https://doi.org/10.36834/cmej.73511>
 34. Nisonger TE. The “80/20 Rule” and Core Journals. *The Serials Librarian*, 2008;55(1–2): 62–84. <https://doi.org/10.1080/03615260801970774>
 35. Solnyshkina MI, Gatiyatullina GM, Kupriyanov RV, Ziganshina CR. Lexical density as a complexity predictor: the case of Science and Social Studies textbooks. *Research Result Theoretical and Applied Linguistics*. 2023;9(1). <https://doi.org/10.18413/2313-8912-2023-9-1-0-2>
 36. Halliday MAK. *Spoken and written language*. Oxford: Oxford University Press; 1985.
 37. Bajerowska A. Kilka uwag o fachowości przyczynę do rozważań teoretycznych [Some remarks on the professionalism of contributions to theoretical considerations]. *Kwartalnik Neofilologiczny*, 2024; 5–19. (In Polish). <https://doi.org/10.24425/kn.2024.149614>
 38. Istiqomah F, Basthomi Y. Exploring nominalization and lexical density deployed within research article abstracts: A grammatical metaphor analysis. *Englisia: Journal of Language, Education, and Humanities*, 2024;11(2): 14. <https://doi.org/10.22373/ej.v11i2.20390>
 39. Qaiser S, Ali R. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*, 2018;181(1): 25–29. <https://doi.org/10.5120/ijca2018917395>
 40. Li X, Zhang A, Li C, Ouyang J, Cai Y. Exploring coherent topics by topic modeling with term weighting. *Information Processing & Management*, 2018;54(6): 1345–1358. <https://doi.org/10.1016/j.ipm.2018.05.009>
 41. Fu Z, Su Y, Meng Z, Collier N. Biomedical Named Entity Recognition via Dictionary-based Synonym Generalization. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore: Association for Computational Linguistics; 2023. p. 14621–14635. <https://doi.org/10.18653/v1/2023.emnlp-main.903>
 42. Slater LT, Bradlow W, Ball S, Hoehndorf R, Gkoutos

- GV. Improved characterisation of clinical text through ontology-based vocabulary expansion. *Journal of Biomedical Semantics*, 2021;12(1): 7. <https://doi.org/10.1186/s13326-021-00241-5>
43. Kugic A, Pfeifer B, Schulz S, Kreuzthaler M. Embedding-based terminology expansion via secondary use of large clinical real-world datasets. *Journal of Biomedical Informatics*, 2023;147: 104497. <https://doi.org/10.1016/j.jbi.2023.104497>
44. Thießen F, D'Souza J, Stocker M. Probing Large Language Models for Scientific Synonyms. In: *SEMANTICS 2023 EU: 19th International Conference on Semantic Systems, September 20-22, 2023, Leipzig, Germany*. 2023;3510:1–14.
45. Pans M, Madera J, González LM, Pellicer-Chenoll M. Physical Activity and Exercise: Text Mining Analysis. *International Journal of Environmental Research and Public Health*, 2021;18(18): 9642. <https://doi.org/10.3390/ijerph18189642>
46. Pavliuk A, Rohach O, Sheludchenko S, Yefremova N, Boichuk V. Structural and derivational parameters of the sports terminology. *Research Trends in Modern Linguistics and Literature*, 2022;5: 16–31. <https://doi.org/10.29038/2617-6696.2022.5.16.31>
47. Pavliuk IB. Terminological fields of fitness terminology. *Folium*, 2023;(2): 59–65. (In Ukrainian). <https://doi.org/10.32782/folium/2023.2.9>
48. Tsybanyuk O, Mishkulynets O, Komisaryk M, Kuznietsova K, Chuyko H. Genesis of the transformation of terminology in the field of physical education and sports in Romania, historical context. *Conhecimento & Diversidade*, 2023;15(37): 381–403. <https://doi.org/10.18316/rcd.v15i37.10966>
49. Mănescu DC. Big Data Analytics Framework for Decision-Making in Sports Performance Optimization. *Data*, 2025;10(7): 116. <https://doi.org/10.3390/data10070116>
50. Lee Y, Kang JH, Lee S, Oh T, Choi S. The Evolution of Terminology: A Scoping Review of Terms and Concepts Used to Research Sport in the Digital Realm. *Quest*, 2024;76(4): 462–480. <https://doi.org/10.1080/00336297.2024.2357370>
51. Klégr A, Bozděchová I. Sports Terminology as a Source of Synonymy in Language: the Case of Czech. *Revista Alicantina de Estudios Ingleses*, 2019;(32): 163. <https://doi.org/10.14198/raei.2019.32.07>
52. Randelović N, Živković D, Piršl D, Piršl T, Đošić A. Classification of sports terms: Thematic approach. *Fizicko vaspitanje i sport kroz vekove*, 2023;10(1): 1–10. <https://doi.org/10.5937/spes2301001R>
53. Qutab I, Malik KI, Arooj H. Sentiment Classification Using Multinomial Logistic Regression on Roman Urdu Text. *International Journal of Innovations in Science and Technology*, 2022;4(2): 323–335. <https://doi.org/10.33411/IJIST/2022040204>
54. Manning CD, Schütze H. *Foundations of statistical natural language processing*. Cambridge, Mass: MIT Press; 1999.
55. Wang J. Utilizing Text Mining Technology to Enhance English Learners' Vocabulary. *International Journal of Electronics and Communication Engineering*, 2024;11(9): 86–98. <https://doi.org/10.14445/23488549/IJECE-V11I9P109>
56. Lu W, Huang S, Yang J, Bu Y, Cheng Q, Huang Y. Detecting research topic trends by author-defined keyword frequency. *Information Processing & Management*, 2021;58(4): 102594. <https://doi.org/10.1016/j.ipm.2021.102594>
57. Zhang Q, Lu W, Yang Y, Chen H, Chen J. Automatic Identification of Research Articles Containing Data Usage Statements. In: *Knowledge Discovery and Data Design Innovation*, Dallas, Texas, USA: WORLD SCIENTIFIC; 2017. p. 67–87. https://doi.org/10.1142/9789813234482_0004
58. Kim A, Kim SS. Engaging in sports via the metaverse? An examination through analysis of metaverse research trends in sports. *Data Science and Management*, 2024;7(3): 181–188. <https://doi.org/10.1016/j.dsm.2024.01.002>
59. Hammerschmidt J, Calabuig F, Kraus S, Uhrich S. Tracing the state of sport management research: a bibliometric analysis. *Management Review Quarterly*, 2024;74(2): 1185–1208. <https://doi.org/10.1007/s11301-023-00331-x>
60. Shilbury D. A bibliometric analysis of four sport management journals. *Sport Management Review*, 2011;14(4): 434–452. <https://doi.org/10.1016/j.smr.2010.11.005>
61. Yan E, Williams J, Chen Z. Understanding disciplinary vocabularies using a full-text enabled domain-independent term extraction approach. Glanzel W (ed.) *PLOS ONE*, 2017;12(11): e0187762. <https://doi.org/10.1371/journal.pone.0187762>
62. Ellen Riloff, Jessica Shepherd. A Corpus-Based Approach for Building Semantic Lexicons. In: *Second Conference on Empirical Methods in Natural Language Processing*. 1997. P. 117124.
63. Horák A, Baisa V, Rambousek A, Suchomel V. A New Approach for Semi-Automatic Building and Extending a Multilingual Terminology Thesaurus. *International Journal on Artificial Intelligence Tools*, 2019;28(02): 1950008. <https://doi.org/10.1142/S0218213019500088>
64. Buhin Pandur M, Dobša J, Kronegger L. Topic modelling in social sciences: case study of Web of Science. In: *Central European Conference on Intelligent and Information Systems; 2020 Oct; Varaždin, Croatia*; 2020. P. 67–72.
65. Ahmad M, Mahmood AM, Siddique AR. Variation in academic writing: A corpus-based research on syntactic features across four disciplinary divisions. *Novitas-ROYAL (Research on Youth and Language)*, 2023;17(2), 50–65. <https://doi.org/10.5281/zenodo.10015816>
66. Nasser M, Thompson P. Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assessing Writing*, 2021;47: 100511. <https://doi.org/10.1016/j.asw.2020.100511>
67. Bakuuro J. In the Belly of Text Complexity: Unravelling the Nexus between Lexical Density and Readability. *Athens Journal of Philology*, 2024;11(3): 255–274. <https://doi.org/10.30958/ajp.11-3-4>
68. Khatra O, Shadgan A, Taunton J, Pakravan A, Shadgan B. A Bibliometric Analysis of the Top Cited Articles

- in Sports and Exercise Medicine. *Orthopaedic Journal of Sports Medicine*, 2021;9(1): 2325967120969902. <https://doi.org/10.1177/2325967120969902>
69. Staunton CA, Abt G, Weaving D, Wundersitz DWT. Misuse of the term 'load' in sport and exercise science. *Journal of Science and Medicine in Sport*, 2022;25(5): 439–444. <https://doi.org/10.1016/j.jsams.2021.08.013>
70. Francoeur A. Fawcett, Peter (1997) : Translation and Language. Linguistic Theories Explained, coll. «Translation Theories Explained», Manchester (UK), St. Jerome Publishing, 160 p. *Meta: Journal des traducteurs*, 1999;44(3): 514. <https://doi.org/10.7202/002768ar>
71. Kim H, Kim SH, Kim J, Kim EH, Gu JH, Lee D. A keyword-based approach to analyzing scientific research trends: ReRAM present and future. *Scientific Reports*, 2025;15(1): 12011. <https://doi.org/10.1038/s41598-025-93423-5>
72. Mulia Al-Amien M, Hidayati D, Haryadi D. Analysis Of Scientific Article Writing Ability. *International Journal of Educational Management and Innovation*, 2022;3(1): 103–110. <https://doi.org/10.12928/ijemi.v3i1.5555>
73. Mendoza-Muñoz M, Vega-Muñoz A, Carlos-Vivas J, Denche-Zamorano Á, Adsuar JC, Raimundo A, et al. The Bibliometric Analysis of Studies on Physical Literacy for a Healthy Life. *International Journal of Environmental Research and Public Health*, 2022;19(22): 15211. <https://doi.org/10.3390/ijerph192215211>
74. Memon AR, Chen S, To QG, Vandelanotte C. Vigorously cited: a bibliometric analysis of the 100 most cited sedentary behaviour articles. *Journal of Activity, Sedentary and Sleep Behaviors*, 2023;2(1): 13. <https://doi.org/10.1186/s44167-023-00022-8>
75. Li F, Xie W, Han Y, Li Z, Xiao J. Bibliometric and visualized analysis of exercise and osteoporosis from 2002 to 2021. *Frontiers in Medicine*, 2022;9: 944444. <https://doi.org/10.3389/fmed.2022.944444>
76. Arnal-Gómez A, Navarro-Molina C, Espí-López GV. Bibliometric analysis of core journals which publish articles of physical therapy on aging. *Physical Therapy Research*, 2020;23(2): 216–223. <https://doi.org/10.1298/ptr.E10024>

Information about the authors:

Sergii Iermakov; <https://orcid.org/0000-0002-5039-4517>; sportart@gmail.com; Department of Methodologies of Cross-Cultural Practices, Kharkiv State Academy of Design and Arts; Kharkiv, Ukraine.

Georgiy Korobeynikov; (Corresponding Author); <https://orcid.org/0000-0002-1097-4787>; k.george.65.w@gmail.com; Department of Theory and Methodology of International Wrestling, Uzbek State University of Physical Education and Sports (Tashkent region, Chirchik, Uzbekistan); Institute of Psychology, German Sport University Cologne (Cologne, Germany); Department of Combat Sports and Power Sports, National University of Physical Education and Sport (Kyiv, Ukraine).

David Curby; <https://orcid.org/0000-0003-1170-4583>; davcurb@gmail.com; International Network of Wrestling Researchers; Chicago, USA.

Cite this article as:

Iermakov S, Korobeynikov G, Curby D. Assessment of terminological density in scientific publications on physical culture. *Physical Culture, Recreation and Rehabilitation*, 2025;4(2):74–89. <https://doi.org/10.15561/physcult.2025.0203>

This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (<http://creativecommons.org/licenses/by/4.0/deed.en>).

Received: 12.07.2025

Accepted: 15.09.2025; Published: 30.12.2025