Pakistan Journal of Law, Analysis and Wisdom
Volume No. 3, Issue No. 4, April 2024
e-ISSN: 2959-0825, p-ISSN: 2959-0817
http://pjlaw.com.pk

Comparing the Two Lexical Density Measures: The Case of Ten Highest Impact Factor Journals in Applied Linguistics

Shazia Aziz
PhD Scholar, Department of English, Fatima Jinnah Women University, Rawalpindi, Pakistan
19-2144-21-001@eng.fjwu.edu.pk

Dr. Fakhira Riaz
Assistant Professor, Department of English, Fatima Jinnah Women University, Rawalpindi, Pakistan
drfakhirariaz@fjwu.edu.pk

Abstract

This article studied the lexical density of abstracts published in the 10 highest impact factor journals in Applied Linguistics during 3 years i.e., between 2019 to 2021.The purpose was threefold: to compare the lexical densities of the abstracts studied; to compare the findings of the two lexical density measures chosen for the study; and to study the overall use of salient lexical items that reflect the research trends in these journals during this timeline. A corpus of 1172 article abstracts was constructed and analysed using the natural language processing tool Antconc 3.5.9. It was found that the two results found by the measures have no significant correlation and hence, are very different. Moreover, the most common themes based on the word frequencies were found to be language, learning, English, writing, children, and bilingualism. The findings have significant implications for researchers and academic writers in Applied linguistics and related fields, and those involved in materials development for academic writing modules, especially in ESL or EFL context.

Keywords: *Applied Linguistics, Academic Writing, Lexical Density, Frequency, Grammatical Items, Lexical Items.*

1. Introduction

Lexical density is considered to be a dependable measure of productive lexical knowledge and linguistic proficiency in writing, especially academic writing (Bulté & Housen, 2012; Kim, 2014; Lu, 2012). Lexical density—that is, a heavy reliance on content words especially nouns—is seen to be a sign of sophisticated informational prose and condensed academic writing (Biber, 2006; Biber & Grey, 2016; Pietilä, 2015). It can be expressed as the ratio of lexical/content words to ranking clauses or as the ratio of lexical words to all words/tokens. Put differently, lexical density is the degree to which lexical elements are crammed into syntactic structures.

There have been two well-known and popular measures of lexical density among scholars to investigate lexical density employing advanced natural language processing (NLP) tools. They were proposed by Ure (1971) and Halliday (1985). In this study, we investigate how well each of these measures determines the lexical density of condensed academic writing, i.e., the abstracts of research articles published in the top ten impact factor journals related to the field of Applied Linguistics during 2019-2021; and whether the comparative lexical densities of different journals found through both the methods is similar in comparison and follows a similar pattern through the

selected data. Moreover, we explore the frequency of words appearing in the data and examine the most frequently used vocabulary to highlight the themes that are in vogue in these journals during the period studied.

Being the first contact point between the author and the readers, abstracts are 'self-contained microcosms' (Bunton, 1998) of articles wherein the author summarises a whole article in a limited space i.e., between 150–300 words generally. Moreover, impact factor journals present in the Journal Citation Report (JCR) that is issued annually on the basis of citations that the papers published in these journals receive, have special importance among academia as some universities base their promotion and employment decisions on impact factor publications; and it is also one of the indicators of universities' rankings.

2. Theoretical Underpinnings

2.1. Defining Lexical density

The ratio of lexical items, or "content words," to the total number of tokens (N_{lex} / N) in a text or corpus is known as "lexical density," and it was first used by Ure (1971). It is described as "complexity that is typical of written language" by Halliday (1985, p. 62). Lexical density, according to Read (2000), is a feature of written language in which texts contain a more concentrated amount of lexical elements in the form of concepts and information. Nouns and nominalizations have a major role in formal written registers (Biber, 2006; Biber and Gray, 2016), which contributes to the texts' exceptional density.

The definition of a "lexical item" and the distinction between lexical and grammatical elements might affect the lexical density of a given composition (Al-Wahy, 2017). Any selected approach can calculate the compositions' lexical density when it is consistently used over a range of compositions. Typically, oral forms of human communication in the English language have lexical densities around 40%, but written forms characteristically have lexical densities above 40%, according to Ure (1971). According to Stubbs' analysis of historical texts, fictional literature often has a lexical density between 40% and 54%, while non-fiction has a density between 40% and 65% (Crawford, 2007; Warschauer, & Kern (Eds.) 2000; Stubbs 1996). In the words of Halliday (1989, p. 80), "the lexical density of written language is likely to be of the order of twice as high as that for speech." Furthermore, beyond early childhood, written human communication has a tendency of having greater lexical density than spoken human communication (Halliday, 1985; Johansson, 2009). As for different genres, lexical density is typically maximised in instructions, court orders, literature and news read from screen prompts (Castello, 2008; Stubbs, 1986).

Lexical density is measured using two main formulas:

1.The ratio of number of content words (lexical units or lexemes) to the total number of words/functional words (grammatical units) as proposed by Ure's (1971). This is the most common way of measuring lexical density. It can be described as:

$$\text{Lexical Density} = N_{lex} / N*100$$

2. A less commonly explored formula of measuring lexical density is the one proposed by Halliday (1985) whereby the lexical density is found by calculating the ratio of content words to ranking clauses. It is informed by Systemic Functional Linguistics. Clauses are groups of words that form a complete thought. This method of measuring lexical density takes into account the complexity of the sentences in a text, as well as the number of content words.

$$\text{Lexical Density} = N_{lex} / \text{Number of ranking clauses}$$

This formula gives the value of lexical density as a simple ratio, as opposed to the Ures values like a percentage. (Castello, 2008; Halliday, 1985).

There is no one "most authentic" way of measuring lexical density. The best method to use depends on the specific purpose of the measurement. For convenience's sake, most existing studies (e.g., Castello, 2008, p. 53; Yates 1996; and Neumann 2014, p. 76), however, use Ure's formula alone despite admitting the greater reliability of Hyland's formula.

This study, nevertheless, fills this gap by attempting to determine the selected texts' lexical densities using both formulae and compares the results along with analysing the highest frequency.

2.3. Research measuring the lexical density of abstracts

The abstract section of any piece of academic writing, be it a dissertation, thesis or article, is a unique form of academic writing or sub-genre that is characterized by a lexically dense skeleton and a brief summary of the complete text which comprises key findings, significance, contributions of the study, results and implications (Bhatia, 1993; Bitchener, 2009; Gillaerts & Van de Velde, 2010; Pho, 2008), according to a number of notable pieces of research.

Since abstracts are usually the first section of academic texts and are meant to persuade audiences that the remaining work is interesting having significant and reliable results (Bitchener, 2009, p. 11). Abstracts have a unique kind of "persuasive function" (Bunton, 1998; Gillaerts & Van de Velde, 2010). Moreover, the abstracts' must have a good quality because they are used in the indexing and abstracting processes of publishers like Clarivate Analytics and JCR (Journal Citation Report), and as a result, they appear in the Pakistani Higher Education Commission's HJRS (HEC Journal Recognition System).

Owing to this special role, researchers have been interested in abstracts in general for a long time, and many studies have looked at the different structural, linguistic, and stylistic features of the subgenre. For example, Yoneoka and Ota (2017) discovered that compared to low-quality abstracts, high-quality abstracts in publications had longer words, shorter sentences, a higher percentage of noun phrases, and a lower percentage of verb phrases. Previous research (e.g., Egbert & Plonsky, 2015; Pho, 2008) looked for a number of linguistic characteristics in research publication and conference paper abstracts, such as voice, tense, nouns that complement clauses, modal and reporting verbs, stance words, first-person pronouns, and errors.

In short, a lot of research has shown how crucial lexical competency is for academic writing, especially for EFL and ESL students writing theses and dissertations (Engber, 1995; Nasseri & Thompson, 2021, Pietilä, 2015). On the other hand, not enough study has been done on impact factor journals' abstract sections. Therefore, there is a research gap in the measurement of lexical density in impact factor journals, which this work aims to close. Furthermore, there are few studies that compare the lexical density discovered using the two approaches.

3. Research Questions:

A total of 1172 article abstracts was collected and 10 sub corpuses were constructed and analysed to answers the following questions:

Q1. To what extent are the lexical densities of the abstracts published in the top ten impact factor journals different?

Q2. How are the findings of lexical density measures proposed by Ure (1971) and Halliday (1985) different for the abstracts under study?

Q3. What do the most frequent lexical items used in the abstracts reflect about the research trends in these issues?

4. Methodology

Since it is a study employing corpus-based method, the first step was to construct corpora of the required data. The section that follows describes the details of the process of corpus construction:

4.1. Corpus Construction

For the compilation of the corpus, JCR-2021 was consulted and analysed in depth. All the journals that fall under the linguistics or language criterion were highlighted. Then the top ten were selected for corpus construction. When collecting abstracts, it was realized that the journal at No. 2 that is, Language Teaching published only agendas and reflection papers, so it was excluded and the journal next in number was added to make the total number of journals studied 10. All the abstracts of research articles published in the issues during 2019-2021 were downloaded. Abstracts were open access so they were freely downloadable from the journals' websites. While adding abstracts, the following inclusion and exclusion criteria were used.

The impact factors of the journals ranged between 5.741 to 3.532. The average number of words per abstract in the data ranged between 146 to190 words.

Ant Conc 3.5.9 was then used to measure the frequencies and to separate the lexical items from grammatical items. Following Al-Wahy (2017), multi-word proper names (e.g., Hong Kong) were counted as single words. Similarly, hyphens were used to join years and numbers constituting multiple words e.g., six hundred and eight' in order to consider them as single words. The number of lexical items thus found in each corpus are also shown in Table 1.

Whichever approach is employed, measuring lexical density is dependent upon the differentiation between function and content words, or grammatical and lexical items (Halliday, 1989), closed-set and open-set items and (Cruse, 2011). Content words bear the ideational weight of any text and represent meanings that can be grasped mostly independently of the context. Function words carry inadequate meaning outside of their context of usage. Instead of playing a semantic role, they play a grammatical one. Function words connect content words to one another, while content words explain what a text is about (Stubbs 2002, p. 39).

The issue with content-function word dichotomy, according to Al-Wahy (2017) is that it implies that function words have just grammatical meaning, which is not realistic since there are borderline cases in language (e.g., auxiliary verbs in English) that are neither entirely functional nor entirely lexical. Hence, we included auxiliary verbs as lexical items for the purpose of this study.

The content and function words distinction is in fact a matter of degree. As Halliday (1989) propounds, "there is a continuum from lexis into grammar," and hence, there are "intermediate cases" between the two. In reality, there is a "continuous scale of lexicality/grammaticality" (Cruse, 2011).

In order to resolve this issue, Corver and van Riemsdijk (2001, p. 10) propose an additional category for words that combine lexical and function word characteristics: "semi-lexical" terms. Al-Wahy (2017) proposes overcoming the issue by either analysing each borderline case individually to determine its category; and deciding whether to include it in content or function word category; or assigning an intermediate value to such cases. In both situations, Al-Wahy (2017) suggests, results will be approximate. Moreover, we think that from practical aspects, the former would be very time-consuming in cases of very large corpora. Halliday (1989), however, recommends being consistent in whatever one chooses to do. For the purpose of this study, we have regarded borderline cases as function words. In case of words following under multiple categories e.g., 'second' (a number and a verb), 'other' (adjective and a pronoun), we have counted them in the category based on the category of the majority cases as occurring in our data.

Halliday's method, though time consuming due to manual demarcation of clauses involved in it, is a good option to minimize the effects of problems of this kind. Moreover, a lexical density metric according to the ratio of lexical to function words could not apply easily in languages which have the 'function' elements naturally blended with the 'content' lexeme so as to form one inflected word, according to Halliday (1996, p. 104).

According to Halliday (1996), meanings are not articulated in a vacuum but rather in linguistic frames—usually the clauses—that structure how they are presented in the text. Words are packed in larger grammatical units, clauses and sentences, and not within other words, according to Halliday (1989, p. 66) and hence, a ranking clause's lexical density is found out by how many lexical words it contains.

When applied to complete texts, lexical density can be computed by dividing the aggregate number of ranking clauses by the aggregate number of lexical words. In Halliday's formula, lexical density is expressed as a number whose value is proportionate to the text's informative density rather than as a percentage as suggested by Ure's formula.

According to SFG (Halliday, 1994; Halliday & Matthiessen, 2014), clauses fall into two main categories: ranking clauses, which are counted for assessing lexical density, and embedding, or rank shifted clauses, which are not. Whereas embedded clauses, which are found inside longer clauses, have grammatical purposes, ranking clauses do not. Independent clauses, paratactic clauses, and hypotactic clauses are the three types of ranking clauses. To put it briefly, ranking clauses are all non-embedded clauses (Halliday, 1992).

Since ranking clauses' demarcation involved manual tagging, a representative corpus was selected from the data for this part of the study. For this purpose, 1 abstract from each year's data was selected from each journal so as to have 3 abstracts per journal, and 30 in total. Additionally, manual analysis can yield more accurate results and is more suitable than automated counting (Al-Wahy, 2017).

5. Findings and discussions

As described above, Antconc software 3.5.9. was used to find out the word frequencies, lexical items and grammatical items in the data while ranking clauses were marked manually. The findings have been shown in Table 1:

Table 1: Lexical Densities of abstracts using Ure and Halliday's Formulae

S. No.	Journal full name	Lexical Density using Ure's Formula	Lexical Density using Hyland's Formula
1	Applied Linguistics	64%	14.3
2	Computer Assisted Language Learning	64%	14.3
3	Modern Language Journal	65%	17
4	Language Learning	67%	16.6
5	Language Learning & Technology	18%	14.5
6	International Journal of Bilingual Education and Bilingualism	67%	15.5
7	Studies in Second Language Acquisition	66%	11.8
8	Language Teaching Research	64%	16.8
9	Journal of Second Language Writing	65%	15
10	*Bilingualism: Language and Cognition*	68%	12.4

As can be seen in Table 1, the LD value for all the 10 journals' abstracts keeps fluctuating between 64% to 68% as per Ure's formula and hence have little variation except for the 5th journal i.e., Language Learning and Technology which is exceptionally low i.e., 18%. Compared to this, the values according to Halliday's formula keep fluctuating between 11.8 to 17 with a value of 14.5 for Language Learning and Technology.

As for the corresponding values in both the columns, we can see that the values for the first 2 journals viz., Applied Linguistics and Computer Assisted language Learning, remain constant in both the columns i.e., 64% and 14.3 respectively. Whereas, following this, the values don't follow a similar trend in both the columns e.g., for the 3rd journal i.e., Modern Language Journal, the values increase to 65% and 17 respectively but 64 and 16.6% for the 4th journal i.e., Language Learning. From 6th to 7th journal viz., International Journal of Bilingual Education and Bilingualism and Studies in Second Language Acquisition, the value decreased only by 1% i.e., from 67% to 66% according to Ure's method, but corresponding values declined from 15.5 to 11.8 according to Haliday's method. For 8th journal i.e., Language Teaching Research, however, it decreased further by a 1% according to Ure's formula; but it saw a high rise from 11.8 to 16.8 according to Halliday's formula. Similarly, the value according to Ure's formula saw a rise for 9th journal i.e., Journal of Second Language Writing to 65% but a decline to 15 according to Halliday's method. Similarly, the LD value for the 10th journal i.e., Bilingualism: Language and Cognition increased to an all-time high value of 68% according to Ure's method while it decreased to a very low value of 12.4 according to Halliday's method.

As Table 3 shows, in most of the cases, the lexical density of the data sets is greater than 60% which is a very high value and in line with Ure's (1970) contention that the lexical densities of written discourse are greater than 40%. However, the lexical density of abstracts published in Language Learning & Technology is extraordinarily low i.e., 18% according to Ure's Formula; and vice versa for 10th journal which means that both the formulas do not yield comparable results and don't follow similar trends throughout the data sets.

If we compare the values attained using Ure's (1971) formula with those attained by Halliday's (1985) formula, we see that the trend with which the values increase, decrease or fluctuate in the first column does not consistently match with the trends in column 2.

To compare the findings achieved from both methods and their trends statistically, Correlation coefficient was measured. The Pearson correlation coefficient measures the direction and strength of the linear relationship between the variables. Table 2 shows the result of the statistical procedure:

Table 2: Correlations

		Lexical Density using Ure's Formula	Lexical Density using Hyland's Formula
Lexical Density using Ure's Formula	Pearson Correlation	1	.036
	Sig. (2-tailed)		.920
	N	10	10
Lexical Density using Hyland's Formula	Pearson Correlation	.036	1
	Sig. (2-tailed)	.920	
	N	10	10

The correlation between Lexical Density using Ure's (1971) Formula and Lexical Density using Hyland's (1985) Formula is 0.036 which is very low. The p-value associated with this correlation is 0.920 signifying that the correlation observed is not statistically significant. There is a very weak positive linear relationship between the lexical density calculated using Ure's (1971) formula and the lexical density calculated using Hyland's (1985) formula. However, due to the very low correlation coefficient and the high p-value, this relationship is not statistically significant. Therefore, it might be concluded that there is no meaningful association between the two measures. This finding raises concerns regarding both the methods i.e., the ambiguity of lexicality and grammaticality of some items can affect the results, especially while using corpus methodology and studying very large corpora. Secondly, ranking clauses and their demarcation is time taking and hence does not allow for studying large corpora though it yields more dependable results being based on Systemic Functional Linguistics. Many items fall under more than one category and hence corpus methodology alone is not enough if one has to get authentic results and hence, manual double checking of those items is needed to see the context in order to decide under which category one has to count them.

Table 3 below shows the most frequent lexical items found through frequency count using AntConc 3.5.9 in the abstracts published in the selected journals (along with their frequency in parenthesis) during 2019-2021 which reflects the research trends in these journals.

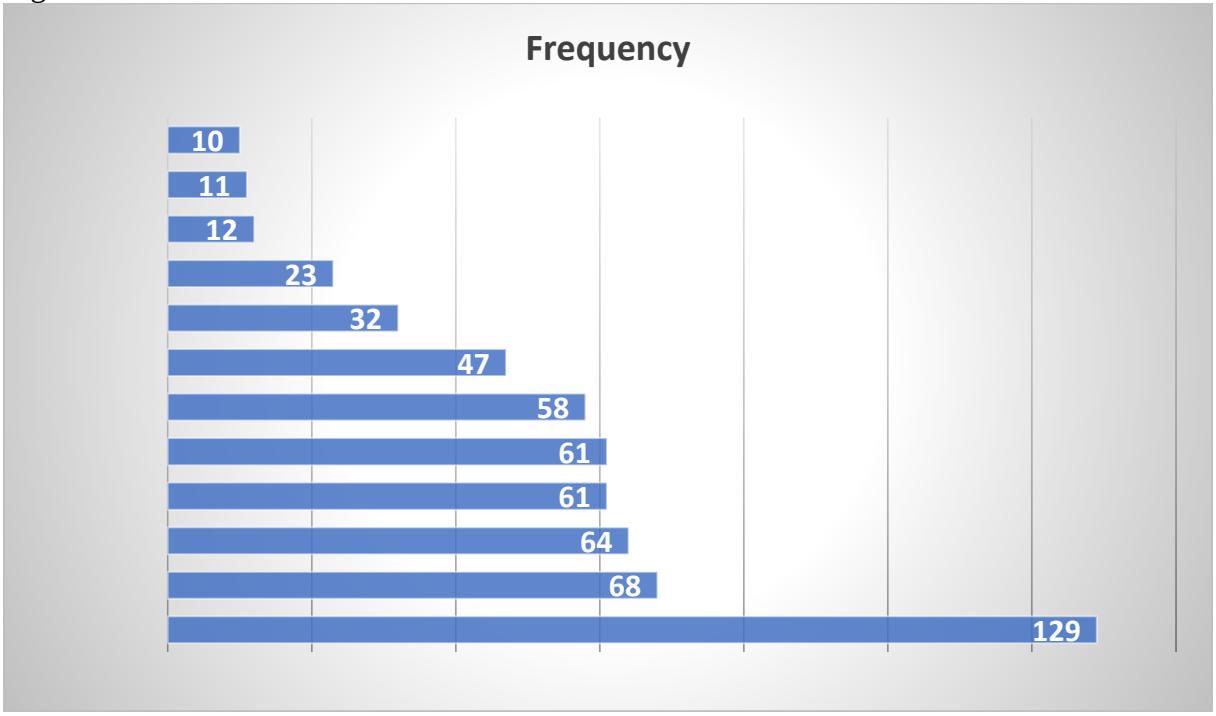
Table 3: Overall Frequencies (/1000 words) of most frequently occurring items in the data

Words	**Frequency**
Language	129
Study	68
L	64
English	61
Learning	61
Learners	58
Students	47
Writing	32
Results	23
Bilingual	12
Research	11
Children	10

As Table 3 shows, the most frequently used words are 'Language', 'study' and 'L' with a frequency of 129, 68 and 64 respectively which is understood as Applied Linguistics as a field addresses issues related to language. 'L' has been counted separately as it was used in both L1 and L2 contexts throughout the data sets. It means the issues deal with L1 and L2 related researches. 'English' standing at number 4 with a frequency of 61 per 1000 words depicts that English is the most commonly being studied language in the issues. It is owing to the fact that English is the dominant language of research, lingua franca, and the international language being spoken and studied most commonly worldwide. The words next three in frequency i.e., 'Learning', 'learners' and 'students' with frequencies of 61, 58 and 47 show that the most common concern of these issues has been teaching and learning English and learners' related matters. 'Writing' and 'results' coming next in frequencies i.e., 32 and 23 per 1000 words depict that writing is the main skill being studied in the context of the data being studied. 'Bilingual', 'research' and 'children' coming next with frequencies of 12, 11 and 10 per 1000 words respectively show that the next important

concern of these issues is bilingualism and children. To sum up, we can say the most common topics among these issues were language; language learning; L1; L2; English; learners/students; writing; bilinguals and bilingualism; and children.

Figure 1



6. Conclusion

This investigation found that different journals publish abstracts with different lexical densities which are generally very high which is in keeping with Ure's (1971) generalization about non-fiction having a density between 40% and 65%. The only exception in our data is abstracts from the journal Language Learning & Technology whose value is extraordinarily low for a written text according to both the formulae i.e., 18% using Ure's formula and 14.5 according to Halliday's (1985).

In this analysis, correlation tests confirmed that the LD measures using both the measures does not yield significantly correlated results. This was perhaps because some abstracts using lesser number of ranking clauses use high number of lexical items and some use very high number of ranking clauses and comparatively lesser number of lexical items. The density using Ure's method has little variation as Ure's formula is based on word level only and abstracts generally follow similar patterns in word usage based on nominalizations and compact forms due to the limited length of the sub-genre.

This discrepancy further implies that either the two approaches assess distinct types of lexical density, or they do not measure the same thing. Because Ure's (1971) measure solely relies on the lexical differentiation between content and function terms, it is single-layered; and Halliday's (1985) approach is more dynamic and multi-layered in comparison as suggested by Al-Wahy (2017). Together with the grammatical and lexical levels, it also makes a deeper distinction between rank-shifted and ranking clauses. By demonstrating the way meanings are given in discrete clauses and the way they contribute to the lexical density of the whole text, it considers the function of smaller units of grammar.

At least in the context of abstracts, this study has demystified Al-Wahy (2017) assumption that the two methods have the tendency of yielding similar results if compared with English texts. The contrasting trends and the low correlation coefficient results suggest that both the methods can't necessarily be considered to measure the same phenomena. Moreover, this study proposes considering the fact that in many lexically dense texts, the same lexical items are used repeatedly or frequently so the diversity should also be considered while measuring lexical density. In this respect, Halliday's (1985) method being based on functional aspect of language, can be said to be more reliable though it needs more manual work and care for accuracy for findings and hence, does not allow for using a large data for analysis through corpus tools. That is the reason that most of the studies use Ure's (1971) method because it can easily be done using corpus tools despite acknowledging the superiority of Halliday's (1985) method in terms of reliability and applicability to various languages. We can say the difference between Ure (1971) and Halliday's (1985) method is similar to the difference between form and function of language. While the former analyses form, the latter analyses function.

However, we propose that some computational experts might think of devising ways and tools to demarcate ranking clauses and enable future researchers to measure lexical density of big data using Halliday's (1985) method for replicating, verifying and validating the results of the scarce existing studies on the topic ensuring greater reliability and avoiding researcher's bias since it is based on a more functional approach to language.

References

Al-Wahy, A. S. (2017). Towards a methodology for measuring lexical density in Arabic. *International Journal of Arabic Linguistics*, *3*(1), 1-33.

Bhatia, V. K. (1993). Analysing genre: Language use in professional settings. London: Longman.

Biber, D. (2006). University language: A corpus-based study of spoken and written registers. Amsterdam: John Benjamins.

Biber, D., & Gray, B. (2016). Grammatical complexity in academic English: Linguistic change in writing. Cambridge: Cambridge University Press.

Bitchener, J. (2009). *Writing an applied linguistics thesis or dissertation: A guide to presenting empirical research*. Bloomsbury Publishing.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, *32*, 21.

Bunton, D. (1998). Linguistic and textual problems in PhD and M.Phil. theses: An analysis of genre moves and metatext. Unpublished PhD Thesis. University of Hong Kong.

Castello, E. (2008). *Text complexity and reading comprehension tests* (Vol. 85). Peter Lang. pp. 49–51.

Corver, N.& van Riemsdijk, H. (2001). Semi-lexical categories. In N. Corver & H. van Riemsdijk (Eds.), Semi-lexical categories: The function of content words and the content of function words (pp. 1–22). Berlin: Mouton.

Crawford, B. B. (2007). The Language of Business Studies Lectures. A Corpus-Assisted Analysis.

Cruse, A. (2011). Meaning in language: An introduction to semantics and pragmatics (3rd ed.). Oxford: Oxford University Press.

Egbert, J., & Plonsky, L. (2015). Success in the abstract: Exploring linguistic and stylistic predictors of conference abstract ratings. *Corpora*, *10*(3), 291-313.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. Journal of Second Language Writing, 4, 139–155.

Gillaerts, P., & Van de Velde, F. (2010). Interactional metadiscourse in research article abstracts. *Journal of English for Academic purposes*, *9*(2), 128-139.

Halliday, M. A. (1992). Language as system and language as instance: The corpus as a theoretical construct. In *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4—8 August 1991* (pp. 61-77).

Halliday, M. A. K. & Matthiessen, C.M.I.M. (2014). Halliday's introduction to functional grammar (4thed.). London: Routledge.

Halliday, M. A. K. (1985). Spoken and written language Victoria. *Australia: Deakin University*.. pp. 61–75 (Chapter 5), 76–91 (Chapter 6).

Halliday, M. A. K. (1989). Spoken and written language (2nd ed.). Oxford: Oxford University Press.

Halliday, M. A. K. (1994). Introduction to functional grammar (2nd ed.). London: Routledge.

Halliday, M. A. K. (1996). Literacy and linguistics: A functional perspective. In J. J. Webster (Series Ed.), The collected works of M. A. K. Halliday. Vol. 9, Language and education (pp. 97–129). London: Continuum.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. Modern Language Journal, 96(2), 190–208.

Nasseri, M., & Thompson, P. (2021). Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assessing Writing*, *47*, 100511.

Neumann, S. (2014). Contrastive register variation: A quantitative approach to the comparison of English and German. Berlin: Mouton.

Pho, P. D. (2008). Research article abstracts in applied linguistics and educational technology: A study of linguistic realizations of rhetorical structure and authorial stance. *Discourse studies*, *10*(2), 231-250.

Pietilä, P. (2015). Lexical Diversity in L2 Academic Writing: a Look at MA Thesis Conclusions. *Lexical issues in L2 writing*, 105.

Read, J. (2000). Assessing vocabulary. Cambridge: Cambridge University Press.

Stubbs, M. (1986). In Malcolm Coulthard (ed.) Lexical density: A technique and some findings. *Talking about text*, 27-48. University of Birmingham: English Language Research.

Stubbs, M. (1996). Text and Corpus Analysis: *Computer Assisted Studies of Language and Culture.* Wiley. pp. 71–73.

Stubbs, M. (2002). Words and phrases: Corpus studies of lexical semantics. Oxford: Blackwell.

Ure, J. (1971). Lexical density and register differentiation. In G. Perren and J.L.M. Trim (eds), *Applications of Linguistics*, London: Cambridge University Press. 443–452.

Warschauer, M., & Kern, R. (Eds.). (2000). *Network-based language teaching: Concepts and practice*. Cambridge university press. pp. 107–108.

Yates, S. J. (1996). Oral and written linguistic aspects of computer conferencing: A corpus-based study. In S. C. Herring (Ed.), Computer-mediated communication: Linguistic, social and cross-cultural perspectives (pp. 29–46). Amsterdam: John Benjamins.

Yoneoka, D., & Ota, E. (2017). Evaluating association between linguistic characteristics of abstracts and risk of bias: Case of Japanese randomized controlled trials. PLOSONE, 12(3), Article e0173526.