国際日本学研究科

修士学位請求論文要旨

〔論文題名〕

Corpus-based Investigation of Lexical and Syntactic Characteristics for Football ESP

国際日本学専攻 博士前期課程	
英語教育学研究領域	
学籍番号	4911245001
氏名	土屋　勇翔

Abstract

Thanks to globalization and advances in information and communication technology (ICT), people in Japan can now interact with other countries more easily than ever before. Information is readily accessible, even to young children, enabling them to quickly obtain a wide range of content that matches their interests. Furthermore, individuals are able to purchase products from overseas and make use of services provided by foreign companies. Consequently, access to global information has become increasingly effortless, leading to greater complexity and diversity in people's interests in Japan. In response to the increasing complexity and diversification of needs in contemporary society, demand has grown within the field of English education for learning English in a more specialized and efficient manner tailored to learners' areas of interest. Such domain-specific approaches to English education are generally referred to as English for Specific Purposes (ESP).

ESP is defined as language education that prioritizes learners' specific needs rather than the acquisition of general English proficiency (Terauchi, 2010; Hutchinson & Waters, 1987). ESP research emerged in the 1960s as a reaction against literature-centered English education and has developed through several stages, including register

analysis, discourse and rhetorical analysis, needs analysis, and genre analysis (Dudley-Evans & St. John, 1998). Early ESP research focused primarily on the grammatical and lexical features of scientific and technical English; however, the field later shifted toward learner-centered approaches and an emphasis on communicative competence. In the 1990s, genre analysis—exemplified by Swales's (1990) Create a Research Space (CARS) model—played a central role in advancing ESP research. Nevertheless, because ESP involves a high degree of specialization across domains, it lacks the well-established and systematically developed teaching materials that characterize English for General Purposes (EGP) and English for Academic Purposes (EAP). As a result, the development of reliable and empirically grounded ESP teaching materials remains a major challenge.

The use of corpora has been regarded as an important data-driven and empirical approach. A corpus is defined as "a collection of electronically stored and searchable texts" (Jones & Waller, 2015, p. 5). Leech et al. (2001) note that corpora can be stored and searched on computers and that computer programs can be used to analyze texts and list the frequency of word occurrences. For instance, by employing corpora, it is possible to identify lexical characteristics specific to particular genres. Ishikawa (2004), for example, compared a general English corpus with a judicial English corpus and

extracted both specialized terms that occur frequently in the legal domain and general words that carry domain-specific meanings. These findings demonstrate the effectiveness of constructing domain-specific vocabulary lists for the development of ESP teaching materials.

Among the many underexplored domains within ESP, the present study focuses on football within the sports domain. Interest in both domestic and international football has increased substantially, supported by expanded media coverage and the global popularity of European leagues, particularly the English Premier League. Because information related to overseas football leagues is disseminated predominantly in English, learners with English proficiency can access a wide range of authentic content, including match reports, transfer news, and analytical commentary. Despite the increasing demand, there has been relatively little ESP research targeting football fans, as most previous studies have focused on professionals such as coaches and players (Nishijo, 2017, 2021).To address this gap, this master's thesis aims to identify the lexical and syntactic characteristics of English football magazines using corpus-based methods and to provide an empirical foundation for the development of football-oriented ESP.

This study builds upon a pilot study conducted by Tsuchiya and Yamamoto

(2025), which attempted to extract vocabulary characteristic of football magazines. In that study, a football corpus was compared with the New JACET 8000 word list (JACET Vocabulary Research Group, 2016), resulting in the identification of 472 football-specific words and the extraction of 169 recommended words for Japanese readers of English football magazines. While Tsuchiya and Yamamoto (2025) successfully identified football-related vocabulary, the study had three notable limitations. First, it relied on a word list rather than a reference corpus for comparison. Second, it did not sufficiently examine the syntactic usage of the extracted vocabulary. Third, it focused primarily on frequency-based analysis without conducting deeper structural investigations. The present study aims to address these limitations by utilizing syntactically annotated corpora and incorporating analyses that focus on syntactic relations.

To address these issues, the use of a corpus annotated with syntactic relations was required. Such a corpus allows for both cross-corpus comparison and syntactic analysis grounded in grammatical structure. Accordingly, this study adopts a framework based on dependency grammar (Tesnière, 1959). Tesnière (1959) also stated that dependency grammar conceptualizes a sentence as a network of words connected through dependency relations and explicitly indicates which word depends on which

other word. In particular, this study employs the Universal Dependencies (UD) framework, which provides cross-linguistically consistent syntactic annotations and facilitates quantitative comparison across corpora.

Two corpora were analyzed in this study. The target corpus, referred to as Tsuchiya and Yamamoto Football corpus (TYF), consists of football-related articles collected from the British football magazine *FourFourTwo*, published between September 2022 and August 2023. After preprocessing procedures, including the removal of non-textual elements and lemmatization, TYF comprised 436,249 tokens and 15,128 types. As a reference corpus, this study employed the English Web Treebank (EWT), a well-balanced corpus of general English on web page texts annotated under the UD framework. To ensure feasibility and comparability, the training portion of EWT, consisting of approximately 217,000 tokens, was selected for analysis.

Using these two corpora, three research questions were formulated to address the limitations identified in Tsuchiya and Yamamoto (2025):

1. Are the keywords identified in Tsuchiya and Yamamoto (2025) also statistically significant in EWT?
2. How are these keywords used in TYF, particularly with respect to their syntactic dependency relations?

3. Are there syntactical differences in the whole text between TYF and EWT?

To answer the first research question, a keyword analysis was conducted by comparing TYF with EWT using AntConc version 4.3.1 (Anthony, 2024). The analysis extracted 238 keywords distinctive to TYF. When these keywords were compared with the 472 characteristic words identified in Tsuchiya and Yamamoto (2025), 73 overlapping words were identified. These overlapping items were defined as the core keywords of the present study. A part-of-speech analysis revealed that nouns constituted the largest category, followed by verbs and adjectives, suggesting that football discourse is noun-centered and strongly content-oriented. Furthermore, a qualitative analysis of these nouns demonstrated that many possess domain-specific meanings that are not readily apparent from general dictionaries. For example, *double* refers to winning both a league title and a cup competition within a single season, while spell often denotes the duration of a player's contract, particularly in the context of loan transfers. These findings indicate that football discourse relies heavily on semantic specialization, underscoring the need for ESP materials that focus on meaning in context rather than isolated word definitions.

To address the second research question, syntactic usage patterns of selected verbs—*stream*, *kick*, *beat*, *defeat*, and *replace*—were analyzed using dependency

relations. These verbs were selected based on their distinctive usage patterns and their relevance to football reporting. The analysis revealed that many of these keywords function predominantly within noun phrases rather than as central predicates of sentences. For instance, *stream* frequently appeared as a nominal element in expressions such as *live stream*, functioning as an object or modifier rather than as a verb. This result suggests that traditional vocabulary learning approaches that rely on one-to-one mappings between words and meanings are insufficient for covering the English knowledge required for football ESP. The limitations of rote vocabulary memorization have also been noted in previous research (Uchida, 2021), and the findings of the present study further highlight the need for ESP curricula that address vocabulary usage in syntactic and discourse contexts.

To reveal the third research question, two indices—Mean Dependency Distance (MDD) and Propositional Idea Density (PID)—were used to compare sentence-level complexity and readability between TYF and EWT. MDD, calculated as the average of all dependency distances within a sentence or a text, is assumed to reflect overall cognitive load during comprehension, such that higher MDD values correspond to increased processing difficulty. Covington (2008) defines PID as the number of propositions within a text sample divided by the total number of words. Contrary to the

initial expectation that TYF would contain more noun-heavy and readable sentences, the results revealed that TYF exhibited longer dependency distances and more complex sentence structures than EWT. Two main factors may account for this result. First, EWT includes not only news articles but also casual texts such as emails, which often contain fragments and non-sentential expressions. These characteristics likely resulted in shorter sentences and lower MDD values in EWT. Second, TYF consists of professionally written journalistic texts intended for reporting purposes, whereas EWT includes texts produced by general users for private and everyday communication. Differences in author roles and communicative purposes may therefore have influenced sentence structure and syntactic complexity.

In conclusion, this master's thesis demonstrates that corpus-based dependency analysis provides a powerful framework for understanding the lexical and syntactic characteristics of football ESP. The study highlights the central importance of vocabulary usage in football ESP instruction and clarifies key considerations for teaching material development in vocabulary-focused pedagogy. The findings and methodology of this study are not limited to the football domain but can be applied to ESP research and educational practice in other specialized fields as well. By overcoming the limitations of previous research and offering both quantitative and

qualitative insights, this study provides a foundation for ESP curriculum development grounded in authentic language use.

References

Anthony, L. (2024). AntConc (Version 4.3.1) [Computer software]. Retrieved july 29, 2024, https://www.laurenceanthony.net/software/antconc/

Covington, M. A. (2008). Idea density: A potentially informative characteristic of retrieved documents. IEEE Southeastcon, 2009, 201–203. https://doi.org/10.1109/SECON.2009.5174076

Dudley-Evans, T., & St. John, M. J. (1998). *Developments in English for Specific Purposes: A multi-disciplinary approach.* Cambridge University Press.

Hutchinson, T., and Waters, A. (1987). *English for specific purposes: A learning-centered approach*. Cambridge University Press.

Ishikawa, S. (2004). Development of an ESP vocabulary list for college law students: Based on the statistical comparison of word frequencies in US Judiciary English Corpus and Freiburg-Brown Corpus. *Journal of the School of Languages and Communication, Kobe University*, *1*, 13–27. https://doi.org/10.24546/00517984

JACET Vocabulary Research Group (2016). *The new JACET list of 8000 basic words*. Kirihara Shoten.

Jones, C., & Waller, D. (2015). *Corpus linguistics for grammar*. Routledge. https://cir.nii.ac.jp/crid/1130287815959979944

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge. https://doi.org/10.4324/9781315840161

Nishijo, M. (2017). Applying interpersonal lexico-grammatical analysis in teaching spoken discourse of English for football coaching. *Ritsumeikan Studies in Language and Culture*, *28*, 245–268. https://cir.nii.ac.jp/crid/1390572174776032640

Nishijo, M. (2021). English education program for Japanese soccer players and coaches seeking career opportunities overseas. *Japan Society for Educational Technology*, *44*, 469–482. https://doi.org/10.15077/jjet.44098

Swales, J. (1990). Genre Analysis: English in Academic and Research Settings. *Cambridge University Press*.

Terauchi, H. (2010). ESP history and definition. In H. Terauchi, H. Yamauchi, J. Noguchi, & S. Sasajima (Eds.), *ESP in the 21st century: ESP theory and application today* (pp. 3–16). Taishukan Shoten. https://ci.nii.ac.jp/ncid/BB04147180.amp

Tesnière, L. (1959). *Eléments de Syntaxe Structurale*. Klincksieck.

Tsuchiya, Y., & Yamamoto, H. (2025). Characteristic words in a football magazine: A corpus study for ESP material development. *JACET-KANTO Journal*, *12*, 84–103. https://doi.org/10.57365/jacetkanto.12.0_84

Ueda, A., & Kanda, M. (2019). Iryo eigo ESP (English for Specific Purposes) gakusyusya no shiten kara no needs analysis [Medical English ESP (English for Specific Purposes): Needs analysis from learners' perspective] *Chiba Prefectural University Health Science Bulletin*, *10*, 110. https://doi.org/10.24624/cpu.10.1_1_110