Assessing Writing 47 (2021) 100511

Contents lists available at ScienceDirect

Assessing Writing

journal homepage: www.elsevier.com/locate/asw

Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences

Maryam Nasseri *, Paul Thompson

Department of English Language and Linguistics, University of Birmingham, Birmingham, B15 2TT, UK

A R T I C L E I N F O

Keywords:
Writing assessment
Academic writing
Lexical proficiency
Lexical density
Lexical diversity
Dissertation abstracts

A B S T R A C T

This study investigated lexical density and diversity differences in English as L1 vs L2 academic writing of EFL, ESL, and English L1 postgraduate students to compare their lexical proficiency in EFL vs. English L1 academic settings. A corpus of 210 dissertation abstracts was analysed using three natural language processing tools [LCA, TAALED, and Coh-Metrix] where the effects of text length and topic were controlled. In doing so, we examined the relationship between 15 lexical indices and the construct-distinctiveness of lexical density and diversity. The measure-testing process also assesses the effectiveness of each measure in a pair/group of closely-related measures (in terms of the quantification methods) in capturing lexical diversity differences of these texts. This is to obtain a small number of unique measures that capture lexical diversity as an indicator of lexical proficiency and to assist future writing researchers in the measure-selection process in the face of a multitude of available measures. The findings have important implications for writing assessment and research on lexical indicators of writing proficiency, materials development in EFL academic settings especially for thesis/dissertation writing modules, and a possible contribution of ESL academic immersion programmes in approximating English L1 and L2 proficiency.

1. Introduction

Lexical density and diversity as two dimensions of lexical complexity and aspects of productive lexical knowledge remain as two of the most reliable indicators of lexical and linguistic proficiency and development of language users in the first and second language as well as writing and academic studies (see e.g., Bulté & Housen, 2012; Lu, 2012). Lexical density is the proportion of lexical/content words to all words/tokens; lexical density, especially a dense use of nouns, is regarded as an indicator of condensed academic writing and advanced informational prose (e.g., in Biber, 2006; Biber & Gray, 2016; Pietilä, 2015) and as a strong predictor of academic writing proficiency (e.g., Kim, 2014). Lexical diversity is the use of a range of diverse words (also known as unique word types) to convey meaning and is regarded as an indicator and predictor of lexical proficiency and development (Gonzalez, 2013; Mazgutova & Kormos, 2015; Yoon, 2017). Lexical density and diversity, although interrelated, can be differentiated in that lexical density seeks to

Abbreviations: CEFR, (Common European Framework of Reference); EAP, English for Academic Purposes; EFL, English as a Foreign Language; ESL, English as Second Language; English, L1 English as the first Language; L1, first language; L2, second language (also subsequent languages); MA, master's; NLP, Natural Language Processing; NS, Native Speakers of English; SLA, Second Language Acquisition; LCA, Lexical Complexity Analyzer; TAALED, Tool for Automatic Analysis of Lexical Diversity; TTR, type-token ratio.

* Corresponding author.
E-mail address: mxn309@bham.ac.uk (M. Nasseri).

https://doi.org/10.1016/j.asw.2020.100511
Received 2 June 2020; Received in revised form 9 November 2020; Accepted 13 November 2020
Available online 11 January 2021

present how densely lexical items are packed into syntactic structures, while lexical diversity is representative of non-repetitious and/or different lexical and grammatical items used in language production, e.g., in a text. Correspondingly, a learner can produce statements with higher lexical density and lower lexical diversity and vice versa (Johansson, 2008).

Traditionally, human raters would qualitatively analyse texts regarding aspects of lexical proficiency (e.g., manual calculation of measures or assessing the overall quality of writing in learner essays or other corpora). With the increasing number of English learners and the availability of large numbers of texts/corpora, however, this is no longer a viable option for assessing large learner corpora. Doró and Pietilä (2015) emphasise the importance of automated essays scoring systems for such assessments; they list studies that show high correlations between the results of these systems and those of human raters. A multitude of measures/indices have been, therefore, proposed to objectively quantify the two constructs of lexical density and diversity using advanced natural language processing (NLP) tools as alternatives to the qualitative analysis of texts by human raters.

Despite the ease with which writing researchers can now parse and examine texts, several additional challenges are acknowledged. A noticeable one is the presence of a myriad of linguistic measures/indices proposed by quantitative linguists and writing assessment researchers to gauge these linguistic constructs, especially lexical diversity. These many measures are formulated as a remedy to the text-length dependency of the ratio of types (unique non-repetitious words) to tokens (all words) as will be discussed in detail in the following sections. Some of these proposed measures are different adaptations of one quantification method for overcoming this problem but to our knowledge, previous works have not investigated a large set of such closely-related indices. In this paper, therefore, we examine the effectiveness of each index in pairs/groups of similarly-calculated indices in capturing lexical diversity differences in condensed forms of academic writing such as the abstracts of theses/dissertations. Obtaining a small set of (unique) measures as such contributes to the measure-selection process of writing assessment researchers in the face of a multitude of similarly-computed measures.

Swales (2004, p. 99) comments on the master's dissertations as "the most sustained and complex piece of academic writing (in any language) they will undertake". MA dissertations in EFL academic contexts are usually the first important/serious scientific piece of writing for most postgraduate students. In the following sections, we argue that abstracts are self-contained microcosms of these dissertations in which a student has to describe and summarise a whole research/dissertation in a limited number of words, usually between 200–300 words. In this article we will also assess postgraduate academic writing proficiency differences in a specialised corpus to investigate how students with different English language backgrounds in English L1 vs. L2 academic contexts use this limited space regarding the diversification of lexis and how densely the lexis is produced in their texts. The two mentioned objectives of this study have important implications for writing assessment research especially regarding advanced English academic writing and writing proficiency differences.

2. Literature review

2.1. *Lexical density and diversity in corpus, SLA, and academic studies*

There is a long-standing line of research that examines the relationship between lexical complexity constructs (and measures) and linguistic proficiency and development as is indicated by the works in the following sections.

Though the scope, objectives, and designs of these studies differ, positive correlations are often reported between higher values of these lexical constructs (and the measures that objectively quantify them) and the learners' (or the sample corpus) holistically assessed quality of written and spoken discourses, vocabulary knowledge, and linguistic proficiency (Crossley & McNamara, 2010; Engber, 1995; Jarvis, 2002; Laufer & Nation, 1995; Lu, 2012; Zareva, Schwanenflugel, & Nikolova, 2005) which prompted many programmers to take these indices as reliable quality indicators of test performance and as parameters in automated text analysis (e.g., Graesser, McNamara, Louwerse, & Cai, 2004; Lu, 2012).

These constructs have been addressed in first and second language development. Lexical density and diversity, for instance, were used as factors to analyse a learner corpus to track the lexical development of 10-year-olds through university (Johansson, 2008). Similarly, Durán, Malvern, Richards, and Chipere (2004) set out to track the lexical diversity development of thirty-two English L1 children across ten different ages and they found a significant developmental trend. They also showed that lexical diversity can be used as an indicator of ESL/EFL development of learners aged 18−30.

A number of proficiency-related SLA (Second Language Acquisition) studies have also been carried out using various measures/indices that represent and quantify these two constructs. For example, Treffers-Daller, Parslow, and Williams. (2016) employed several measures of lexical diversity to discriminate between essays of 179 adult ESL learners in different CEFR (Common European Framework of Reference) proficiency levels. The learners in their cross-sectional comparative study wrote timed essays of 200–300 words. School students in Olinghouse and Wilson's (2013) cross-sectional study demonstrated higher lexical diversity in narrative texts than in informative and persuasive genres of texts. Other studies reported that proficient L2 and L1 writers use a wider range of vocabulary (e.g., Crossley & McNamara, 2010; Jarvis, 2002; Laufer & Nation, 1995; Wolfe-Quintero, Inagaki, & Kim, 1998 among others).

Finally, a number of studies on academic writing addressed the effect of one or several of these lexical measures on writing quality, linguistic proficiency, and development. For instance, a corpus of college essays written by Korean students was analysed by Kim's (2014) cross-sectional study to distinguish between proficiency levels based on CEFR. Kim used the measures that are also investigated in Lu's (2012) study of the quality of transcribed oral narratives, and reported that lexical density and one lexical diversity measure were strong predictors of academic writing proficiency. However, Pietilä (2015) who analysed the conclusion sections of MA dissertations by English L1 vs. L2 students did not find any differences between these groups regarding lexical density and diversity. In a comparative academic writing study, Gregori-Signes and Clavel-Arroitia (2015) analysed lexical density and lexical diversity in the written corpus of three groups of university students to detect the shortest time span for the development of lexical knowledge in undergraduate writing. Similarly, lexical density and diversity scores of university major students' (Hungarian EFL undergraduates) essays were compared to their productive vocabulary tests (Doró, 2008). Two lexical diversity measures along with the vocabulary size (measured by word frequency means) were also used to analyse 104 ESL and 68 NS university students' academic writing (Gonzalez, 2013) where lexical diversity showed a greater effect on writing scores than vocabulary size, and NS's lexical proficiency was proved to be significantly higher than the ESL group. In Mazgutova and Kormos's (2015) study of an academic writing immersion programme for ESL students in the UK, lower proficiency ESL students significantly improved in all measures of lexical diversity, and both proficiency levels improved in verb variation over the course of one semester. Bulté and Housen (2014), however, did not find any significant results concerning lexical diversity as an indicator of writing proficiency of ESL students over the course of one semester in an academic language programme (an intensive EAP course).

There are, however, two main issues regarding the studies in this area. One is the variability in research design e.g., sample size, learner vs. academic corpus, text genres and registers, learner English language backgrounds, and the number and type of lexical measures included in the studies. This makes the comparison of the results difficult and the generalisability of the findings problematic. Even so, studies which systematically analysed academic writing genres and sub-genres using a large set of these lexical measures were few and far between (e.g., Pietilä, 2015). The second issue is a dearth of large-scale measure-testing investigations using many indices that can inform language practitioners in either SLA and academic contexts with the measure-selection process especially considering similarly-calculated measures. The multiplicity of measures (especially lexical diversity measures) that have been offered in the literature makes it difficult for writing researchers and practitioners to navigate through and choose a few unique indices for writing assessment especially regarding different academic writing genres and registers, and the effect of English language backgrounds and academic contexts.

2.2. Lexical density and diversity: definitions, measures, and quantification methods

The term 'lexical density' is believed to have been introduced by Ure (1971) who described it as the ratio of the number of lexical items (also called 'content words') to the total number of words/tokens (N_{lex} / N). Lexical density is regarded by Halliday (1985, p. 62) as "the kind of complexity that is typical of written language". Read (2000), likewise, proposes that lexical density is a characteristic of written language where texts represent a more concentrated proportion of lexical items in the form of information and ideas. Among text types, the density of formal texts like academic writing is shown to be higher than that of informal texts (e.g., in Biber, 2006). As an instance, the process of nominalisation reduces the grammatical words and contributes to higher lexical density. Consequently, texts with characteristics such as a dense use of nouns are more informative and can be regarded as a characteristic of the academic genre and advanced writing (Biber & Gray, 2010; Biber, 2006, 2016). Biber and Gray (2010, p. 2) also emphasise that "these styles are efficient for expert readers, who can quickly extract large amounts of information from relatively short, condensed texts". Lexical density as an indicator of the information content of a text is particularly relevant to abstracts as distinct registers of writing and sub-genre of academic writing because of their condensed nature and word limit that obliges writers to express key ideas concisely.

The findings of several studies also indicate that lexical density values are higher in higher proficiency levels and English L1s. For instance, this index showed significant between-proficiency level differences of the academic argumentative essays of EFL students in Kim's (2014) study and is shown as a strong predictor of academic writing proficiency. In Gregori-Signes and Clavel-Arroitia's (2015) study also, students at higher levels of proficiency produced texts with higher lexical density values. Linnarud (1975) found that lexical density of English texts written by English native speakers has higher values than those written by Swedish speakers. Malvern, Richards, Chipere, and Durán (2004) argue that because lexical density is a token-token ratio, it is not affected by sample size.

The construct of lexical diversity or variation is often defined as the variety or range of different words in a text (Johansson, 2008; Housen, Bulté, Pierrard, & Van Daele, 2008; Malvern et al., 2004) or to put it precisely, "phonologically-orthographical different word forms" that is representative of the size of vocabulary knowledge (Housen et al., 2008, p. 3). Capturing new and diverse words, however, poses technical problems because as the length of discourse (e.g., a text) increases, the probability of new words occurring decreases. This is tightly related to the notion of TTR or Type-Token ratio: as the text length increases, the total number of words or tokens increases, while the number of new types (or new lexical words) does not increase with the same ratio. This is due to the repetitive nature of the function or grammatical words vs. content or lexical words. This problem has been extensively discussed in the literature (e.g. in McCarthy & Jarvis, 2010). Consequently, the texts are usually truncated to have a uniform length in comparative studies; many mathematical and computational alternatives have been also offered to capture the types in the face of increasing (grammatical) tokens when comparing texts with different length. In this study, we describe 14 most-frequently reported measures/indices as indicators and predictors of lexical proficiency (differences) and development. These selected indices are theoretically and empirically motivated by the studies in the previous section and following paragraphs; however, the literature on some of the similar measures is not conclusive and the studies that have been conducted in academic settings usually have used only one or a few lexical diversity indices. Here we divide these indices into six pairs or groups of closely-related measures based on their quantification methods to examine (in the analysis section of this study) the effectiveness of each measure in each pair or group in capturing lexical diversity differences of texts produced by English L1 vs. L2.

The first pair calculates different versions of the Number of Different Words or NDW (i.e., types) and are labelled as NDWERZ and NDWESZ in Lu (2012). They are recommended by Malvern et al. (2004) and Lu (2012) as standardised indices that, unlike the simple number of types, are not affected by text length. Both indices select 10 random sub-samples of 50 words in a text to get the averages of NDW. The sub-samples of the former measure include a random but standard number of words from the sample and the latter measure's sub-samples include a standard number of consecutive words, but the starting point is randomly selected. The two indices showed a strong correlation as well as significant differences between proficiency levels in Lu's (2012) study of transcribed oral narratives of university EFL learners. The number of different words was also shown to be the strongest predictor of writing proficiency in Kim's (2014) academic essays of EFL learners with different proficiency levels based on CEFR.

The second group comprises the three word-string-based measures of MSTTR (Mean Segmental TTR; Johnson, 1944 cited in Lu, 2012), MATTR (Moving Average TTR; Covington & McFall, 2010), and MTLD (Measure of Textual Lexical Diversity; tested by McCarthy & Jarvis, 2010) which are reported to be robust to variations in sample size and text-length (see e.g., Jarvis, 2013 and McCarthy & Jarvis, 2010). MSTTR averages the TTRs from all fixed-size segments of the texts; in Lu (2012) this index is found to have a high correlation with test-takers' rankings and there were significant between-proficiency level differences based on this measure as well. MATTR computes the lexical diversity of a text by assigning a moving window and estimates the type-token ratios for fixed-length successive windows. This moving window is a feature which distinguishes it from the MSTTR measure and which allows the words in a text to be successively calculated, not just fixed successive chunks or segments. MTLD is computed as "the mean length of word strings that maintain a criterion level of lexical variation" (McCarthy & Jarvis, 2010, p. 381), e.g., maintaining a type-token ratio of 0.72 in the mentioned study. Koizumi (2012) reported that this measure is not affected by text length changes for very short texts ranging from 50 to 200 tokens; he recommends a minimum of 100 tokens per text for using this measure. In Gonzalez (2013), MTLD was proved to be a significant indicator of writing proficiency which distinguished between English L1 vs. ESL academic writing. Treffers-Daller et al. (2016) also concluded that MTLD is a strong predictor of proficiency in ESL essays scores.

The next pair is the two logarithm-based measures of LOGTTR and UBER. The former is also called Bilogarithmic TTR or Herdan's C (Herdan, 1960). According to Malvern et al. (2004, p.27), it is conceptually based on the fact that "the rate of increase of types with increasing token count will be proportional to the TTR for any given value of N". Uber's U index (Dugast, 1978) scales down the ratio of tokens to the division of tokens to types. This measure showed significant correlations with test-takers' rankings as well as significant between-proficiency level differences in Lu (2012). Jarvis's (2002) study of adolescent narrative texts (which were shorter than 500 words) also found that U accurately models the actual TTR curve and is an optimal model of lexical diversity of whole texts.

Fourth, we will investigate the two variants of D measures, namely Vocd-D and HD-D which are reported to capture unique lexical information and are recommended to be suitable for analysing the lexical diversity of short texts (McCarthy & Jarvis, 2010). The D measure was first proposed by Malvern and Richards (1997) to obtain the lexical variation by capturing the rate of decrease of TTR by finding the best-fitting curve of TTR; higher token curves are the results of greater diversity in the text. Curve fitting is a process of using a mathematical function which can fit all or a specified number of data points (e.g., on a curve) in its best possible way. The Vocd-D variant of D is a probabilistic method which does so by random sampling of words in a text without replacement; it randomises the order of words before measuring the TTR segment. The Vocd-D measure in Lu's (2012) study of oral narratives (transcribed texts ranging from 105 to 476 words) showed significant correlation with test-takers' rankings as well as significant differences between proficiency levels. However, Koizumi (2012) found that this measure is text-length dependent for very short texts up to 150 tokens. Yoon (2017) recorded Vocd-D as a significant predictor of L2 writing proficiency. Vocd-D is also shown to differentiate between NS and ESL academic writing in Gonzalez (2013) but not in Pietilä (2015). HD-D is another variant which calculates "for each lexical type in a text, the probability of encountering any of its tokens in a random sample of 42 words drawn from the text" (McCarthy & Jarvis, 2010, p. 383). The sum of all the probabilities for all lexical types in the text is used as an index of the text's lexical diversity. McCarthy and Jarvis (2010, p. 384) comment that Vocd-D and HD-D have different scales (see Section 5.3 below). They also found that the values and the correlations of these two measures are affected by text register variations, with Vocd-D having more distinguishing power in

Table 1
Lexical density and 14 lexical diversity measures and their quantification methods.

Measure	Label	Quantification Method
Lexical Density (Ure, 1971)	LD	N_{lex} / N
Number of Different Words-type I	NDWERZ	Means of NDW for 10 random sub-samples of 50 words
Number of Different Words-type II	NDWESZ	Means of NDW for 10 random sub-samples of 50 consecutive words with random starting points
Bilogarithmic TTR or Herdan's C (Herdan, 1960, cited in Tweedie and Baayen, 1998)	LOGTTR	Log T / Log N
Uber's U (Dugast, 1978)	UBER	$\mathrm{Log}^2 N/\mathrm{Log}(N/T)$
Mean Segmental TTR (Johnson, 1944 cited in Lu, 2012)	MSTTR	Mean of TTRs for 50-word segments/word strings
Moving Average TTR (Covington & McFall, 2010)	MATTR	Mean of TTRs for the fixed-length successive windows of word strings
Measure of Textual Lexical Diversity (McCarthy, 2005)	MTLD	The mean length of consecutive word strings that maintain a TTR of 0.72, words/factors method
Verb Variation-type I	VV1	T_{verb} / N_{verb}
Corrected Verb Variation I	CVV1	$T_{verb} / \sqrt{2N_{verb}}$
Verb Variation-type II	VV2	T_{verb} / N_{lex}
Lexical Variation	LV	T_{lex} / N_{lex}
Noun Variation	NV	T_{noun} / N_{lex}
The adapted D (Malvern & Richards, 1997) measure of lexical diversity; Vocd-D by Malvern et al. (2004)	VOCD	Random sampling of words for TTR segments, curve-fitting method
Hypergeometirc D (McCarthy and Jarvis, 2007)	HDD	Sum of lexical probabilities based on random samples of 42 words

NDW = Number of Different Words, TTR = Type Token Ratio, N = the number of tokens, T = the number of types, lex = lexical items, Log = logarithm.

register variations. In the present study, we will examine these differences in short academic writings.

The fifth category includes the three verb diversity indices of VV1, CVV1, and VV2 as they showed between-proficiency-level differences in Lu (2012) study of the quality of transcribed oral narratives by university students as well as in Mazgutova and Kormos (2015) academic writings of ESL groups. The quantification methods of these indices are provided in Table 1. Granger and Paquot (2009, p. 193) reported that lexical verbs play important roles in "expressing personal stance, reviewing the literature, quoting, expressing cause and effect, summarising and contrasting" in academic writing. They, for example, found that EFL learners use a limited range of verbs and underuse most "academic verbs … that express rhetorical functions" (p. 210). Paquot (2019) also reported that the values of CVV1 index increase across CEFR-based proficiency levels and VV2 in non-adjacent levels in the academic writings of EFL groups.

The two indices of lexical word variation (LV) and noun variation (NV) were lastly investigated in this study as the last pair in order to find out if the proportion of noun types and lexical types are significantly different in English L1 vs L2 texts. These two measures share the same denominator and both are classed as lexical variation based on TTR of word classes in Lu (2012) with strong correlations. Biber (2006) and Biber and Gray (2016) illustrate that written university registers heavily rely on nouns and nominalisations. Paquot (2019) also reported that the values of LV index increased in non-adjacent proficiency levels of three groups of EFL academic writers.

2.3. Research on the abstract section as a distinct sub-genre of academic writing

Several notable studies consider the abstract section of a thesis/dissertation or research article to be a distinct sub-genre of academic writing which is characterised by a lexically dense outline, and a summary of the whole thesis/article which reflects brief introductory statements, methodology, major findings, and the significance, implications, and contributions of the study and results (Bhatia, 1993; Bitchener, 2010; Gillaerts & Van de Velde, 2010; Pho, 2008) and as Bunton (1998) puts it, as a microcosm of the thesis and "a self-contained piece of discourse" representing "some of the best writing of the author" (p. 72).

The abstract is the first part of most academic texts and is supposed to invite readers to continue reading by persuading them that the rest of the work is interesting, and the results are reliable and significant (Bitchener, 2010; Bunton, 1998; Gillaerts & Van de Velde, 2010); Bitchener (2010, p.11) calls this the "persuasive function" of abstracts. Furthermore, the quality of the abstracts is important as they appear in the abstracting and indexing of publishers especially in English e.g., the indexing of thesis/dissertation abstracts by ProQuest (PQDT A&I).

For many years, researchers' attention has been drawn to abstracts in general, and various works have investigated different linguistic, stylistic and structural characteristics of this sub-genre. Yoneoka and Ota (2017), for instance, found that high-quality abstracts in articles contain longer words, shorter sentences, a larger proportion of noun phrases and smaller proportion of verb phrases than low-quality abstracts; however, they did not find any difference between low and high-quality abstracts in terms of lexical diversity. Hyland and Tse (2005) investigated the frequencies as well as forms and functions of 'evaluative that' in the research article, MA dissertations and PhD theses abstracts written by English L2 writers. Other works investigated abstracts in research articles and conference papers for various linguistic features of tense, voice, modal and reporting verbs, stance words, nouns, that-complement clause, errors, and first-person pronouns, (e.g., Egbert & Plonsky, 2015; Pho, 2008). Nevertheless, the description of the abstract section of master's dissertations is an underinvestigated area, especially regarding the differences in lexical density and diversity in abstracts produced by both English L1 and L2 students. This is also to consider the role of the word limit in obliging abstract writers to express key ideas concisely and the range of non-repetitious words that manifest in this limited space.

3. Methodology, research aims and questions

Numerous works have underlined the importance of lexical proficiency in academic writing, specifically thesis and dissertation writing, particularly for EFL and ESL learners (Engber, 1995; Pietilä, 2015), in this study's case to compare the English proficiency needed to earn an MA in EFL as well as the English-medium contexts. There is a dearth of previous research on postgraduate academic writing proficiency differences compared to the prevalence of investigations on the general/argumentative essays of undergraduate students. To our knowledge, no study has so far conducted a systematic analysis of a large set of lexical measures in sub-genres of master's dissertations in a discipline-specific corpus; most previous academic writing studies vary considerably in their research designs while examining one or a few of such measures to gauge proficiency differences of English L1 vs. L2 students as well as textual differences. In this cross-sectional comparative study, we focus on MA dissertation abstracts to examine the diversification of lexis and how densely English L1 vs. L2 writers use the lexical words in an abstract to provide key information about the study as concisely as possible within the word limit. Taking the English language background of students (English L1, EFL, ESL) as the independent variable, and lexical density and 14 diversity measures (as specified in Table 1), each as a dependent variable, this study is, therefore, designed to answer the following questions in the order of presenting the results:

Q1. To what extent are the patterns i.e., the use and amount of of lexical units (e.g., nouns, verbs, lexical types and tokens) different in the abstracts of English L1 academic writers vs. their EFL and ESL peers?

Q2. To what extent are the measures in each pair/group of similarly-quantified measures correlated based on an academic writing corpus of abstracts, and to what extent can the construct-distinctiveness of lexical density and diversity be supported based on the academic data?

Q3. To what extent do the ESL students who have been studying in the UK academic setting produce more lexically-complex abstracts than their EFL peers who have been studying English in a non-English-speaking context? And to what extent do the ESL students' performances approximate the English L1 group considering the effect of the shared academic setting?

Furthermore, this study seeks to contribute to the measure-selection process of future studies with similar research design by asking:

Q4. Which lexical diversity measures in pairs/groups of similarly-calculated measures (as discussed in Section 2.2) better capture the differences between groups with various English backgrounds (English L1, EFL, ESL) and academic contexts?

4. Method

4.1. Corpus

This study analyses lexical density and diversity measures in the Abstract section of master's dissertations, as a distinct sub-genre of academic writing, in applied linguistics and other EFL-related disciplines written by EFL (English as a Foreign Language), ESL (English as a Second Language), and English L1 students. The EFL students are all Iranian nationals with various L1s. They studied in various universities in Iran with a centralised curriculum. The data for the EFL group was collected from different geographical regions to include different ethnic and L1 backgrounds. ESL students have various L1s and ethnic backgrounds, and moved from non-English-speaking countries (i.e., originally EFL students) and resided in the UK only for the duration of their postgraduate studies often as part of academic immersion programmes. They have submitted their dissertations to various universities in the UK. English L1 students are all British nationals (born and educated in the UK) who have studied and submitted their dissertations to various universities in the UK. All students in the three groups submitted their dissertations within eight years prior to the commencement of this study. An academic learner corpus was drawn from students who were contacted to share copies of their dissertations along with certain demographic information such as age, nationality, and language background for sorting the data and ensuring the homogeneity in each group. The corpus comprises 70 abstracts in each group - a total of 210 texts. Texts in the corpus range between 175 and 300 words. However, the total word count in each group is truncated to 15,400 with a mean of around 220 words per group. To maximise corpus representativeness care has been taken to include as similar number of texts as was possible in terms of the topic (different sub-disciplines of applied linguistics as the dominant subject area) in each group as shown in Table 2. Therefore, the effects of text length and topic are controlled in this study. Students reassured us that editing and proofreading of the abstracts were minimal and mainly regarded the grammar/spelling check.

4.2. Measure selection and data analysis

Lexical density and diversity measures that were described in detail in 2.2 were investigated as indicators and predictors of academic writing proficiency (differences) and development as recommended by the mentioned scholars. Three natural language processing tools, Lexical Complexity Analyzer (LCA, described and implemented in Lu, 2012), TAALED (beta version 1.2.4, Kyle, 2018), and Coh-Metrix (version 3.0; Graesser et al., 2004) were used to analyse these indices as outlined in Table 1. LCA analysed the ten measures of LD, NDWERZ, NDWESZ, LOGTTR, UBER, VV1, CVV1, VV2, LV, and NV; TAALED analysed the four indices of MSTTR, MATTR, HDD and MTLD, and finally, the Vocd-D measure was analysed by Coh-Metrix. To ensure the validity of the comparisons of the results across the three analysers, care has been taken to make the tokenisation, tagging, and lemmatisation processes of the input texts consistent [1]. Measures in both LCA and TAALED are calculated using lemma forms. The Vocd-D measure in Coh-Metrix is calculated based on word forms; to make it conform with the other tools, lemmatised files are used as input to Coh-Metrix.

4.3. Statistical procedures

Although the data for the three groups showed a fairly normal distribution, this study adopts the robust statistics method of bootstrapping recommended by Larson-Hall (2016). Instead of relying solely on the dichotomous (significant vs. non-significant) nature of p-values in conventional statistics (i.e., NHST) to interpret the results, the results of between-group differences are also accompanied by confidence intervals and effect sizes. The confidence intervals (CIs) are calculated based on the 95 % Bca (bias--corrected and accelerated) bootstrapped CI format with 2000 replicates. Omnibus one-way ANOVAs were calculated by fitting a series of general linear models; based on that, post-hoc multiple comparison tests of Tukey HSD were run to get the mean differences and confidence intervals. The point estimate effect sizes of Cohen's d was calculated based on the pooled standard deviation as the

Table 2
The distribution of sub-disciplines of applied linguistics across groups.

	TEFL/ ELT	First/Second Language Acquisition	Discourse Analysis	Corpus-based Studies	Linguistics[a]	Socio-linguistics	Cognitive Linguistics
EFL	25	18	14	7	3	2	1
ESL	22	22	7	10	4	3	2
English L1	19	15	18	13	1	3	1

–The classification is based on the dominant subject area; some studies are inter-disciplinary in nature.
[a] Linguistics dissertations included phonology, phonetics, lexical, and syntactic studies.

standardiser; Hedges' g was recommended by Gerlanc and Kirby (2013) and Larson-Hall (2016) as an unbiased and more conservative CI estimator. The interpretation of effect sizes will be based on the guidelines of Plonsky and Oswald (2014) which treat 0.4 as small, 0.7 as medium and 1 as large effect sizes. Correlation coefficients (r) were obtained via the Pearson correlation test. To offset the possible increase in the type I error rate due to multiple comparisons, the stricter Bonferroni-corrected new alpha level of 0.003 (15 measures / 0.05 conventional alpha level = 0.003) was applied to interpret the significance level of the between-group-difference tests. The statistical procedures were all carried out using the R statistical and programming language (version 3.6.0; R Core Team, 2013).

5. Results and discussions of the findings

5.1. *Lexical profiles of the EFL, ESL, and English L1 groups*

Table 3 presents the frequency counts of lexical production units for the three groups. Word tokens are all instances of words; word types are the unique (i.e., non-repetitive) tokens. Lexical words (e.g., lexical tokens and types) follow the same principle but are lexical/content words rather than grammatical/function words. As noticed, the EFL group produced significantly lower amounts of word types and lexical types compared with the other two groups. This is directly reflected in the results of the TTR-based measures as indicated by Table 5.

5.2. *The relationship among lexical density and diversity indices*

The results of the correlation test as presented in Table 4 confirm the construct-distinctiveness of lexical density and diversity as indicated by weak correlations between lexical density (LD) and various lexical diversity measures (all categories) with the exception of a moderate but negative (inverse correlation) between LD and LV ($r = -0.55$). This construct-distinctiveness result is also supported by Lu's (2012) correlation results of an academic EFL learner corpus of transcribed oral narratives. This also shows that this construct-distinctiveness is independent of the mode of language and genre.

Most of the measures in each category are moderately or strongly correlated. The exception is the LV and NV indices (category 6) that show a weak correlation coefficient of 0.4; this suggests that noun types do not comprise a great proportion of all lexical types in this corpus of abstracts. Verb-based indices in category 5 (VV1, VV2, CVV1), on the other hand, show moderate to strong correlations as indicated by this table. These two categories that are classed in one general category of 'lexical variation based on the TTR of word classes' in Lu (2012) show weak correlations based on this study's corpus of MA abstracts and should indeed be regarded as separate categories as Lu's (2012) correlation results also suggest. A comparison of VV2, NV, and LV with the same denominator, however, indicates that there were more noun types than verb types in all lexical types (i.e., r^2 or the square of correlation coefficients of these pairs that explain the proportion of their shared variances).

The strongest correlation is recorded for the two logarithm-based measures of LOGTTR and UBER (category 3) with $r = 0.9$, followed closely by the Vocd-D and HD-D indices (category 4) with $r = 0.89$. This latter finding compared to the results of McCarthy and Jarvis (2010) shows the effect of text register variations on the correlations of Vocd-D and HD-D as they mentioned. The two measures of NDWERZ and NDWESZ (category 1) which are based on the number of different words also showed a strong relationship with $r = 0.74$ but less strong as the previous categories. Finally, very high correlations are noticed for the measures in category 2 between the measures of MSTTR, MATTR, and MTLD which are based on word strings or word-segments: MSTTR vs. MATTR = 0.89, MSTRR vs. MTLD = 0.85, and MATTR vs. MTLD = 0.88. Categories 1 and 2 which are generally based on word segments have higher correlations as expected.

5.3. *Lexical density and diversity in the abstracts of dissertations: between-group differences*

As is demonstrated in Table 5, the ESL and English L1 groups performed similarly in the production of lexically dense and diverse texts as indicated by ten of the indices and significant differences with larger effects were found only in comparisons which include the EFL group. The descriptive statistics and the post-hoc comparison tests of mean differences, confidence intervals and their respective effect sizes of these ten measures all witness that the EFL group produced the least lexically dense and diverse texts. However, none of the verb-based nor the LV and NV indices showed any between-group differences in the mean values of these lexical variations based on the TTR of word classes, indicating the production of similar amounts of diverse nouns and verbs by these students. This also suggests that TTR transformations (e.g., log-based measures of LOGTTR and UBER) as well as the word-string and word-segment-based measures better capture the differences between these groups with different English backgrounds in the production of lexically diverse texts. Unlike the findings of Mazgutova and Kormos (2015) on academic writing of low and high-level ESL groups where verb-variation was a differentiating feature of between-proficiency levels, in this study, the three postgraduate groups performed similarly regarding

Table 3
Lexical profiles of the EFL, ESL, and English L1 groups.

	Word Tokens sum (mean)	Word Types sum (mean)	Lexical Tokens sum (mean)	Lexical Types sum (mean)
EFL	15,400 (223)	7837 (111.95)	397 (5.6)	266 (3.8)
ESL	15,400 (221)	8157 (116.52)	553 (7.9)	338 (4.8)
English L1	15,400 (220)	8145 (116.35)	538 (7.6)	337 (4.8)

Table 4
Correlations among lexical density and diversity measures.

	ld	ndwerz	ndwesz	logttr	uber	msttr	mattr	mtld	vv1	cvv1	vv2	lv	nv	vocd	hdd
ld	1														
ndwerz	-0.005	1													
ndwesz	-0.03	0.74	1												
logttr	-0.02	0.73	0.65	1											
uber	0.02	0.78	0.69	0.92	1										
msttr	-0.04	0.69	0.76	0.61	0.68	1									
mattr	-0.01	0.72	0.87	0.64	0.70	0.89	1								
mtld	0.03	0.70	0.75	0.65	0.73	0.85	0.88	1							
vv1	-0.03	0.08	0.09	0.05	0.08	0.09	0.12	0.17	1						
cvv1	0.13	0.09	0.06	0.02	0.07	0.10	0.09	0.13	0.88	1					
vv2	-0.22	0.009	-0.007	0.06	0.03	0.04	0.04	0.08	0.65	0.70	1				
lv	-0.55	0.15	0.17	0.26	0.23	0.21	0.18	0.16	0.33	0.17	0.33	1			
nv	-0.11	0.06	0.05	0.15	0.14	0.04	0.04	0.04	0.02	-0.03	-0.20	0.42	1		
vocd	0.05	0.83	0.75	0.73	0.83	0.74	0.76	0.77	0.11	0.11	0.02	0.15	0.09	1	
hdd	0.01	0.87	0.79	0.75	0.80	0.82	0.86	0.83	0.11	0.10	0.007	0.16	0.11	0.89	1

Table 5
Significant between-group differences, multiple comparisons and effect sizes.

			Tukey HSD group differences			ANOVA		
Measure	Descriptive Mean & (SD)	Group Comparisons	Mean difference & [95% Bca CIs]	Effect size cohen's d	Bootstrapped Effect size [95% Bca CIs] Hedges' g	F	Pr(>F)	Sig.
ld	EFL 0.02 (0.02) ESL 0.04 (0.02) NS[a] 0.03 (0.02)	ESL-EFL NS-EFL	0.011 [0.002,0.019] 0.009 [0.001, 0.017]	0.52 0.51	[0.20, 0.80] [0.16, 0.81]	5.72	0.003	**
ndwerz	EFL 37.35 (1.93) ESL 38.61 (1.77) NS 38.53 (1.98)	ESL-EFL NS-EFL	1.257 [0.499, 2.014] 1.177 [0.419, 1.934]	0.67 0.6	[0.30, 1.03] [0.25, 0.96]	9.62	<0.001	***
ndwesz	EFL 36.83 (2.20) ESL 38.58 (2.08) NS 38.24 (2.37)	ESL-EFL NS-EFL	1.742 [0.857, 2.628] 1.401 [0.515, 2.287]	0.81 0.61	[0.39, 1.15] [0.23, 0.96]	12.12	<0.001	***
logttr	EFL 0.87 (0.02) ESL 0.88 (0.02) NS 0.88 (0.02)	ESL-EFL NS-EFL	0.010 [0.002, 0.018] 0.010 [0.002, 0.018]	0.54 0.54	[0.14, 0.89] [0.19, 0.86]	6.6	0.001	**
uber	EFL 18.53 (2.40) ESL 20.19 (2.66) NS 20.42 (3.31)	ESL-EFL NS-EFL	1.662 [0.538, 2.875] 1.887 [0.764, 3.010]	0.65 0.65	[0.28, 1.01] [0.33, 0.96]	9.39	<0.001	***
msttr	EFL 0.71 (0.04) ESL 0.74 (0.04) NS 0.75 (0.04)	ESL-EFL NS-EFL	0.031 [0.014, 0.047] 0.043 [0.026, 0.059]	0.72 1.04	[0.37, 1.06] [0.68, 1.41]	20.37	<0.001	***
mattr	EFL 0.71 (0.04) ESL 0.74 (0.04) NS 0.75 (0.04)	ESL-EFL NS-EFL	0.035 [0.018, 0.051] 0.038, 0.022, 0.055]	0.87 0.92	[0.49, 1.18] [0.52, 1.30]	18.76	<0.001	***
mtld	EFL 48.36 (10.84) ESL 58.92 (13.34) NS 60.39 (14.87)	ESL-EFL NS-EFL	10.55 [5.31, 15.78] 12.02 [6.78, 17. 26]	0.86 0.92	[0.52, 1.18] [0.55, 1.21]	17.49	<0.001	***
vocd	EFL 65.89 (16.07) ESL 78.32 (19.05) NS 79.82 (19.18)	ESL-EFL NS-EFL	12.42 [5.17, 19.67] 13.92 [6.68, 21.17]	0.7 0.78	[0.34, 1.02] [0.42, 1.11]	12.41	<0.001	***
hdd	EFL 0.75 (0.04) ESL 0.78 (0.04) NS 0.78 (0.03)	ESL-EFL NS-EFL	0.029 [0.015, 0.043] 0.030 [0.016, 0.044]	0.81 0.87	[0.41, 1.16] [0.48, 1.22]	16.94	<0.001	***

a. English L1 is labelled as NS.
–Only the lexical measures which showed between-group differences for at least one pair of comparison are included in this table. The Bonferroni-corrected new alpha level of 0.003 is applied to significant results in the Sig. column. The significant codes:

Annotation	p-value range	Significance level
***	[0, 0.001]	0.001
**	[0.001, 0.01]	0.01

–The degrees of freedom for all F statistics are 2 and 207.

the production of various verbs as measured by the three verb-based indices.

Medium effects are observed for the ESL-EFL and NS-EFL comparisons regarding lexical density. Likewise, in Kim's, 2014 study of college essays, LD distinguished between three proficiency levels of basic, intermediate, and advanced. These two findings are in sharp contrast with Lu's (2012) study of the quality of transcribed oral narratives and the studies cited in Wolfe-Quintero et al. (1998) where LD did not differentiate between proficiency levels, indicate different profiles of lexical density in written vs. spoken data as well as written learner data vs. academic writing.

Both indices based on the number of different words in this study recorded significant differences between the ESL-EFL and NS-EFL groups; however, NDWESZ captured differences with larger effect sizes and CIs and much lower p-value of the F statistic. Similarly, Kim's 2014 results indicate that NDW was the strongest predictor of L2 writing proficiency levels as well as the findings of Lu (2012) where it proved a useful predictor of the quality of ESL's oral discourse across proficiency levels. Lu (2012) also found significant between-level differences in these two measures and significant correlations with test-takers' rankings. The combined results may suggest that these indices can capture differences regardless of the mode of data (spoken vs written) and text genres and registers.

All three measures of MSTTR, MATTR, and MTLD in category 2 also recorded meaningful differences between the mean values in the mentioned comparison sets, with large effects, large F statistics. The MATTR and MTLD results show very similar effect sizes for the two comparisons both in point estimates and confidence intervals. Gonzalez (2013) showed a similar result where MTLD was a significant differentiator between English L1 and ESL academic writing scores and an indicator of writing proficiency. The MSTTR measure in this study, however, shows a larger effect for the NS-EFL comparison with a CI reaching up to 1.41 as well as a larger F value of the overall between-group differences.

Both log-based measures as indicated in Table 4 also show significant between-group differences for the mentioned comparison sets with the UBER leading to slightly larger effect sizes, with confidence intervals reaching up to 1. Lu (2012) however, did not find any significant between-level differences in the mean values of LOGTTR in timed spoken data. Even though these two measures are highly correlated ($r = 0.9$), they use different log-based approaches. LOGTTR merely changes the shape of the curve of TTR and alters the scale, while UBER is found by Jarvis (2002) to provide more accurate fits for the TTR curve vs. token curves. Jarvis (2002) also found significant differences in short English narrative writings of EFL school learners with different L1 groups with regard to the UBER index.

The two indices in category 4, namely Vocd-D and HD-D also captured between-group differences in the mentioned sets with medium to large effect sizes and CIs and significant F statistics at <0.001. Between the two, the HD-D index captured more differences leading to slightly larger effect sizes and, with confidence intervals reaching up to 1.22 for the NS EFL comparison. Lu (2012) found significant between-level differences in the D measure (the vocd utility of CLAN, see the discussion in Lu, 2012) and in the Gonzalez (2013) study, the Vocd-D measure differentiated between the NS and ESL academic writing. Bulté and Housen (2014), however, did not find the Vocd-D measure to be a significant indicator of writing development. This might be due to the nature of data and research design in these two studies. Besides, the HD-D output, as McCarthy and Jarvis (2010) describe, is simply the sum of probabilities (of drawing tokens of a type) while the Vocd-D converts the sum of probabilities of word occurrences (using random sampling) into TTR and then to a final D value, which is a more complex estimation and may lead to "greater fluctuations" for texts of high diversity (p. 390).

6. Concluding remarks: revisiting research questions, implications, and recommendations

This study investigated the differences in English L1 vs. L2 (both EFL and ESL) academic writing regarding the extent to which they produce lexically dense and diverse abstracts. Dissertation abstracts are regarded both as distinct text registers and as sub-genres of theses/dissertations with a rhetorical structure that reflects the rest of a research dissertation.

The results show that the EFL group produced the least lexically dense and diverse texts compared to the other groups. Taking the selected measures as indicators of lexical/linguistic proficiency (differences) as recommended by previous scholars, this result may suggest a lexical proficiency gap between the EFL and the other groups as well. It is premature to derive a definite conclusion regarding this proficiency difference in the entirety of dissertations; however, materials developers in the EFL academic settings such as Iran could explicitly address the concepts of lexical density and diversity in thesis/dissertation writing modules and demonstrate how a lexically dense abstract includes more informative words (i.e., the content words) and therefore, teaching the strategies to use this limited space to convey all relevant information e.g., by using nominalisations and longer/elaborated noun phrases, appositive noun phrases (see e.g., Biber, 2006; Biber & Gray, 2010, 2016). Gonzalez (2013), for instance, argues that lexical proficiency measures can inform instructors to examine "if particular aspects of vocabulary knowledge are being overlooked in learner compositions" (p.15). Doró and Pietilä (2015) equally underline the importance of NLP tools in assessing writing and lexical proficiency and development of learners by researchers and instructors.

The abstracts of the ESL group produced comparable statistics to those of the English L1 students regarding the diversification of lexis and how densely their abstracts were represented with lexical words. This could be due to the immersion postgraduate programmes where ESL students study, live and communicate in an English-speaking country, in this study's case, in the UK, and attend the same postgraduate programmes as their English L1 peers. These results point to the possible contribution of academic immersion programmes on the linguistic proficiency of ESL students (for a detailed discussion, see Mazgutova & Kormos, 2015). Mazgutova and Kormos (2015) for instance, believe that EAP programmes need to explicitly address various lexical features of academic genres. As this study only investigated short academic texts as abstracts, further studies using different academic writing corpora and gauging the effect of academic immersion programmes are needed to confirm the effectiveness of such programmes and the role of shared academic context with English L1s. This is necessary because most research in this area has mainly focused on undergraduate EAP programmes.

We also recommend examining the effect of short intensive thesis/dissertation writing EAP immersion programmes in bridging the lexical proficiency gap between EFL and English L1 postgraduate students.

Perhaps an important implication for writing research and assessment is the need to investigate a large set of lexical diversity indices that capture texts' range of non-repetitious words from different angles/with different formulas and examining the effectiveness of each measure in pairs/groups of closely-related measures based on their quantification methods. In this analysis of a corpus of MA dissertation abstracts, correlation tests confirmed the construct-distinctiveness of lexical density and diversity and showed that the measures in the specified categories of pairs/groups of similar measures in terms of the quantification methods are indeed highly correlated. A series of general linear model tests accompanied by Tukey multiple comparison tests then revealed that lexical density (LD) and the lexical diversity measures of NDWESZ, UBER, MSTTR (and MATTR) and HD-D better capture academic writing differences of the three groups with larger effect sizes and confidence intervals. Among these measures, the two measures of MSTTR and HD-D were the most effective ones in capturing the lexical diversity differences of abstracts of English L1 vs. EFL students and therefore, are recommended to future writing researchers for studies with similar research designs. Using a smaller set of indices is particularly beneficial for writing researchers considering the variety of measures proposed in the literature to capture proficiency or text differences, some of which are mathematical or computational alternatives of others. It is also hoped that future researchers new to this field can navigate through these indices and make informed decisions about the use of each based on their research objectives. Furthermore, automated writing assessment using any of these recommended indices is more beneficial when dealing with large corpora and frequent assessments; in local settings, however, these can complement teachers'/researchers' judgment of writing quality, e.g., to track the lexical performance/development of students or in comparative proficiency studies with English L1 data (see e.g., the discussion in Doró & Pietilä, 2015).

Future research could also address the limitations of this study by incorporating other genres and sub-genres of academic writing into the design, by including other English L1 students besides the British ones and other EFL students beside Iranian ones, by considering the effect of the L1 of EFL and ESL students on the production of lexically dense and diverse texts, and by comparing students' texts with those of expert writers/scholars. It will be beneficial to duplicate this study or to conduct a study with similar designs with a larger sample size, in an ideal situation where a researcher has access to a larger collection of dissertations. It would be also interesting to examine whether similar patterns would emerge from postgraduate students in speaking proficiency tests using the same set of lexical measures, e.g., to examine the relationship between the two production modes and the findings of researchers like Halliday (1985) and Lu (2012). Investigating the effects of the topic (e.g., sub-disciplines of linguistics), pedagogic interventions and explicit teaching of such lexical measures in experimental studies, and the use of other computational systems could also reveal other significant aspects of linguistic proficiency.

7. Note

1. The tools use the PTB (based on the Penn Treebank project) style of tokenisation, taggers with tagging models trained on the Penn Treebank Tagset (see Marcus, Santorini, & Marcinkiewicz, 1993 for more details), and morpha class style of lemmatisation and Morpha (version 2003; Minnen, Carroll, & Pearce, 2001) for morphological processing.

CRediT authorship contribution statement

Maryam Nasseri: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Validation, Visualization, Writing - original draft, Writing - review & editing. **Paul Thompson:** Supervision, Writing - review & editing.

Acknowledgements

We would like to thank the two anonymous reviewers for their valuable suggestions and comments.

References

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. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes, 9*, 2–20.
Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge: Cambridge University Press.
Bitchener, J. (2010). *Writing an applied linguistics thesis or dissertation: A guide to presenting empirical research*. Basingstoke: Palgrave MacMillan.
Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, V. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Investigating complexity, accuracy and fluency in SLA* (pp. 21–46). Amsterdam: John Benjamins.
Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing, 26*, 42–65.
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.
Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics, 17*(2), 94–100.
Crossley, S. A., & McNamara, D. S. (2010). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 35*(2), 115–135.
Doró, K. (2008). *The written assessment of the vocabulary knowledge and use of english majors in Hungary* (Doctoral Dissertation). Graduate School in Linguistics: University of Szeged.
Doró, K., & Pietilä, P. (2015). Researching vocabulary in L2 writing: Methodological issues and pedagogical implications. In P. Pietilä, K. Doró, & R. Pípalová (Eds.), *Lexical issues in L2 writing* (pp. 11–26). Newcastle upon Tyne: Cambridge Scholars Publishing.
Dugast, D. (1978). Sur quoi se fonde la notion d'étendue théoretique du vocabulaire? *Le francais moderne, 46*, 25–32.

Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics, 25*(2), 220–242.
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.
Gerlanc, D., & Kirby, K. (2013). *BootES: Bootstrap effect sizes. R package version 1.01*.
Gillaerts, P., & Van de Velde, F. (2010). Interactional metadiscourse in research article abstracts. *Journal of English for Academic Purposes, 9*, 128–139.
Gonzalez, M. (2013). *The intricate relationship between measures of vocabulary size and lexical diversity as evidenced in non-native and native speaker academic compositions* (Doctoral Dissertation). College of Education and Human Performance: University of Central Florida.
Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods Instruments & Computers, 36*, 193–202.
Granger, S., & Paquot, M. (2009). Lexical verbs in academic discourse: A corpus-driven study of learner use. In C. Maggie, D. Pecorari, & S. Hunston (Eds.), *Academic writing: At the interface of corpus and discourse* (pp. 193–214). London: Continuum.
Gregori-Signes, C., & Clavel-Arroitia, B. (2015). Analysing lexical density and lexical diversity in university students' written discourse. *Procedia, Social and Behavioral Sciences, 198*, 546–556.
Halliday, M. A. K. (1985). *Spoken and written language*. Geelong, Vic: Deakin University Press.
Herdan, G. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. The Hague: Mouton.
Housen, A., Bulté, B., Pierrard, M., & Van Daele, S. (2008). Investigating lexical proficiency development over time – The case of Dutch-speaking learners of French in Brussels. *French Language Studies, 18*, 1–22.
Hyland, K., & Tse, P. (2005). Hooking the reader: A corpus study of evaluative that in abstracts. *English for Specific Purposes, 24*, 123–139.
Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19*, 57–84.
Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63*, 87–106.
Johnson, W. (1944). Studies in language behavior: I. A program of research. Psychological. *Monographs, 56*, 1–15.
Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Lund Working Papers in Linguistics, 53*, 61–79.
Kim, J. Y. (2014). Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study. *English Teaching, 69*(4), 27–50.
Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction, 1*(1), 60–69.
Kyle, K. (2018). *TAALED: Tool for the automatic analysis of lexical diversity (beta version 1.2.4, Python-based)*.
Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and r* (2nd ed.). Oxon: Routledge.
Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307–322.
Linnarud, M. (1975). *Lexis in free production: An analysis of the lexical texture of Swedish students' written work*. University of Lund, Department of English: Swedish-English Contrastive Studies. Report n. 6.
Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *Modern Language Journal, 96*(2), 190–208.
Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan, & A. Wray (Eds.), *Evolving models of language* (pp. 58–71). Clevedon, UK: Multilingual Matters.
Malvern, D., Richards, B., Chipere, N., & Dur´an, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills, England: Palgrave MacMillan.
Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313–330.
Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for academic purposes programme. *Journal of Second Language Writing, 29*, 3–15.
McCarthy, P., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. Language. *Testing, 24*(4), 459–488.
McCarthy, P., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381–392.
P. McCarthy, (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). Unpublished PhD dissertation, University of Memphis.
Minnen, G., Carroll, J., & Pearce, D. (2001). Applied morphological processing of english. *Natural Language Engineering, 7*, 207–223.
Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing, 26*, 45–65.
Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research, 35*(1), 121–145.
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 M.A. thesis conclusions. In P. Pietilä, K. Doró, & R. Pípalová (Eds.), *Lexical issues in L2 writing* (pp. 105–125). Newcastle upon Tyne: Cambridge Scholars Publishing.
Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912.
R Core Team. (2013). *R: A language and environment for statistical computing*. URL. Vienna, Austria: R Foundation for Statistical Computing http://www.R-project.org/.
Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.
Swales, J. (2004). *Research genres: Explorations and applications*. Cambridge: Cambridge University Press.
Treffers-Daller, J., Parslow, P., & Williams, S. (2016). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*. amw009:1-27.
Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and Humanities, 32*, 323–352.
Ure, J. (1971). Lexical density and register differentiation. In G. E. Perren, & J. L. M. Trimm (Eds.), *Applications of linguistics: Selected papers of the 2nd International congress of applied linguists* (pp. 443–452). London: Cambridge University Press.
Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Honolulu: University of Hawaii Press.
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.
Yoon, H. (2017). Linguistic complexity in L2 writing revisited: Issues of topic, proficiency, and construct multidimensionality. *System, 66*, 130–141.
Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency—Variable sensitivity. *Studies in Second Language Acquisition, 27*, 567–595.

Dr. Maryam Nasseri has received her doctoral degree from the University of Birmingham where she has investigated lexical and syntactic complexity of postgraduate academic writing in learner corpora using NLP and various statistical modelling methods such as machine learning methods. She has been a book reviewer for Routledge, an examiner for the UK Linguistics Olympiad since 2016, and has received multiple grants such as ISLE2020.

Dr. Paul Thompson is a reader in applied corpus linguistics and the deputy director of the Centre for Corpus Research at the University of Birmingham. He has extensively presented at and published in international conferences and journals on corpus-based analyses of academic writing.