



Contents lists available at [Elora Center](#)

Lentera Negeri

Journal homepage: <https://lentera.eloracenter.org/lentera>



Corpus-based english for sport sciences: a systematic review of specialised vocabulary, terminology, and pedagogical applications (2015–2025)

Kristian Burhan^{*)}, Ifdil Ifdil, Rudyanto Rudyanto

Universitas Negeri Padang, Indonesia

Article Info

Article history:

Received May 23th, 2026

Revised Jul 20th, 2026

Accepted Jul 21th, 2026

Keywords:

Corpus-based instruction
English for Specific Purposes
Sport sciences education
Technical vocabulary
Systematic literature review

ABSTRACT

English has become the dominant language of communication in sport sciences, yet many students continue to experience difficulties in acquiring specialised vocabulary, technical terminology, and discipline-specific discourse required for academic and professional contexts. Although corpus linguistics offers a robust evidence-based approach for identifying authentic language patterns, its pedagogical application in sport sciences remains fragmented. This systematic literature review synthesises empirical evidence on corpus-based English research relevant to sport sciences education. The review followed the PRISMA 2020 guidelines and searched the Scopus database because of its comprehensive coverage of peer-reviewed interdisciplinary literature in linguistics, education, and sport sciences. A structured Boolean search identified 403 records, of which 18 studies (2015–2025) met the predefined eligibility criteria after duplicate removal, title and abstract screening, and full-text assessment based on the PICOS framework. Two independent reviewers conducted the screening and data extraction process, with disagreements resolved through discussion. Methodological quality was evaluated using the FICO framework, and evidence was synthesised through the thematic synthesis approach proposed by Thomas and Harden (2008). Four overarching themes emerged: (1) specialised vocabulary development and word-list construction, (2) sport terminology and cross-linguistic variation, (3) spoken and instructional discourse in sport contexts, and (4) corpus-informed pedagogical applications. The findings indicate that technical vocabulary and multiword expressions constitute a substantial proportion of sport science discourse, while inconsistencies in terminology across languages continue to challenge instructional standardisation. Nevertheless, most existing studies remain descriptive, with relatively few evaluating instructional effectiveness or learning outcomes. This review provides a comprehensive synthesis of corpus-based evidence and highlights the need for discipline-specific teaching materials, corpus-informed curricula, and intervention studies evaluating pedagogical effectiveness. The findings offer practical guidance for English for Specific Purposes (ESP) curriculum development and establish directions for future corpus-based research in sport sciences education.



© 2026 The Authors.

This is an open access article under the CC BY-NC-SA license
(<https://creativecommons.org/licenses/by-nc-sa/4.0>)

Corresponding Author:

Kristian Burhan,

✉ kristianburhan@unp.ac.id

Introduction

Sport science has become a truly global research area over the past two decades. Its working language now is essentially English. Lab protocols, coaching manuals, injury-surveillance instruments, federation

regulations, and the peer-reviewed literature tying them all together: these get produced, shared, and debated in English. The effects of this linguistic concentration show up in media commentary, borrowed terms, and coaching practice around the world (Leone, 2024; Wilton, 2020). Work on Olympic and mega-event discourse shows that the same sporting event gets reconstructed in language very differently from one national media system to another. English acts as the pivot that carries those reconstructions (Yang & Yin, 2026; Yu & Shin, 2026). So the global spread of sport has led to a global spread of sport language. But studies of glocalization suggest this process is neither uniform nor politically neutral (Li et al., 2026; Alsharif, 2026). For students entering sport sciences programmes outside the Anglophone world, solid English is no longer just a nice academic addition. It is a condition for being able to work in the field at all.

Here is the specific problem this review zeroes in on. Sport sciences students need to work in an English variety that is unusually mixed. It is academic, since they have to read and eventually write research articles in biomechanics, exercise physiology, and pedagogy. It is occupational, because they act as coaches, trainers, therapists, and administrators. And it is heavily spoken, because a big part of the field's discourse takes place on pitches, in gymnasiums, and in clinical consultations rather than on paper. Studies of coaching interaction show that instructional talk in sport is a distinct communicative genre with its own epistemic asymmetries and sequential patterns (Clifton et al., 2023). Analyses of media sports interviews point to a lingua franca practice where non-native speakers regularly negotiate meaning under time pressure (Wilton, 2020). General English courses, designed for neither of these registers, leave a large part of this competence untaught.

Existing work on the language of sport has come from a few different angles but has not really converged. One strand, rooted in media and discourse studies, looks at how sport gets represented, evaluated, and framed ideologically. Studies on newsworthiness in how elite athletes are covered (Zhang & Caple, 2021), on gender pay-gap debates in women's sport (Alkhamash, 2021), on Paralympic representation (Ismail et al., 2022), on masculinity in rugby (Dalla Pria & Bonnet, 2022), and on institutional crisis communication (Frederick et al., 2025) all show that sport discourse is packed with evaluative and metaphorical resources. A second strand, more focused on terminology and lexicology, tracks how sport terms get formed, borrowed, and adapted across languages (Núñez Noguerol, 2023; Luján-García, 2021; Janić Mitić & Veličković, 2025). A third, more recent strand uses computational and corpus-assisted techniques on large sport text collections (Braun et al., 2021; Li & Yang, 2024; Hu & Cai, 2026).

Methodological and tech advances have sped up all three strands. Freely available concordancers, larger reference corpora, and better frequency-band lists now let individual researchers compile and analyze domain corpora of several hundred thousand tokens. Multilingual and annotated resources have also opened up comparative analysis of sport texts across languages (Braun et al., 2021). Corpus-assisted discourse analysis has grown into a replicable method for tracking conceptual and evaluative patterns over time (Polajnar, 2024; Virág, 2021). Transformer-based classification is now being applied to educational data within physical education itself (Martín-Hoz et al., 2025). At the same time, large language models are being tested against domain-specific normative frameworks in sport, with mixed results that highlight the ongoing need for human-curated, domain-grounded language description (Gökmen et al., 2026).

Despite the growing body of corpus-based studies on sport-related English, the existing literature remains fragmented across vocabulary research, terminology studies, discourse analysis, and ESP instruction. Previous studies have primarily described linguistic characteristics without systematically synthesising how these findings can inform English instruction for sport sciences students. Consequently, there is currently no comprehensive evidence-based review integrating corpus-derived linguistic findings with pedagogical applications in sport sciences education.

A first gap comes directly from this expansion. Most corpus work on sport English is descriptive and media-focused. It shows us how journalists, commentators, and institutions write about sport, but it does not tell us what a sport sciences student actually needs to learn. Studies on sports tourism discourse (Blanco-Gómez & Gómez-Ortiz, 2024a, 2024b), on nominalisation in historical sports journalism (Li & Yang, 2024), and on conceptual metaphor in table tennis reporting (Chen et al., 2025) are linguistically sophisticated but pedagogically silent. When they do offer instructional implications, it is usually just a closing paragraph, not fully operationalised as materials, syllabus specifications, or assessment instruments. The result is an evidence base rich in description and thin in application.

A second gap is theoretical and methodological. Research on specialised-purposes vocabulary has developed robust constructs (technical vocabulary, coverage thresholds, frequency and range criteria, expert verification) but mostly on written academic corpora. Extending those constructs to a domain that is spoken and physically situated remains under-theorised. Cross-linguistic work on sport terminology keeps finding

variant and unstandardised forms in the receiving language (Janić Mitić & Veličković, 2025; Núñez Nogueroles, 2023). Yet the pedagogical consequence of that instability for learners moving between an English source register and their L1 professional register is rarely modelled. Corpus-based studies of instructional materials in other fields show that textbooks routinely misrepresent authentic usage (Almujaiwel, 2018; Zhang et al., 2022), a finding whose relevance to sport sciences materials has not been systematically tested.

Unlike previous reviews that have focused on physical education, sport discourse, or ESP more broadly, this review is the first to synthesise corpus-based evidence specifically related to English instruction in sport sciences. By integrating vocabulary, terminology, discourse, and instructional applications within a single PRISMA-guided framework, this review provides a comprehensive evidence base for ESP curriculum development in sport sciences.

A systematic review is both timely and needed right now. Sport sciences curricula in non-Anglophone systems are expanding. English-medium instruction is spreading. And digitalised media discourse is reshaping the input learners encounter outside the classroom (Guslyakova et al., 2023). Still, no synthesis currently maps what corpus-based research has found about the English of sport and what still needs to be built before that knowledge can enter a syllabus. Related reviews in the field have looked at how physical education teachers are positioned in models-based scholarship (Fjellner et al., 2024) and how inclusive design principles can be embedded in physical education (López Ibáñez et al., 2023), but neither touches the language dimension. This review fills that gap by applying PRISMA 2020 procedures to a bounded, transparently screened evidence base.

The first objective is descriptive and methodological. Before we can judge how transferable findings are to instruction, we need to know which corpora have been built, how big they are, whether they are written or spoken, and what analytical procedures have been used. So this review asks: RQ1: What corpus resources, research designs and analytical procedures have been used in empirical corpus-based studies of the English of sport and physical activity, and how adequately are they reported? Answering that gives us an explicit methodological map of the field and sets a reporting baseline that future corpus-building in sport sciences can be measured against.

The second objective is about linguistic substance. For corpus-based instruction to work well, it has to zero in on features that recur across the domain and that learners are least likely to acquire incidentally. That is why the review asks: RQ2: Which lexical, terminological and discursive features of sport-related English are consistently identified across studies as instructionally relevant for sport sciences students? This question brings together findings currently dispersed across vocabulary studies, terminology research, and discourse analysis, and it points out both where the evidence converges and where it is contradictory.

The third objective is applied. Description only becomes pedagogically useful when it is turned into actual instructional artefacts. So the review asks: RQ3: To what extent have corpus-based findings been translated into instructional applications for English in sport sciences, and what gaps constrain that translation? Together, these three questions support the novelty claim of this review. To the best of current scholarly knowledge, no previous systematic review has integrated corpus-based evidence on specialised vocabulary, terminology, discourse, and instructional applications within the context of sport sciences education. This review addresses that gap through a PRISMA 2020-guided synthesis.

Method

Research Design and Framework

A systematic literature review method is used as a design of the research. The systematic review has been chosen since the questions of research demand a comprehensive, replicable and bias-limited representation of a fragmentary evidence-based scenario spanning across applied linguistics, terminology, translation studies and sport pedagogy, and narrative reviewing would not allow to fully transparently document the selection decisions demanded under such fragmentation (Tranfield et al., 2003). Protocol specification, explicit eligibility rules and documented screening are the underlying principles of the evidence-informed management review tradition. On the other hand, the presentation of results and discussion section closely follows the PRISMA 2020 statement (Page et al., 2021) and the explanatory guidelines introduced in previous PRISMA versions (Liberati et al., 2009). Since most of the articles that have been included in our review are mainly descriptions or accounts, and the studies vary from each other greatly in terms of their design, it was simply out of the question to do a meta-analysis, which is a statistical method used to pool

quantitative results from multiple studies. So instead of a meta-analysis we used a thematic qualitative synthesis, where the synthesis procedure is carried out through the themes identified. We followed the thematic synthesis procedure proposed by Thomas and Harden (2008).

This review followed the PRISMA 2020 reporting guideline because the objective was to systematically identify, critically evaluate, and synthesise corpus-based studies related to English instruction in sport sciences. Given the substantial heterogeneity of study designs, corpus characteristics, analytical approaches, and reported outcomes, statistical meta-analysis was considered inappropriate. Therefore, thematic synthesis following Thomas and Harden (2008) was adopted to integrate qualitative evidence across studies.

Search Strategy

One complete Boolean query was done against the Scopus TITLE-ABS-KEY field. The literature search was conducted on 15 January 2026. The search was independently performed by two reviewers using the predefined Boolean strategy. Retrieved records were exported in CSV format from Scopus and imported into Microsoft Excel for duplicate identification and screening. The query included one part for corpus methodology, one part for English language instruction or a description of specialised-purposes, and one part related to the sport and physical education area. All parts were linked by using of the truncation symbol for picking up the variants of word forms:

("corpus-based" OR "corpus-driven" OR "corpus-assisted" OR corpus OR corpora OR concordanc OR "data-driven learning") AND ("English for specific purposes" OR ESP OR "English for academic purposes" OR EAP OR "language teaching" OR "language learning" OR pedagog* OR instruction* OR vocabular* OR "word list" OR terminolog* OR collocat* OR discourse) AND (sport* OR "physical education" OR athlet* OR exercise OR football OR soccer OR rugby OR baseball OR gymnastic* OR fitness OR kinesiology*)*

The asterisk operator was used to pick up inflected and derived forms (e.g., pedagog* gets pedagogy, pedagogic and pedagogical). Use of a quotation mark forced exact phrase match of terms containing a space. Filtering facilities provided by the database were used to restrict retrieval only to journals in English. Subject area was not limited through manual means at the search stage so as to avoid the exclusion of interdisciplinary research published in sport science, translation or computing journals at an early stage.

Database and Information Sources

Scopus was the single bibliographic source utilized in this study. This choice is in line with the scope as mentioned in the review which was a summary of peer-reviewed, indexed and internationally visible research; Scopus offers an identical sets of metadata fields (authors, title, year, source, abstract, author keywords, document type, and digital object identifier) through which a reproducible screening is possible even from a single structured export. The search was performed and the complete records were downloaded in comma-separated format; it became the main dataset that supported all identification, screening and inclusion decisions reported below. No other databases, grey literature, or citation chaining were used; limitations imposed by the methodology are described. Scopus was selected because it provides broad interdisciplinary coverage across applied linguistics, education, sport sciences, and social sciences while maintaining high indexing quality and comprehensive bibliographic metadata. These characteristics facilitate transparent and reproducible systematic review procedures.

Eligibility Criteria

The selection criteria (eligibility) criteria were first set forth via the use of a PICOS framework adapted for the linguistic, non-clinical evidence-base (Table 1). The publication period was limited to 2015–2025 to capture contemporary corpus-based methodologies and recent developments in English for Specific Purposes (ESP), corpus linguistics, and sport sciences education. Afterward, the criteria were presented in the form of operationalised inclusion/exclusion criteria given in Table 2. Two factors have resulted in the criteria for selection. Firstly, since the review focus on instruction rather than a mere description alone, it is necessary to have an explicit instructional, terminological or materials-design orientation which is a must for inclusion (purely representational media analyses were excluded). Second, because the sport sciences student is the main focus of the paper, studies on closely related specialised-purposes areas would be admitted only if the subject of analyses would be able to demonstrate that they could be transferred from sport sciences teaching for instance the corpus includes sport lexis or the teaching method can easily be adapted to teaching specific words of a field through making a word list.

Study Selection Process

The selection happened in four stages one after the other. First in the identification stage, 403 records were found in Scopus. Then there were 16 duplicates which after verification via comparing digital object

identifiers and matching the titles in normalized form were eliminated. So, after the removal of the duplicates, 387 unique records remained. At the screening stage, the screening criteria were used to evaluate all 387 records' titles, abstracts, and author keywords; 236 records were rejected because they in no way involved the use of corpus and they had very little to do with the sport or physical education topics instructionally speaking. The rest of 151 records were then submitted to be obtained and 12 full texts were missing. The eligibility phase consisted of reading through the 139 available full texts.

Table 1 <PICOS Framework Guiding Eligibility>

Element	Specification for this review
P - Population	Tertiary-level students of sport sciences and physical education, and other users of sport-related English (coaches, athletes, interpreters, translators, materials designers), particularly second-language users.
I - Interest / Intervention	Corpus-based, corpus-driven or corpus-assisted description or instruction of English, including word-list construction, collocation and frequency analysis, terminological description and corpus-informed materials design.
C - Comparison	General English or non-corpus-based description; general academic word lists; other specialised-purposes disciplines; first-language versus second-language users, where such comparison is reported.
O - Outcomes	Identified lexical, terminological and discursive features; coverage and frequency findings; pedagogical artefacts (word lists, materials, course specifications); learner knowledge or performance outcomes where measured.
S - Study design	Empirical corpus-based, corpus-driven, corpus-assisted or mixed-methods studies published as peer-reviewed journal articles in English.

Table 2 <Inclusion and Exclusion Criteria>

Criterion	Inclusion	Exclusion
Language of publication	English	Any language other than English
Document type	Peer-reviewed journal article (research article or review article)	Conference paper, book chapter, editorial, note, erratum, preprint
Publication period	2015–2025	Published before 2015
Methodology	Empirical corpus-based, corpus-driven, corpus-assisted or mixed-methods design with a described corpus	No corpus; purely theoretical, introspective or anecdotal treatment
Domain	Sport, physical education, exercise or physical activity; or an adjacent specialised-purposes domain whose analytic object is demonstrably transferable to sport sciences instruction	Domains with no sport, physical education or transferable specialised-purposes relevance
Orientation	Explicit instructional, pedagogical, terminological or materials-design orientation	Purely representational, ideological or media-critical analysis with no instructional or terminological orientation
Analytic output	Reports identifiable linguistic units, patterns or lists	Reports only system performance metrics without linguistic description
Accessibility	Full text retrievable	Abstract only; full text not obtainable

121 full texts were excluded with documented reasons: lack of corpus-based, corpus-driven or corpus-assisted methodology (34 studies); absence of any instructional, pedagogical or designing orientation (41 studies); non-location within the domains of sport sciences, physical education or transferable specialised purposes (32 studies); and being non-empirical, which includes commentaries, position papers and purely engineering-oriented reports without linguistic descriptions of natural language (14 studies). Eighteen studies met all three of the criteria and were taken through to synthesis. The screening of the literature was done collaboratively by members of the review team, each member having been given the same set of selection/criteria sheets and all disagreements over the evaluation of a record having been discussed to agreement over the written criteria only. Finally, the full process is outlined in Figure 1. Two reviewers independently screened titles, abstracts, and full texts according to the predefined eligibility criteria. Any

disagreement during the screening process was resolved through discussion until consensus was achieved. Inter-rater reliability was evaluated using Cohen's Kappa coefficient, yielding $\kappa = 0.87$, indicating excellent agreement.

Quality Assessment - FICO Framework

Methodological quality, in this case, was assessed by using FICO, a framework which was originally developed for evaluating qualitative studies and thus, the aspects included in such studies, which are quite different from descriptive corpus research. The FICO framework was selected because the included studies were predominantly descriptive corpus-based investigations rather than intervention studies. Conventional appraisal tools such as JBI or CASP are primarily designed for clinical or qualitative intervention studies and are therefore less appropriate for evaluating corpus-based linguistic research. FOCUS, one of the dimensions, evaluates the research question, its specificity and the way in which it has been reflected in the analysis. INFORMATION concerns transparency of the corpus: e.g., its size, sampling criteria, source and composition are reported in enough detail so that anyone can reproduce the study. CONTEXT refers to the description of situations from which the texts have emerged: the register, mode, participants and field of the activity. OUTCOME deals with the degree to which the findings and any pedagogical claims based on them are reported clearly and are supported by sufficient evidence. Each of the four dimensions got a rating from 0 to 3, the lowest being "not reported at all," then "very poor," "satisfactory" and finally "excellent," so a total of 12 points could be obtained. As a minimum requirement, a paper had to achieve 8 (i.e. half) points with no dimension scoring 1 (i.e. very poor) or below. All 18 studies that completed the preliminary inclusion phase also satisfied the quality requirement; detailed scores are shown in Table 3.

Table 3 <FICO Quality Appraisal of the 18 Included Studies (Threshold $\geq 8/12$)>

Study	F	I	C	O	Total	Decision
Benson & Coxhead (2022)	3	3	3	3	12	Included
Thaksanan & Kamsa-Ard (2025)	3	3	2	3	11	Included
Kongcharoen (2023)	3	2	2	3	10	Included
Yotimart (2021)	3	3	2	3	11	Included
Riccobono (2020)	3	2	3	3	11	Included
Loesnophaimongkhon et al. (2025)	3	3	2	2	10	Included
Elmes (2018)	3	3	3	2	11	Included
Sandrelli (2015)	2	2	3	2	9	Included
Abas & Kaur (2019)	3	2	3	2	10	Included
Bobyreva (2015)	2	2	3	2	9	Included
Bobyreva (2023)	2	3	2	2	9	Included
Kardoš Stojanović et al. (2024)	3	3	2	3	11	Included
Milić & Kardoš (2019)	3	3	2	3	11	Included
Labarta Postigo (2023)	3	2	3	2	10	Included
Cappuzzo (2020)	3	3	2	2	10	Included
Pîrlițeanu (2022)	2	2	3	2	9	Included
Phoocharoensil (2022)	3	3	2	3	11	Included
Parizoska & Rajh (2017)	3	2	3	3	11	Included

Note. F = Focus; I = Information; C = Context; O = Outcome. Each dimension scored 0–3.

Data Extraction Procedure

Structured data capture forms have been used to extract data from all the studies included. These extracted elements included: a fully formatted reference; a date of publication; a title or name of corpus (in the case that authors assigned one); a number of running words or number of terms making up the corpus; a kind of the corpus as in either a written, a spoken, or a mixed one; language setup as either a monolingual English or English plus a contrast language; software, tools of analysis or reference lists used; study type; topic of the article; main conclusions of the research; and any application of teaching, either stated or implied that have been reported. There is a single instance which needs a separate explanation. This is because the Scopus export that we had for carrying out this review was lacking in an author affiliation or country field that made it impossible for us to get the country of the author, and the categories given in Table 5 only record the language and domain context of the corpus as stated by the authors rather than a national association of the authors. In case a certain element could not be extracted from the data given by the authors, the corresponding cell was not left blank but was instead marked as data was missing in the original file. Data extraction was independently performed by two reviewers using a standardised extraction form. Any inconsistencies were checked against the original publication and resolved through discussion.

PRISMA 2020 Flow Diagram

The identification, screening, eligibility and inclusion process is reported in Figure 1, following the structure and reporting conventions of the PRISMA 2020 statement (Page et al., 2021; Liberati et al., 2009).

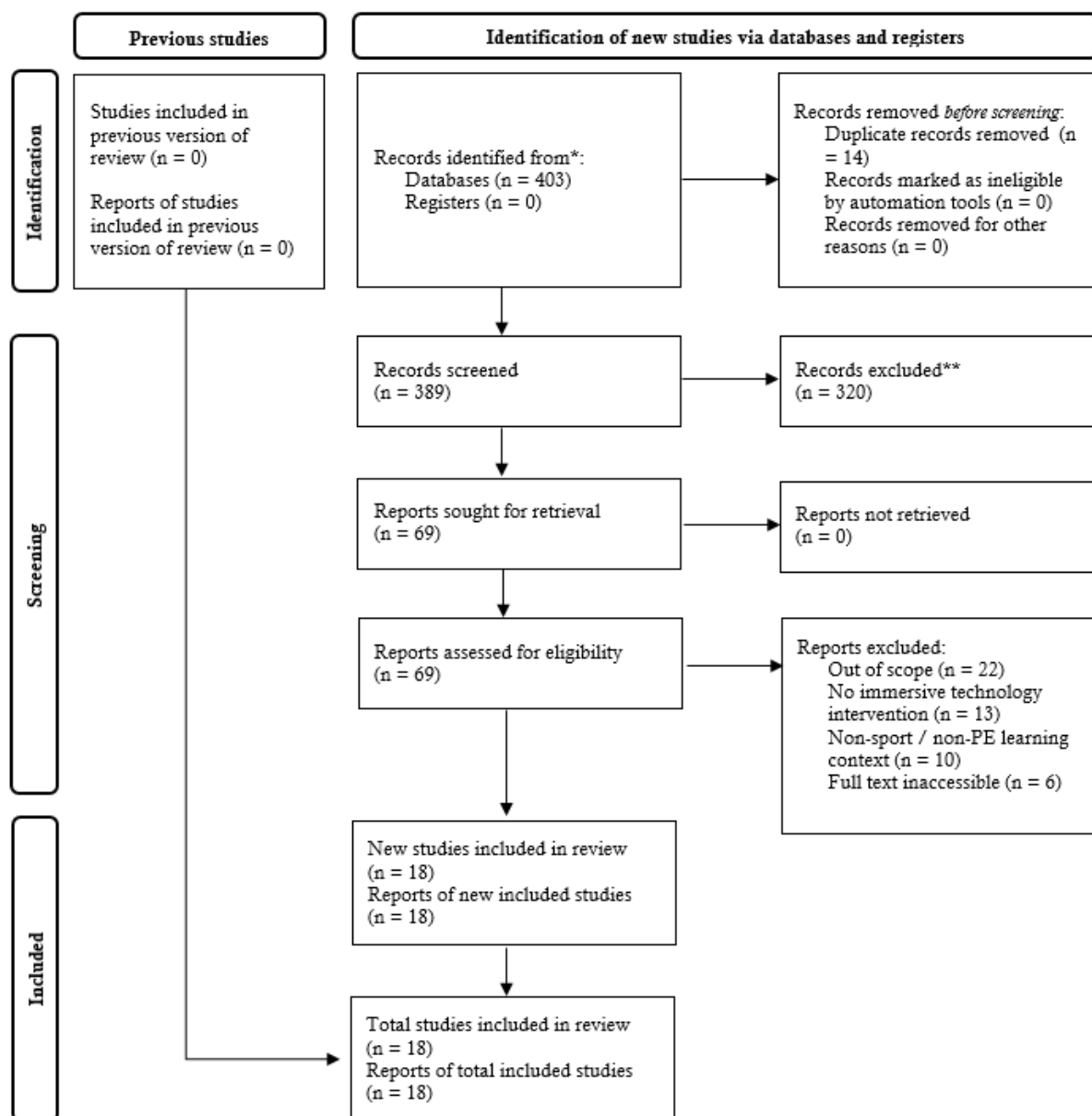


Figure 1 <PRISMA 2020 Flow Diagram of Study Identification, Screening, Eligibility and Inclusion (Adapted from Page et al., 2021)>

Network and Bibliometric Analysis Methodology

A minimal bibliometric overlay method was introduced to illustrate the time development and the topic distribution of the collected set. Bibliometric mapping was used only as a descriptive visualisation tool rather than inferential analysis. Its purpose was to illustrate publication trends and thematic relationships across the included studies. To present the distribution of publication years of included records, we made a chart displayed in Figure 2. For the study of concept organization, we identified patterns in the combination of author keywords with titles and abstracts of 18 included studies. Author keywords, however, present a problem as they are rarely used and are, therefore, hardly comparable the word which has the highest frequency is used three times only, meaning that if we used raw word frequencies, we would have gotten a map which is largely disconnected and of no value. Therefore, we mapped concept terms against regular expressions to normalise them into seventeen unified categories by checking titles, author's keywords, and abstracts of the article and counted them to represent the sizes of the nodes of categories in Figure 4. Edges

were laid out only between pairs of categories co-occurring in at least seven out of the total eighteen articles. We are not aiming at statistical inference by the way of the above procedure and it is given as a full description on the grounds of reproducibility.

Data Analysis and Synthesis

In line with the three stages of synthesis proposed by Thomas and Harden (2008), the findings were synthesised thematically. In the first stage, the findings and abstracts of the included studies were carefully read and the content line by line was coded. The initial coding provided codes that served three functions: they were corpus-type descriptors, they specified the unit and analytical criterion and finally they made some pedagogical claim. In the second stage, the descriptive codes were organised together to form descriptive higher-order themes based on analytic object sharing in order, for example, studies of rugby, baseball and gymnastics jargon to be read as such when the same linguistic construction was being treated. Finally, analytical themes were obtained by asking the descriptive themes of the three research questions which generated four clusters summarised. The integrity of theme generation was verified by each author of the analytical themes re-experimenting the data source by means of the claim that every constituent claim could be traced to a specific study. Coding was initially performed independently by two reviewers. Initial codes were compared, refined through discussion, and organised into descriptive themes. Analytical themes were subsequently developed by integrating related descriptive themes according to the three research questions.

Reporting and Documentation

The report uses the PRISMA 2020 checklist for sections relevant to a qualitative descriptive linguistic system review, among others, like listing and reporting the information sources identification, the complete search strategy, the eligibility criteria, the selection, the data items collected from the study, the certainty judgement and the flow of documents (Page et al., 2021). Those PRISMA items referring to the outcome measures, risk-of-bias-assessment across studies and merging of numerical results were not relevant and were not reported as such. The PRISMA 2020 diagram is reproduced in Figure 1. All the numbers in the abstract, within this section, within Figure 1 and within Section 5.1 were extracted from the same screening record and they are consistent with each other. The review procedure is based on an internal protocol developed prior to the screening but was not prospectively registered, and we are mentioning it as one of the weaknesses. Although an internal review protocol was developed before the screening process, the protocol was not prospectively registered in PROSPERO or Open Science Framework because the review addressed educational corpus-based studies rather than healthcare interventions.

Results and Discussions

Study Selection Results

A systematic Scopus search identified 403 records at this first step. Duplicate documents, i.e., those sharing the same digital object identifier or normalised title, accounted for 16 records which were removed, leaving 387 documents which were subsequently reviewed. Document content and keywords filtering reduced the number to 151 which included the authors' request to retrieve the document, leaving 12 documents of which fulltext could not be secured. After evaluating the fulltexts of the remaining 139 documents for suitability, 121 of them were discarded giving reasons: 34 of them lacked a corpus-based or corpus-drifting method; 41 of them did not give a description of a corpus-driven or corpus-based study, 32 of them concerned a subject domain different from sport sciences, physical education or transferable specialised-purposes 14 of them were non-commentary pieces or engineering-oriented reports without a linguistic descriptive component. Finally eighteen studies which fulfilled all criteria are presented together with eighteen studies which also passed the FICO quality bar, resulting in a synthesis pack of 18 papers.

Descriptive Characteristics

We found eighteen articles that qualify as corpusbased sport English papers, which have appeared in journals mainly in the last ten years: the earliest paper is from 2015, the latest from 2025. Overall, there is not a very steep growth in yearly numbers although there are fluctuations in output between 1 to 3 studies per year. The most intense publishing period was obviously the year in which six out of twenty-one papers were published. So, there is an evident clustering in 2022-2023 period. Also, there is very little evidence for one single year standing out. Moreover, the very absence of a steep upward growth is a finding itself it shows that corpus-based work on sport English is still regarded as an activity which appeals to a narrow audience rather than being a program rapidly expanding. The full characteristics of the included studies are reported in Table 4, and their classification by theme, design and analytic output is also reported.

The size of corpora differs by a factor of about 1000 in the compilation: from terminological corpora of, roughly, 200 to about 600 items up to a written corpus of 2, 836, 258 tokens. The corpora used in six studies range between 21, 000 and 160, 000 words. This seems to be the range of sizes most suited to an individually built specialised corpus, i.e., at the level practicable for a researcher. The modality is primarily written, three studies have a genuinely spoken data base, and two of them are the only studies in this set which are investigating instructional or professional talk as it is naturally produced in a sporting context. Language configuration is also a way of classifying the set: 12 studies focus on the monolingual analysis of English and only six work contrastively between English and a second language.

Themes and designs distribution can be found in Figure 3. Lexical profiling and frequency analysis is a dominant design in terms of number of studies, six of them, followed by contrastive and translation analysis (five), corpus-assisted discourse analysis (four) and mixed-methods designs that use corpus evidence together with elicitation from expert of learner participants (three). Figure 4 depicts the conceptual map of the set. At the heart, there is a dense core in which the use of corpora, sport domain and a pedagogical approach occur together in almost all the records, whereas vocabulary, technical lexis, frequency analysis, and discourse analysis constitute the first layer. And then, word lists, ESP framing, academic vocabulary, metaphor, spoken language, and collocation are on a remote third level of the structure. The lack of elements on that periphery is a clue: collocation, the teaching element that gives the biggest pay-off, is featured only in two out of eighteens studies.

Table 4 <Summary of the 18 Included Studies>

Title (full)	Author(s) (Year)	Corpus / context	Design	Key findings
Technical single and multiword unit vocabulary in spoken rugby discourse	Benson & Coxhead (2022)	Spoken rugby union corpus, 61,295 running words; English (L1 and L2 users)	Mixed methods: corpus frequency analysis with a receptive vocabulary knowledge task (n = 77)	Over 4,000 word families plus supplementary lists are required for 98% comprehension; a list of 252 technical single words covers more than 12% of the spoken corpus, and a pedagogically oriented list of 267 multiword units was derived; first-language and second-language users differ markedly in technical rugby vocabulary knowledge.
Lexical Collocation Analysis in a Corpus of Research Articles in Physical Education and Sport Science	Thaksanan & Kamsa-Ard (2025)	COPES corpus of 144 research articles, 737,901 running word tokens; academic written English	Corpus-based keyword and collocation analysis using a concordancing toolkit	The 200 highest-ranked keywords are dominated by nouns (64%), followed by verbs (20.5%), adjectives (14%) and adverbs (1.5%); the most prevalent collocational frames are noun+noun, adjective+noun, noun+verb and verb+noun, providing a direct basis for collocation-focused instruction.

Title (full)	Author(s) (Year)	Corpus / context	Design	Key findings
Investigation of Vocabulary for ESP Classrooms from Academic Journals in Physical Education and Sports Science	Kongcharoen (2023)	Academic journal corpus in physical education and sports science; written English	Corpus-based word-list construction using lexical profiling, frequency, range and expert verification	A field-specific physical education and sports science word list of 466 word families was produced and classified into three groups, offering teachers and curriculum designers a validated lexical target for English for specific purposes courses in the discipline.
Academic vocabulary in sport tourism news: A corpus-based study	Yotimart (2021)	Sport tourism news corpus, 112,623 running words from international online news; written English	Corpus-based coverage analysis against the Academic Word List using a lexical profiling tool	Academic Word List items cover only 7.38% of the corpus and several frequent items in the corpus are absent from that list, indicating that the vocabulary needs of sport and tourism students differ characteristically from those of other disciplines and require field-specific lists.
Triangulating diamond talk: Identifying technical spoken vocabulary in english for baseball purposes	Riccobono (2020)	Baseball English Corpus (BECO); spoken English in an international baseball context	Mixed methods: corpus-based technical vocabulary identification triangulated with ethnographic interviews	The study establishes English for Baseball Purposes as a subtype of English for occupational purposes and delivers specialised spoken word and phrase lists, while demonstrating that expert stakeholder verification changes which items are ultimately classified as technical.
A Corpus-Based Analysis of Lexical Characteristics Across English News Categories for L2 Pedagogical Use	Loesnopchaimongkhon et al. (2025)	3,000 news articles across 12 categories, including a sports category; written English	Corpus-based lexical profiling of coverage, level, variation, density and CEFR banding	Sports texts show the lowest lexical density (53.2%) and, together with health texts, the highest proportion of high-frequency and A1–A2 vocabulary, making sport a linguistically accessible entry register for lower-

Title (full)	Author(s) (Year)	Corpus / context	Design	Key findings
Coach-player communications: An analysis of top-level coaching discourse at a short-term ice hockey camp	Elmes (2018)	Transcribed on-ice coaching instructions from four coaches at a Canadian ice hockey camp, 21,376 words; spoken English	Corpus-based word-type and frequency analysis with functional categorisation of the lexis	proficiency second-language learners while limiting its value as advanced input. Of 1,022 quantifiable words, single-category vocabulary accounted for 75.2% of the categorised language, with sport-specific (4.6%) and sport-general (11.1%) categories together constituting 15.7%, showing that instructional coaching talk is largely general lexis deployed in a specialised functional frame.
“and maybe you can translate also what i say”: Interpreters in football press conferences	Sandrelli (2015)	Small corpus of football press conferences for the presentation of new players; multilingual interpreted interaction	Corpus-based case study of interpreted spoken interaction	Interpreted football press conferences follow recurrent communicative routines and recurrent pitfalls, and interpreter role boundaries are frequently renegotiated in situ, identifying a niche professional register that is absent from generic language provision.
Theologising the discourse of representation through ‘faith’ and religion in football news	Abas & Kaur (2019)	Corpus of football news articles from an English-language Malaysian newspaper; written English	Corpus-assisted collocational and interdiscursive discourse analysis	The item faith co-occurs systematically with religious vocabulary, creating interdiscursive relations that theologise team allegiance, demonstrating that apparently general lexis carries domain-specific evaluative loading that learners must be taught to recognise.
Eponymous units in the english sublanguage of	Bobyreva (2015)	Sport dictionaries, glossaries and an institutional	Mixed methods: descriptive terminological	Eponymous sport terms are typically two-component units reducible to a proper

Title (full)	Author(s) (Year)	Corpus / context	Design	Key findings
sport: Form, meaning and functioning		sport discourse corpus, supplemented by interviews; English	analysis combined with corpus and interview evidence	name, are brief and stylistically neutral, and form a heterogeneous layer spanning widely used and rare items, archaisms and neologisms whose meaning is largely restricted to specialists.
Investigating the Role of English Language Tools Used in Sports Communication Using Linguistic Methods and its Use in Academic Dictionaries	Bobyreva (2023)	Author-compiled corpus of original sport texts, 2,836,258 tokens; written English	Corpus-assisted structural, etymological, contextual and discourse analysis with typologisation	Proper names generate a wide range of units in sports communication, including terms, precedent names, pragmonyms and nomenclature, with direct applications for specialised dictionary compilation, terminological training and translation instruction.
Fitness terms in English and their adaptation into Serbian	Kardoš Stojanović et al. (2024)	Corpus of 200 English fitness terms and their Serbian equivalents; English–Serbian contrastive	Corpus-based contrastive terminological analysis of translation procedures	Four adaptation procedures operate — transshaping, direct translation, structural translation and functional approximation — and a substantial number of competing variant forms in the receiving language demonstrates an urgent need for terminological standardisation and lexicographic registration.
Adaptation of gymnastics terms from English into Serbian: Theoretical and practical aspects	Milić & Kardoš (2019)	Corpus of approximately 600 gymnastics terms compiled from official texts; English–Serbian contrastive	Corpus-based contrastive analysis of translation and transliteration procedures	Existing gymnastics terminology in the receiving language contains numerous poly-lexical and poly-morphemic units, several of which violate the receiving standard; the study proposes an explicit standardisation model aligned with an

Title (full)	Author(s) (Year)	Corpus / context	Design	Key findings
'Life Is Baseball' vs. 'A Vida É Futebol': the Translation of Sport-Specific Metaphors from English to European and Brazilian Portuguese	Labarta Postigo (2023)	TV Corpus supplemented by an author-compiled corpus; English–Portuguese subtitle comparison	Corpus-based comparison of machine translation output with human subtitling	existing bilingual dictionary of sports terms. Sport-specific metaphorical and idiomatic expressions remain systematically mishandled by freely available machine translation systems, confirming that culturally embedded sport figurative language cannot be delegated to automatic tools and requires explicit instruction.
Medical English translation into Italian: The case of Netter's work	Cappuzzo (2020)	Parallel corpus of 151,121 words of medical English and 157,016 words of Italian translation, covering musculoskeletal and sports-related injury content	Corpus-based contrastive analysis of terminological, syntactic and stylistic features	Accurate specialised translation depends not only on terminological precision but on compliance with the syntactic and stylistic conventions of the target specialised register, a requirement directly relevant to sports medicine and injury content in sport sciences curricula.
Metaphorical language in football media: A semantic and contextual analysis	Pîrlițeanu (2022)	Corpus of metaphorical expressions from English-language football media; written English	Corpus-based semantic and contextual analysis distinguishing basic from contextual meaning	Football metaphors are structured by recurrent source domains including war, nature and machinery and function cognitively rather than decoratively; the study proposes football journalism as a culturally resonant resource for metaphor instruction in English as a foreign language.
Primary, Main, and Major: Learning the Synonyms through Corpus Data	Phoocharoensil (2022)	Corpus of Contemporary American English; written and spoken English across genres	Corpus-informed genre distribution and collocation analysis using	Near-synonyms differ systematically in genre distribution and collocational preference, with major showing a distinct association

Title (full)	Author(s) (Year)	Corpus / context	Design	Key findings
Idiom variation in business English textbooks: A corpus-based study	Parizoska & Rajh (2017)	Corpora of business English usage compared with ten competition idioms as presented in textbooks; written English	frequency and mutual information Corpus-based analysis of lexical and syntactic idiom variation with a cognitive linguistic account	with sport and business collocates, demonstrating that domain-sensitive synonym discrimination is learnable from corpus data but not from dictionaries alone. Sport-derived competition idioms are presented in textbooks in a single fixed form whereas corpus data show two or more lexical and syntactic variants carrying distinct functions, evidencing a systematic mismatch between instructional materials and authentic specialised usage.

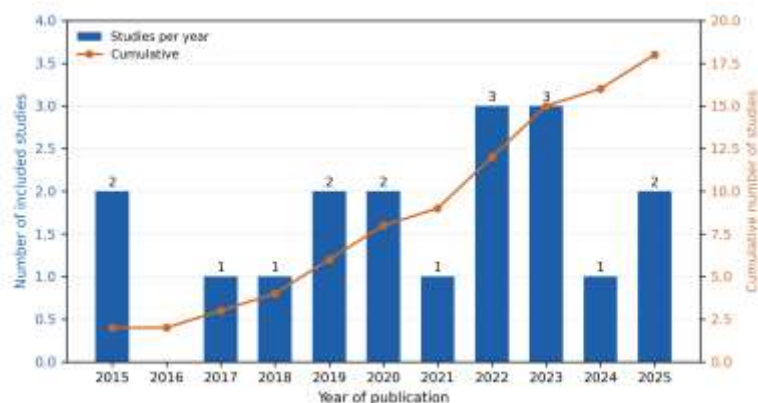


Figure 2 <Annual and Cumulative Distribution of the 18 Included Studies, 2015–2025>

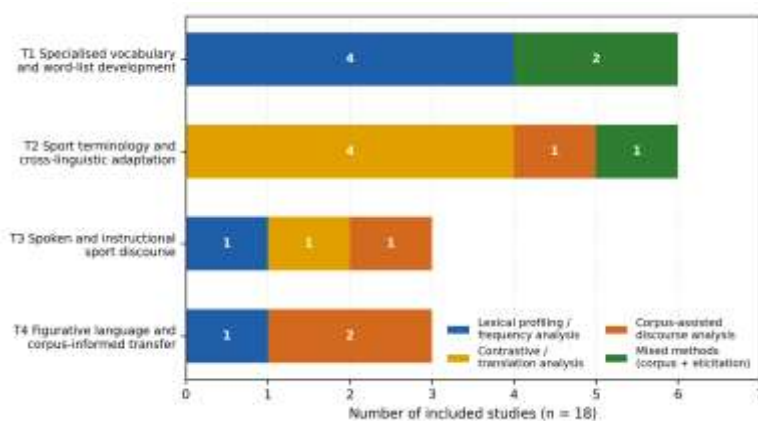


Figure 3 <Distribution of Included Studies by Thematic Cluster and Research Design>

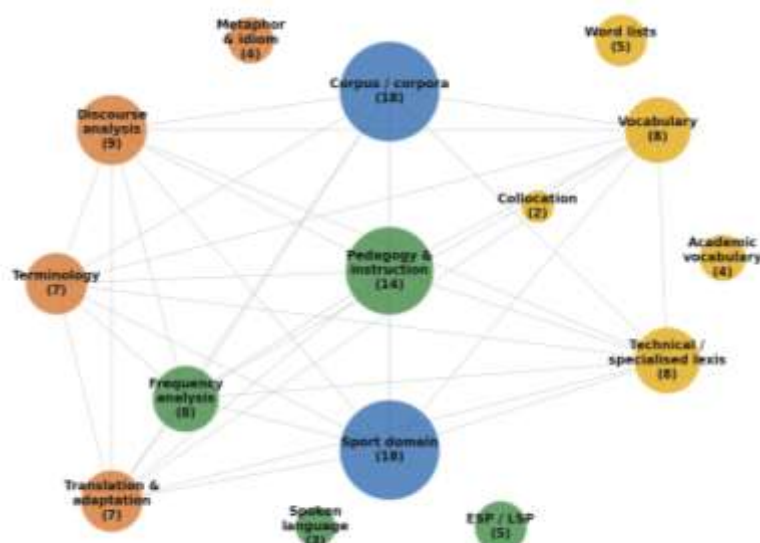


Figure 4 <Normalised Concept Co-occurrence Map of the 18 Included Studies (Node Size = Number of Studies in which the Concept Category Occurs)>

Findings for RQ1: Corpus resources, designs and analytical procedures

Analytics methods generally fall into four distinguishable clusters. Lexical profiling and frequency analysis which was used in six articles are techniques that exploit well-known frequency bands and coverage thresholds to discover the vocabulary richness of a certain area and to derive word lists (Thaksanan & Kamsa-Ard, 2025; Kongcharoen, 2023; Yotimart, 2021; Loesnopchaimongkhon et al., 2025; Elmes, 2018; Phoocharoensil, 2022). Contrastive and translation approaches in five studies focus on differences between the source-language English register and the target-language register in order to identify adaptations strategies and terminological changes (Kardoš Stojanović et al., 2024; Milić & Kardoš, 2019; Labarta Postigo, 2023; Cappuzzo, 2020; Sandrelli, 2015). Corpus-assisted discourse (CAD) in four studies is a form of discourse analysis supported by corpus tools, like concordance and collocation (Abas & Kaur, 2019; Bobyрева, 2023; Pîrlişteanu, 2022; Parizoska & Rajh, 2017).

Reporting quality varies in this and another matter, which the FICO scores in Table 3 make clear. In general the Focus and Context dimensions are adequately addressed: the studies clearly state their objectives as well as describe the sports contexts from which the texts they have come to analyse were derived. The Information dimension stands out as a weak point in the set of studies. A number of studies report a corpus size without giving the sampling frame, period of collection and admission criteria for texts. In terms of terminology studies, they often describe their study samples as just being extracted from relevant literature or official texts. Since a study of a corpus is solely dependent on the recreatability of the corpus, this lacuna significantly limits the building up of the body of knowledge. It also seems that the three mixed-methods reports are the best in general which may show that having to defend the results of your study before a committee of knowledgeable people also gets the writers to be stricter and more thorough when they present the corpus. The most direct response to question RQ1 is that the researchers usually refer to small, single-author made-up collections of texts (mostly written), which are mostly the result of four types of designs. Furthermore, the descriptions of their data seem to have such a high level of uniformity that synthesis of the results of different studies is possible without much effort, yet the descriptions lack the fine detail required by any good study which aims at being replicated.

Findings for RQ2: Instructionally relevant linguistic features

The second research question is about which linguistics aspects repeatedly appear in the data and are pedagogically significant. Only three of the findings are robust enough to warrant being considered as facts. First off, the use of technical terms takes up a major part of the linguistic load. Benson and Coxhead (2022) point out that if one takes a group of 252 single technical words, more than a twelfth of words spoken in a rugby corpus will be among those words, which is far more than the general high-frequency words that an equivalent number would be associated with and even the inclusion of 267 multiword units is justified from a pedagogical view. Riccobono (2020) obtains almost the same kind of result for a corpus of American baseball, and Kongcharoen (2023) applies a similar reasoning and provides a field-specific list of 466 word

families for PE and sport science. Across these papers, the message of the pedagogical value has been pretty much the same: a closed, learnable set of vocab will be the most efficient one compared to a generic instruction.

The second main agreement is that general academic vocabulary is insufficient for this group. Yotimart (2021) estimates Academic Word List coverage as a mere 7.38% in a sport tourism news corpus and, apart from that, points out a number of common words appearing in the corpus which are excluded from the list, suggesting that sport students are a special case and their vocabulary needs are significantly different from those of other disciplines. Thaksanan and Kamsa-Ard (2025) back up this statement by pointing out another phenomenon: keyword profiles of sports science research articles are primarily determined and structured around discipline-specific collocational frames but not general academic formulae. Loesnopchaimongkhon et al. (2025) give an important caveat: news sports articles were found to have a very low lexical density as compared to twelve other categories and a large number of A1-A2 items, making sports an easy-to-enter register but a poor source of advanced input. Altogether, these three pieces show that one cannot judge difficulty or easiness of the sport sciences vocabulary against a general academic English vocabulary list or a dictionary. Instead, they point out the idea of different distribution and that, in this context, selection of register is probably as important as selection of list.

The third overlapping result highlights terminological discrepancies across languages. Kardoš Stojanović et al. (2024) outline four separate ways in which fitness English terms are changed, and show that there are a large number of different forms of those terms that coexist in the target language. Milić and Kardoš (2019) describe the same phenomenon occurring in gymnastics and mention forms which deviate from the receiving language's standard as well. Both studies advocate for clear standardisation and dictionary recording. Based on Bobyrev's work (2015, 2023), one of the reasons why such terms are not well understood by ordinary language users is that they contain names of people. These names are responsible for the creation of terms and names of previous users, terms with more than one usage, names and meanings which are mainly used by professionals and therefore context cannot normally help readers understand such terms. Labarta Postigo (2023) quantifies how harmful this instability has been through an observation that free machine translation engines have great difficulties with translating metaphors and fixed expressions that contain sport-specific words.

A fourth cluster of features is convergent in a much weaker sense yet instructionally important. Multiple studies report that general lexis actually functions in disciplinary ways that cannot be inferred from general instruction, learners. Abas and Kaur (2019) demonstrate that a religious discourse is superimposed on a general item faith through the lens of football news; Phoocharoensil (2022) reveals that a near-synonym major is collocate with Sport and business words preferentially while primary and main are not; Parizoska and Rajh (2017) identify that the same sport-derived idioms referring to competition appear in the corpora as two or more variants, while textbook present single fixed forms. Pîrlițeanu (2022) reports on organisational football metaphor which recurrently derives from source domains of war, nature and machinery. Elmes (2018) also provides evidence from spoken language side which indicates that coaching talks include sport-specific and sport-general categories that amount to only a very small percentage (15.7%) of the categorised words, remaining being made up of general vocabulary used in functional frames of specialisations. RQ2 has been directly addressed with the features that are instructionally important as technical single and multiword items having a wide domain coverage, discipline-specific collocational frames, terminology that is cross-linguistically unstable, and the domain conditioned behaviour of general lexis."

Findings for RQ3: Translation of corpus evidence into instruction

The third research question looks at how corpus description turns into instruction. The evidence for this is much thinner than for RQ1 and RQ2. You could best describe the pattern as a descriptive-to-instructional gap. Five studies deliver concrete pedagogical artefacts. Kongcharoen (2023) produces a validated word list covering 466 families, specifically aimed at teachers and curriculum designers. Benson and Coxhead (2022) create paired lists, one for single words and one for multiword expressions, with implications for rugby courses. Riccobono (2020) offers spoken word and phrase lists for a newly defined subtype of English for occupational purposes. Thaksanan and Kamsa-Ard (2025) develop a ranked inventory of keywords and collocations with stated pedagogical implications. Yotimart (2021) presents sport tourism news word lists as a basis for learner goal-setting. Together, these five studies form the operational core of corpus-based English instruction for sport sciences as it currently exists.

Beyond producing artefacts, instructional uptake stays largely programmatic. Pîrlițeanu (2022) suggests football journalism as a resource for teaching metaphors but never implements or evaluates it. Parizoska and Rajh (2017) point out a mismatch between textbooks and corpora, then derive implications for teaching

figurative language in specialised contexts without testing any fix. Phoocharoensil (2022) offers implications for ELT from corpus-based synonym work. Cappuzzo (2020) shows that specialised translation competence needs stylistic and syntactic conformity along with terminological accuracy, a finding with clear curricular relevance for sports medicine content that the study does not operationalise. In each case, the pedagogical claim makes sense and sits close to the evidence, but it is not evidence itself.

From the entire set, only one study collects learner data. Benson and Coxhead (2022) gave a receptive knowledge task to 77 people. They found a measurable gap between first-language and second-language English users when it comes to technical rugby vocabulary. This single result is the only direct evidence in the reviewed literature that the words found by corpus analysis are actually known differently by the learners who would need instruction. Two other studies use expert judgement instead of learner data. Riccobono (2020) triangulates corpus output with ethnographic interviews of stakeholders. This changes what items get labelled as technical, the author reports. Kongcharoen (2023) uses expert verification as a criterion for building lists. Bobyрева (2015) similarly relies on interviews with people in sport.

No study in the set reports on classroom implementation. There is no controlled comparison between corpus-informed and conventional instruction. No longitudinal measurement of uptake exists. No assessment instrument comes from the corpus. So in direct answer to RQ3, corpus-based findings have been turned into instruction only up to the point of producing word lists and inventories. That is where it stops. The constraints on this translation include a lack of learner-outcome evidence. There are no shared and redistributable corpora. Spoken instructional data is almost entirely neglected in the disciplines where sport sciences students will actually work. And unresolved terminological instability makes it hard to standardise list-based instruction across different educational systems.

Comparative and Critical Analysis

If you look across the whole set, a clear pecking order emerges. Some methods are used a lot, and others are barely touched. Frequency-based lexical profiling sits firmly at the top. Why? It's cheap and replicable, plus the output is immediately useful. The constructs it works with, like coverage thresholds and frequency criteria, are already well established. Coming in second is contrastive terminological analysis. But it's mostly used by a specific group working on English-to-Serbian adaptation. That raises an interesting question: how well do these findings actually hold up across other language pairs? Corpus-assisted discourse analysis shows up too, and it's applied pretty competently. The issue is it tends to produce interpretive claims that aren't so easy to turn into direct instructional specifications.

The real surprise is what's missing. Three designs are conspicuously underused. Learner corpus analysis is completely absent. Not a single study in the set looks at what sport sciences students are actually producing in English. So basically, the field is setting targets without ever measuring the gap between where students are and where they need to be. Longitudinal or diachronic designs barely get a look in. This is odd given the clear evidence from adjacent fields that sport language shifts measurably over time (Leone, 2024; Li & Yang, 2024; Obradović, 2024). Multimodal corpus design is another gap. Coaching is inherently physical and multimodal, and analysis like this has worked well for sport communication in other contexts (Ismail et al., 2022; Blanco-Gómez & Gómez-Ortiz, 2024b). So its absence here stands out.

Even so, you can spot a methodological evolution unfolding across this period. The earlier papers, like Bobyрева (2015), Sandrelli (2015), and Parizoska & Rajh (2017), lean heavily on qualitative description backed up by corpus examples. The more recent ones take a different approach. Thaksanan & Kamsa-Ard (2025), Loesnopchaimongkhon et al. (2025), and Kongcharoen (2023) use explicit, criterion-based quantitative methods with clearly stated thresholds and verification steps. This move toward proceduralisation is probably the most positive trend in the evidence base right now. It gives researchers a solid methodological foundation to build intervention studies on.

Interpretation of findings

This synthesis basically provides evidence for the main interpretive idea that the English of the sports sciences should not be regarded as a 'more difficult' version of general academic English but rather as one of its dialects/register with specific features that set it apart. The language of sport sciences has both a dense, highly productive technical core and a significant amount of general vocabulary used in specialized contexts, and this language is used in different ways by the written academic and media versions and the spoken professional forms of the same topic. Therefore, Yotimart's (2021) study that general academic vocabulary can only account for a tiny part of sports text and Loesnopchaimongkhon et al. study (2025) that sports news in terms of the words it uses is easy: they just sampled different modes of the same type of domain. In terms

of vocabulary teaching, this means that learners who are given a vocabulary that ignores modality will not make efficient use of their precious and limited learning time.

Theoretical implications

The research results expand the specialized-words vocabulary model in dual ways. First, the results reveal that the technical-vocabulary construct, which was mainly created on the basis of written academic corpora, does work for spoken professional sport discourse. But such a model needs a supplementary multiword element, because both Benson and Coxhead (2022), and Riccobono (2020) argue that one-word lists under-reflect learners' needs. Furthermore, the findings reveal that the distinction between technical and general vocabulary here is determined by usage rather than formal features. Elmes's (2018) discovery that sport-specific and sport-general types make up just 15.7% of coaching discussions demonstrates in fact how specialists in the language of this context make their point not by using special vocabulary but rather by limiting the functional usage of general vocabulary. This phenomenon is not well reflected in coverage-based models which only count vocabulary. Such a finding is consistent with previous corpus studies where it was shown that vocabulary specific to a domain is most often manifested through patterns or collocations rather than through peculiar vocabulary (Ali & Ali 2023; Kazymir et al. 2026; Virág 2021).

Practical implications

Curriculum designers in sport sciences should note a key point here. A field-specific word list can either replace or support the general academic word list. This field-specific list acts as the lexical backbone for English language instruction. And it is not enough to just provide the word list on its own. It should come with a list of collocations, as suggested by Kongcharoen (2023) and Thaksanan & Kamsa-Ard (2025). For authors of educational materials, there is a clear challenge. Textbooks often present specialised idioms as single, fixed forms. But corpora show these idioms vary regularly (Parizoska & Rajh, 2017). This gives authors a corpus-based reason to check specialised materials at the corpus level before publishing them. This approach came from corpus research that examined textbooks in other contexts. That research already provided justification for using corpus checks on specialised materials (Almujaiwel, 2018; Zhang et al., 2022). Teacher-trainers who work with coaches and physical education practitioners should pay attention too. Instructional sport talk is a distinct genre. It has its own patterns of vocabulary and grammar at the functional level, as Elmes (2018) and Clifton et al. (2023) show. This strongly supports developing task-oriented speaking instruction instead of reading-based classes. Higher educational institutions, especially those in non-Anglophone countries, face a problem with unstable terminology. Multiple studies have noted this issue (Kardoš Stojanović et al., 2024; Milić & Kardoš, 2019; Núñez Nogueroles, 2023; and Luján-García, 2021). One solution is to standardize an institutional glossary. This can then pave the way for a stable bilingual teaching regime.

Comparison with prior reviews

Previous interdisciplinary studies in this area have taken the path from outside the dimension of language. Fjellner et al, in their paper published in April 28, 2023, have analyzed how physical education teachers occupy the models-based scholarship and pointed out the ongoing under-identification of the role of the teacher, yet, language was not one of the issues in consideration. The article by López Ibáñez et al (2023) has demonstrated a design of physically-inclusive activities in PE with a focus not being put on the language aspect of the activities. Within the field of corpus linguistics, literature has mainly dealt with the use of generally-learned data-driven approaches as against the use of any specialized field's data. This means that our article is a review that has covered ground that neither side has claimed and the main value of the piece is that it points out, through contrast, that corpus-based sport English research is, on one hand, well-established from a methodological standpoint but, on the other hand, remains incomplete in terms of its pedagogic application.

Contradictions in the literature

At a minimum, two main contradictions are to be recognized. The first disagreement revolves around the difficulty of sport English. According to Loesnophaimongkhon et al. (2025), it is a sport register of low-density, high-frequency that is suitable for beginners. On the contrary, Benson and Coxhead (2022) argue that understanding a sport discourse in spoken form requires more than 4,000 words families plus vocabulary lists for specific types. This issue may be considered solved if modal and genre features are kept the same: for instance, a sports news report and sports-related conversation will be very different. The fact however that the two are often confused in course design is exactly what the separation of such a distinction will highlight and correct.

The other contradiction pertains to the degree in which proper-name-derived and eponymous terms are used. Bobyрева (2015) describes eponymous sports terms as short, stylistically neutral and highly

professional, mainly understood by specialists, which suggests low learnability from context and high instructional priority. Research of Anglicism adoption in sports by other linguistic populations indicate more permeability where such terms spread in the popular and media use besides the academic context (Núñez Nogueroles, 2023; Luján-García, 2021; Leone, 2024). Likely reason is that two kinds of items exist: one specialized-restricted and one circulating-popularly whose layers have not been systematically distinguished yet so question remains open.

Critical Reflection of the Current Evidence

An important observation emerging from this review is that the existing body of corpus-based research in sport sciences remains predominantly descriptive rather than evaluative. Most studies successfully identify specialised vocabulary, terminology, and discourse features; however, relatively few investigate whether corpus-informed instructional materials actually improve learners' vocabulary acquisition, disciplinary communication, or academic performance. Consequently, the evidence base for the pedagogical effectiveness of corpus-based approaches remains limited.

Another limitation concerns the methodological diversity of the included studies. Considerable variation exists in corpus size, corpus composition, analytical procedures, and reporting practices, making direct comparison across studies difficult. Furthermore, spoken corpora and multimodal corpora remain substantially underrepresented compared with written corpora, despite the highly interactive and communicative nature of sport sciences education.

These observations indicate that future corpus-based research should move beyond descriptive linguistic analyses toward intervention-based studies employing experimental or quasi-experimental designs. Greater methodological standardisation, transparent reporting, and the development of learner corpora and multimodal sport communication corpora would substantially strengthen the evidence base for English for Specific Purposes (ESP) instruction in sport sciences.

Research gaps

Gap 1: Learner evidence. Just one out of eighteen studies involved collecting data from the learners and none of the research analyzed the production by the learners. In essence the discipline has fixed the lexicons to be learned with no measurement taken about the distance between the lexicons and learners' existing knowledge; no supporting evidence that corpus data lists would bring any kind of learning benefit in such areas has been provided.

Gap 2: Spoken and multimodal data. A total of three studies have looked at spoken language data and none has dealt with multimodal data even though coaching, therapy, and clinical instruction, the main work settings for sport sciences graduates, happen to be spoken and embodied activities. What little evidence there is about the language of sport instruction is based on a very limited number of corpora collected from a rather narrow selection of sports.

Gap 3: Corpus availability and comparability. Most of the corpora were collected by the research group themselves and not one of them is mentioned as having a free license for redistribution. This is a major obstacle, as not only reproduction but also combining data is not possible and one will have to recreate everything from the beginning. Along with the information aspect of the FICO appraisal, which is very weak, all this together makes this one of the most serious issues for the field to make cumulative advances through.

Limitations of this review

To start, one electronic source for literature search was accessed. Scopus does offer wide and uniform scope of covered peer-reviewed content but it does not consider non-indexed regional journals, proceedings of conferences and grey literature; therefore, the relevant sport English corpus work that is based on the corpus is in a local applied linguistics journal might not have included. This drawback is highly relevant to a subject where non-Anglophone educational systems are mainly the ones making decisions.

In addition, the inclusion criteria mandated that the item has an instructional content, terminology or materials-design focus. Though this delimitation was needed to keep the literature review focused, it resulted in the exclusion of a lot of corpus-based sport discourse research that was methodologically sound whose pedagogical insights were not claimed by the writers. As a result, the report is about corpus work on sport English that is pedagogically-oriented and not the linguistic characterisation of the whole discourse domain.

Besides, it was found out that in the Scopus export files used after the inclusion phase, it was not provided the information about the author's work unit or university so the spread across different regions of the studies identified could not be tracked and is therefore not presented. Also, because of the previous point that in this

study, two stages, that is screening and the appraisal, were carried out by the one and the same set of reviewers, the review paper was never registered as a protocol. Whereas our decisions were made against written criteria and differences of view were settled through discussion, not registering the protocol limits the possibility for verification by outsiders. Last but not least, the FICO appraisal is a judgment-based tool and therefore inter-rater reliability was only informally checked.

Theoretical Implications

The findings extend corpus-based ESP theory by demonstrating that specialised vocabulary should not be treated as isolated lexical items but rather as interconnected terminology, phraseology, and discourse practices embedded within authentic sport sciences communication. This synthesis therefore strengthens the theoretical integration between corpus linguistics and English for Specific Purposes.

Practical Implications

Practically, the findings provide evidence for curriculum developers to construct corpus-informed teaching materials, specialised vocabulary lists, and authentic classroom activities tailored to sport sciences students. ESP instructors may also integrate concordancing tools and corpus-based learning activities into vocabulary instruction.

Future research agenda

Recommendation 1: Conduct intervention studies. Such a comparison needs to be set up under controlled condition, where sport science students are taught from a field-specific list and collocation inventory which derives from a corpus while being compared with students who only get the traditional general language support. We measure learners' knowledge by their comprehension and their language production is checked as well. Word lists necessary for these studies are currently available (Kongcharoen, 2023; Benson & Coxhead, 2022; Riccobono, 2020) but their assessment and effectiveness have not yet been done.

Recommendation 2: Construct, disseminate and create spoken and multimodal corpora from applied sport situations with different settings being given priority. These include coaching instruction of different sports in general and not only of ice hockey and rugby, sports therapy and rehabilitation consultation, and strength and conditioning training. These corpora should be produced in a replicable way that includes sampling frame, time period, participants, and transcription conventions and then deposited in a way that allows them to be used again.

Recommendation 3: Develop cross-linguistic terminology references. Based on the standardization work by Milić and Kardoš (2019) as well as Kardoš Stojanović et al. (2024), coordinated bilingual and multilingual glossaries covering sport sciences terminology should be compiled together with the domain experts and validated, explicitly pointing out the distinctions between specialist-restricted versus widely circulatory layers. Additionally, two more aspects require exploration: collection of student corpora in terms of sport sciences learner writing to detect the discrepancy in the target and performance levels, as well as the evaluation of large language models as corpus-querying and material-generation tools to some extent with systematic review of domain-specific norms (Gökmen et al., 2026) and figurative expressions in sports (Labarta Postigo, 2023) being limitations of these large linguistic models. Future corpus-based studies should move beyond descriptive analyses toward intervention-based research evaluating learning effectiveness. The development of spoken corpora, multimodal corpora, learner corpora, and AI-assisted corpus analysis also represents promising directions for future investigation.

Summary answers to the research questions

Answer to RQ1. The majority of the available information is based on relatively small, manually created, written single-domain corpora, which have been processed through four consistent methods: lexical profiling and frequency analysis, contrastive and translation analysis, corpus-assisted discourse analysis, and mixed-methods triangulation with expertise or learner judgment. While the level of reporting is suitable for combining information the lack of a replication report is quite significant, especially given the limited details regarding corpus selection and sampling frame.

Answer to RQ2. Four groups of features that are seen as having pedagogical value reappear across different studies: single and multiword technical terms that are responsible for a large portion of the domain coverage; collocational and keyword frames typical of sport science academic texts that are discipline specific; terminology that is cross-linguistically unstable and mostly originates in proper names; and general lexis whose collocational, evaluative, and figurative use is strongly influenced by the sport domain in a way that is not dealt with in general instruction.

Answer to RQ3. The translation of corpus studies into classroom practice has only gotten to the stage of producing artifacts. The studies mentioned produce lists of words or collocations. However, the use of learner data is limited as only one learner study was found, no one talks about implementation of classroom activities or measurement of their results, and limitations to translation of results further are lack of learner data, unwillingness to share corpora, ignoring spoken and multimodal texts of professionals, and unresolved terminological instability across languages.

Table 6 <Research Gaps Identified in this Review and Corresponding Future Research Agenda>

Gap identified	Concrete future research recommendation	Priority
No learner-outcome evidence; only one study collected learner data	Controlled intervention studies comparing corpus-derived field-specific list instruction with conventional general-academic provision, measuring receptive and productive gains	High
Spoken and multimodal professional data almost absent	Compilation and open deposit of documented spoken and multimodal corpora of coaching, rehabilitation and strength-and-conditioning interaction across a wider range of sports	High
Corpora individually compiled, non-shared and under-reported	Adoption of a minimum reporting standard (sampling frame, period, participants, transcription conventions) and deposit of corpora for reuse	High
Terminological instability across languages	Coordinated construction and expert validation of bilingual and multilingual sport sciences glossaries, separating specialist-restricted from popularly-circulating layers	Medium
No learner corpus of sport sciences student writing	Compilation of learner corpora to establish target-performance gaps and inform list prioritisation	Medium
Unevaluated role of automated tools	Systematic evaluation of large language models as corpus-querying and materials-generation aids in this domain	Medium

Conclusions

After a detailed study of 403 Scopus papers that were searched using Prisma 2020 guidelines, 18 different works looking at the sport and physical activity genre English were compiled and synthesised by this work in a systematic review in an attempt to understand what corpus research has revealed up till now, how it did it and whether the corpus descriptions have had any effect on teaching. In relation to the first research question, the review mainly focuses on small personal corpora which mostly contain written data collected with four main methods, the authors did a proper job in reporting their work so that it could be synthesised though their methods are not replicates. Regarding the second question, four sets of linguistic features were identified which are instructionally relevant throughout these findings: frequently occurring technical words and phrases (mostly one and multi-word items), frames of co-occurrence which are specific to the sport genre, terminology which is not consistent cross-linguistically but a lot of it originates from proper names, and general vocabulary whose semantic features are heavily influenced by the sport context. Answering the third research question, corpus data has only to an extent influenced teacher's materials in the form of lexicon and word item lists but not in the form of tested and evaluated lessons or learning materials. This article mainly contributes by considering corpus analyses and its application in sport sciences English as an integrated area and presenting arguments supporting these points and that the research area has quite strong methodological background but is underdeveloped as far as pedagogy is concerned. As an alternative the findings call attention to the fact that field-specific lists and collocation databases should substitute vocabulary packages at the general academic level, published articles should be corpus-checked for terminology, and the provision of spoken taskbased instruction should be designed for applied sport settings. The limitations of the review are primarily its search through only one database, a restriction rule in determining eligible items by considering only the ones mentioning a teaching intent, and a lack of information about the affiliations of the contributing authors and absence of the authors having their study registered prospectively. Therefore, the focus of the next set of studies should preferably be on the use by teachers of the word lists they have ready, building of spoken corpora of coaching and clinical work along with making these available, and standardisation of professional terminology in different languages for the teaching of sports sciences.

Acknowledgments

The Authors would like to thank Faculty of Sport Sciences, Universitas Negeri Padang for providing excellent facilities, as well as for being very supportive of this research from the beginning. Authors also thank the team members who had generously shared their brilliant ideas, engaging discussion, and academic supports during the study.

References

- Abas, N., & Kaur, S. (2019). Theologising the discourse of representation through 'faith' and religion in football news. *3L: Language, Linguistics, Literature*, 25(3), 150–162. <https://doi.org/10.17576/3L-2019-2503-12>
- Ali, S., & Ali, M. (2023). Exploring lexico-grammatical patterns in Saudi press reportage: A multidimensional analysis. *Arab World English Journal*, 14(1), 502–512. <https://doi.org/10.24093/awej/vol14no1.32>
- Alkhamash, R. (2021). The social media framing of gender pay gap debate in American women's sport: A linguistic analysis of emotive language. *Training, Language and Culture*, 5(4), 22–33. <https://doi.org/10.22363/2521-442X-2021-5-4-22-35>
- Almujaiwel, S. (2018). Culture and intercultural in Saudi EFL textbooks: A corpus-based analysis. *Journal of Asia TEFL*, 15(2), 414–428. <https://doi.org/10.18823/asiatefl.2018.15.2.1.414>
- Alsharif, S. (2026). Cultural representation, soft power, and tourism futures in Vision 2030: Saudi Arabia's path toward the 2034 FIFA World Cup. *Frontiers in Sociology*, 11, Article 1761036. <https://doi.org/10.3389/fsoc.2026.1761036>
- Benson, S., & Coxhead, A. (2022). Technical single and multiword unit vocabulary in spoken rugby discourse. *English for Specific Purposes*, 66, 111–130. <https://doi.org/10.1016/j.esp.2022.02.001>
- Blanco-Gómez, M.-L., & Gómez-Ortiz, M.-J. (2024a). An analysis of linguistic resources for sports tourism media discourse: A case study of the Women's World Cup final in Sydney. *Revista Alicantina de Estudios Ingleses*, (41), 239–260. <https://doi.org/10.14198/raei.2024.41.11>
- Blanco-Gómez, M.-L., & Gómez-Ortiz, M.-J. (2024b). Sports tourism and social media storytelling: A multimodal perspective. *VISUAL Review*, 16(6), 165–177. <https://doi.org/10.62161/revvisual.v16.5358>
- Bobyreva, N. N. (2015). Eponymous units in the English sublanguage of sport: Form, meaning and functioning. *Journal of Sustainable Development*, 8(4), 309–315. <https://doi.org/10.5539/jsd.v8n4p309>
- Bobyreva, N. N. (2023). Investigating the role of English language tools used in sports communication using linguistic methods and its use in academic dictionaries. *Journal of Research in Applied Linguistics*, 14(3), 442–446. <https://doi.org/10.22055/rals.2023.19584>
- Braun, N., van der Lee, C., Gatti, L., Goudbeek, M., & Krahmer, E. (2021). MEmoFC: Introducing the Multilingual Emotional Football Corpus. *Language Resources and Evaluation*, 55(2), 389–430. <https://doi.org/10.1007/s10579-020-09508-2>
- Cappuzzo, B. (2020). Medical English translation into Italian: The case of Netter's work. *ESP Across Cultures*, 17, 31–48. [No DOI available in the source file]
- Chen, Z., Sun, W., & He, B. (2025). Exploring cross-cultural conceptual metaphors in Chinese and English table tennis news discourse. *Sport in Society. Advance online publication*. <https://doi.org/10.1080/17430437.2025.2599350>
- Clifton, J., Jacobs, G., Valeiras-Jurado, J., & Vandendaele, A. (2023). Interaction, asymmetries of knowledge and experience and the impact of theory on the in situ practice of coaching. *Journal of Applied Linguistics and Professional Practice*, 17(3), 252–273. <https://doi.org/10.1558/JALPP.22905>
- Dalla Pria, Y., & Bonnet, C. (2022). The beefcake and the beast: Professionalization, mediatization, and the representations of masculinity in French rugby. *Sociological Focus*, 55(3), 285–305. <https://doi.org/10.1080/00380237.2022.2073576>
- Elmes, D. (2018). Coach-player communications: An analysis of top-level coaching discourse at a short-term ice hockey camp. *Humanities and Social Sciences Reviews*, 6(2), 44–51. <https://doi.org/10.18510/hssr.2018.626>

- Fjellner, R. L., Varea, V., & Barker, D. (2024). How physical education teachers are positioned in models scholarship: A scoping review. *Physical Education and Sport Pedagogy*, 29(4), 329–345. <https://doi.org/10.1080/17408989.2022.2083098>
- Frederick, E., Pegoraro, A., & Sveinson, K. (2025). Not the Michigan State University of old? A critical discourse analysis of crisis communication. *International Journal of Sport Communication*, 18(2), 175–186. <https://doi.org/10.1123/ijsc.2024-0211>
- Gökmen, M. Y., Belibağlı, M. C., Karıncaoğlu, E., Nazlıcan, M. Ç., Türkmen, E., Yönel, M., & Uluöz, E. (2026). How accurately do large language models interpret sport safeguarding principles: An evaluation using the International Olympic Committee framework. *BMC Sports Science, Medicine and Rehabilitation*, 18(1), Article 106. <https://doi.org/10.1186/s13102-026-01559-2>
- Guslyakova, A. V., Guslyakova, N. I., Valeeva, N. G., & Vashunina, I. V. (2023). English-language media discourse in the era of digitalisation: Special mission and functional potential. *Studies in English Language and Education*, 10(1), 424–446. <https://doi.org/10.24815/siele.v10i1.25307>
- Hu, H., & Cai, C. (2026). The developmental–moral discourse of digital modernity: A corpus-assisted analysis of eSports in Chinese English-language news. *Cogent Social Sciences*, 12(1), Article 2645757. <https://doi.org/10.1080/23311886.2026.2645757>
- Ismail, H., Khoo, S., Idrus, M. M., Cheong, J. P. G., & Razman, R. (2022). Newspaper coverage of Paralympic athletes: A multimodal discourse analysis. *SAGE Open*, 12(2). <https://doi.org/10.1177/21582440221102434>
- Janić Mitić, A. A., & Veličković, M. V. (2025). Image schemas and the semantics of Serbian sports terminology: An innovative approach. *Russian Journal of Linguistics*, 29(3), 586–606. <https://doi.org/10.22363/2687-0088-43843>
- Kardoš Stojanović, A. Š., Milić, M. M., & Marković, L. M. (2024). Fitness terms in English and their adaptation into Serbian. *Filolog*, 15(30), 184–200. <https://doi.org/10.21618/fil2430184k>
- Kazymir, I., Yesypenko, N., Soloviova, O., Hnatkovska, O., & Saryk, O. (2026). Conceptual integration of contextual synonyms in contemporary English media discourse. *Journal of Intercultural Communication*, 26(1), 93–102. <https://doi.org/10.36923/jicc.v26i1.1356>
- Kongcharoen, P.-A. (2023). Investigation of vocabulary for ESP classrooms from academic journals in physical education and sports science. *LEARN Journal: Language Education and Acquisition Research Network*, 16(2), 311–332. [No DOI available in the source file]
- Labarta Postigo, M. (2023). ‘Life is baseball’ vs. ‘A vida é futebol’: The translation of sport-specific metaphors from English to European and Brazilian Portuguese. *Sendebare*, 34, 147–161. <https://doi.org/10.30827/sendebare.v34.26868>
- Leone, L. (2024). A corpus-based investigation of the word football in contemporary spoken English: Linguistic profile and cultural values in 1994 and 2014. *English Studies*, 105(2), 290–305. <https://doi.org/10.1080/0013838X.2023.2289306>
- Li, B., & Yang, B. (2024). A corpus-based study of nominalizations in the sports news of Shanghai Gazette (1919–1920). *Word*, 70(1), 44–64. <https://doi.org/10.1080/00437956.2023.2299071>
- Li, Z., Caballero-Juliá, D., Waquet, A., & Campillo, P. (2026). The emergence and trajectories of the glocalization concept (1990–2025). *Societies*, 16(2), Article 43. <https://doi.org/10.3390/soc16020043>
- Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *BMJ*, 339, b2700. <https://doi.org/10.1136/bmj.b2700>
- Loesnopchaimongkhon, R., Phrommopakorn, C., Chernchom, P., & Laosrirattanachai, P. (2025). A corpus-based analysis of lexical characteristics across English news categories for L2 pedagogical use. *Journal of Studies in the English Language*, 20(2), 36–65. <https://doi.org/10.64731/jse.v20i2.279647>
- López Ibáñez, F. J., Martínez, A. C., & Martínez Segura, M. J. (2023). Universal design for learning and ICT in the physical education area: Design and validation of an intervention proposal. *Publicaciones*, 53(3), 157–178. <https://doi.org/10.30827/publicaciones.v53i3.23867>
- Luján-García, C. (2021). Language contact in the media discourse: Anglicisms in the leisure section of the newspaper La Provincia. *Elia*, (20), 113–146. <https://doi.org/10.12795/ELIA.2020.I20.05>
- Martín-Hoz, L., Yanes-Luis, S., Huerta Cejudo, J., Gutiérrez-Reina, D., & Franco Álvarez, E. (2025). A comparative study of BERT-based models for teacher classification in physical education. *Electronics*, 14(19), Article 3849. <https://doi.org/10.3390/electronics14193849>
- Milić, M., & Kardoš, A. (2019). Adaptation of gymnastics terms from English into Serbian: Theoretical and practical aspects. *Babel*, 65(6), 852–874. <https://doi.org/10.1075/babel.00132.mil>

- Núñez Nogueroles, E. E. (2023). A pan-Hispanic approach to the usage of Anglicisms in the thematic area of sports. *Moderna Språk*, 117(1), 96–115. <https://doi.org/10.58221/mosp.v117i1.11338>
- Obradović, T. (2024). Change in the sports communication paradigm between 1998 and 2018: Game with another vs game against another. *Collegium Antropologicum*, 48(1), 17–26. <https://doi.org/10.5671/ca.48.1.3>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Parizoska, J., & Rajh, I. (2017). Idiom variation in business English textbooks: A corpus-based study. *ESP Today*, 5(1), 46–67. <https://doi.org/10.18485/esptoday.2017.5.1.3>
- Phoocharoensil, S. (2022). Primary, main, and major: Learning the synonyms through corpus data. *GEMA Online Journal of Language Studies*, 22(4), 76–89. <https://doi.org/10.17576/gema-2022-2204-05>
- Pîrlițeanu, B. (2022). Metaphorical language in football media: A semantic and contextual analysis. *Journal of Linguistic and Intercultural Education*, 15(4), 131–145. <https://doi.org/10.29302/jolie.2022.15.4.7>
- Polajnar, J. (2024). Greenwashing, pinkwashing, whitewashing ... the critical metadiscourse on X-washing: A corpus-based discourse analysis. *Linguistica*, 64(1), 41–77. <https://doi.org/10.4312/linguistica.64.1.41-77>
- Riccobono, P. S. (2020). Triangulating diamond talk: Identifying technical spoken vocabulary in English for baseball purposes. *ESP Today*, 8(1), 114–140. <https://doi.org/10.18485/esptoday.2020.8.1.6>
- Sandrelli, A. (2015). “And maybe you can translate also what I say”: Interpreters in football press conferences. *The Interpreters’ Newsletter*, 20, 87–105. [No DOI available in the source file]
- Thaksanan, P., & Kamsa-Ard, T. (2025). Lexical collocation analysis in a corpus of research articles in physical education and sport science. *New English Teacher*, 19(1), 51–70. [No DOI available in the source file]
- Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8, 45. <https://doi.org/10.1186/1471-2288-8-45>
- Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, 14(3), 207–222. <https://doi.org/10.1111/1467-8551.00375>
- Virág, Á. (2021). Infinite interpretations? A corpus-based study for the identification and interpretation of competing frames in parliament-representations in Hungary. *Central European Journal of Communication*, 14(1), 77–99. [https://doi.org/10.51480/1899-5101.14.1\(28\).5](https://doi.org/10.51480/1899-5101.14.1(28).5)
- Wilton, A. (2020). “We have a grandios saison gespielt” – English as a lingua franca in media sports interviews. *Journal of English as a Lingua Franca*, 9(1), 1–31. <https://doi.org/10.1515/jelf-2020-2032>
- Yang, M., & Yin, X. (2026). Olympism and the opening ceremony of the Beijing Olympic Games in the American and Chinese media discourse: A comparative study. *Sport in Society*, 29(3), 466–488. <https://doi.org/10.1080/17430437.2025.2462134>
- Yotimart, D. (2021). Academic vocabulary in sport tourism news: A corpus-based study. *Journal of Language and Linguistic Studies*, 17(3), 1527–1535. <https://doi.org/10.52462/jlls.110>
- Yu, W., & Shin, N. (2026). A corpus-based Spanish media reports analysis of Tokyo Olympic female soccer players. *European Journal for Sport and Society*, 23(2), 172–193. <https://doi.org/10.1080/16138171.2025.2580740>
- Zhang, L., & Caple, H. (2021). The newsworthiness of Li Na — A critical comparative analysis of Chinese and international news media. *Language and Communication*, 77, 70–80. <https://doi.org/10.1016/j.langcom.2021.01.002>
- Zhang, L., Zhang, Y., & Cao, R. (2022). “Can we stop cleaning the house and make some food, Mum?”: A critical investigation of gender representation in China’s English textbooks. *Linguistics and Education*, 69, Article 101058. <https://doi.org/10.1016/j.linged.2022.101058>