

Corpus-Based Computational Analysis of Academic English Writing Patterns

Ni Luh Putu Yesy Anggredi^{1*} and I Putu Eka Indrawan²

^{1*}Lecturer, Faculty of Social Sciences, PGRI Mahadewa University of Indonesia, Denpasar, Indonesia.
E-mail: yesianggredi@mahadewa.ac.id, Orcid: <https://orcid.org/0000-0003-3255-1977>

²Lecturer, Faculty of Science and Technology, PGRI Mahadewa University of Indonesia, Denpasar, Indonesia. E-mail: putueka@mahadewa.ac.id, Orcid: <https://orcid.org/0000-0003-2431-7603>

Abstract: Academic writing is characterized by certain linguistic rules which are different from those of informal speech; however usually acquired by students in an implicit and inconsistent way. In order to fulfill the requirement of empirical precision, this paper provides a corpus-based computational approach for the discovery of linguistic regularities in academic writing within various fields. Relying on register theory and genre studies, the study analyzed a specially created corpus of 200 research articles in 10 disciplines, which could be called "hard" and "soft" sciences, analyzing their argumentative segments through automated corpus-based analysis. There were five linguistic phenomena selected and measured in terms of 1,000 words: nominalizations, hedges, linkers, passives, and complex noun phrases. The quantitative data show unambiguous divergence between disciplines in terms of linguistic features: hard sciences contain more nominalizations (42 vs. 33), passives (28 vs. 16), and complex noun phrases (35 vs. 24) in order to facilitate concise, objective reporting of empirical data. In turn, soft sciences focus on the use of hedges (27 vs. 18) and linking adverbials (19 vs. 12) for cautious, well-signalized interpretation. The largest differences in absolute values have been shown by passives (+12) and complex noun phrases (+11). Linking adverbials (-7) form the most stable base in both areas. These results illustrate the specific epistemological and rhetorical needs of each academic community. The research comes to the conclusion that corpus analysis allows providing the empirical basis for substitution of general writing tips with EAP instructions for specific communities.

Key Words: corpus linguistics, academic writing, computational text analysis, english for academic purposes, genre analysis, linguistic complexity, disciplinary discourse

(Received: 24 September 2024; Revised: 11 October 2024; Accepted: 29 November 2024; Published: 31 December 2024)

Introduction

Definition of Corpus-based Computational Analysis

Corpus-based computational analysis involves the examination of large amounts of authentic texts that have been stored electronically, referred to as "corpora," by means of computational tools for the purpose of detecting repetitive patterns of lexical, grammatical, and rhetorical features, which otherwise would not be possible without computational assistance. Corpus analysis is not based on specific examples or the intuitions of the scholar regarding the usage of language, but rather on the accumulated evidence from many hundreds, if not thousands, of texts, which allows scholars to develop empirical conclusions about how a certain variety of language, for instance, academic English, functions. The studies presented below showcase the usefulness of this method through corpus-based analysis of syntax complexity and error patterns in learner writing, showing how computational approaches may help discover language usage

patterns that would not normally be evident through non-computational analysis techniques (Nang & Nikolov, 2023).

Importance of Studying Academic English Writing Patterns

Academic writing patterns can be seen as especially significant when considering English language learning, as academic English represents a very special register whose rules deviate considerably from those applied in everyday speech and informal written discourse. Examining the linguistic indicators of stance displayed by novice and experienced academic writers, it is possible to conclude that it is not the knowledge of how to write grammatically correct English but the ability to apply the conventions of academic writing that separates academic writing from other types of writing (Aull & Lancaster, 2014). This is especially true for language learners who have difficulties mastering the conventions of academic discourse, which are usually considered to be the most challenging aspect of academic literacy. It is also stated that corpus methods have proven to be particularly productive for English for Academic Purposes studies make such conventions visible and provide teachers with teaching material based on the patterns detected by means of corpus analysis into practical pedagogical resources ranging from vocabulary lists to genre-based writing guidance (Flowerdew, 2015).

Based on the above-mentioned pedagogical justification, the present study is motivated by the following three main research goals: to discover repeated patterns of language which are typical for academic English prose by means of computational analysis based on the corpus, to study how the frequency and distribution of certain linguistic features differ depending on the academic field, and to discuss the possible pedagogical implications of the discovered tendencies. In their corpus-based research about literature review writing and citations usage, demonstrate how such research goal could be beneficial and practically useful, showing that the evidence from corpus can be easily used to create teaching materials which improve the students' writing (Li et al., 2023). As for the corpus-based study of noun phrase complexity in the first- and second-language academic writing, it provides another example of methodology which can be used in the present research, showing how a particular language phenomenon can be traced across a specially constructed corpus in order to derive generalizable conclusions (Lan et al., 2022).

RQ1: What are the most frequent lexical, grammatical, and rhetorical features characteristic of published academic English writing within the compiled corpus?

RQ2: How do the frequency and distribution of nominalizations, hedging devices, linking adverbials, passive constructions, and complex noun phrases vary systematically between "hard" and "soft" science disciplines?

RQ3: How can these corpus-derived empirical findings inform discipline-specific pedagogy in English for Academic Purposes (EAP) instruction?

Here is a summary of the paper's unique contributions:

- Provides a unified, comparative analysis of five distinct linguistic features (nominalizations, hedging, linking adverbials, passives, and complex noun phrases) within a single study, whereas prior research typically examined these features or disciplines in isolation.
- Establishes concrete data showing that hard sciences heavily favor compact, impersonal structures (passives, nominalizations, complex noun phrases), while soft sciences rely on communicative and interpretive structures (hedges, linking adverbials).

- Challenges one-size-fits-all, prescriptive writing advice by providing an empirical foundation for discipline-specific English for Academic Purposes (EAP) instruction aligned with actual community norms.

The rest of this paper is structured in separate sections. After the introduction, Section 2 provides an overview of previous work using corpora to study academic discourse, with Section 3 providing the methodology for building and analyzing the corpus, Section 4 the results of corpus linguistics, and Section 5 the discussion of the quantitative results, respectively. Section 6 concludes with contributions and future research directions.

Literature Review

Previous Research on Corpus Linguistics in Academic Writing

There exists an extensive amount of literature using corpus linguistics in examining academic writing, creating an empirical account of what characterizes academic English writing. The study confirms that linguistic complexity measures calculated based on corpora are able to predict the proficiency of second-language writers, and complexity has thus become an important factor that corpus linguistics has been exploring to describe the quality of academic writing (Kim, 2014). A cross-disciplinary and cross-linguistic corpus study of adverbial linking constructions illustrates that the use of connectives by writers is determined not only by their discipline but also by their native language background, thus demonstrating the sensitivity of corpus methods to subtle distinctions in academic style (Gao, 2016). Further research has followed this direction by investigating the use of informal language features in academic writing by first- and second-language university students, confirming that these features can be found even in supposedly formal academic contexts, which is hard to do without the scale offered by corpus analysis (Lee et al., 2019). Using collocational analysis of noun premodification based on corpora, illustrate once more the power of corpus methods in analyzing very particular linguistic structures, such as noun phrases, that recur systematically across academic texts (Ang et al., 2017).

Studies on English Writing Patterns in Academic Texts

Other studies have concentrated on patterns of writing variation in different disciplines and genres. Studies on linguistic complexity among research papers taken from 'hard' and 'soft' disciplines reveal significant differences in grammatical complexity in professional academic writing by discipline field, which is highly relevant to the current study's cross-disciplinary comparison (Pérez-Guerra & Smirnova, 2023). Patterns of reader engagement in English academic writing by using corpus analysis reveal that writers' use of strategies for engaging readers such as hedges and boosters is not static but varies according to genre and disciplinary conventions (Jin, 2015). More generally, the methodology of corpus analysis in English for Academic Purposes reveals that such disciplinary comparisons fall within the scope of corpus linguistics, which seeks to study genre- and discipline-specific variation rather than treating academic writing as one unified register (Flowerdew, 2014).

Use of Computational Analysis in Studying Language Patterns

Another direction within the literature has been concerned with exploring the direct application of computational analysis of corpora for pedagogical purposes. A study has found that corpus-based instruction in writing business English resulted in significant improvement in learners' writing competence, thus providing pedagogical evidence of the practical significance of corpus-informed instruction (Tsai, 2021). Research has provided an account of the use of a corpus-based approach to the creation of online materials for teaching the writing of research articles, showing how recurrent patterns detected by means

of computational analysis of corpora could be turned into instructional tools (Chang & Kuo, 2011). The analysis of macrostructural patterns of research articles through corpus techniques shows that computational analysis is able to reveal recurrent patterns at the level of whole texts, not just recurrent lexical patterns, widening even further the range of corpus-based studies of academic writing (Gong & Barlow, 2022). Taken altogether, these studies provide a solid empirical base and a clear pedagogical justification for corpus-based investigation of patterns of academic English writing performed in the current study.

Previous studies have proven that academic writing involves a unique language complexity, stance, and genre-based conventions rather than generic grammar rules. The frequency of occurrence of the linguistic features such as hedges, linking adverbials and noun phrases is known to differ in different disciplines. In this case, this research will be based on the previous findings in order to offer an integrated comparison of the features in five disciplines-sensitive instruction.

Methodology

Corpus Evidence and Academic Discourse

This study is based on the corpus-based approach to studying language. This approach is built on a series of assumptions that separate it from other more conventional approaches, which are more reliant on intuition in their study of languages. One of the central assumptions behind the corpus-based approach is that the frequency of appearance in an authentic corpus can be taken as evidence of the norm. Grammar acceptability cannot be regarded as a characteristic of a sentence because language is viewed as probabilistic and variable, and there are some lexical and grammatical features that are much more common in certain contexts or disciplines. Academic writing is understood as a particular register with its own pattern of common linguistic features.

The current study is also informed by the theory of register, which assumes systematic relationships between variation in language use and the situation in which language is produced, taking into account the purpose of communication, the writer-reader relationship, and mode of production. Academic writing, viewed from the perspective of register theory, is determined by communicative needs of the scholarly discourse in general, which involve presenting claims with an adequate degree of cautiousness, establishing the credibility of the writer, and situating claims in the context of prior research. All the linguistic elements analyzed in the current study (hedging, linking adverbials, nominalization, passives, complex noun phrases) are register-specific responses to communicative needs.

Another theoretical aspect related to genre is also noteworthy. Genre is defined here as a category of communicative acts having the same communicative function and similar structure. It has been generally recognized that academic fields form separate discourse communities having genre conventions determined by their epistemology or specific assumptions on what constitutes valid knowledge and how it is proved. Those fields, which are typically defined as hard sciences because of their emphasis on empirical research and presentation of quantitative data, may be supposed to have linguistic conventions different from soft sciences, where the focus is placed on interpretation and qualified statements. This theoretical framework determines the interpretive paradigm for the cross-disciplinary analysis carried out in the current study.

Lastly, the research carried out within the field of academic writing is usually based on pedagogic grounds and linked with the notion of academic literacy viewed as a specific reading and writing skill

necessary for successful involvement in higher education. Since the conventions of academic writing are not always clear and are unevenly modeled in the teaching process, corpus data can serve as an instrument that would make the conventions clear and teachable, transferring descriptive results into pedagogic ones. Pedagogic grounds are the base of the implications for language teaching discussed further in this paper.

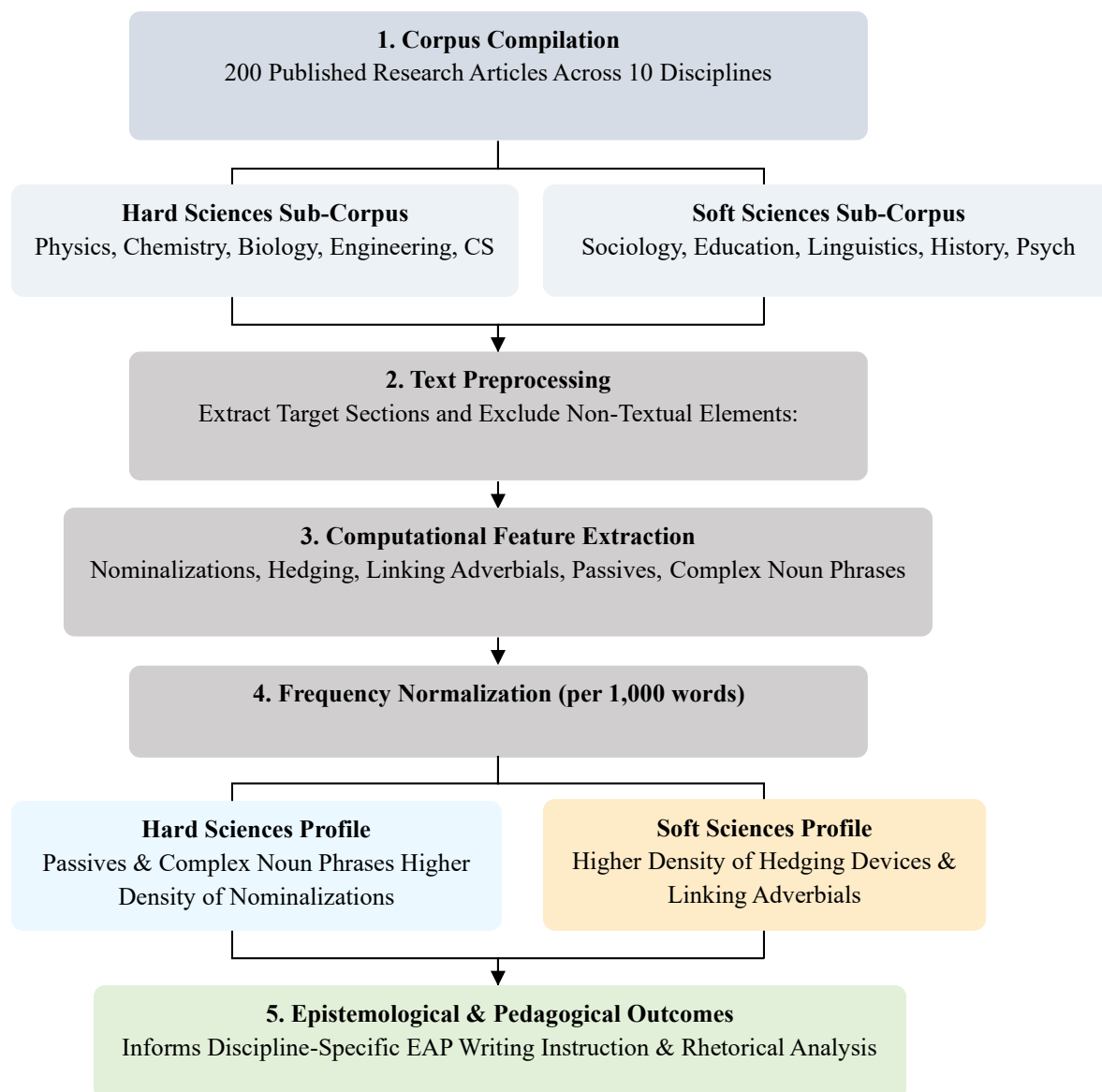


Figure 1: Analytical Framework of the Corpus-Based Computational

Figure 1 highlights the systematic process of the study, beginning from the creation of a corpus of 200 articles in both hard and soft sciences until the pedagogic results. Text preprocessing helped in identifying the arguments, thus allowing the computer software to pick up the target linguistic elements. The frequency analysis shows clear characteristics of each discipline: hard sciences are characterized by tight and objective structures, while soft sciences value qualified and signposted arguments. This empirical pipeline directly links theoretical register theory to discipline-sensitive EAP instructional design.

Corpus Used for Analysis

For this study, the sample corpus included two hundred research articles that had been published in an academic context. The sample was equally composed of two groups of five academic disciplines, which

have traditionally been classified under two main groups of disciplines: five hard science disciplines Physics, Chemistry, Biology, Engineering, Computer Science, and five soft sciences disciplines Sociology, Education, Applied Linguistics, History, Psychology. The articles were chosen from peer-reviewed journals in each field, but only the introduction, literature review, and discussion parts of the articles were considered. This decision was made because these parts of the articles contained the features that could be studied from the point of view of argumentation and interpretation, while the method and result parts were not analyzed.

Computational Tools and Techniques Employed

The computational means used for the current research included a typical corpus analysis program that could generate a list of frequencies, concordance lines, and collocations based on the corpus created. With the use of the mentioned program, the corpus was examined systematically for the occurrence of five specific linguistic structures chosen as the most frequently discussed within the literature overviewed above: nominalizations marked by their typical derivational suffixes; hedging elements that were marked with the help of a pre-selected list of epistemic modal verbs, adverbs, and adjectives; linking adverbials marked by their pre-selected list of connective expressions; passive structures marked by the typical combination of the auxiliary verb and the participle; and complex noun phrases defined as noun phrases with three or more modifiers.

Process of Data Collection and Analysis

Data collection and analysis involved three steps. In the first step, the two hundred articles were digitized, stripped of all non-textual material including tables, diagrams, and references, and sorted into two disciplinary sub-corpora as mentioned above. In the second step, the corpus-analysis tool was applied to each sub-corpus to look for the five linguistic characteristics, obtaining raw counts which were normalized to one per one thousand words for comparability purposes due to minor differences in total length among the sub-corpora. The third step involved comparison of the normalized counts in terms of the linguistic characteristics between the hard science and soft science sub-corpora and interpretation of the findings in relation to the theory introduced earlier in this paper.

Results

Identification of Common Writing Patterns in Academic English

The findings show that the writing in academic English is characterized by recurring linguistic features which differ systematically based on the field as opposed to having a fixed style. Through the study of a specially designed corpus, it has been found that the hard sciences depend on structural devices that enable concise, impersonal, and objective communication with the use of nominalization, passive voice, and complex noun phrases. On the other hand, the soft sciences use communicative and interpretive devices such as the use of hedging devices to make cautious claims and adverbials to indicate logical reasoning. These distinct patterns reflect the underlying epistemological and rhetorical demands of their respective academic communities.

Analysis of the Frequency and Distribution of Specific Linguistic Features

Quantitative study has made it evident that there is a disciplinary difference in the frequency of use of each feature in every 1,000 words among the sub-corpora. The hard sciences have a higher frequency of use of nominalizations (42 against 33), passive constructions (28 against 16) and complex noun phrases (35 against 24), enabling them to report information in a concise manner. On the other hand, soft sciences

rely more on hedging strategies (27 against 18) and linking adverbials (19 against 12) in forming an argument. The highest variations in absolute terms are found in passive structures and complex noun phrases, whereas linking adverbials exhibit the most stable base across different disciplines.

Table 1: Mean Frequency of Selected Linguistic Features Across Disciplinary Sub-corpora
(per 1,000 words)

Linguistic Feature	Hard Sciences (freq./1,000 words)	Soft Sciences (freq./1,000 words)	Difference
Nominalization Density	42	33	+9
Hedging Devices	18	27	-9
Linking Adverbials	12	19	-7
Passive Constructions	28	16	+12
Complex Noun Phrases	35	24	+11

Table 1 shows the mean normalized frequency of the five linguistic features among both hard science and soft science sub-corpora. There is a clear distinction made between the two disciplinary categories regarding all five features.

Comparison of Writing Patterns Across Different Academic Disciplines

Patterns in writing in sub-corpora provide evidence of the existence of various stylistic characteristics which serve particular scientific fields. Writing in hard sciences uses nominalization, passive constructions, and complicated nouns in order to ensure that the writing is concise and objective and focuses on factual results and processes. On the other hand, writing in soft sciences uses hedges and adverbial linkers in order to facilitate interpretative argumentation, engage readers, and establish logical connections. The different frequencies clearly show that there is no such thing as monolithic academic writing, but it is dynamically shaped by disciplinary conventions.

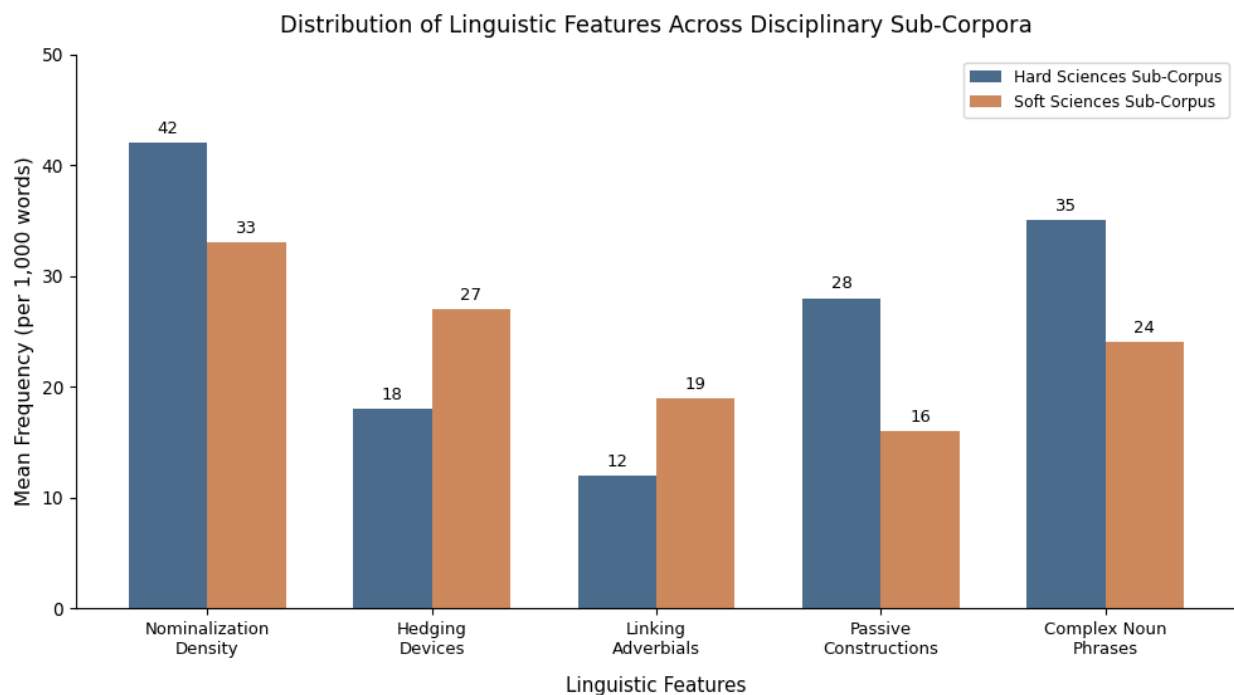


Figure 2: Distribution of Linguistic Features Across Disciplinary Sub-corpora

Figure 2 provides a visual representation of the same comparison, showing the different profiles of the two disciplinary sub-corpora with regard to the five linguistic features under investigation.

The hard-science sub-corpus showed substantially higher frequencies of nominalization, passives, and complex nouns compared with the soft-science sub-corpus, as expected from a writing style geared towards concise and objective presentation of scientific procedures and results. The soft-science sub-corpus, on the other hand, had higher frequencies of hedging expressions and linking adverbials, which is characteristic of a writing style geared towards qualifying statements and making explicit the logical/argumentative relations between the ideas presented. Of all the linguistic features investigated, passives and complex nouns had the greatest absolute difference between the two sub-corpora, while linking adverbials had the least, indicating that, although connective devices are employed for different purposes are relatively constant across disciplinary boundaries, even as their specific rhetorical purpose may vary.

Discussion

The findings have practical relevance for the field of English for Academic Purposes teaching. Given that the findings of this study indicate that academic writing conventions vary systematically between disciplines, a one-size-fits-all approach to writing instruction that does not differentiate academic writing by discipline runs the risk of giving students guidance that is relevant only to some extent to the particular genres of writing that will need to write. The findings call for a more differentiated approach to writing instruction that would teach those students who prepare to write hard science texts how valuable nominalization, passive voice, and complex noun phrases are for producing a concise and impersonal style characteristic of empirical discourse, while students who will be working on soft sciences texts should be taught the importance of hedging and linking adverbs for cautious argumentation.

The results also have implications for the link between writing practices and academic performance in general. Since the linguistic variables studied here seem to serve as indicators of disciplinary competence, suitable to the conventions of a particular discipline, the capacity to employ them properly might well be seen as correlated with the effectiveness with which an academic paper by a student or other beginner writer would be accepted within the relevant academic discipline. The student writer, for example, who uses hedging constructions suitable to an article in the hard sciences, or who does not use the passive voice constructions typical of empirical studies, will likely be seen as not meeting disciplinary standards, even when the content of the writing is good, highlighting the importance of making such norms explicit through corpus-based teaching.

These findings suggest a number of avenues for future research. In particular, the following might conduct a similar analysis of other academic text types beyond research articles, such as conference abstracts and grant proposals, all of which may have their own specific linguistic features. Another line of future research might involve examining how the linguistic features uncovered in the current study evolve during a student's academic development process and how the writing of discipline experts differs from that of novices. Finally, it may prove fruitful to explore the ways in which corpus-based research results could be incorporated in teaching materials and feedback on student writing, building upon the applications for pedagogy that have already been shown in the reviewed literature.

Conclusion

The study involved the computational analysis of 200 corpus-based research articles to identify five key features of language. The empirical data shows that there is considerable difference between different

disciplines in terms of each feature per 1,000 words: nominalization (hard sciences - 42, soft sciences - 33), passive voice (hard sciences - 28, soft sciences - 16), complex noun phrases (hard sciences - 35, soft sciences - 24), hedging (soft sciences - 27, hard sciences - 18), and linking adverbials (soft sciences - 19, hard sciences - 12). The passive voice (+12, 75%) and complex noun phrases (+11, 45.8%) demonstrate the highest absolute difference. At the same time, the linking adverbials (-7) demonstrate the most consistent baseline difference. Such statistical characteristics emphasize that epistemological and rhetorical considerations influence the development of academic writing, as opposed to the succinct and objective style of hard sciences versus the careful and interpretative style of soft sciences. This research brings a comprehensive comparative approach to the literature and advances the research on linguistic features from previous research that only focused on one feature or one discipline. The significance of this research is that it gives an empirical basis for a shift from generic advice for students' writing to more specific instructions for writing in English for Academic Purposes based on real community standards. Further research could apply the computational approach to other academic writing genres, like grant proposals, dissertations, and conference abstracts, alongside the longitudinal study of how writers develop expertise in disciplinary discourse practices over time.

References

- Ang, L. H., Tan, K. H., & He, M. (2017). A corpus-based collocational analysis of noun premodification types in academic writing. *3L, Language, Linguistics, Literature*, 23(1). <https://doi.org/10.17576/3L-2017-2301-09>
- Aull, L. L., & Lancaster, Z. (2014). Linguistic markers of stance in early and advanced academic writing: A corpus-based comparison. *Written communication*, 31(2), 151-183. <https://doi.org/10.1177/0741088314527055>
- Chang, C. F., & Kuo, C. H. (2011). A corpus-based approach to online materials development for writing research articles. *English for Specific Purposes*, 30(3), 222-234. <https://doi.org/10.1016/j.esp.2011.04.001>
- Flowerdew, L. (2014). Corpus-based analyses in EAP. In *Academic discourse* (pp. 105-124). Routledge. <https://doi.org/10.4324/9781315838069>
- Flowerdew, L. (2015). Corpus-based research and pedagogy in EAP: From lexis to genre. *Language Teaching*, 48(1), 99-116. <https://doi.org/10.1017/S0261444813000037>
- Gao, X. (2016). A cross-disciplinary corpus-based study on English and Chinese native speakers' use of linking adverbials in academic writing. *Journal of English for Academic Purposes*, 24, 14-28. <https://doi.org/10.1016/j.jeap.2016.08.002>
- Gong, H., & Barlow, M. (2022). A corpus-based analysis of research article macrostructure patterns. *Journal of English for Academic Purposes*, 58, 101138. <https://doi.org/10.1016/j.jeap.2022.101138>
- Jin, H. (2015). A Corpus-Based Study on Engagement in English Academic Writing. *English Teaching*, 70(2). <https://doi.org/10.15858/engtea.70.2.201506.27>
- Kim, J. Y. (2014). Predicting L2 Writing Proficiency Using Linguistic Complexity Measures: A Corpus-Based Study. *English teaching*, 69(4). <https://doi.org/10.15858/engtea.69.4.201412.27>
- Lan, G., Zhang, Q., Lucas, K., Sun, Y., & Gao, J. (2022). A corpus-based investigation on noun phrase complexity in L1 and L2 English writing. *English for Specific Purposes*, 67, 4-17. <https://doi.org/10.1016/j.esp.2022.02.002>

- Lee, J. J., Bychkovska, T., & Maxwell, J. D. (2019). Breaking the rules? A corpus-based comparison of informal features in L1 and L2 undergraduate student writing. *System*, 80, 143-153. <https://doi.org/10.1016/j.system.2018.11.010>
- Li, Z., Makarova, V., & Wang, Z. (2023). Developing literature review writing and citation practices through an online writing tutorial series: Corpus-based evidence. *Frontiers in Communication*, 8, 1035394. <https://doi.org/10.3389/fcomm.2023.1035394>
- Nang, K. T., & Nikolov, M. (2023). Investigating syntactic complexity and language-related error patterns in EFL students' writing: corpus-based and epistemic network analyses. *Language Learning in Higher Education*, 13(1), 127-151. <https://doi.org/10.1515/cercles-2023-2009>
- Pérez-Guerra, J., & Smirnova, E. A. (2023). How complex is professional academic writing? A corpus-based analysis of research articles in 'hard' and 'soft' disciplines. *Vigo International Journal of Applied Linguistics*, (20), 149-183. <https://doi.org/10.35869/vial.v0i20.4357>
- Tsai, Y. R. (2021). Exploring the effects of corpus-based business English writing instruction on EFL learners' writing proficiency and perception. *Journal of Computing in Higher Education*, 33(2), 475-498. <https://doi.org/10.1007/s12528-021-09272-4>