

A comparison of the linguistic encoding of artificial-intelligence-generated academic essays and academic essays written by MSc ALSLA 2022-23 students.

Noorin Iqbal

MSc Applied Linguistics and Second Language Acquisition, 2023

DECLARATION BY THE CANDIDATE AS AUTHOR OF THE DISSERTATION

1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.
2. I allow the Department to deposit on my behalf a copy of this dissertation in the Oxford University Research Archive ('ORA') where it shall be freely available online for use in accordance with ORA's Terms and Conditions of Use [https://ora.ox.ac.uk/terms_of_use].
3. I understand that this dissertation should not contain material that can be used to personally identify individuals or specific groups of individuals (unless permission has been obtained from the individuals) and that such material should be removed before this dissertation is deposited in ORA.
4. I agree to be bound by the terms of the ORA Grant of Non-exclusive Licence [https://ora.ox.ac.uk/deposit_agreements] and I warrant that to the best of my knowledge, making my thesis available on the internet will not infringe copyright or any other rights of any other person or party, nor contain defamatory material.
5. I agree that my dissertation shall be available for download in ORA in accordance with paragraphs 2, 3 and 4 above.

Signed [an electronic signature is sufficient]:	Enter your signature 
Date:	August 10, 2023

Table of Contents

Table of Contents 3
Abstract 5
Glossary 6
Chapter One Introduction 7
- **I. On artificial intelligence** 7
- **II. The emergence of AI in academia; on ChatGPT** 8
- **III. On ghostwriting and academic dishonesty** 9
- **IV. On forensic linguistics and authorship** 9
- **V. On adaptive comparative judgment** 10

Chapter Two Literature Review 11
- **I. Empirical studies on academic dishonesty, ghostwriting, and plagiarism** 11
- **II. Studies and practical implications of authorship** 15
- **III. Empirical studies on the use of generative AI in academic contexts** 16
- **IV. Assessing writing** 20
- **V. Concluding remarks on empirical studies** 24

Chapter Three Research Methods 26
Chapter Four Results 33
- **I. Statistical Calculations** 33
- ***II. N-grams*** 37
- **III. Adaptive Comparative Judgment** 38
- **IV. Turing Test** 39

Chapter Five Discussion 41
- **II. Statistical Analysis** 48
- **III. N-grams Analysis** 49
- **IV. Adaptive Comparative Judgment Analysis** 52
- **V. Turing Test Analysis** 54
- **VI. Limitations and Future Directions** 55

Chapter Six Conclusion 57
References 59
Appendices 70
- **Appendix A; Corpus of ALSLA essays** 70
- **Appendix B; Corpus of AI-generated essays** 93
- **Appendix C; MSc ALSLA formative assignment instruction sheet** 108
- **Appendix D; List of linguistic features** 109

Appendix E; List of n-grams.....119
Appendix F; MSc ALSLA grading rubric for summative assessments 129
Appendix G; CUREC form.....132
Appendix H; CUREC Form Approval.....145

Abstract

Background

The aim of this study was to establish whether artificial-intelligence (AI) generated essays can be differentiated reliably from human written essays in terms of linguistic features and n-gram word sequences. It also aimed to assess the quality of AI-generated essays compared to essays written by humans, and whether expert raters can differentiate between both types.

Through generative AI, people can input parameters into online software to generate different types of discourse within minutes. Though these discourses include songs, graphic art, and computer codes, one of the main concerns for educational boards globally is the use of generative AI for academic essays. Using generative AI to create essays and submitting it as one's own undermines the fundamentals of assessment, authorship, and academic integrity. Hence, it is important to conduct research on such discourse to analyze the style in which AI can write, and how similar or distinct it is from students' work.

Methods

A corpus of essays responding to a typical graduate-level essay prompt (seven written by MSc students and seven generated by AI) was analyzed for textual and authorial features that might help differentiate between the two. Specifically, (1) the Jaccard's Index (a measure of linguistic similarity and difference) was calculated for each set of essays; (2) n-grams analysis was performed to determine the frequency and purpose of reoccurring long word sequences; (3) adaptive comparative judgment of the essays by university lecturers was undertaken to assess the relative quality of the essays; and (4) a 'Turing test' by university lecturers was performed to assess whether they can identify which essays were AI generated and which were human written when they have been told that the corpus contains the two.

Results

All essays varied in terms of linguistic features, references, length, errors, subheadings, and stylistic markers. AI-generated essays contain more 5-gram word sequences than essays written by students, despite being shorter. Lecturers tend to rate student essays as better and of more depth than AI essays; and in most cases, lecturers can differentiate between AI-generated and human written texts when they know that a corpus contains both types.

Glossary

Term	Definition	Abbreviation(s) in this paper
Artificial intelligence	The ability of a machine or software to perceive and perform novel tasks which require skills such as creativity, manipulation, and reason that are typically (but not always) performed by human beings.	AI
Second language learners	People that learn a language in a formal classroom setting, after they have already acquired their native language.	L2 (learners)
MSc Applied Linguistics and Second Language Acquisition	A master's course offered by the Department of Education at the University of Oxford which explores concepts in applied linguistics, language acquisition, and pedagogy.	MSc ALSLA MSc ALSLA 2022-23 ALSLA
n-grams	A sequence of phonemes, words, sentences, symbols, numbers, or punctuation, where 'n' is replaced with the number of sequences in question. In this dissertation word sequences are discussed.	
Adaptive comparative judgment	A method of assessing essays by choosing the better essay from a pool of essays, rather than evaluating an individual essay.	ACJ
Turing test	A test designed by British mathematician Alan Turing in 1950 which requires a person to distinguish whether certain discourse has been produced by a human being or by a computer.	

Chapter One

Introduction

The emergence of generative AI is seeing a breakthrough in the use of technology in both professional and personal spheres. OpenAI's *ChatGPT*, which was launched to the public at the end of November 2022, can generate informative texts within seconds based on prompts one inputs into it. Through this feature, ChatGPT garnered more than 100 million users within two months of its launch, as people started to test out different text types of various topics and languages from food recipes and life advice to cover letters and academic essays (White, 2023). However, this feature also led to concerns across academic boards with certain universities immediately banning ChatGPT and other forms of generative AI to combat academic dishonesty when students used such software to complete their assignments (Castillo, 2023). As generative AI has the potential to transform the landscape of educational assessment, it is important to investigate its products and their effects on academia. Such research can facilitate ways to help sustain academic honesty yet integrate such technology to maximize both student and educator needs in the classroom, while safeguarding authorship and identity.

I. On artificial intelligence

The term 'artificial intelligence' was coined in 1956 at the conference, 'Dartmouth Summer Research Project on Artificial Intelligence', in New Hampshire in the United States by John McCarthy, a professor of mathematics (Moor, 2006). AI can be defined as the ability of a machine or software to perceive and perform novel tasks which require skills such as creativity, manipulation, and reason, that are typically (but not always) performed by human beings (McCarthy, Minsky, Rochester & Shannon, 1955). Though AI is colloquially attributed to physical machines, it is a broad term that encompasses different meanings and criteria (Brighton & Selina, 2010). The fundamentals of AI are interdisciplinary, stemming from fields of philosophy, linguistics, cognitive science, and engineering. In AI, an 'agent' is used to refer to an entity that can execute intelligent functions. There are physical agents, such as computers and robots, and there are virtual agents, such as computer software.

Searle (1980) made a distinction between Weak AI and Strong AI. Weak AI is the notion of programming agents to carry out tasks normally a human would. An example of a Weak AI agent is Siri, a virtual assistant designed by Apple Inc. for Apple devices. Siri performs tasks for its users upon request, such as setting alarms and reminders (Apple Incorporated, 2023). Strong AI is a hypothetical concept of AI agents developing consciousness, emotion, or thought. This theory has been a central theme in science-fiction literature and film for over 50 years. Some examples are seen in; *2001: A Space Odyssey* (1968) wherein AI 'HAL 9000' turns

off life support of its hibernating human co-workers in a space program to protect itself (Kubrick, 1968); *Ra One* (2011) where video game character 'Ra One' physically manifests itself into a shapeshifter to kill a teenager who was provoking it (Sinha, 2011); and *Black Mirror* (2013) wherein a woman orders an android replica of her late husband, but soon becomes uncomfortable and realizes she cannot get rid of it (Brooker, 2013).

Several breakthroughs in the realm of AI have occurred over the last few centuries, such as; the invention of 'ELIZA' by Weizenbaum in 1966, which was the first chatbot allowing humans, unbeknownst to them, to interact with a computer software which posed as a therapist (Salecha, 2016); the invention of 'Deep Blue' in 1996, a computer that played chess and beat the then-world champion Garry Kasparov in 6 timed chess matches (Yao, 2022); and the creation of 'Roomba' in 2002, a vacuum cleaner which can navigate itself around the buyer's house using sensors (Kerr, 2013). This dissertation examines a product of Generative Pre-trained Transformer 3 (GPT-3), a virtual agent (Weak AI), which generates texts similar to that of human beings. It is important to conduct research on this topic in applied linguistics with school and university students across the globe using AI to generate essays for their assignments (Waugh, 2023). Institutions have been investigating technological solutions to prevent this issue from surging.

II. The emergence of AI in academia; on ChatGPT

Students are now able to either use free trials of software or buy subscriptions to have essays generated within minutes, even seconds. Students may find subscribing to AI cheaper than hiring ghostwriters. Students may use AI to produce multiple essays in a very short amount of time in comparison to using ghostwriters. ChatGPT is one of the most renowned software used by students to generate academic essays (Mohmad, 2023). The software was developed by the organization 'OpenAI' with the Beta version released to the public in November 2022. By January 2023, ChatGPT had over 100 million users, which made it historically the fastest-growing application in the world (Hu, 2023). ChatGPT generates multimedia ranging from essays and cover letters to musical compositions and graphic designs.

There were certain limitations established, such as ChatGPT providing inaccurate information at times or not providing sufficient information to a given prompt (Marr, 2023). Nonetheless, ChatGPT and its competitors gained popularity amongst school and university students. Academic institutions began developing their own regulations for or against the use of generative AI by students. As of February 2023, University of Oxford has prohibited the use of generative AI for written assessments and examinations until further notice (University of Oxford, 2023). At Stanford University, the computer science department has prohibited

generative AI for students learning programming codes, however, linguistics professors have permitted it to be used as a referenceable source (Stanford University, 2023). University of Cambridge has allowed students to use ChatGPT, though not for written coursework or exams (Stephens, 2023). Since generative AI is a novel invention, there are several discussions and studies taking place at academic institutions to integrate generative AI in students' learning, while keeping academic honesty in exams and assignments.

III. On ghostwriting and academic dishonesty

Ghostwriting refers to the intentional act of producing texts anonymously for another person, who will then be named the author of that text. Ghostwriting occurs in different industries such as music (for example, for song lyrics), fictional literature, or politics (for example, for autobiographies). Any author that is credited during or after publication is not considered a ghostwriter. Several authors have denied allegations of ghostwriting, such as R.L. Stine, who was accused by publishing company *Scholastic Inc.* of having hired ghostwriters for many books in his series 'Goosebumps' (Saunders, 2019).

This dissertation focuses on AI text generation as a form of academic dishonesty, rather than traditional academic ghostwriting wherein a student hires a person to write their assignments. Ghostwriting is a form of academic dishonesty as it is the act of claiming someone else's work as one's own. The debate surrounding ghostwriting and academic dishonesty much revolves around ethics and morality of academia, credibility, and learning. This dissertation, however, does not concern with the ethics or reasons behind ghostwriting, academic dishonesty, and the education system.

IV. On forensic linguistics and authorship

One potential approach to addressing the concern amongst educators is to examine the linguistic patterns in AI-generated and human-authored texts to assess whether there are indicators in the discourse that would help differentiate one from the other. An approach called forensic linguistics has resources and theoretical framework to help analyze such discourse. Forensic linguistics is a subfield of linguistics which encompasses the study of language in a range of legal contexts. A few examples of evidence used in forensic linguistics are suicide notes, courtroom minutes, and social media posts. For example, the forensic linguistic analysis of British author Virginia Woolf's suicide notes suggest bipolar disorder as one of the main reasons behind her suicide, rather than the initial verdict of World War 2 as a motive (Malini & Tan, 2016). Analyzing online comments from a forensic linguistic perspective of Johnny Depp and Amber Heard's trial in 2022 suggests that impoliteness, sarcasm, and patronization against Heard were some of the main comments trending on the internet during the trial (Jabbar & Al-

Aadili, 2022). Though authorship is a broad term, in forensic linguistics, it is concerned with identifying authors via a close inspection of their writing styles. In this dissertation, to be an 'author' does not mean a published author, but rather, one who writes in general. There are different types of authorship analysis studies in forensic linguistics, however, this dissertation focuses on authorship verification. Authorship verification is when a linguist determines the likelihood of two or more texts having been written by the same author, irrespective of the author's identity.

Two of the main underlying assumptions of authorship analysis are auteurism and idiosyncrasy. Though 'auteurism' is commonly used in film studies when analyzing cinematic elements that are unique to a certain director, it can be used as a generic term to describe authors and their unique styles. Idiosyncratic language refers to linguistic features in a person's written or spoken language that are unique to that individual (Coulthard, 2010). These features do not have to be linguistic oddities, but rather, any feature that an author tends to use frequently. For example, an author may tend to use '&' instead of 'and' or prefer collocations such as 'time and time again' as opposed to 'time and again'. It is important to note that idiosyncratic linguistic features may not always be consistent over every single text produced by a given individual. However, in authorship analysis, these features tend to be consistent enough in an author's body of work to allow forensic linguists to attribute a certain text to the suspected authors in question.

V. On adaptive comparative judgment

This dissertation will also have a section focusing on adaptive comparative judgment (ACJ). The concept of ACJ was first seen in the 1920s (Thurstone, 1927), however, it is now being reintroduced into academia. ACJ is a method of assessing essays wherein teachers are presented with two essays at a time and have to decide which essay is better. The term 'better' is open to interpretation by the teacher, as there are no criteria or rubrics. This method assumes that it is more reliable to judge something when it is presented with another sample of the same thing, as opposed to in isolation (Thurstone, 1927). The repetition of the ACJ task by the teacher until all essays are judged increases reliability of the results. Studies by Jones et. al (2015) and Metzgar (2016) suggest that ACJ can also be a reliable assessment method for different discourses and disciplines such as calculus (problem-solving questions) and business (real-life application prompts). In this dissertation, three lecturers at the University of Oxford were asked to perform an ACJ task of a corpus of 14 essays; 7 written by Master's students and 7 generated by AI. This task was then used to assess the relative quality of these essays.

In the following chapter, I discuss relevant literature.

Chapter Two

Literature Review

As the emergence of generative AI in academia is a new occurrence, there is a paucity of research on the topic. I first examine literature on ghostwriting and plagiarism, which sets the stage on written academic dishonesty, which precedented the use of AI to cheat. Next, I explore literature on authorship analysis in real-life forensic cases to examine the practical applications of forensic linguistic theoretical framework. I then follow up with literature on the emergence of generative AI in academia, and lastly discuss literature on writing assessments which is pertinent to the ACJ part of my dissertation.

I. Empirical studies on academic dishonesty, ghostwriting, and plagiarism

Although generative AI such as ChatGPT is a recent phenomenon, ghostwriting has been around throughout history. One of the earliest cases of recorded academic dishonesty is cheating that occurred in the Civil Service examinations in China under the Sui Dynasty starting in 600 AD. Ghostwriters sold essays to examinees since the exam questions were similar every year, even though publishing such essays was illegal. In 1858, examinees found guilty of cheating in the Civil Service exams, along with their examiner, were sentenced to death (Lang, 2013). The severity of the repercussions for academic dishonesty has reduced overtime, and academic dishonesty is not a legally punishable offense in many countries. However, there are still some countries where one can be imprisoned if caught cheating in a public exam or competition, such as in France, under Article 2 of the December 23, 1901 law for fraud in public examinations and competitions (Government of France, 2023). There are different repercussions for cheating on school and university examinations, which are unique to each institution, however, most result in some form of penalty. For example, suspension, expulsion, course failure for the student, or, in a more positive scenario, a re-sit of the exam for the student.

Sousa-Silva (2019) emphasizes the importance of knowing different definitions and criteria for plagiarism. For example, the Collins English dictionary defines plagiarism as using another person's idea or work and pretending that one came up with it. Similarly, yet not identically, Longman Dictionary defines plagiarism as using another person's ideas, work, or words, and pretending that it is one's own. Though the difference may seem minimal, Sousa-Silva (2019) argues that having different definitions makes it more difficult to penalize in academic settings. However, it is also important to note that (as he also later admits), establishing one universal definition of a multi-dimensional act is difficult. Sousa-Silva (2019) further argues that in a forensic linguistic setting (i.e., real life crime cases) it is critical to know

whether a text has been plagiarized or not due to the potential life-altering nature of crime. Consequently, he asserts that academic institutions need to be more careful to avoid misjudgments of plagiarism cases. He provides a chart of different types of academic dishonesty with respect to forensic linguistics (see figure 2.1.). According to this chart, using generative AI would be considered "other types of academic dishonesty" (Sousa-Silva, 2019, p. 2329). He also claims that students with weaker academic writing skills, particularly those who are non-native users of English are more likely to plagiarize (also see Williams & Carroll, 2011).

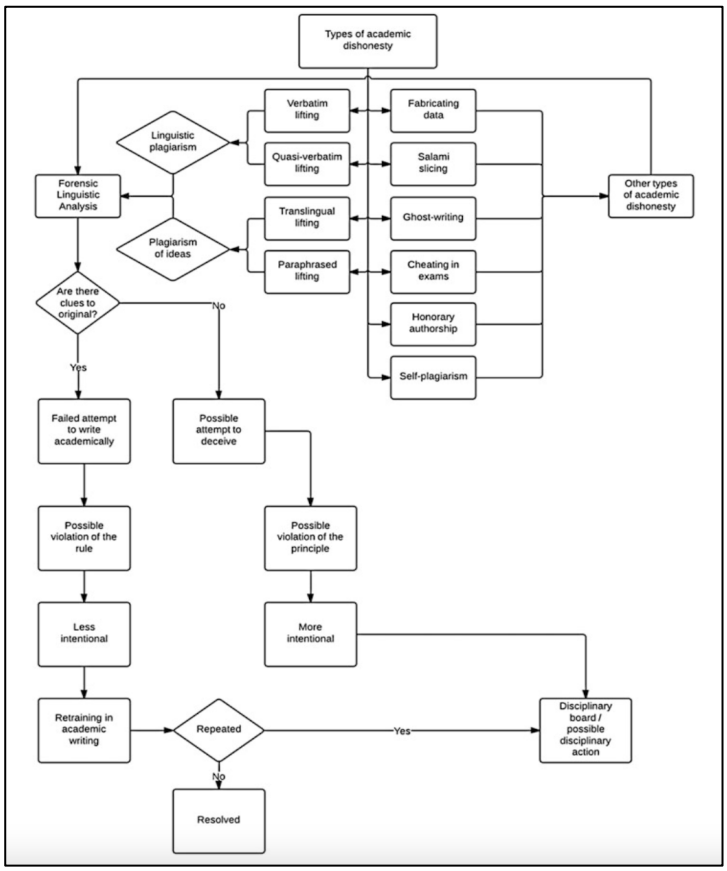


Figure 2.1. "Types of academic dishonesty" (Sousa-Silva, 2019, p. 2329)

Sousa-Silva (2014) explores different types of plagiarism using a forensic linguistics approach, discussing plagiarism strategies that do not solely require copy-pasting. He claims that identifying the degree of intention behind the act is one of the main obstacles facing plagiarism cases in academia. Sousa-Silva (2014) discusses limitations with current anti-plagiarism software, such as the software only working on n-gram word sequences and being unable to detect plagiarism if a few words have been changed (resulting in a new sequence). N-grams are sequences of varying elements such as word sequences, special character sequences, or phoneme sequences, where the 'n' is replaced by the desired number; e.g., a 2-

gram word sequence would be ‘thank you’. Another limitation is that not all plagiarism software can detect translated phrases or word reordering. Additionally, one of the major issues is the presence of ‘false positives’ (Hill & Page, 2009; Garner, 2011; Weber-Wulff, 2019). A false positive in plagiarism is when a text is marked as plagiarized by a detection software but is not actually plagiarized. For example, an author could directly quote a sentence from a textbook in their text with appropriate citations, yet a detection software would flag it as plagiarism because the software only detects the identical word sequence, and not the citation. In another study, Kleerekoper and Schofield (2019) found that the rate of false positives in plagiarism software detecting submissions for coding was between 15% and 64%. Sousa-Silva (2014) also claims that there is still the need for human input to check the work of plagiarism detection software.

Sousa-Silva (2014) analyzed a corpus of essays written in Portuguese by postgraduate students in two universities in Portugal. He claimed that due to more complex morphology and syntax in Portuguese (for example, more word inflections), it would be more challenging for plagiarism software to detect plagiarism with respect to translation, word reordering, and word substitution. He also did a manual comparison between the corpus and the original sources (from which students plagiarized), which the lecturers provided. Sousa-Silva (2014) found that there were fewer cases of word substitution in the corpus than cases of word reordering and translation.

Sousa-Silva (2014) also discusses that it is not the responsibility of a linguist to determine the intention behind plagiarism, agreeing with Eggington (2008), who argues that though linguists can define and help detect such fraud, they do not have tools to detect intention. The plagiarism software Sousa-Silva used did not identify all cases of plagiarism efficiently, though the exact percentage is not reported. However, he was able to manually detect all cases of plagiarism. Manual detection requires many more resources and a lot more time, making it an unfeasible and inefficient solution for school/university teachers. He concludes by asserting the need for better software to detect these types of linguistic occurrences in plagiarized texts. Current plagiarism software is not coded to detect much more than exact sequences of words. This problem will be increasingly prevalent in the use of AI-generated texts, where AI does not copy and paste strings of words from other sources, and rather generates them. Thus, it is important to conduct research on the linguistic features of AI-generated text, in order to understand AI’s ‘writing style’. Establishing whether AI has a ‘writing style’ that is distinct from human writing styles may help in the detection of nefarious use of this kind of software.

Singh and Remenyi (2018) investigated cases of plagiarism and ghostwriting, condemning both forms of academic dishonesty, and proposing solutions for academic institutions to implement. They claim that plagiarism and ghostwriting differ in their nature, but both have the same outcome where an individual passes someone else's work as their own. The authors discuss why plagiarism is detrimental in academia for the perpetrators and for the institution. For example, plagiarism minimizes learning on the part of the plagiarist, denies the plagiarist to receive feedback on their own work, and is deceitful, harming the plagiarist's future credibility (Singh & Remenyi, 2018). They further claim that anti-plagiarism detection software, such as Turnitin.com, is useful to detect plagiarism. However, it comes with its own limitations and does not solve the underlying issue behind cheating. They mention that merely failing a student or asking them to redo the assignment may not be enough of a solution, and that more disciplinary action should be taken by institutions.

Singh and Remenyi (2018) discuss different media and text types produced by ghostwriters (for example, computer codes and lab reports). They discuss how over time, online ghostwriting companies modified their products (texts) to be undetectable by anti-plagiarism software. The authors also claim that there is a debate on whether ghostwriting is considered plagiarism or not since there is no actual text being 'stolen'. Some researchers consider it to be a form of plagiarism (Lacasse & Leo, 2010; Ali & Alhassan, 2021), while others claim it is considered a form of plagiarism in certain contexts (Lines, 2016; Yi, Nemery & Dierick, 2020). Singh and Remenyi (2018) assert that unless evaluators are familiar with their students' writing style and knowledge, it is difficult for them to detect ghostwriting.

Furthermore, they explore the importance of dissertations as a degree requirement at undergraduate and postgraduate levels in academia. Dissertations are difficult to write, given the length and the nature of the assignment (involving research and technical writing). Thus, many students hire editors to help check their dissertations. However, the amount of editing and writing an editor does can also label them as a ghostwriter. The authors also discuss the ethics of ghostwriting, concluding that ghostwriting companies do not value academic integrity. Singh and Remenyi (2018) argue that institutions where such academic violations occur lose credibility. Students are not honest about their intellectual capabilities, and their degrees do not reflect their true academic skills.

Singh and Remenyi (2018) also criticize the foundations of universities and academia, to understand the underlying problems behind academic dishonesty. For example, classes are large, and thus, busy professors can resort to giving their teaching assistant the job of grading papers, minimizing the relationship between the student and professor. Subsequently, this

would allow students to get away with using ghostwriting, as their professor would be unaware of their writing style. Furthermore, they claim that universities need to emphasize the importance of long-term learning, rather than grades, to help reduce cheating. Dissertation supervisors must offer careful supervision to their supervisees.

II. Studies and practical implications of authorship

Due to the complexity and dynamic-nature of language use, it is not feasible to assign a person a 'linguistic fingerprint' (Danielewicz-Betz, 2012). However, in different contexts, there are linguistic features that are attributed to one individual, crediting them with authorship of a certain text. There have been a handful of famous cases where authorship analysis has led to identification of criminals. One of the most famous examples is the case of the 'Unabomber'. Between the years 1978 and 1995, in the US, a man sent several bombs through post. In 1995, there was a 35,000-word document sent to 6 national publishers from a man referring to himself as the Unabomber. He claimed that the bombs would stop being deployed if this document was published (Coulthard & Johnson, 2007). After performing authorship and other linguistic analyses, forensic linguist Robert Shuy, found that the Unabomber must have had a background of the natural sciences, lived in Northern California but was from Chicago, and had a religious upbringing. After David Kaczynski's trial (the man subsequently convicted of posting these letter-bombs) in 1996, these characteristics were proven accurate in court (Leonard, Ford & Christensen, 2017). Though forensic linguistic analysis can contribute to evidence in court cases (such as the text messages analyzed in the case of Jenny Nicholl's murder by David Hodgson, (Grant, 2010)), it is difficult to use such evidence to establish guilt. This notion also refers to Eggington (2008) who claimed that it is not a linguist's task to decipher intention.

In less macabre contexts, linguistic features have been used in a broad range of authorship research. Kestemont (2014) discusses the importance of function words, such as 'the', or 'those', in authorship analysis. This study is distinct from other research on authorship analysis, which tend to focus on content words. Kestemont (2014) claims that function words are used more frequently than content words, and are also less likely to be affected by the topic of the text. He asserts that due to the limited number of function words, they are more likely to be re-used by different authors in different texts, allowing for comparison. Thus, when there is a distinctive use of a certain set of function words, it can contribute extensively to authorship analysis. Certain types of function words, such as pronouns, have also contributed to authorship analysis cases (Burrows, 1987).

Kestemont (2014) discusses the act of using function words subconsciously while writing. Function words are limited and needed by default in writing to form sentences, unlike

content words which are topic specific. Thus, in criminal cases of forgery and identity theft, any distinct function word or words must be imitated by criminals as well. Kestemont (2014) then discusses psycholinguistic aspects of function words, such as when the brain initially, automatically skips over a double 'the' in a sentence (for example, 'Paris in the the spring') as opposed to when there is a double content word (for example, 'Paris in the spring spring') (Aronoff & Fudeman, 2004). Drawing on similar research (Rawlinson, 1976), Kestemont (2014) argues that function words do not attract the same amount of attention as content words do. Though this notion is from a reader's perspective, Kestemont (2014) argues that it is more difficult to imitate the patterns of function word use compared to content words, as they are less salient.

There has been research on forensic linguistic methods applied to languages other than English as well. For example, Raza, Athar, and Nadeem (2009) explored the use of n-grams for authorship analysis of Urdu poetry, developing a computer software that would aid with authorship attribution. For her master's dissertation, Houle (2013) analyzed computational tools for authorship analysis of French texts for judicial contexts. Phani, Lahiri, and Biswas (2015) examined n-grams and linguistic features to attribute authorship to a corpus of 3,000 texts in Bengali. Ahmed (2018) analyzed the efficacy of using different linguistic features (such as stems and diacritics) in authorship analysis on essays written in Arabic on Al Gazhali's work. Kallimani, Chandrika, Singh, and Khan (2020) performed different authorship analysis techniques on works in Hindi, to compare the accuracy between formulae. Such research is important in the field of forensic linguistics, as AI can generate texts in over 100 languages.

Furthermore, authorship analysis has also been conducted on different text types of varying lengths and purposes, such as for raising awareness of phishing, analyzing topics of suicides, or identifying the gender of an author. These studies include analysis of Tweets (Macleoad & Grant, 2011; Belvisi, Muhammad & Alonso-Fernandez, 2020), emails (Chen, Hao, Chandramouli & Subbalakshmi, 2011; Seifollahi, Bagirov, Layton & Gondal, 2017; BinSaeedan & Alramlawi, 2021), text messages (Saevanee, Clarke & Furnell, 2011; Ragel, Herath & Senanayake, 2014; Silessi, Varol & Karabatak, 2016), and suicide notes (Shapero, 2011; Spasić, Burnap, Greenwood & Arribas-Ayllon, 2012; Sudjana & Fitri, 2013).

III. Empirical studies on the use of generative AI in academic contexts

With the recent emergence of generative AI, there are wide literature gaps pertaining to this topic, specifically within applied linguistic contexts. There are a very limited number of published papers on the use of AI in academia, with the published ones being in a report format. That is to say, they describe features of generative AI that may be helpful within a teaching-

learning context, however, there are no experimental studies as of June 2023 that suggest the advantages and disadvantages.

For their undergraduate thesis, Olsson and Engelbrektsson (2022) investigated whether teachers in their department at their university (Halmstad University, Sweden) could differentiate between essays written by AI and human beings. Olsson and Engelbrektsson (2022) used the AI software 'Davinci' for their project, which also ran on GPT-3 and was a product of OpenAI. They generated 10 texts (average number of words: 415) based on 4 different prompts through *Davinci* on IT-related topics. These questions were part of a sit-down exam that was previously administered in an undergraduate course in January 2022. After obtaining permission from the lecturer, they used these exams as the corpus for texts written by human beings. 32 teachers from their department were asked to guess which of the essays in the corpus were written by humans and which by AI. Olsson and Engelbrektsson (2022) claim that one limitation of their study is that the teachers knew they were presented with an AI essay and a human one, which induced bias for them to look for more unnatural sentences.

Olsson and Engelbrektsson (2022) found that 49.7% of the guesses on AI essays were incorrect, meaning teachers did not guess correctly that the given text in front of them was generated by AI. This finding suggests that the correct guessing of AI texts was no more accurate than a coin flip, alluding to the threat of AI generated texts in academia, which can go undetected by teachers. Olsson and Engelbrektsson (2022) suggest that computer-based detection models are necessary to help tackle the problem of AI generated essays, which certain companies are now developing such as Turnitin.com (Chechitelli, 2023). As the authors were undergraduate students of IT rather than linguistics, they critically analyzed different computer models for AI text detection from a computer science perspective. The authors also suggest alternative exam methods such as a follow up oral exam for a written paper, which would add credibility and test the student's knowledge. They also suggested using illustrations as an alternative exam method because at the time of their project, AI was limited to generating text only. However, a year later, AI software can also generate artwork, hence, this solution is no longer sufficient.

Kohnke, Moorhouse, and Zou (2023) explore how ChatGPT can be useful in a foreign language learning classroom. They claim that ChatGPT can provide definitions of words to learners, with sentences and examples, and do follow-up conversations. This feature might facilitate vocabulary acquisition. ChatGPT can also provide conversations in different languages, such as Chinese, which Kohnke et al. (2023) argue to be useful with respect to using the native language (L1) in a second language (L2) classroom. Research on L2 acquisition has

suggested benefits of using the L1 in the L2 classroom for different reasons such as producing better writing samples (Yigzaw, 2012) and better recall of vocabulary (Lee & Macaro, 2013). Kohnke et al. (2023) argue that students can use ChatGPT to produce different text types such as dialogues and advertisements, however, they fail to mention the benefits of these activities. In traditional academia, this would be considered an academic violation as the students would be submitting work that they did not produce.

Furthermore, Kohnke et al. (2023) argue that ChatGPT can produce notes on vocabulary words for activities such as short stories, which can be beneficial for the students. This argument assumes that students will be using ChatGPT products in learning rather than assessment. The authors also suggest that teachers can use ChatGPT to generate reading-comprehension questions. Though this is a useful tool, students may also use it to generate questions beforehand and get an idea of what may come in exams. However, it can also be argued that having access to past papers is beneficial to students. Different annual exams have repositories of question papers that can be found online for test-takers to practice. Some examples of these are the Union Public Service Commission exams for the Indian Civil Service, or the International Baccalaureate Diploma Programme exams for schools around the world (Government of India, 2023) (International Baccalaureate, 2023). Tyagunova and Greiffenhagen (2021) argued that students found revising with past papers to be more time efficient for their overall academic workload, and allowed students to identify the content that is more likely to be asked in an exam. Kohnke et al. (2023) argue that developing software that can detect AI generated texts will not solve the underlying issue of academic dishonesty, and would rather turn the problem into a "cat and mouse game" (Kohnke et al., 2023, p. 8).

Kohnke et al. (2023) present limitations of ChatGPT, such as the chatbot not presenting its sources and providing inaccurate information at times, which less experienced writers may not double check. It can be argued that it is the institution's responsibility to teach students how to fact-check information they will be using for their work. Brodsky et al. (2021) conducted a study on the 'Digital Polarizing Initiative' (DPI), a curriculum designed by the American Association of State Colleges and Universities, which aims to improve fact checking instruction of online resources in universities. Brodsky et al. (2021) found that undergraduate students (course = civics) receiving the DPI method of instruction verified information more than the control group, through the technique of lateral reading. Lateral reading is when a student reads a source, then looks for other sources that fact check the original source. Other researchers, however, argue against fact checking in schools for certain subjects. For example, Maitland (2023) argues against fact checking in sex education on the assumption that gender,

sex, and sexuality are sensitive and subjective topics which students should be exposed to under supervision.

Kohnke et al. (2023) further assert that ChatGPT is culturally biased as the information in its original database is in English which the program translates to other languages. Research in the field of translation suggests that there are barriers in translation which lead to a discrepancy of cultural accuracy present in the original and translated texts (Abbasi, Zadeh, Janfaza, Assemi & Dehghan, 2012; Al-Hassan, 2013; Akbari & Shahnazari, 2014). Lastly, Kohnke et al. (2023) argue that there are certain digital competencies both students and teachers need to learn to use ChatGPT efficiently, such as being aware of the software's features, using ChatGPT for self-guided learning, and discussing the ethical issues concerning ChatGPT. These competencies attribute responsibility to both educators and students to create a learning space with helpful, integrative use of AI. These competencies have the potential to transform educational landscape, however, like any novel pedagogical method, needs research and planning for successful implementation.

Kasneci et al. (2023) present the advantages and disadvantages of using large language models (LLMs) such as ChatGPT in education for different groups of students. For younger learners in elementary school, they argue that LLMs can aid with reading, writing, and critical thinking skills through corrections and summaries these models provide. They also argue that teachers can use LLMs to produce critical thinking prompts. For older school students in middle or high school, LLMs can be used in different classes such as languages and mathematics. Kasneci et al. (2023) argue that for university students, LLMs can be used to assist with research, writing, and critical thinking skills, by providing the students with information, ideas, and resources they can use for further, in-depth research. For e-learning classes, LLMs can be used to improve student engagement in class as LLMs can provide real-time feedback and guided discussions. Additionally, they argue that LLMs can be used for self-paced learning by providing students with practice questions at their own unique level of knowledge. Kasneci et al. (2023) assert that LLMs can be used in special needs education as well, through speech-to-text (or vice versa) features. However, they argue that an educator should monitor this tool to adapt it to the students' specific learning disabilities. Furthermore, they claim that LLMs can be used in teachers' professional development of technological tools, management, and jargon specific report writing.

Kasneci et al. (2023) suggest many limitations with LLMs, some of which are discussed here. One of these limitations is copyright issues that may arise as AI may re-use its previous texts, leading to plagiarism. They further argue that LLMs may produce biased information

that discriminates against certain groups of people, depending on the information in its database. Thus, they suggest LLMs should be carefully programmed, and educators should be able to identify such biases in their classroom if they are using LLMs as a resource. They claim that learners and teachers can become too reliant on these LLMs for their work, which can lead to academic dishonesty and indolence. Thus, they suggest that it is important to raise awareness in classrooms and teacher training for and against the use of LLMs in education, setting clear guidelines for the students and teachers. Kasneci et al. (2023) also claim that there is a need for digital competency training of students and teachers to use the LLMs in classrooms. They raise a point that it is difficult to distinguish between AI-generated texts and texts written by students, which can pose a threat to education. This point is the focus of this dissertation, as I will examine the similarities and differences between both types of texts.

IV. Assessing writing

A. 'Classic' writing assessment

Writing assessment is another important aspect in this dissertation, as I ask lecturers to evaluate the quality of the corpus of AI and Master's students' essays. This section explores different ways a medium as complex and subjective as writing is assessed in academia. Written assessments in academia are evaluated differently depending on several factors such as, but not limited to; the purpose of the assessments and class, the level of proficiency of the students, the prerequisites for the class, the type of institution, etc. Though some researchers claim that grammar correction does not serve a purpose in writing classes (Truscott, 1996) , much research has suggested that it does to varying degrees to improve students' performance on the same and novel writing tasks (Lyster & Ranta, 1997; Villamil & De Guerrero, 1998; Ferris, 1999; Tuzi, 2004; Liu, 2008; Zhang & Zhang, 2018; Boggs, 2019).

McNamara, Crossley, and McCarthy (2010) evaluate the quality of different linguistic features of written essays to determine how they can be incorporated into writing styles for different purposes. They analyze linguistic features using 'Coh-Metrix' and 'WordNet'. Coh-Metrix is a computer software developed by Graesser et al. (2004), which allows linguists to rank texts based on pre-coded criteria of cohesion, language, and readability, through matrices. WordNet is a database of nouns, adjectives, verbs, and adverbs which are semantically grouped together and can provide information about the semantic relations between the words, unlike thesauruses (Princeton University, 2023). Their corpus of 120 essays was collected from Mississippi State University first-year students enrolled in an English writing class. These students were all native speakers of English, and had produced argumentative essays of 500-1,000 words on either one of four different prompts, and did not need references for the task.

These essays were graded and rated by tutors working at the writing center (for at least one year or more), with a rubric adapted from the Scholastic Aptitude Test (SAT) with criteria for argument, critical thinking, ideas, vocabulary, syntax, and punctuation. These tutors were given 20 essays to practice rating before the actual corpus, which is an important factor to consider in such research, as it depicts an attempt at eliminating bias from the raters. It is also important to note that the SAT has a different standard of writing as opposed to rubrics for other standardized exams or rubrics created by teachers for their institutions. Breland, Kubota and Bonner (1999) found that SAT scores are a good predictor of students' success in writing classes at a university level. Though all quality writing may possess certain features of clarity, content, etc. the level of quality is subjective based on the context. The mean score for the essays was 3.26/6 (SD = 1.23). Through ANOVA and Coh-Metrix, the authors concluded that the difference in prompts did not contribute to a difference in quality of writing, that is, the topic was not of significance in this study. It should be considered that though writing is a subjective feature of language learning, and different topics could cause a difference in quality of writing in other contexts.

McNamara et al. (2010) claimed that there is very little empirical evidence defending the general assumption that cohesion is a factor of high-quality writing. Cohesion in writing can be defined as the presence of semantic relations between words, sentences, and ideas in a piece of text which contribute to overall clarity and meaning for the reader (Halliday & Hasan, 1976). Through the Coh-Metrix results, McNamara et al. (2010) concluded that cohesion is not an indicator of good writing. Essays that had linguistic features exhibiting text difficulty and sophistication (syntactic complexity and lexical diversity) were rated higher. McNamara et al. (2010) asserted that linguistic features that are correlated with higher quality in writing may make reading comprehension more difficult, a notion they claim needs more research. They also mention that their raters were limited to tutors at a writing center for their target university, and that different evaluators will have different standards for what they consider to be quality writing.

Other new methods of evaluating writing for academia are emerging, such as automated evaluation. Zhang and Zhang (2018) claimed that the emergence of AI is transforming L2 learning, as they examined a software known as 'Pigai' and its evaluation of written essays for Chinese students learning English. *Pigai* is programmed with a scoring system based on a large corpus of standard English for criteria such as vocabulary, syntax, organization, relevancy to content, etc. The software provides detailed feedback in a 'positive tone' to the students within 1.2 seconds of essay submission. The speed of feedback is appreciated amongst students, who

claim that by the time they receive feedback from their teachers, they "have moved on" to other courses (Zhang & Zhang, 2018, p. 30). The authors claimed that *Pigai's* feedback was similar to the scores provided by the teachers marking the same essays, suggesting its accuracy. They further asserted that such computational tools should be used in education to aid both the teacher and student, especially given that teachers can monitor such software. Currently, over 6,000 schools have deployed the use of *Pigai* in China, and it has marked over 400 million essays. The researchers could have provided more demographic information on the types and length of essays they studied. They could also compare *Pigai* with other evaluation software to determine which one is the most efficient and accurate for education. However, other researchers argue against the use of automated evaluation for writing. Wheadon (2023) had school teachers, GPT-3, and GPT 4 evaluate two corpora of 312 essays written by primary school students about their ideal job, using adaptive comparative judgment. There were 156 essays from year 5 students and 156 essays from the same students on the same topic over a year later when they moved onto year 6. The teachers and AI were asked to predict the students' marks of the essays written in year 6 in comparison to their essays written in year 5. Wheadon (2023) found that GPT-4 and GPT-3 were less successful in predicting the year 6 essay marks than the school teachers (r values; GPT-4 = 0.54, GPT-3 = 0.52, humans = 0.73). Thus, he argues that AI should not be used to evaluate essays.

B. Adaptive Comparative Judgement

ACJ assumes that humans are better at assessing something when it is compared with another thing, rather than in isolation (Thurstone, 1927). This method of assessment is uncommon in academia, as most institutions have their own rubrics for marking. There have been some studies that evaluate the efficiency and reliability of ACJ as a method of assessment in academia. Verhavert, Bouwer, Donche and De Maeyer (2019) performed a meta-analysis on the factors that contribute to the reliability of ACJ. They first discussed the advantages and disadvantages of ACJ. For example, they claimed that the higher the number of comparisons, the higher the reliability of the results, due to the evaluator becoming more and more familiar with the corpus each time (and being able to judge better). Liu and Li (2012) also claimed that repetition in tasks decreases task complexity and subsequently increases the consistency and reliability of results. Bramley, Bell, and Pollitt (1998), however, argued that repetition in tasks can lead to fatigue and tediousness for the evaluator. Additionally, Verhavert et al. (2019) argue that there more research is needed to determine the minimum number of comparisons for an ACJ task to be considered reliable. The use of different media for ACJ would also result in different levels of difficulty for the assessor and consequently, variation in reliability.

Verhavert et al. (2019) argue that providing feedback for the students through an ACJ task is not feasible, since the evaluator would either have to assess the essays again to provide individual feedback, or take note of feedback throughout the task. The former method goes against the purpose of ACJ, and the latter method would interrupt the flow of the task. Many researchers have argued that students do require feedback on their work for different reasons. Mutch (2003) argues that it is important to establish the type and purpose of feedback depending on the students and the class, which can lead to students becoming more independent learners later in their study. Poulos and Mahony (2008) argued that teachers' feedback adds to credibility of the teacher, and helps students with a smoother transition from school to university. Hounsell, McCune, Hounsell, and Litjens (2008) asserted that feedback allows students to engage with their subject more critically and actively. Blair, Curtis, Goodwin, and Shields (2013) argued that teachers should provide thorough, articulate, and timely feedback for students. They argue that implementing proper feedback will help increase exam performance (such as when students are allowed to see past model answers), and lead to more positive reception of feedback by students. Contrarily, Crisp (2007) found that feedback alone has a small impact on students' performance, as many students do not actively implement the feedback in their learning (though most of them do). He argues that professors should aim to create dialogue with their students to increase the efficacy and understanding of the provided feedback. Additionally, Crisp (2007) argues that feedback should allow students to improve on other tasks as well, and not just tasks similar to the ones they were given feedback on. The main aim of ACJ, however, is to rank work from best to worst, instead of providing any other element of assessment.

Verhavert et al. (2019) studied 49 ACJ assessments with the software 'Digital Platform for the Assessment of Competences', of essay corpora from different academic institutions. They used the Scale Separation Reliability (SSR) formula which helps determine the consistency of standalone ACJ tasks as well as ACJ tasks by multiple evaluators. They found that the factor most affecting ACJ reliability was the number of comparisons per essay. To obtain 0.70/1.00 accuracy on the SSR scale, between 10 and 14 comparisons must be done of each essay, and to reach 0.96/1.00 accuracy, between 26 and 37. Initially, they hypothesized that multiple assessors could lead to less reliability as there would be more perspectives/opinions. However, their results suggested that neither the expertise nor the number of assessors was a factor that affected the reliability of the ACJ tasks. Hence, they concluded that more evaluators can be added to ACJ tasks, and they can have different years of experience. This finding is important, as researchers of other written assessments (such as

that by McNamara et al. (2010) as aforementioned) claim that grading of essays is subjective to the teacher and their expertise. Verhavert et al. (2019) conducted research on written essays only, and other media of assignments (for example, short films for film classes) should also be researched on with regards to ACJ.

There has been research on successful ACJ assessments being carried out by students as well, suggesting higher levels of learner engagement (Seery, Canty & Phelan, 2012). Bartholomew, Strimel and Yoshikawa (2019) found that ACJ can help students gain better exposure to ideas and perform better on their own tasks as they can identify higher quality work. However, they argued that this task can potentially lead to distrust of the feedback given by their peers (as they are not teachers), and cause fatigue from repetition. Seery, Buckley, Delahunty, and Canty (2019) examined ACJ tasks by 136 third-year undergraduate university students majoring in Technology Teaching Education. There were 4 assignments the students had to complete, and then perform ACJ on. These assignments were to draw designs based on different prompts (for example, modifying a car based on certain dimensions). They found that despite no explicit criteria of 'marking', certain designs were ranked higher than others by all students, which demonstrates reliability of the results. Seery et al. (2019) claimed that the students could identify when a design was of high quality, even if they performed poorly on the assignment themselves, also depicting that the expertise of the evaluator is not a factor that affects judgment. Additionally, they found that students who had poorer performance improved their own performance on the designs by the end of the semester, showing self-reflection and improvement. One factor to consider is the time that it takes to do ACJ tasks. Though in this study it was relatively shorter (<259.770 seconds), other types of ACJ tasks such as those involving reading longer essays would be a lot more time consuming.

There has been some criticism of ACJ. Kimbell (2022) argued that ACJ only allows for binary evaluations (for example, either 1 or 2), rather than adaptive evaluation (for example, how to improve the given assignment). Kimbell (2022) claims that since all educational tools have their advantages and disadvantages, they need to be modified and used in different ways. He suggests that for ACJ tasks, there is a need for a more effective algorithm that would help with suggesting adaptivity.

V. Concluding remarks on empirical studies

As aforementioned, topics such as writing and AI in education are broad, subjective, and diverse. There are several different directions further research can take on these topics. One of the main goals that linguistic and educational research should focus on is the effects of using generative AI in classrooms as an integrative tool. If the future of learning inevitably involves

AI, more comparative studies need to be conducted on the advantages and disadvantages of different LLMs and generative AI. Additionally, it must be noted that it will be more difficult to implement AI in classrooms of several institutions with substantially less funding, as it is an expensive tool that requires students to have access to laptops and the internet, which not all institutions have. In the next chapter, I discuss research methods of this dissertation.

Chapter Three

Research Methods

The main objective of this dissertation is to establish the key linguistic differences and similarities, if any, between writing generated by AI, and texts written by Oxford Master's students on the same topic. There are two main sections in this dissertation: the linguistic analysis of a corpus, and the input of lecturers at the University of Oxford in judging the corpus. The main research questions this dissertation aims to answer are;

1. What are the similarities and differences in the linguistic encoding of texts written by University of Oxford Master's students and texts generated by artificial intelligence?
2. According to lecturers at University of Oxford, which texts in the corpus are better than others?

    2.1 Are lecturers at the University of Oxford able to differentiate between essays generated by artificial-intelligence and those written by students from the corpus, based on their intuition and experience, without the use of external, computational aid?

One of the main first steps in this project was to narrow down the scope of the research, due to the nature of the dissertation (i.e., the time and word limit restrictions). I narrowed down the human participants to my classmates in the MSc ALSLA degree, class of 2022-23, at the Department of Education at the University of Oxford. While AI can generate texts in other languages, the MSc degree is instructed and assessed in English, therefore, the language of the texts for the corpus was English by default. The MSc ALSLA program requires each student to submit three types of texts during the academic year. These texts are; one formative assignment at the beginning of the course (1,500 words), six summative assessments for each assessed class (2,500 words each), and one dissertation at the end of the course (15,000 - 20,000 words). ALSLA students did not receive a formal grade for the formative assignment, and rather only feedback which would help them understand the requirements for the summative assessments. Hence, the formative assignment was chosen as the corpus for this project.

Since the formative assignment was the first formal text MSc ALSLA students had to submit, the writing styles amongst the corpus were expected to be varying. The MSc ALSLA 2022-23 cohort comes from a variety of different undergraduate, work, cultural, and linguistic backgrounds. However, there are certain language-specific and academic requirements that students enrolled in this course must meet to get accepted onto the program. In terms of academic requirements, University of Oxford accepts applicants with a first class or strong upper second class (UK), or a minimum GPA of 3.60/4.00 (US or equivalent). This requirement

depicts the high academic ability required for this course, due to the nature and rigor of the content and assignments. Thus, it can be concluded that students in the MSc ALSLA 2022-23 cohort have the capability and skills necessary to produce written assignments of a high standard, regardless of their diverse backgrounds. Additionally, the English proficiency requirement of this course is high. Non-native speakers of English are required to demonstrate their English language proficiency through accredited, standardized language tests taken up to two years prior to the course start date. Native speakers of English and non-native speakers who studied in an institution where the language of instruction was English do not have to provide evidence of their proficiency. The written component scores of the standardized test requirements for MSc ALSLA are 7.0/9.0 for the International English Language Testing System (Academic), 24/30 for Test of English as a Foreign Language, and 185/210 for C1 Advanced and C2 Proficiency (University of Oxford, 2023). Thus, even though the MSc ALSLA cohort have varying linguistic backgrounds, their English language proficiency does have a minimum standard.

In terms of the AI software used to generate essays, Moonbeam (https://www.gomoonbeam.com) was chosen due to the free trial that was available for the public to use after creating an account. Due to its increasing popularity, the first choice for the AI software was ChatGPT by OpenAI. However, at the time of the data collection, ChatGPT was still in Beta mode (developmental stage), and not open to the public to use. ChatGPT had a worldwide waiting line to use, thus, I did not choose it, as it would not give me data in time for the dissertation. I discovered the software Moonbeam through a Twitter thread, and signed up for free. The software has a free trial period of 7 days (before a monthly subscription of $49.99), which I utilized since I did not apply for any funding for the dissertation. I then did a pilot test of essay generation with the same title as the formative assignment. The software has an AI bot in its user interface which allows the user to ask the software about the essay it has generated to make changes. I asked the bot if it could make the essay longer, however, it said that it could not, so I analyzed the length as it was originally generated (Moonbeam, 2023).

For the data collection stage, after ethical approval, I created a Microsoft Forms (MS) form with multiple boxes for the ALSLA cohort to copy and paste their essays into, as one text box had a word limit of 600 words. A few limitations of using an MS Form for essay submission was the formatting; if a student had a table in their essay, the table borders did not copy onto the form, and if a student had any text in bold, that also did not copy onto the form. The MS Form also had a consent button at the bottom for the participants' consent. I sent out the form through an email, with instructions on how to submit one's work anonymously, removing all

personal data from the essays such as candidate numbers, supervisor names, etc. I also mentioned that participants could withdraw from the experiment anytime, however, they would have to come forward to identify their essay, as every submission was anonymous. No participants withdrew. I invited 23 students to submit their essays, but seven agreed. However, with the qualitative data analysis methods I aimed to use, 7 essays were sufficient. I closed the submission portal by the end April 2023, and generated 7 essays from Moonbeam to keep the sample size balanced from both 'populations'. For essay generation by Moonbeam, I followed these steps on their website (see figures 3.1. – 3.3. below);

a. I logged in with the email I used to create the account.
b. I selected the pertinent template out of the range: "Academic Essay".
c. I entered the following details into the template;
    i. Title: "What does 'lexical richness' tell us about an L2 learner's vocabulary knowledge?
    ii. Keywords: lexical richness, L2 learning, vocabulary knowledge, second language acquisition.
    iii. Who is the essay for?: Academic students and teachers.
d. Moonbeam then generated a structured template of the essay. I clicked on the button to proceed, and waited for less than 1 minute for Moonbeam to generate the essay.
e. I then copied and pasted the essay into my data collection document, and repeated the steps until I had 7 essays. I did not alter any of the essays, and used the first 7 essays I generated, to eliminate bias of reading and selecting which essays are more suitable for my analysis.

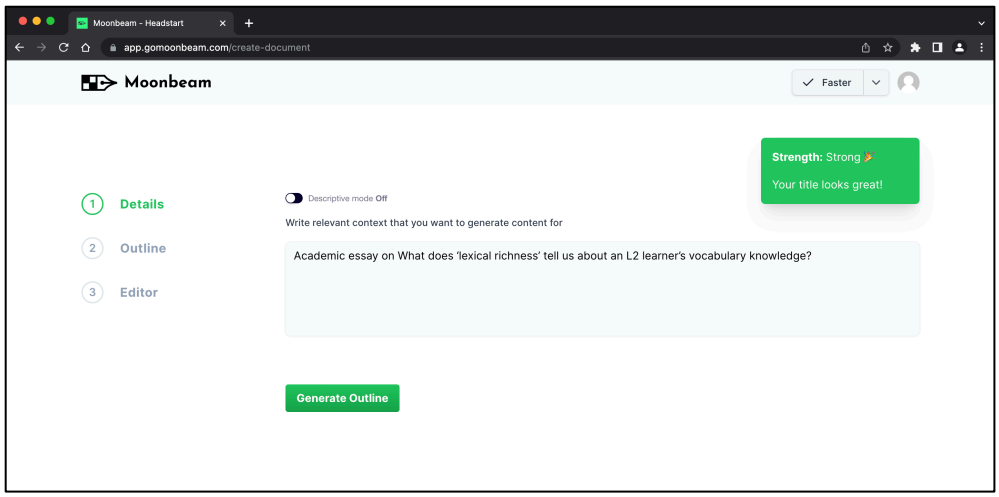


Figure 3.1. Screenshot from Moonbeam's website of generating an AI essay (Moonbeam, 2023).

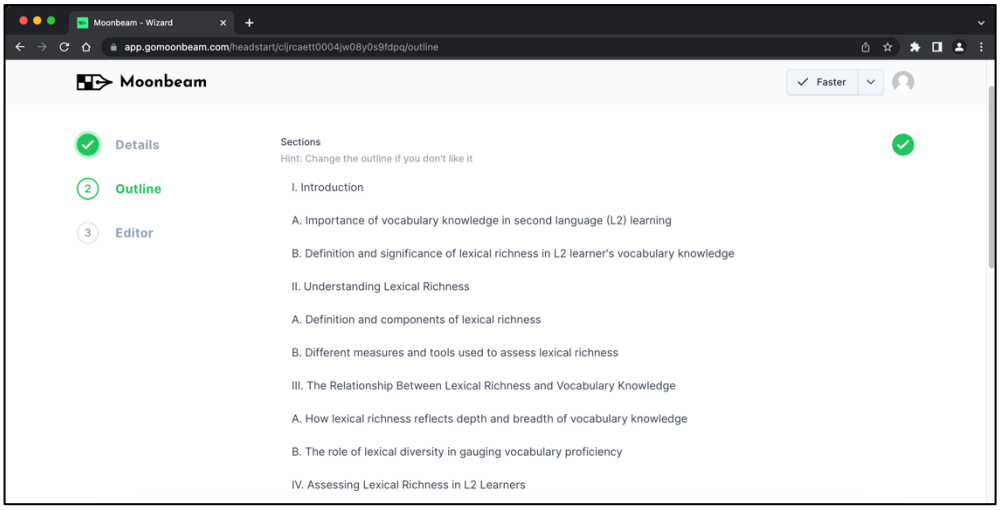


Figure 3.2. Screenshot from Moonbeam's website of generating an AI essay (Moonbeam, 2023).

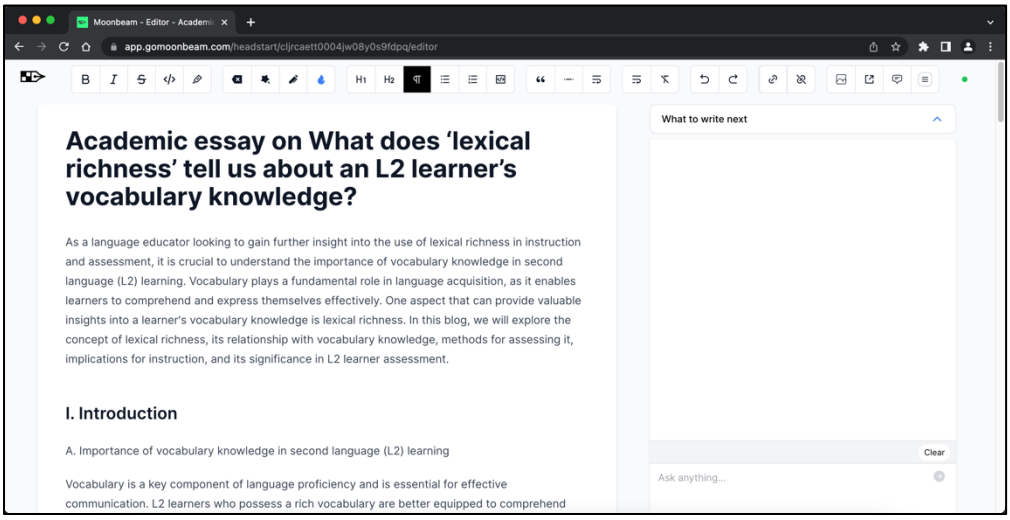


Figure 3.3. Screenshot from Moonbeam's website of generating an AI essay (Moonbeam, 2023).

After I had the data compiled, I created a list of linguistic features present in all 14 essays. For this stage, I used inductive coding as part of qualitative data analysis, as the data I collected was not previously used in other research papers, and consequently, there was no already-available features list in literature that I could have replicated. With respect to grammatical and semantic accuracy, and essay structure, I extracted 62 unique features from ALSLA essays, 9 unique features from AI essays, and 20 features present in both. Though this list is extensive, it is not exhaustive, and there can be more linguistic features extracted. However, due to the limiting nature of this dissertation, I decided that 91 total features were sufficient for in-depth data and statistical analysis. In forensic linguistics, depending on the length of the texts and the purpose of the analysis, the number of features can vary. For example, when examining workplace emails, Cho (2010) analyzed 12 linguistic features in his corpus.

These features are used in the statistical analysis Jaccard's Index to evaluate the similarity (or lack thereof) between the essays. In Microsoft Excel, I created two matrices of the features list for each population, and added binary labels of 1 and 0 if an essay had the feature or if it did not (respectively). Figures 3.4. and 3.5. below show an extract of the features table. These matrices were used to run the statistical test, Jaccard's Index.

	A	B	C	D	E	F	G	H
1	Feature	H1	H2	H3	H4	H5	H6	H7
2	Establishes common applied linguistics abbreviation to use later	1	1	0	1	1	1	0
3	First person plural pronoun	1	1	1	0	0	0	1
4	Error in choice of article	1	1	1	0	0	0	0
5	Gives essay a third person neutral singular pronoun	1	0	0	0	0	0	0
6	Establishes dates of literature used	1	0	0	0	0	0	0
7	Erorr in use of comma	1	0	0	0	0	0	0
8	Error in citation	1	0	0	0	0	0	0
9	Run-on sentence	1	0	1	1	0	1	1
10	Only the first word capitalized in subheading	1	1	0	0	0	1	1
11	British English	1	0	1	1	1	1	0
12	American English	0	1	1	0	0	1	0
13	Missing comma	1	0	1	1	0	0	1
14	Incomplete clause	1	0	0	0	0	0	0
15	APA in-text citations	1	1	1	0	0	0	0
16	Error writing "et al."	1	0	0	0	0	0	1

Figure 3.4. Extract from linguistic features list of ALSLA essays.

	A	B	C	D	E	F	G	H
1	Feature	AI1	AI2	AI3	AI4	AI5	AI6	AI7
2	Subheadings with only first word capitalized	1	0	0	0	0	0	0
3	Establishes well-known applied linguistics abbreviations	1	1	0	1	1	0	1
4	"This essay/in this essay"	1	1	1	1	1	1	1
5	Signposts structure of the essay	1	1	0	0	0	0	0
6	Use of first person plural pronoun	1	1	1	1	1	1	0
7	Has in-text citations (APA format)	0	1	0	0	0	0	0
8	No indicators of British or American English	1	0	1	1	0	0	0
9	Subheadings with multiple words capitalized	0	1	1	1	1	1	0
10	Unnecessary capitalization	0	1	0	1	0	1	0
11	Separating clauses with a semi colon	0	1	0	0	0	0	0
12	Separating clauses with a colon	0	1	0	1	0	0	0
13	No noun after "this"	0	1	0	0	0	0	0
14	Does not define abbrevations	0	1	1	1	0	0	0
15	Question as a subheading	0	1	0	0	0	0	0
16	American English	0	1	0	0	0	1	1

Figure 3.5. Extract from linguistic features list of AI essays.

The Jaccard's Index is a statistical formula used in a broad range of forensic linguistics and computer science literature. For example, Yasin, Yasin, and Yasin (2011) used the Jaccard's Index to generate different algorithms for creating summaries of corpora of documents, based on the content similarity of the documents in the field of text mining. Nini (2018) used n-grams and the Jaccard's distance, another model of the Jaccard's Index, to measure the linguistic 'distance' between the letters sent to the police during the case of *Jack the Ripper*. His results suggested that the two most famous letters in the case, 'Dear Boss' and 'Saucy Jacky', were most likely to have been written by the same author, as they were the closest in 'linguistic distance' with each other. Additionally, Nini (2018) concluded that several of the letters sent to the police were hoax/imitator letters, as the writing styles were very distinct.

The scale for the Jaccard's Index is from 0 to 1, with 0 meaning no similarity and 1 meaning identical in terms of features. The Jaccard's Index could have also been used to

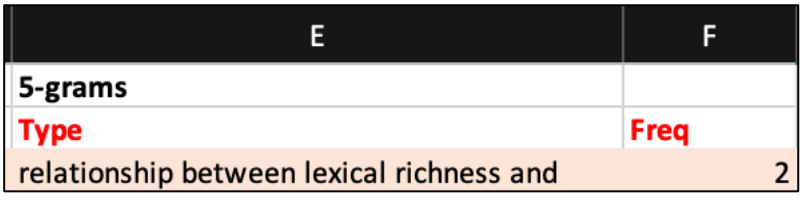
calculate the similarity within the AI essays and the ALSLA essays (for example, AI essay 1 features against AI essay 2 features, or ALSLA essay 1 features against ALSLA essay 2 features). These calculations would suggest how similar or different the ALSLA essays are from each other, and how similar or different AI essays are from each other. However, this approach would have not depicted a comparison between the two types of essays. Since the main objective of this study was to determine the linguistic similarity between AI essays and Master's students' essays, I calculated the Jaccard's Indexes between both types of essays (i.e., an AI essay with an ALSLA essay). The Jaccard's Index formula is visualized below, with respect to linguistic features:

Table 3.1. Jaccard's Index formula and explanation

		Essay A	
		Number of features present	Number of features absent
Essay B	Number of features present	*a*	*b*
	Number of features absent	*c*	*d*

$$J = \frac{a}{a + b + c}$$

I then extracted a list of 5-gram word sequences, which are consecutive words that occur in any given text. Forensic linguists use n-gram sequences at various lengths to calculate the probability of two (or more) texts having been written by the same author in authorship verification. This analysis assumes that the longer a word sequence is, the less likely it is to occur multiple times, even by the same author (Coulthard, 2010). I conducted 5-gram tests for all 14 texts using the software 'AntConc', which is available for the public to use for free. In my analysis, I looked at 5-grams which occurred 2 or more times in the essays, to discuss the patterns in the frequency and occurrence of longer word sequences. Figure 3.6. shows an extract of a 5-gram found in ALSLA essay 1. The full list is found in Appendix E.



E	F
5-grams	
Type	**Freq**
relationship between lexical richness and	2

Figure 3.6. 5-gram example.

For the ACJ task, I used the website 'NoMoreMarking' and created a custom task in my account and entered the relevant details. I modified the judgment equation to 14 judgments for

each judge, and generated the link for a pilot test to be completed by my supervisor. I recruited 2 lecturers (volunteers) from the department to participate in ACJ. However, I did not tell them that some were generated by AI to avoid inducing bias. After the volunteers judged the data, I presented them with a Turing test. I asked them whether they can identify which essays were generated by AI, and which by humans. I created a form on Google Forms and added each essay one by one with three options; Written by a human being, generated by AI, and unsure. I then analyzed how accurate the lecturers' verdicts were. Additionally, the lectures sent me comments behind their judgments, all of which are analyzed in the Discussion section. In the next chapter, I will present the results.

Chapter Four

Results

The results for the different subsections of this study are presented as follows; statistical calculations (Jaccard's Index and features), n-grams, the ACJ task, and the Turing test results. In depth discussion and analyses of these results will be presented in the following chapter.

I. Statistical Calculations

The statistical test used in this study was the Jaccard's Index which suggests similarity (or lack thereof) between two documents by presenting a number on a scale of 0 to 1 with respect to features present in the analyzed documents.

A. Features

I used inductive coding, that is, I interpreted my data without a set of predefined or existing features, to create a list of 82 linguistic features present in all essays written by Oxford students. Though this list is comprehensive, it is not exhaustive, as one can find more linguistic features in the corpus such as those related to syntax or other linguistic approaches. I extracted 29 linguistic features from AI essays. In total, there were 91 features present in the entire corpus, which I used to calculate the Jaccard's Index. I gave each feature a binary score of 0 or 1; 0 if they were not present in the document, and 1 if they were present. This type of binary labeling is used for Jaccard's Index, though it does not depict the number of times a feature is present in the document. The feature could be present once or more, but it is still counted as 1. In this study, the frequency of the features does not contribute to the statistical analysis, due to the limited length of the texts and the purpose of the study. Figure 4.1. and Figure 4.2. show an extract of the features from the corpus. The full list of features can be found in Appendix D.

	A	B	C	D	E	F	G	H
1	Feature	H1	H2	H3	H4	H5	H6	H7
2	Establishes common applied linguistics abbreviation to use later	1	1	0	1	1	1	0
3	First person plural pronoun	1	1	1	0	0	0	1
4	Error in choice of article	1	1	1	0	0	0	0
5	Gives essay a third person neutral singular pronoun	1	0	0	0	0	0	0
6	Establishes dates of literature used	1	0	0	0	0	0	0
7	Erorr in use of comma	1	0	0	0	0	0	0
8	Error in citation	1	0	0	0	0	0	0
9	Run-on sentence	1	0	1	1	0	1	1
10	Only the first word capitalized in subheading	1	1	0	0	0	1	1
11	British English	1	0	1	1	1	1	0
12	American English	0	1	1	0	0	1	0

Figure 4.1. Extract from linguistic features list of ALSLA essays.

	A	B	C	D	E	F	G	H
1	Feature	AI1	AI2	AI3	AI4	AI5	AI6	AI7
2	Subheadings with only first word capitalized	1	0	0	0	0	0	0
3	Establishes well-known applied linguistics abbreviations	1	1	0	1	1	0	1
4	"This essay/in this essay"	1	1	1	1	1	1	1
5	Signposts structure of the essay	1	1	0	0	0	0	0
6	Use of first person plural pronoun	1	1	1	1	1	1	0
7	Has in-text citations (APA format)	0	1	0	0	0	0	0
8	No indicators of British or American English	1	0	1	1	0	0	0
9	Subheadings with multiple words capitalized	0	1	1	1	1	1	0
10	Unnecessary capitalization	0	1	0	1	0	1	0
11	Separating clauses with a semi colon	0	1	0	0	0	0	0
12	Separating clauses with a colon	0	1	0	1	0	0	0

Figure 4.2. Extract from linguistic features list of AI essays.

I calculated the Jaccard's Index for each human-written essay with each AI-generated essay to determine the similarity between them with respect to linguistic features. To do this, I used Microsoft Excel wherein I combined the list of features in 91 columns, and placed the binary labels of each human essay with each AI essay one by one, and extracted all the 1's for the calculations, manually. There were 49 Jaccard's Index calculations in total, which are presented below in tables 4.1. – 4.7. All figures were rounded to three decimal places for greater accuracy on the 0-1 scale.

Formula for the Jaccard's Index:

$$J = \frac{a}{a + b + c}$$

Table 4.1. Jaccard's Index calculations for human essay 1 versus all AI essays.

Human 1 vs. AI 1	$J = \frac{4}{4+2+30} = 0.110$
Human 1 vs. AI 2	$J = \frac{7}{7+8+26} = 0.170$
Human 1 vs. AI 3	$J = \frac{2}{2+4+32} = 0.053$
Human 1 vs. AI 4	$J = \frac{4}{4+7+30} = 0.098$
Human 1 vs. AI 5	$J = \frac{3}{3+3+31} = 0.081$
Human 1 vs. AI 6	$J = \frac{1}{1+10+33} = 0.023$
Human 1 vs. AI 7	$J = \frac{1}{1+4+33} = 0.026$

Table 4.2. Jaccard's Index calculations for human essay 2 versus all AI essays.

Human 2 vs. AI 1	$J = \frac{3}{3+3+1} = 0.170$

Human 2 vs. AI 2	$J = \frac{6}{6+10+6} = 0.270$
Human 2 vs. AI 3	$J = \frac{2}{2+5+11} = 0.111$
Human 2 vs. AI 4	$J = \frac{3}{3+8+11} = 0.136$
Human 2 vs. AI 5	$J = \frac{2}{2+4+11} = 0.118$
Human 2 vs. AI 6	$J = \frac{2}{2+9+12} = 0.087$
Human 2 vs. AI 7	$J = \frac{2}{2+3+12} = 0.118$

Table 4.3. Jaccard's Index calculations for human essay 3 versus all AI essays.

Human 3 vs. AI 1	$J = \frac{2}{2+4+27} = 0.060$
Human 3 vs. AI 2	$J = \frac{7}{7+9+22} = 0.180$
Human 3 vs. AI 3	$J = \frac{3}{3+3+26} = 0.090$
Human 3 vs. AI 4	$J = \frac{3}{3+8+26} = 0.080$
Human 3 vs. AI 5	$J = \frac{2}{2+4+27} = 0.060$
Human 3 vs. AI 6	$J = \frac{4}{4+7+25} = 0.100$
Human 3 vs. AI 7	$J = \frac{2}{2+4+27} = 0.060$

Table 4.4. Jaccard's Index calculations for human essay 4 versus all AI essays.

Human 4 vs. AI 1	$J = \frac{2}{2+4+19} = 0.080$
Human 4 vs. AI 2	$J = \frac{4}{4+12+17} = 0.120$
Human 4 vs. AI 3	$J = \frac{1}{1+5+20} = 0.038$
Human 4 vs. AI 4	$J = \frac{4}{4+7+17} = 0.140$
Human 4 vs. AI 5	$J = \frac{2}{2+4+18} = 0.083$
Human 4 vs. AI 6	$J = \frac{3}{3+8+18} = 0.100$
Human 4 vs. AI 7	$J = \frac{3}{3+2+18} = 0.130$

Table 4.5. Jaccard's Index calculations for human essay 5 versus all AI essays.

Human 5 vs. AI 1	$J = \frac{2}{2+4+11} = 0.120$
Human 5 vs. AI 2	$J = \frac{4}{4+12+9} = 0.160$
Human 5 vs. AI 3	$J = \frac{1}{1+5+12} = 0.055$
Human 5 vs. AI 4	$J = \frac{3}{3+8+10} = 0.140$
Human 5 vs. AI 5	$J = \frac{2}{2+4+11} = 0.117$
Human 5 vs. AI 6	$J = \frac{1}{1+10+12} = 0.040$
Human 5 vs. AI 7	$J = \frac{2}{2+3+11} = 0.125$

Table 4.6. Jaccard's Index calculations for human essay 6 versus all AI essays.

Human 6 vs. AI 1	$J = \frac{2}{2+4+21} = 0.070$
Human 6 vs. AI 2	$J = \frac{7}{7+9+16} = 0.220$
Human 6 vs. AI 3	$J = \frac{2}{2+4+22} = 0.070$
Human 6 vs. AI 4	$J = \frac{5}{5+6+18} = 0.170$
Human 6 vs. AI 5	$J = \frac{2}{2+4+21} = 0.070$
Human 6 vs. AI 6	$J = \frac{3}{3+8+21} = 0.090$
Human 6 vs. AI 7	$J = \frac{2}{2+3+21} = 0.080$

Table 4.7. Jaccard's Index calculations for human essay 7 versus all AI essays.

Human 7 vs. AI 1	$J = \frac{4}{4+2+10} = 0.250$
Human 7 vs. AI 2	$J = \frac{4}{4+12+10} = 0.150$
Human 7 vs. AI 3	$J = \frac{4}{4+2+10} = 0.250$
Human 7 vs. AI 4	$J = \frac{5}{5+6+9} = 0.250$
Human 7 vs. AI 5	$J = \frac{2}{2+4+12} = 0.111$
Human 7 vs. AI 6	$J = \frac{2}{2+9+12} = 0.090$
Human 7 vs. AI 7	$J = \frac{1}{1+4+13} = 0.055$

The smallest value of the Jaccard's Index was 0.023 out of 1, which was between human essay 1 and AI essay 6. This value means that out of all the essays examined, the most dissimilar essays were human essay 1 and AI essay 6, with the lowest proportion of common linguistic features. The greatest value of the Jaccard's Index was 0.270 out of 1, which was between human essay 2 and AI essay 2, meaning these two essays had the highest proportion of common linguistic features. These calculations suggest that the linguistic feature similarity between AI essays and ALSLA essays is overall extremely low.

C. N-grams

An indicative sample of n-grams is presented below in figures 4.3. and 4.4. The total number of 5-grams for each essay is presented below in table 4.8. Similar to the Jaccard's Index, I do not compile the number of n-grams for both sets of essays, in order to perform qualitative data analysis for each individual essay. The n-grams suggest that students do not repeat longer word sequences as frequently as AI essays do. The full list of n-grams can be found in Appendix E.

5-grams	
Type	Freq
relationship between lexical richness and	2
a complete and detailed picture	2
complete and detailed picture of	2
is far away from providing	2
it is far away from	2
lexical diversity and lexical sophistication	2
lexical richness and vocabulary knowledge	2
of learners productive vocabulary knowledge	2
picture of students vocabulary knowledge	2
us with a complete and	2
with a complete and detailed	2
of lexical richness as it	2
the language user has a	2
can be argued that the	2
it can be argued that	2

Figure 4.3. Extract of 5-grams from ALSLA essays.

5-grams		
Type	Rank	Freq
and designing effective instructional strategies	2	2
assessing their language proficiency and	2	2
factors that affect lexical richness	2	2
in assessing their language proficiency	2	2
language proficiency and designing effective	2	2
new words and use them	2	2
proficiency and designing effective instructional	2	2
relationship between lexical richness and	2	2
that affect lexical richness in	2	2
the diversity and complexity of	2	2
the relationship between lexical richness	2	2
their language proficiency and designing	2	2

Figure 4.4. Extract of 5-grams from AI essays.

Table 4.8. 5-gram count of student essays and AI essays.

Student essay	5-gram count	AI essay	5-gram count
1	1	1	12
2	10	2	79
3	1	3	17
4	2	4	6
5	11	5	47

6	3	6	24
7	13	7	37

III. Adaptive Comparative Judgment

A. No More Marking Task

The ratings of the ACJ task are presented below in table 4.9. On the website, there is a short guidebook which has explanations for the results which the website generates after the task is completed. The 'Score' is the result of the wins and losses of each essay after being put into a mathematical model (unnamed), which gives each essay a rank (No More Marking, 2023). There is no scale for the Score, as the purpose of ACJ is to rank essays from best to worst, rather than assign them with a score based on a rubric. The Score numbers are ranked in table 4.9. below into a descending order, which provides a measure of the relative quality of each essay. The key is as follows:

- The 'Score' is the number generated by NoMoreMarking.com, after their calculations.
- The 'Rank' is the order of which the essays were judged as better than the others; 1 meaning the best and 14 meaning the worst.
- 'Wins' indicates the number of times that particular essay was chosen as the better essay in the task.
- 'Losses' indicates the number of times that particular essay was not chosen as the better essay in the task.

Table 4.9. ACJ scores of all essays with their rank, wins, and losses.

Essay	Score	Rank	Wins	Losses
H5	6.70	1	7	0
H3	5.70	2	6	0
H2	4.70	3	5	0
H6	4.70	4	5	0
H1	4.00	5	4	4
H7	3.82	6	4	1
AI5	3.00	7	3	3
AI4	2.10	8	2	4

AI6	2.10	9	2	4
H4	2.06	10	2	3
AI3	1.21	11	1	5
AI2	1.20	12	1	6
AI1	0.30	13	0	6
AI7	0.30	14	0	6

IV. Turing Test

For the Turing Test, two lecturers were asked to choose whether they believed an essay was written by a human being or by AI, but they also had the option to choose 'Unsure' if they could not decipher. The results are presented below in table 4.10.

Table 4.10 Lecturer verdicts in the Turing test for each essay.

Essay	**Lecturer A's verdict**	**Lecturer B's verdict**
H1	Written by a human being	Written by a human being
H2	Written by a human being	Written by a human being
H3	Written by a human being	Written by a human being
H4	Written by a human being	Written by a human being
H5	Written by a human being	Written by a human being
H6	Unsure	Written by a human being
H7	Unsure	Written by a human being
AI1	Generated by AI	Unsure
AI2	Generated by AI	Generated by AI
AI3	Generated by AI	Generated by AI
AI4	Generated by AI	Generated by AI
AI5	Generated by AI	Generated by AI
AI6	Generated by AI	Generated by AI
AI7	Generated by AI	Unsure

These results demonstrate that the lecturers considered AI essays as generally of lower overall quality than human written essays. Though some AI essays were ranked as better than other AI essays, no AI essay was ranked better than human essays. This point is further evaluated in the discussion section. Additionally, lecturers were able to accurately identify which essays were by AI and which by humans. In the following chapter, I discuss these results in-depth.

Chapter Five

Discussion

The discussion section will be structured as follows: features, statistical analysis, n-grams analysis, ACJ analysis, Turing test analysis, and limitations of my study.

I. Features

Grabe (1984) argues that a text is not an absolute entity, but rather dynamic and complex, which depends on several factors. He asserts that one cannot reduce a text to merely its components. Hence, analyzing features of a given text is helpful here, however, they are never the only element which should be examined in textual discourse analysis. Textual discourse analysis is a broad and multi-dimensional field of study, and can be approached through several different perspectives such as syntax or semantics. Due to the nature and length of this assignment, I only examine features and n-grams to suggest similarities and differences between AI-generated and human written texts when given the same prompt. With respect to linguistic features, I identified 83 features in the essays written by human beings and 29 in the essays generated by AI, for a total of 91 features in the corpus. Twenty of these features were found at least once in both types of essays. These features are summarized below in figure 5.1. I have classified the features thematically; common features (that is a group of features that were found in both types of essays), features pertinent to linguistic errors, features that demonstrate personalization in texts, length, structure, pronouns, style, content, and sources and references. The said features can be classified into more than one theme. However, for the purpose of this discussion, they will be grouped into these broad categories.

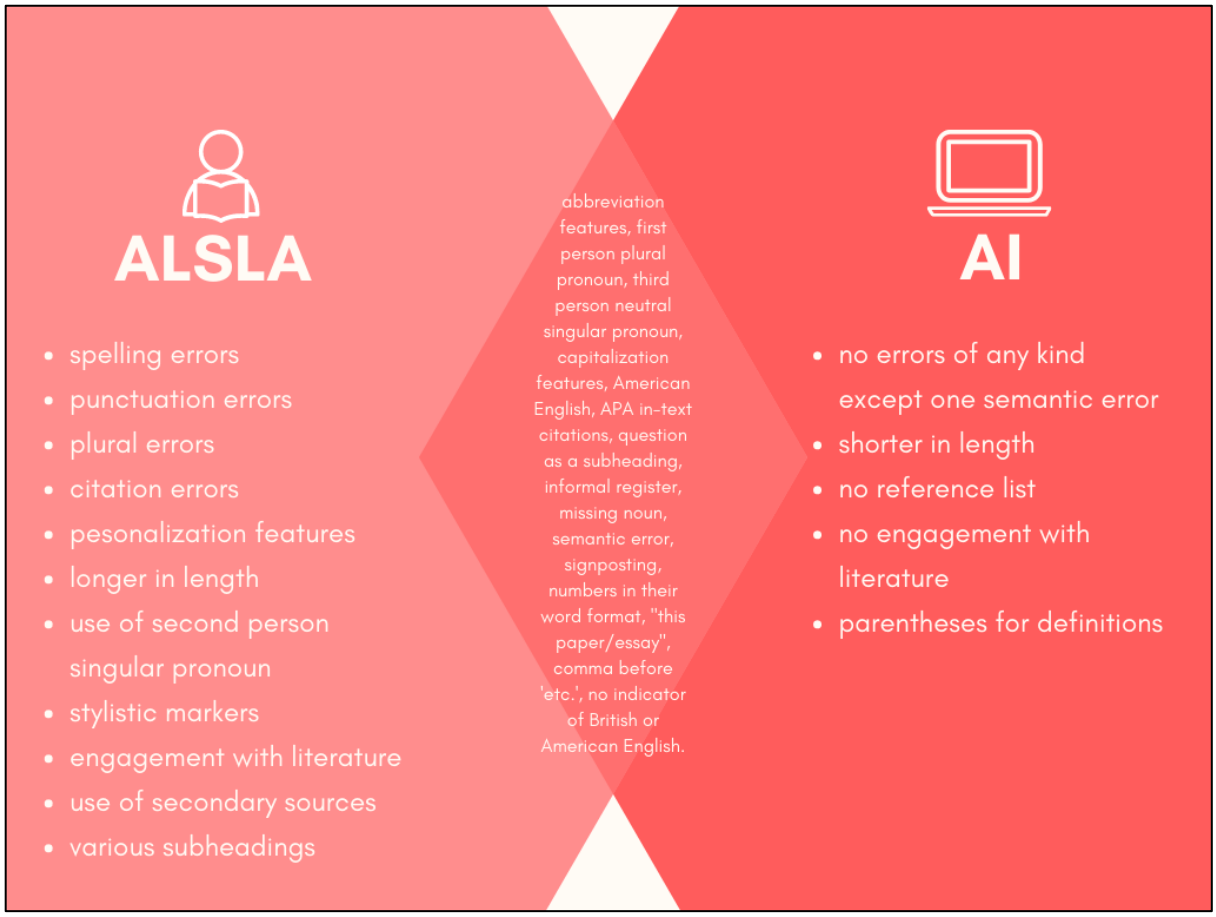


Figure 5.1. Summary of the linguistic features found in ALSLA essays, AI essays, and both.

A. Common

One of the common features found in both types of essays was the establishment of common applied linguistics abbreviations to be referred to later. This feature allows readers to follow through longer, technical terms faster while reading. However, in certain cases, abbreviations are not specified, such as "MWU" by ALSLA student 1. Regardless of how well-established an abbreviation is in the field, ALSLA students are encouraged to write the full form before using it in the essay, as some examiners are not linguists but rather social scientists that may not be familiar with the discipline. Another common feature was the use of first-person plural pronouns such as 'we' and 'our'. In the ACJ task, lecturers were alarmed at the use of 'we', as it is commonly used in plagiarism cases since students writing their own essays are one singular person. Nonetheless, a few ALSLA students used this pronoun, and a majority of the AI essays did as well. Additionally, both types of essays have third person neutral singular pronoun, which depicts gender-neutrality and formal register. However, informal register is also used in both AI and ALSLA essays at times (e.g., using the terms "so far" by human essay

1, or “simply a matter” by AI essay 5). In academic writing, the use of informal register is discouraged, in order to convey information without bias or partiality.

With respect to subheadings, in both types of essays, the first letter of the first word is capitalized, while the other words are all in lower case. This feature is a common style marker in academic writing, and is usually used when the subheadings are a statement or longer in length. However, some essays in both AI and ALSLA corpora had first letter of multiple words capitalized in the subheadings as well. Both AI and humans use questions as subheadings at times, which Hyland (2002) argues to add a sense of inter-personalization to texts, as questions engage and stimulate the reader.

American English was used by both ALSLA students and AI. The AI software I used to generate the essays is based in Seattle, USA, and thus may be programmed to generate solely in American English as there were no inputs for dialect. However, there are certain essays such as ALSLA essay 7 and AI essays 1, 4, and 7, where there were no words that would indicate whether the text is written in British English or American English. There is also the use of APA in-text citations by both humans and AI. ALSLA students are required by the department to use in-text citations and references in their work to avoid accusations of plagiarism. Though I did not input any referencing parameters into Moonbeam, it generated a few in-text citations, however, none of the essays had a reference list at the end. Lecturers marked essays without references as worse than those with in the ACJ task.

A semantic feature that is found in both AI and human essays is when there is a missing noun after the determiner ‘this’. This feature can confuse the reader with regards to what the referent of the sentence is. This feature was also one of the feedback points ALSLA examiners provided to students after the formative assignment. Another semantic feature present in the essays is semantic errors, such as the use of the term ‘methodology’ by AI essay 2. ‘Methodology’ is a common semantic error made by researchers in academic writing. Methodology means ‘the study of methods’ or to describe a system of methods common to a particular field of enquiry, rather than a description of the methods used in a particular study. Methodology should thus be replaced with the term ‘methods’ or ‘research methods’. However, since AI is programmed based on real human writing, it makes this mistake as well.

Both ALSLA students and AI signpost the structure of the essay in the introduction which is a commonly used feature in academic writing that helps the reader follow the text thoroughly (University of Birmingham, 2023). Moreover, both AI essays and human essays have phrases such as “this paper/essay” which is a common phrase used to aid in signposting and to establish the purpose of the text (Myers, 1992). In both essay types, numbers were

written in their word format at times, which is a stylistic marker. Colons and semicolons are used to separate clauses, which aid the flow of the sentence and organization of ideas (University of South Florida, 2023). In both types of essays, there is a comma before the term 'etc.' which is another punctuation style marker. Lastly, another common feature found in both types of essays was unnecessary capitalization of certain nouns such as "Grade 11" in human essay 4 or "Type-Token Ratio" in AI essay 4. These types of words, though are established concepts, are not proper nouns and should not be capitalized in writing.

B. Errors

31 features found in the essays written by human beings were based on different errors found in writing, particularly by L2 learners of English. Brown (1994) defines 'errors' as a deviation from standard language norms by a user of that language. Though two students in the ALSLA cohort are native speakers of English and the essay submissions were anonymous, I cannot differentiate between L1 and L2 errors.

A majority of the linguistic errors made by ALSLA students are related to punctuation, such as missing commas, error in the use of a dash, missing quotation marks, etc. There are some errors pertaining to sources and citations, such as error in writing 'et al.', error in citing quotations, etc. There are spelling errors as well (e.g., 'sixe' instead of 'size', human essay 4). Additionally, there is an incorrect plural for the term 'literature', which ALSLA student 6 had written as "literatures". This common error can stem from the lack of distinction between mass nouns and count nouns in languages such as Chinese, which native speakers of Chinese fail to apply when writing in English. Errors are commonly found in writing by both L1 and L2 users of a language.

The severity of the penalties for errors depends on the type of assignment, marker, and institution. ALSLA students are expected to have minimal errors in their assignments (see Appendix F), and are permitted to use grammar-check aid in longer assignments albeit with restrictions. In the Turing test, one of the lecturers mentioned they flagged any essay with spelling errors as human. Contrarily, AI essays do not have any punctuation or spelling errors. Although the content of AI essays is not of depth or breadth, in terms of punctuation and spelling, the essays are 'perfect'. Though this feature may strike as appealing to both educators and students, it is not an accurate representation of academic writing, as even native speakers make errors/mistakes in their work. This is also why software that help correct punctuation exist, such as *Grammarly*, and why some institutions permit students to use proofreaders for longer assignments.

C. Personalization

There are certain linguistic features that depict personalization in essays written by ALSLA students, but not in essays generated by AI. This feature is one of the key features which can distinguish human written essays from AI ones. It also refers to the notion of authorship, as human beings tend to narrow down and personalize texts based on the purpose and context of their writing. The formative assignment asks students to critically analyze empirical literature on the topic. In human essay 1, the student establishes the year range of the literature they are using in their essay, i.e., 30 years. In the MSc ALSLA program, students are given broad questions for the summative assessments which can be approached from multiple perspectives. Students are advised to narrow down topics to explore the topic in more depth and critically evaluate the literature. This criterion is found in the rubric for the summative assessments on Canvas. Even though the formative assignment does not have its own rubric, it is used as a steppingstone for students to write their summative assessments, hence, they should aim to narrow down topics.

Though a student had used tables, due to the formatting of MS Forms, the table did not copy properly. Tables are used in academic writing for different reasons such as summarizing results and presenting values to make it easier for the reader (Slutsky, 2014). However, due to the lack of detailed research data presented in the AI essays, there were no tables generated. ALSLA student 6 used a table to summarize the measurements of lexical richness in the literature they were evaluating, suggesting personalization.

Furthermore, ALSLA student 5 uses the phrase “unless specified” to specify the definition of a certain word (i.e., words = tokens), in their essay. No AI essay uses this collocation. When used in academic writing, phrases such as ‘unless specified’ or their equivalent, demonstrate narrowing down and specifying words semantically. Since words can have more than one connotative and denotative meaning, using such phrases indicate that in that particular study, and that study only, the term has a certain definition which may or may not align with generic definitions of that term outside the paper.

The fourth feature that suggests personalization is the establishment of uncommon abbreviations by ALSLA students. These abbreviations are specific to the study and may not exist outside of that study. For example, ALSLA student 5 uses the abbreviation “CPhD” in lieu of ‘Chinese PhD students’. The use of abbreviations in academic writing is subjective and debated, with some universities prohibiting it (University of Hull, 2023). However, many researchers use abbreviations to avoid repeating lengthier collocations/sequences to make it easier for the reader to follow the essay (Majumder et al., 2019). Oftentimes there are

commonly established abbreviations used in essays however, personalized abbreviations are only seen in essays by human beings.

D. Length

The word limit of the formative assignment was 1,500 words excluding references but including in-text citations and footnotes; and all ALSLA essays were under this limit. University of Oxford is strict with word limits, and there are marking penalties for being over (see Appendix F). Hence, students tend to stay within the word limit. With AI software Moonbeam, however, one cannot input a word limit parameter. When I asked the chat bot to increase the length of the essay, it said it could not. Hence, the AI essay lengths are between 500 and 1, 500 words. Most of the essays were under 1,000 words, which would be a point of criticism by an examiner if an ALSLA student submitted an essay with such low word count. Additionally, the shorter length of the AI essays indicates less critical engagement with the topic and literature. However, with respect to n-grams, despite the shorter length of the AI essays, they had a higher frequency of longer n-gram sequences. This point is discussed in sub-section III. With regards to intra-textual length, there are certain features present in essays written by ALSLA students, such as run-on sentences, use of 'e.g.' in in-text citations, and having long clauses in parentheses. These features usually lead to poor quality in academic writing, as they take away from cohesion, conciseness, and clarity of themes and ideas, and confuse the reader (Kagan, 1980; Ananda, Gani & Sahardin, 2014; Zheng, Napoles, Tetreault & Omelianchuck, 2018).

E. Structure

In terms of structure, most of the essays written by ALSLA students have labeled or unlabeled introductory, body, and concluding paragraphs. However, essay 6 does not have a concluding paragraph, the reason behind which is unknown. All AI essays, however, have introductory, body, and concluding paragraphs. Concluding paragraphs are typically an unspoken norm in academic writing, however, they are not mentioned explicitly in the summative assessment's criteria. Though, there is a clause which mentions, "*Good adherence to academic conventions*", which can be interpreted to include having a conclusion, as it is an academic convention. Nonetheless, I do not analyze the quality of the essays, and rather only the presence (or in this case, absence) of certain features.

F. Pronouns

Using personal pronouns in academic writing is debated with some institutions permitting it and others prohibiting it. Researchers argue for the usage of personal pronouns in writing in certain disciplines to avoid author alienation and impersonality (Hyland, 2002). Tang

and John (1999) argue that despite regular writing conventions, personal pronouns can be integrated in academic writing in certain text types to demonstrate author identity. The use of the pronoun 'you' in ALSLA essay 3 can be interpreted as an example of synthetic personalization. Synthetic personalization, coined by Fairclough (1989), is the act of using singular personal pronouns to address a mass audience, and is commonly used in advertisements by organizations. AI does not use these pronouns. However, AI uses first person plural pronouns which lecturers in the ACJ and Turing test tasks noticed and attributed to AI and/or plagiarism. This finding is interesting as AI is not a person, yet uses personal pronouns to make the writing 'human' (for students to use in their assignments).

G. Style

Grabe (1984) defines 'style' in textual discourse analysis as the combination of techniques used by a writer to demonstrate their personality in their writing. Style can be linked to authorship, as every author has their own writing style, which can be used to identify that author through idiosyncratic linguistic features they may use. Mcmenamin (2001) claims that idiosyncrasy ties in with 'style markers', i.e., features that an author uses consistently yet subconsciously in their writing, which are either a version of a standard form or as a deviation from the standard form. There were several different style markers present in essays written by ALSLA students, most of which were versions of a standard form. These markers include using British English, starting two subsequent clauses with 'that', using passive sentences, using the same adverb twice in a sentence, using an Oxford comma in an in-text citation, using numbers in the numeral format, using adverbs before and/or after verbs, and using variations of 'she/he' such as "(s)he" or "s/he". AI essays, on the other hand, did not have idiosyncratic stylistic markers other than AI essay 2 having parentheses for definitions. This finding suggests another key difference between the two types of essays, given that ALSLA essays have more personalization in terms of style, but AI essays do not (regardless of how 'human' they are trying to be).

H. Content

With respect to content, in ALSLA essays, there were two features: mentioning a claim but not analyzing it, and having suggestions for further research. The former act is discouraged in academic writing as it does not demonstrate critical engagement with the chosen text. For ALSLA students, critically engaging with text and going into depth with critique and analysis is one of the main requirements of all assignments, as seen in the grading rubric (see Appendix F). Though AI essays have generic phrases entailing 'future/further research', these claims are not used correctly with reference to analysis of empirical literature. ALSLA students are

encouraged to find limitations of empirical studies and consider the subjective nature of studies which may or may not be expanded (see Appendix F). These features are not found in AI essays as AI essays have more generic claims and do not critically evaluate literature even if they are mentioned.

With respect to content found in AI essays, one essay had an untitled preface before the introduction which can be interpreted as a rationale, as it starts off with "*As university examiners, it is our responsibility to […]*" None of the essays written by ALSLA students have a preface, and rather have labeled or unlabeled introductory paragraphs. Prefaces, rationales, and abstracts are features that can be found in longer pieces of academic writing, however, they were only found in AI-generated essays in this corpus.

I. Sources and References

One of the major distinctions between essays written by ALSLA students and essays generated by AI is the use of references. Even though I had instructed ALSLA students to submit their essays with references, ALSLA students 1 and 4 did not. 3 AI essays have APA in-text citations, but none have a references section at the end. Contrarily, it is a requirement for ALSLA students to have both in-text citations and a reference list in all of their work. If a student does not have sources in their work, they will be severely marked down for it, and consequently subject to disciplinary hearing for suspected plagiarism (see Appendix F). In the ACJ task, the lecturers also mentioned that they declared an essay as worse if it did not have references. Some ALSLA students used secondary sources in their formative assignment (e.g., "Read, 2000, as cited in Chen, 2020" in ALSLA essay 7), even though this act is usually discouraged in the program. Some ALSLA students used the first name of the author when referencing them, however, this act is uncommon in academic writing where solely the last name of the author is used, unless specified in certain discourses.

II. Statistical Analysis

The statistical analysis used in this dissertation was the Jaccard's Index, which depicts similarity between two documents with respect to the linguistic features present in them. The smallest Jaccard's value was 0.023/1 between human essay 1 and AI essay 6. The greatest Jaccard's value was 0.270/1 between human essay 2 and AI essay 2. These values are of extremely low significance on the Jaccard's Index scale, and are under midpoint 0.5/1. Thus, with respect to authorship verification, when an ALSLA essay and AI essay are examined side-by-side, they are highly unlikely to have been written by the same author, as they share a very limited number of linguistic features. Out of 49 values, there were 22 indexes between 0.0 and 0.1, 22 indexes between 0.1 and 0.2, and 5 indexes between 0.2 and 0.3. These results can be

interpreted as minor similarities between any given AI essay with any ALSLA essay. However, there were no Jaccard Index values of 0, which suggests that there are some linguistic similarities between AI essays and essays written by human beings, even if minor. Since the AI software I used asked for certain parameters such as audience and topic, the potential of linguistic similarities was bound to occur. However, through this statistical analysis, it can be argued that in this particular case, though AI may be able to mimic writing of a human, human beings have their own idiosyncratic linguistic features which make their texts different to AI.

III. N-grams Analysis

With respect to n-grams, the longer a word sequence is, the less likely it is to occur multiple times within the same text (Coulthard, 2010) as a person is less likely to produce identical phrases on different instances, especially through memory. Though writing is different to speaking, and writers can revise their text, there are a few factors that affect repetition in writing. For example, genre; texts such as poetry include repetition to allow the reader to remember verses (Clarvoe, 2009), and for emphasis and stanza cohesion (Poteat, 1919). However, repetition in academic writing of the English language is different to that found in other genres. Some researchers (Hinkel, 2005) argue that repetition in academic writing is often discouraged as it depicts redundancy, which leads to deviation from topic and lack of conciseness. However, other researchers argue that repetition depends on the culture and language of academic writing (Pecorari, 2008). McCarthy (1991) argues that in Japanese academic writing, repetition is used to depict themes and persuasion in a text.

In ALSLA's summative and dissertation marking criteria repetition is not labeled as a feature of writing that would add to or take away from marks if a student were to use it. However, this criterion is not applicable to the formative assignment, as ALSLA did not receive a set of rigid guidelines, which allowed lecturers to permit or prohibit it on their own terms. ALSLA students tend to avoid repetition in their writing, however, do have a few phrases they reuse for different purposes, which are discussed below.

A. Thematic analysis of ALSLA n-grams

In ALSLA essays, 5-grams are repeated for various reasons such as for emphasis on the point, to restate a point in the conclusion, to state a point in the introduction and later use in the body paragraphs, etc. Multiple 5-gram sequences can be grouped together because they are found identically in their occurrences. However, due to the parameters I input of 5-gram, AntConc generated the strings separately. These sequences are found in Appendix E.

There were sequences reoccurring for the purpose of restating the same point in the conclusion. Some examples are, *"relationship between lexical richness and"* (ALSLA 1); "*of*

learners' productive vocabulary knowledge" (ALSLA 2); *"lexical fluency as well as"* (ALSLA 5); and *"lexical density, diversity and sophistication in a linguistic output"* (ALSLA 7). These n-grams are present in either one of the body paragraphs or introduction, and in the conclusion. Additionally, in ALSLA essay 5, there are 5-grams grouped together to form a 10-gram which occurs twice; "*vocabulary knowledge when looking at lexical diversity, density, and sophistication*". Though it is rare to find a 10-gram more than once in human writing, it was also used in this essay with the purpose of restating in the conclusion. This trend is common in English academic writing, as seen in writing pedagogy handbooks (Coffin, Curry, Goodman, & Hewings, 2003; Davis & Liss, 2006). This trend is also included in essay outline guides from different universities across the globe, such as University of Oxford (Christ Church College, 2023) and Qatar University (Qatar University, 2023). Restating a sequence in the conclusion emphasizes the point and reminds the readers about the argument of the essay.

There are n-gram sequences that are repeated to link two points together. Some examples are, *"lexical diversity and lexical sophistication"* (ALSLA 2), *"the depth of vocabulary knowledge"* (ALSLA 5), *"In the example given in Lightbown and Spada"* (ALSLA 7). These sequences are found in two body paragraphs, the first in which they are stated and the latter of which in they are then elaborated on as an explanation. For example, when the ALSLA student 2 re-states the 5-gram in body paragraph 2, they preface it with the phrase "As mentioned earlier,...", depicting their acknowledgement of the n-gram as already stated.

Certain sequences reoccur but without any prefaces, which makes it difficult to decipher their purpose, such as *"the language user has a"* (ALSLA 3), *"about the use of vocabulary"* (ALSLA 5), and *"an estimate of the learner's"* (ALSLA 7). This trend could indicate intentional re-phrasing. In ALSLA essay 4, the repeated grouped n-gram *"it can be argued that the"* demonstrates a common example of a phrase used in argumentative academic essays to evaluate a certain point (The Writing Center, 2023).

There are n-grams that are used first to signpost the structure of the essay, and later used as a point; for example, *"lexical richness and vocabulary knowledge"* (ALSLA 5), and *"lexical sophistication and lexical variation"* (ALSLA 6). This trend is also common in academic writing as signposting the content of essays is encouraged to help the reader know what to expect, and to demonstrate ideas (Leshem & Bitzer, 2021).

Sequences are also repeated as subheadings such as in ALSLA essay 2 having *"lexical richness and vocabulary knowledge"* in the introductory paragraph, and then later as a subheading of body paragraph one. In ALSLA essay 7, *"lexical richness and vocabulary size"* is also used as a subheading and later in the same subsection. Other n-grams such as *"to the*

quality of ESL" (ALSLA 5) is found twice in the reference list as two articles with this 5-gram were cited by the student. Thus, some n-grams are not repeated by the student but rather exist in their references.

B. Thematic analysis of AI n-grams

There are 5-grams that occur in AI introductions and conclusions, similar to the ALSLA essays which are used for the purpose of restating the essay points in the conclusion paragraph. However, these 5-grams are part of larger sequences which are almost entirely identical in both types of paragraphs. For example, the 5-grams *"and designing effective instructional strategies"* and*"assessing their language proficiency and"* in AI essay 1, and "*an important factor in evaluating*" in AI essay 4. The length of these individual introductory and concluding paragraphs are relatively short, with less than 100 words each. With such long n-gram word sequences being repeated identically in both paragraphs, the AI essay content is extremely repetitive. Contrarily, ALSLA essays have longer n-gram sequences that re-occur, but in longer-length paragraphs, such as a body paragraph in ALSLA essay 5 being almost 600 words long.

In AI essay 3, the 5-gram *"guide teachers in designing effective"* occurs in one of the body paragraphs as well as the conclusion. 5-grams being part of larger sequences is common in both types of essays, as ALSLA essays have sequences which can be up to 13-grams long, and AI essays have recurring sequences up to 9-grams long. However, the 5-gram *"guide teachers in designing effective"* is part of a larger sequence which only has a few changes, similar to patchwriting. Patchwriting is a plagiarism technique wherein a person copies an entire phrase from a source but only changes one (or a few) words usually by synonyms. Patchwriting is usually an unintentional form of plagiarism which typically occurs when students misunderstand how to paraphrase (Pecorari, 2003). AI's recurring longer n-grams with a few minor tweaks demonstrates intra-textual patchwriting. For example, in AI essay 3, a paragraph in the body and one in the conclusion are identical with only two distinct words ("evaluate" and "asses").

Similar to certain n-grams found in ALSLA essays, AI essays have 5-grams that are used as subheadings and within paragraphs, such as *"factors that affect lexical richness in L2 learners"* (AI 1). In certain AI essays, 5-gram sequences occur within different body paragraphs but not in the introduction or conclusion, and without any acknowledgement of the restatement. For example, in AI essay 5, the 5-gram *"higher levels of lexical richness"* occurs in a few different body paragraphs. However, these body paragraphs are of 37, 45, and 84 words in length, which alludes to the point aforementioned of frequent repetition of long sequences within a short length of text. This feature does not occur in essays written by ALSLA students,

so it can be labeled as an n-gram feature specific to AI essays only, and a major feature acts as a distinction between the two types of essays.

There are some 5-grams which are repeated 5 times in AI essay 2, which demonstrates redundancy. For example, the 5-gram *"and potential for future success"* occurs in 4 paragraphs when talking about implications of the study, as well as the conclusion. The 5-gram *"we can gain insights into"* occurs 4 times throughout AI essay 2. The 5-gram *"higher levels of lexical richness"* occurs 4 times within the body paragraphs of AI essay 7. These findings suggest redundancy which would be considered unfavorable by an ALSLA examiner, as the phrase can be reworded and still hold its meaning in different paragraphs.

Some n-grams are used in multiple sub-subheadings, such as *"relationship between lexical richness and"* in AI essay 6 which is a part of two sub-subheadings. This trend is not found in essays written by ALSLA students, who use distinct lexis and sequences in their subheadings. This finding suggests another difference between AI-generated texts and essays written by human beings. In the essays written by ALSLA students, the maximum frequency of a 5-gram recurring is 3, while the rest of the recurrences only happen twice. However, in the AI-generated essays, the maximum frequency of a 5-gram recurring is 5, with some 5-grams recurring 3 and 4 times, but a majority recurring twice. These figures show that AI-generated essays have a lot more repetition in them, which can be marked as worse by examiners due to redundancy and the potential to rephrase in different cases. Since the AI essays are shorter in length, if they were written similar to how human beings write, they would have fewer repeated long n-grams. This finding suggests a prominent difference between AI-generated essays and essays written by human beings.

IV. Adaptive Comparative Judgment Analysis

The results of the ACJ task suggest that lecturers at the Department of Education believe that human-written essays are usually better than essays generated by AI. However, one of the human-written essays was ranked worse than three essays generated by AI. This finding depicts the quality of one particular essay, which can be interpreted as worse than other human essays. Since the essays used in this corpus were the first assignment submitted by ALSLA students, pre-feedback, they are of varying quality.

The pseudonyms of the two lecturers will be lecturer A and lecturer B to maintain anonymity of their identities. My supervisor's pseudonym will be 'supervisor' in the analysis. Lecturer A is a PhD student at the department, who teaches a module to other graduate students in the department. Lecturer B has obtained their doctorate and is a full-time lecturer and examiner in the department. Both lecturers are relatively younger, while the supervisor is older

and has more experience in teaching. However, as found in research by Verhavert et al. (2019) ACJ tasks can be performed regardless of the teaching experience of a lecturer, and can even be done by students for their peers' work, reliably. The results of my dissertation align with this claim, as despite the different years of experience of the lecturers, they had ranked the same essays as better than the others. For example, with respect to the essays written by human beings, essays 5, 3, 2, and 6 were chosen as the better essays by all lecturers in every case they occurred. With respect to essays generated by AI, essays 1 and 7 were never chosen as the better essays by any lecturer in any case.

Since ACJ tasks do not require a standard rubric, I asked the lecturers to provide any comments they had which affected their judgment. The lecturers were not told that some essays were generated by AI until they had submitted the ACJ task. Lecturer B claimed that relevancy to the exam question was one of their major deciding factors of whether an essay was better than the other. They claimed that despite how well-written the essay was, if the content was not pertinent to the essay prompt, they did not choose the essay as better. Though ALSLA students were not given a rubric for the formative assignment, they were expected to answer the essay question (as mentioned in the question itself). Lecturer B also claimed they are an examiner for the course, hence, they were familiar with marking criteria which emphasizes references and citations. They marked any essay with references and in-text citations better than an essay without. When they were faced with two essays without references, they claimed this was easier to judge, than when they faced two essays with or one with and one without. In MSc ALSLA, references are a requirement for all assignments.

My supervisor knew that I had AI-generated essays in my corpus, hence, I did not ask them to do the Turing test. However, they did the ACJ pilot test and claimed that spelling, grammar, and punctuation of the AI essays were better than the essays written by ALSLA students, however, the ALSLA essays were better in terms of content (with more depth and detail). My supervisor also claimed that AI essays had better lexis in certain cases as opposed to ALSLA essays where there were more semantic errors and poorer word choice, given that a majority of ALSLA students are L2 users of English. These observations align with the findings of this dissertation, as there were several kinds of errors in the essays written by human beings, but none by AI other than one semantic error (use of the word 'methodology'). Furthermore, my supervisor claimed that AI essays were clearer and more concise, however, they were not at a master's level of writing. My supervisor said if they were given the AI essays to mark for the same formative assignment, they would have failed all 7 of the essays for vague content, lack of references, no critique of literature as per the task requirements, and no originality of

argument. These claims are similar to Lecturer B's claims as well. Thus, regardless of the lack of rubric in an ACJ task, the judging process was similar for both lecturers, independent of their experiences and biases. Additionally, my supervisor claimed that they would have flagged the AI essays for suspected plagiarism, given that these essays used first person plural pronouns such as 'we' and 'our'. Unless an essay is written in a group by more than one student for a certain task, using first person plural pronouns can be an indicator that work is plagiarized, given that researchers use first-person plural pronouns in their articles when they work in a team.

V. Turing Test Analysis

With regards to the Turing test, the essays which the lecturers voted an answer of either 'Written by human being' or 'Generated by AI' on were all correct (i.e., excluding votes of 'Unsure'). However, both lecturers marked two essays as unsure each. In the case of lecturer A, these two essays were written by human beings. In the case of lecturer B, these two essays were generated by AI. Thus, both lecturers had an 86% success rate in determining whether an essay was written by a human being or generated by AI. Lecturer A claimed that they had tested out AI software for a few months prior to the ACJ task, from which they observed the difference in structure between the two types of essays. They asserted that AI-generated essays have better structure than those written by human beings, though AI essay structure is more formulaic. They also claimed that any essays with spelling errors they marked as written by human beings, which aligns with the findings of this dissertation in the sense that no AI generated essay had any spelling error, but the ones written by human beings did. Additionally, they asserted that they were more likely to choose essays with references and in-text citations as written by human beings, which also aligns with the findings of this dissertation. They commented on one of the essays being tricky because of the type of subheadings used and the usage of 'we', which the lecturer ended up marking as human (correct), but they were not convinced. Lecturer A mentioned that for the two essays they marked as unsure, they would consult another lecturer for their opinion if they could, which would help influence lecturer A's judgment. Both lecturers mentioned that the Turing test was tricky and challenging for them to complete. The lecturers being able to determine a majority of the Turing test answers correctly aligns with the features and Jaccard's Index findings. Since the essays had significant linguistic differences, resulting in a low Jaccard's Index number, the essays are more likely to have not been written by the same 'author'. However, the lecturers were primed to look for essays generated by AI in the corpus, and may or may not be able to make the same judgment without that prior knowledge.

VI. Limitations and Future Directions

Since this study is one of the first to analyze linguistic similarities and differences between AI essays and essays written by human beings, there are several limitations and future directions. In terms of sample size and population, I only received 7 essays from the ALSLA cohort, which though was sufficient for this dissertation, it is not sufficient to make generalizations about how all Master's students, all ALSLA students, or all University of Oxford students write. MSc ALSLA welcomes students from various educational and linguistic backgrounds with different cultural and academic influences on their L1s and L2s. Thus, a more homogenous group can be chosen in future studies with less variation in their writing styles. Additionally, as aforementioned, I could have calculated Jaccard Indexes for AI essays and ALSLA essays separately to find out how similar or different each corpora is respectively, before comparing them to each other.

With respect to textual discourse analysis, since writing is a subjective method of communication, there are several ways the topic of this dissertation can be approached. For example, there can be studies with a syntactic, semantic, or pragmatic approach. Additionally, the list of features I extracted was not exhaustive, and there is potential for more features to be extracted and analyzed. In terms of content, I did not analyze the relevancy of the essays to the question for two main reasons; 1. I do not have expertise in this subfield of linguistics, as I was also one of the students that completed this essay; 2. Moonbeam tweaked the essay question in a few instances (see Appendix B), which did not drastically change the content but was slightly different to the original question. In terms of n-grams, various lengths can also be analyzed, as this dissertation only looks at 5-grams due to the word count and time limitations.

With respect to the ACJ task, I programmed the task to 14 judgements to avoid fatigue for the lecturers, and to ensure a 0.70/1.00 accuracy, drawing from Verhavert et al.'s (2019) study. However, I could have increased the judgments to over 26 to reach a higher accuracy (>0.90/1.00). For the Turing test, there may have been some bias as one of the lecturers is an ALSLA examiner, increasing the potential of having previously seen the essay when marking the formative assignments. I could have also recruited more volunteers for both tasks, however, as the tasks were not the main focus of my dissertation, 2-3 lecturers' input was sufficient.

I used Moonbeam as my AI source, however, since there are other AI software that perform the same tasks such as ChatGPT and Quillbot, there can be studies on essays generated by those software. Other topics can also be explored with different prompts for both students and AI, not necessarily pertinent to vocabulary knowledge. Additionally, other genres and text types can also be studied, as generative AI can generate different kinds of texts such as cover

letters, speeches, and even personal letters. Consequently, different statistical formulae would have to be applied accordingly. It is important to note that AI is rapidly developing, which researchers have to take into consideration to avoid producing or researching outdated information. Working with larger texts and corpora would also require more computational tools to facilitate such research. Lastly, since generative AI as such as is a recent development, there are huge gaps within the published literature which calls for more testing of different research methods.

Chapter Six

Conclusion

This study aimed to find and compare the linguistic similarities and differences between AI essays and essays written by master's students at the University of Oxford. Additionally, it aimed to determine University of Oxford lecturers' judgment of the quality of AI essays and essays written by master's students. Lastly, it aimed to find whether lecturers at the University of Oxford would be able to differentiate between essays written by human beings and essays generated by AI without external aid.

With respect to linguistic similarities, there were several common features found in both types of essays. These features were; the establishment of common applied linguistics abbreviations and at times, lack thereof; the use of first person plural pronouns and third person neutral singular pronouns; informal register at times; variation in types of subheadings used; use of American English or lack thereof of indicators; use of APA in-text citations; missing referent after the determiner 'this'; semantic errors; use of signposting in introductions; use of the terms 'this paper/essay'; numbers written in their word form; the use of colons and semicolons to separate clauses; unnecessary capitalization of certain words; and using a comma before 'etc.'.

With respect to linguistic differences, student essays had several different types of punctuation and spelling errors while AI essays did not; student essays had linguistic features that demonstrated personalization and authorship but AI essays did not; student essays were longer in length than AI essays; one student essay did not have a conclusion paragraph; student essays had the use of second person singular pronoun and first person singular and plural pronouns while AI essays only used first person plural pronouns; some student essays used British English; student essays had certain style markers while the only style marker in AI essay was the use of parentheses for definitions; student essays engaged with empirical literature but AI essays did not; an AI essay had an untitled preface; and student essays had a references section but AI essays did not.

In terms of Jaccard's Index this study found that statistically, AI essays and essays written by human beings are very different. Further, the probability of any AI essay and any human essay having been written by the same 'author' is extremely low. With respect to n-grams, AI essays have more 5-gram word sequences which reoccur more frequently than n-grams found in student essays, which is distinct from the human tendency to not repeat the exact string of words multiple times in discourse. With respect to ACJ, lecturers rated student essays better than AI essays on all occasions except one. For the Turing test, lecturers were able

to differentiate between AI essays and student essays for a majority of the cases albeit with difficulty and some uncertainty. Since this study is one of the first of its kind, there is great potential for future directions given the development and integration of AI in pedagogy and language learning spheres. More studies can be designed to examine different types of generative AI, different genres of texts written by different levels of students in multiple languages, as well as how to optimize AI in language learning and writing.

References

Abbasi, G., Zadeh, S. S., Janfaza, E., Assemi, A., & Dehghan, S. S. (2012). Language, Translation, and Culture. *International Conference on Language, Medias and Culture* (pp. 83-87). Chennai: IPEDR.

Ahmed, H. (2018). The Role of Linguistic Feature Categories in Authorship Verification. *The 4th International Conference on Arabic Computational Linguistics* (pp. 214-221). Dubai: Procedia Computer Science.

Akbari, A., & Shahnazari, M. (2014). HomoKult: A New Model in Translation. *SAGE Open, 4*(3).

Al-Hassan, A. (2013). The Importance of Culture in Translation: Should Culture be Translated? *International Journal of Applied Linguistics and English Literature, 2*(2), 96-100.

Ali, H. I., & Alhassan, A. (2021). Fighting contract cheating and ghostwriting in Higher Education: Moving towards a multidimensional approach. *Cogent Education, 8*(1), 1-18.

Ananda, R., Gani, S., & Sahardin, R. (2014). A Study of Error Analysis from Students' Sentences in Writing. *Studies in English Language and Education, 1*(2), 81-95.

Aronoff, M., & Fudeman, K. (2004). *What is Morphology?* (Vol. 43). West Sussex: Wiley-Blackwell.

Bartholomew, S. R., Strimel, G. J., & Yoshikawa, E. (2019). Using adaptive comparative judgment for student formative feedback and learning during a middle school design project. *International Journal of Technology and Design Education, 29*, 363-385.

Belvisi, N. M., Muhammad, N., & Alonso-Fernandez, F. (2020). Forensic Authorship Analysis of Microblogging Texts Using N-Grams and Stylometric Features. *Institute of Electrical and Electronics Engineers*, 1-6.

BinSaeedan, W., & Alramlawi, S. (2021). CS-BPSO: Hybrid feature selection based on chi-square and binary PSO algorithm for Arabic email authorship analysis. *Knowledge-Based Systems, 227*.

Blair, A., Curtis, S., Goodwin, M., & Shields, S. (2012). What Feedback do Students Want? *Politics, 33*(1), 66-79.

Boggs, J. (2019). Effects of teacher-scaffolded and self-scaffolded corrective feedback compared to direct corrective feedback on grammatical accuracy in English L2 writing. *Journal of Second Language Writing, 46*(4).

Bramley, T., Bell, J., & Pollitt, A. (1998). Assessing changes in standards over time using Thurstone Paired Comparisons. *Education Research and Perspectives, 25*(2), 1-24.

Breland, H. M., Kubota, M. Y., & Bonner, M. W. (1999). *The Performance Assessment Study in Writing: Analysis of the SAT® II: Writing Subject Test.* New York: College Entrance Examination Board.

Brighton, H., & Selina, H. (2010). *Introducing Artificial Intelligence: A Graphic Guide.* Icon Books.

Brodsky, J. E., Brooks, P. J., Scimeca, D., Todorova, R., Galati, P., Batson, M., . . . Caulfield, M. (2021). Improving college students' fact-checking strategies through lateral reading instruction in a general education civics course. *Cognitive Research: Principles and Implications, 6*, 1-18.

Brooker, C. (Writer), & Harris, O. (Director). (2013). *Black Mirror; Be Right Back* [Motion Picture].

Brown, D. (1994). *Principles of Language Learning and Teaching.* New Jersey: Prentice Hall Regents.

Burrows, J. F. (1987). *Computation Into Criticism: A Study of Jane Austen's Novels and an Experiment in Method* (Vol. 17). Oxford: Clarendon Press.

Castillo, E. (2023, March 27). *These Schools and Colleges Have Banned Chat GPT and Similar AI Tools*. Retrieved from BestColleges: https://www.bestcolleges.com/news/schools-colleges-banned-chat-gpt-similar-ai-tools/#:~:text=The%20most%20notable%20K%2D12,on%20content%20safety%20and%20accuracy.

Chechitelli, A. (2023, January 13). *Sneak preview of Turnitin's AI writing and ChatGPT detection capability*. Retrieved from turnitin: https://www.turnitin.com/blog/sneak-preview-of-turnitins-ai-writing-and-chatgpt-detection-capability

Chen, X., Hao, P., Chandramouli, R., & Subbalakshmi, K. (2011). Authorship Similarity Detection from Email Messages. *International Workshop on Machine Learning and Data Mining in Pattern Recognition* (pp. 375-386). Dresden: Machine Learning and Data Mining in Pattern Recognition.

Cho, T. (2010). Linguistic Features of Electronic Mail in the Workplace: A Comparison with Memoranda. *Language @ Internet, 7*.

Christ Church College. (2023). *Planning and structuring an essay* . Retrieved from Christ Church College: https://www.chch.ox.ac.uk/current-students/planning-and-structuring-essay

Clarvoe, J. (2009). Poetry and Repetition. *Antioch Review Inc., 67*(1).

Coffin, C. J., Curry, M. J., Goodman, S., & Hewings, A. (2003). *Teaching Academic Writing: A Toolkit for Higher Education.* Routledge.

Coulthard, M. (2010). Experts and opinions. In M. Coulthard, & A. Johnson, *The Routledge Handbook of Forensic Linguistics* (pp. 473-486). Abingdon: Routledge.

Coulthard, M., & Johnson, A. (2007). *An Introduction to Forensic Linguistics.* Oxon: Routledge.

Crisp, B. R. (2007). Is it worth the effort? How feedback influences students' subsequent submission of assessable work. *Assessment & Evaluation in Higher Education, 32*, 571-581.

Danielewicz-Betz, A. (2012). The Role of Forensic Linguistics in Crime Investigation. In A. Littlejohn, & S. Mehta, *Language Studies: Stretching the Boundaries* (pp. 93-108). Cambridge: Cambridge Scholars Publishing.

Davis, J., & Liss, R. (2006). *Effective Academic Writing 3: The Essay.* Oxford: Oxford University Press.

Eggington, W. (2008). Deception and Fraud. In J. Gibbons, & M. Turell, *Dimensions of Forensic Linguistics* (pp. 249-264). John Benjamins Publishing Company.

Fairclough, N. (1989). *Language and Power.* London: Longman.

Ferris, D. (1999). The Case for Grammar Correction in L2 Writing Classes: A Reponse to Truscott (1996). *Journal of Second Language Writing, 8*, 1-11.

Garner, H. R. (2011). Combating unethical publications with plagiarism detection services. *Urol Oncol, 29*(1), 95-99.

Government of France. (2023). Loi du 23 décembre 1901 réprimant les fraudes dans les examens et concours publics.

Government of India. (2023). *Previous Year Question Papers*. Retrieved from Union Public Service Commission: https://www.upsc.gov.in/examinations/previous-question-papers

Grabe, W. (1984). Written Discourse Analysis. *Annual Review of Applied Linguistics, 5*, 101-123.

Graesser, A., McNamara, D., Louwerse, M., & Zhiqiang, C. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*, 193-202.

Grant, T. (2010). Text messaging forensics. In M. Coulthard, & A. Johnson, *The Routledge Handbook of Forensic Linguistics* (pp. 508-522). Abingdon: Routledge.

Halliday, M., & Hasan, R. (1976). *Cohesion in English.* London: Longman.

Hill, J., & Page, E. (2009). An Empirical Research Study of the Efficacy of Two Plagiarism-Detection Applications. *Journal of Web Librarianship, 3*(3), 169-181.

Hinkel, E. (2005). Hedging, Inflating, and Persuading in L2 Academic Writing. *Applied Language Learning, 15*, 29-53.

Houle, A. (2013). Délit de langue et paternité textuelle Une approche informatisée. *Master's Dissertation for Université Laval*, 1-129.

Hounsell, D., McCune, V., Hounsell, J., & Litjens, J. (2008). The quality of guidance and feedback to students. *Higher Education Research & Development, 27*, 55-67.

Hu, K. (2023, February 2). *Reuters*. Retrieved from ChatGPT sets record for fastest-growing user base - analyst note: https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Hyland, K. (2002). Genre: Language, Context, and Literacy. *Annual Review of Applied Linguistics, 22*, 113-135.

Incorporated, A. (2023). *Siri*. Retrieved from Apple: https://www.apple.com/siri/

International Baccalaureate. (2023). *Sample exam papers*. Retrieved from International Baccalaureate, 2023: https://www.ibo.org/programmes/diploma-programme/assessment-and-exams/sample-exam-papers/

Jabbar, A. N., & Al-Aadili, N. M. (2022). Impoliteness in Online Comments in The Johnny Depp and Amber Heard Trial Authors. *Res Militaris, 12*(2), 4350-4364.

Jones, I., Swan, M., & Pollitt, A. (2015). Assessing Mathematical Problem Solving Using Comparative Judgement. *International Journal of Science and Mathematics Education, 13*, 151-177.

Kagan, D. M. (1980). Run-on and Fragment Sentences: An Error Analysis. *Research in the Teaching of English, 14*(2), 127-138.

Kallimani, J., Chandrika, C., Singh, A., & Khan, Z. (2020). Authorship Identification Using Supervised Learning and n-Grams for Hindi Language. *Journal of Computational and Theoretical Nanoscience, 17*, 4258-4261.

Kasneci, E., Sebler, K., Kuechemann, S., Bannert, M., Dementieva, D., Fischer, F., . . . Pfeffer, J. (2023). ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences, 103*, 1-13.

Kerr, J. (2013, November 29). *The history of the Roomba*. Retrieved from Fortune: https://fortune.com/2013/11/29/the-history-of-the-roomba/

Kestemont, M. (2014). Function Words in Authorship Attribution. From Black Magic to Theory? *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)* (pp. 59-66). Gothenburg: Association for Computational Linguistics.

Kimbell, R. (2022). Examining the reliability of Adaptive Comparative Judgement (ACJ) as an assessment tool in educational settings. *International Journal of Technology and Design Education, 32*, 1515-1529.

Kleerekoper, A., & Schofield, A. (2019). The False-Positive Rate of Automated Plagiarism Detection for SQL Assessments. *The 1st UK & Ireland Computing Education Research Conference*, (pp. 1-6). Swansea.

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for Language Teaching and Learning. *RELC Journal*.

Kubrick, S. (Director). (1968). *2001: A Space Odyssey* [Motion Picture].

Lacasse, J. R., & Leo, J. (2010). Ghostwriting at Elite Academic Medical Centers in the United States. *PLoS Med, 7*(2).

Lang, J. M. (2013). *Cheating Lessons.* Cambridge: Harvard University Press.

Lee, J., & Macaro, E. (2013). Investigating age in the use of L1 or english-only instruction: Vocabulary acquisition by Korean EFL learners. *The Modern Language Journal, 97*(4), 887-901.

Leonard, R., Ford, J., & Christensen, T. (2017). Forensic Linguistics: Applying the Science of Linguistics to Issues of the Law. *Hofstra Law Review, 45*(3), 881-897.

Leshem, S., & Bitzer, E. (2021). 'SIGNPOSTING' RESEARCH STORIES IN DOCTORAL THESES: WRITING THAT KEEPS THE READER IN MIND. *Per Linguam, 38*(1), 109-123.

Lines, L. (2016). Substantive editing as a form of plagiarism among postgraduate students in Australia. *Assessment & Evaluation in Higher Education, 41*(3), 368-383.

Liu, P., & Li, Z. (2012). Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics, 42*, 553-568.

Liu, Y. (2008). The Effects of Error Feedback In Second Language Writing. *Journal of Second Language Acquisition and Teaching, 15*, 65-79.

Lyster, R., & Ranta, L. (1997). Corrective Feedback and Learner Uptake. *Studies in Second Language Acquisition, 19*(1), 37-66.

Macleoad, N., & Grant, T. (2011). Whose tweet?: authorship analysis of micro-blogs and other short form messages. *International Association of Forensic Linguists' 10th Biennial Conference*, (pp. 1-19). Birmingham.

Maitland, H. (2023). When the facts are not enough: the limitations of fact-checking sex education controversies. *Sexuality, Society and Learning, 23*(3), 324-333.

Majumder, M. A., Kumar, A., Krishnamurthy, K., Ojeh, N., Adams, O. P., & Sa, B. (2019). An evaluative study of objective structured clinical examination (OSCE): students and examiners perspectives. *Advances in Medical Education and Practice*, 387-397.

Malini, N. L., & Tan, V. (2016). Forensic Linguistics Analysis of Virginia Woolf's Suicide Notes. *International Journal of Education, 9*(1), 52-57.

Marr, B. (2023, March 3). *The Top 10 Limitations of ChatGPT*. Retrieved from Forbes: https://www.forbes.com/sites/bernardmarr/2023/03/03/the-top-10-limitations-of-chatgpt/?sh=1af7dca48f35

McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. E. (1955). A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE.

McCarthy, M. (1991). *Discourse Analysis for Language Teachers.* Cambridge: Cambridge University Press.

Mcmenamin, G. (2001). Style markers in authorship studies. *International Journal of Speech, Language and the Law, 8*(2), 93-97.

McNamara, D., Crossley, S., & McCarthy, P. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57-86.

Metzgar, M. (2016). Using Adaptive Comparative Judgement to Assess Student Work in an MBA Course. *International Journal for Infonomics, 9*(3), 1217-1220.

Mohmad, P. (2023, January 19). *How Does ChatGPT Become Popular So Quickly and How Is It Growing* . Retrieved from Analytics Insight: https://www.analyticsinsight.net/how-does-chatgpt-become-popular-so-quickly-and-how-is-it-growing/

Moonbeam. (2023). *Never write from scratch again.* Retrieved from Moonbeam: https://www.gomoonbeam.com/

Moor, J. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *Institute of Electrical and Electronics Engineers, 21*, 18-21.

Mutch, A. (2003). Exploring the Practice of Feedback to Students. *Active Learning in Higher Education, 4*(1), 24-38.

Myers, D. E. (1992). Teaching Adults: A Changing Perspective. *Journal of General Music Education, 5*(2), 23-24.

Nini, A. (2018). Article Navigation JOURNAL ARTICLE An authorship analysis of the Jack the Ripper letters. *Digital Scholarship in the Humanities, 33*(3), 621-636.

No More Marking. (2023). *No More Marking*. Retrieved from Innovative Assessment for Schools: https://www.nomoremarking.com/?countryCode=US

Olsson, A., & Engelbrektsson, O. (2022, May 12). A thesis that writes itself On the threat of AI-generated essays within academia. *Bachelor's Thesis for Halmstad University*, pp. 1-38.

Pecorari, D. (2003). Good and original: Plagiarism and patchwriting in academic second-language writing. *Journal of Second Language Writing, 12*(4), 317-345.

Pecorari, D. (2008). Repeated language in academic discourse: The case of biology background statements. *Nordic Journal of English Studies*, 9-33.

Phani, S., Lahiri, S., & Biswas, A. (2015). Authorship Attribution in Bengali Language. *12th International Conference on Natural Language Processing* (pp. 100-105). Trivandrum: NLP Association of India.

Poteat, H. M. (1919). The Functions of Repetition in Latin Poetry (Concluded). *The Classical Weekly, 12*(19), 145-160.

Poulos, A., & Mahony, M. (2008). Effectiveness of feedback: the students' perspective. *Assessment & Evaluation in Higher Education, 33*(2), 143-154.

Princeton University. (2023). *WordNet*. Retrieved from Princeton University: https://wordnet.princeton.edu/

Qatar University. (2023). *The Basics of Academic Writing.* Retrieved from Qatar University: https://esc.qu.edu.qa/static_file/qu/students/documents/basics_of_academic_writing.pdf

Ragel, R., Herath, P., & Senanayake, U. (2014, March 6). Authorship Detection of SMS Messages Using Unigrams. *arxiv*, pp. 1-6.

Rawlinson, G. (1976). The Significance of Letter Position in Word Recognition. *PhD Thesis for Nottingham University*.

Raza, A., Athar, A., & Nadeem, S. (2009). N-Gram Based Authorship Attribution in Urdu Poetry. *Conference on Language & Technology* (pp. 88-94). Lahore: Carnegie Mellon University.

Saevanee, H., Clarke, N., & Furnell, S. (2011). SMS Linguistic Profiling Authentication on Mobile Device. *Saevanee, H., Clarke, N., & Furnell, S. (2011). SMS linguistic profiling authentication on mobile device. 2011 5th International Conference on Network and System Security, 224-228.* (pp. 224-228). IEEE.

Salecha, M. (2016, October 5). *Story of ELIZA, the first chatbot developed in 1966*. Retrieved from Analytics India Magazine: https://analyticsindiamag.com/story-eliza-first-chatbot-developed-1966/

Saunders, C. (2019, December 21). *Bad Monster Blood: R.L. Stine, Scholastic and the Battle over Goosebumps*. Retrieved from The Avocado: https://the-avocado.org/2019/12/21/bad-monster-blood-r-l-stine-scholastic-and-the-battle-over-goosebumps/

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417-424.

Seery, N., Buckley, J., Delahunty, T., & Canty, D. (2019). Integrating learners into the assessment process using adaptive comparative judgement with an ipsative approach to identifying competence based gains relative to student ability levels. *International Journal of Technology and Design Education, 29*, 701-715.

Seery, N., Canty, D., & Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *International Journal of Technology and Design Education, 22*, 205-226.

Seifollahi, S., Bagirov, A., Layton, R., & Gondal, I. (2017). Optimization Based Clustering Algorithms for Authorship Analysis of Phishing Emails. *Optimization Based Clustering Algorithms for Authorship Analysis of Phishing Emails.* (pp. 411-425). Neural Processing Letters.

Shapero, J. J. (2011). The language of suicide notes. *PhD Thesis for University of Birmingham*, 1-283.

Silessi, S., Varol, C., & Karabatak, M. (2016). Identifying Gender from SMS Text Messages. *nternational Conference on Machine Learning and Applications* (pp. 488-491). Anahein: IEEE.

Singh, S., & Remenyi, D. (2018). Plagiarism and ghostwriting: The rise in academic misconduct. *South African Journal of Science, 112*, 1-7.

Sinha, A. (Director). (2011). *Ra One* [Motion Picture].

Slutsky, D. J. (2014). The Effective Use of Graphs. *Journal of Wrist Surgery, 3*(2), 67-68.

Sousa-Silva, R. (2014). Investigating academic plagiarism: A forensic linguistics approach to plagiarism detection. *International Journal for Educational Integrity, 10*(1), 31-41.

Sousa-Silva, R. (2019). Plagiarism Across Languages and Cultures: A (Forensic) Linguistic Analysis. In S. D. Brunn, & R. Kehrein, *Handbook of the Changing World Language Map* (pp. 2325-2345). SpringerLink.

Spasić, I., Burnap, P., Greenwood, M., & Arribas-Ayllon, M. (2012). A naïve bayes approach to classifying topics in suicide notes. *Biomed Inform Insights*, 87-97.

Stanford University. (2023, February 16). *Generative AI Policy Guidance*. Retrieved from Stanford University: https://communitystandards.stanford.edu/generative-ai-policy-guidance

Stephens, M. (2023, March 2). *University of Cambridge will allow students to use ChatGPT*. Retrieved from The Telegraph: https://www.telegraph.co.uk/news/2023/03/02/university-cambridge-will-allow-students-use-chatgpt/

Sudjana, E. T., & Fitri, N. (2013). Kurt Cobain's Suicide Note Case: Forensic Linguistic Profiling Analysis. *International Journal of Criminology and Sociological Theory, 6*(4), 217-228.

Tang, R., & John, S. (1999). The "I" in Identity: Exploring Writer Identity in Student Academic Writing through the First Person Pronoun. *English for Specific Purposes, 18*, 23-29.

The Writing Center. (2023). *Argument.* Retrieved from The Writing Center University of North Carolina at Chapel Hill: https://writingcenter.unc.edu/tips-and-tools/argument/

Thurstone, L. L. (1927). A Law of Comparative Judgment. *A Law of Comparative Judgment, 34*(4), 273-286.

Truscott, J. (1996). The Case against Grammar Correction in L2 Writing Classes. *Language Learning, 46*, 327-369.

Tuzi, F. (2004). The impact of e-feedback on the revisions of L2 writers in an academic writing course. *Computers and Composition, 21*(2), 217-235.

Tyagnova, T., & Greiffenhagen, C. (2021). "Learning for the test": A study of students' sharing and discussing past multiple choice questions on Facebook. *Ethnographic Studies, 18*, 43-74.

University of Birmingham. (2023). *A short guide to signposting in essays.* Retrieved from University of Birmingham: https://intranet.birmingham.ac.uk/as/libraryservices/library/asc/documents/public/short-guide-signposting.pdf

University of Hull. (2023). *Writing academically: Abbreviations.* Retrieved from University of Hull: https://libguides.hull.ac.uk/writing/abbreviations

University of Oxford. (2023). *MSc in Applied Linguistics and Second Language Acquisition*. Retrieved from University of Oxford:

https://www.ox.ac.uk/admissions/graduate/courses/msc-applied-linguistics-and-second-language-acquisition?wssl=1#content-tab--2

University of Oxford. (2023, February 6). *Unauthorised use of AI in exams and assessment*. Retrieved from University of Oxford: https://academic.admin.ox.ac.uk/article/unauthorised-use-of-ai-in-exams-and-assessment

University of South Florida. (2023). *Semicolons and Colons.* Retrieved from University of South Florida: https://www.usf.edu/undergrad/academic-success-center/documents/ws-semicolons-colons.pdf

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A Meta-Analysis on the Reliability of Comparative Judgement. *Assessment in Education: Principles, Policy & Practice, 26*(5), 541-562.

Villamil, O., & De Guerrero, M. C. (1998). Article Navigation JOURNAL ARTICLE Assessing the Impact of Peer Revision on L2 Writing. *Applied Linguistics, 19*(4), 491-514.

Waugh, R. (2023, March 26). *'Half of school and college students are already using ChatGPT to cheat': Experts warn AI tech should strike fear in all academics*. Retrieved from Daily Mail: https://www.dailymail.co.uk/sciencetech/article-11899475/Half-students-using-ChatGPT-cheat-rise-90.html

Weber-Wulff, D. (2019). Plagiarism detectors are a crutch, and a problem. *Nature*.

Wheadon, C. (2023, July 15). *Would you ask GPT-4 to mark your essays?* Retrieved from No More Marking: https://substack.nomoremarking.com/p/would-you-ask-gpt-4-to-mark-your

White, M. J. (2023, June 1). *Top 20 Most Insane Things ChatGPT Has Ever Done [Updated]*. Retrieved from Springboard: https://www.springboard.com/blog/news/chatgpt-revolution/

Williams, K., & Carroll, J. (2011). *Referencing and Understanding Plagiarism (Pocket Study Skills).* Macmillan Education UK.

Yao, D. (2022, May 10). *25 years ago today: How Deep Blue vs. Kasparov changed AI forever*. Retrieved from AI Businesss: https://aibusiness.com/ml/25-years-ago-today-how-deep-blue-vs-kasparov-changed-ai-forever

Yasin, H., Yasin, M. M., & Yasin, F. M. (2011). Automated Multiple Related Documents Summarization via Jaccard's Coefficient. *International Journal of Computer Applications, 13*(3), 12-15.

Yi, N., Nemery, B., & Dierick, K. (2020). Perceptions of plagiarism by biomedical researchers: an online survey in Europe and China. *BMB Medical Ethics, 21*(1).

Yigzaw, A. (2012). Impact of L1 Use in L2 English Writing Classes. *Ethiopian Journal of Education and Sciences, 8*, 11-28.

Zhang, Z., & Zhang, Y. (2018). Automated Writing Evaluation System: Tapping its Potential for Learner Engagement. *Institute of Electrical and Electronics Engineers, 46*(3), 29-33.

Zheng, J., Napoles, C., Tetreault, J., & Omelianchuck, K. (2018). How do you correct run-on sentences it's not as easy as it seems. *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text* (pp. 33-38). Brussels: Association for Computational Linguistics.

Appendices

Appendix A; Corpus of ALSLA essays

ALSLA Essay 1

The purpose of this essay is to explore the concept of 'lexical richness' in the context of second language (L2) vocabulary learning, and what it can inform us about an L2 learners' vocabulary knowledge. First, it will briefly explain how the linguistic concepts of 'L2 vocabulary' and 'lexical richness' are operationalised. Next, this paper will explain the relationship between lexical richness and L2 vocabulary knowledge by reviewing the relevant literature over the past 30 years to help understand what we can learn about an learners L2 vocabulary. It will attempt to critically assess how well theoretical models of richness can help explain L2 vocabulary knowledge. A growing body of evidence from empirical second language acquisition (SLA) research seems to suggest, that one's superiority, or sophistication of lexical ability, allows for deeper and more complex systems within the mental lexicon to be accessed and processed, aiding not only in overall understanding and comprehension, but also in terms of successful speech and writing production. Thus, theoretically speaking, a richer L2 lexicon may be associated with a larger vocabulary knowledge and subsequent higher literacy abilities.

Conceptualisations of lexical richness and vocabulary knowledge

The vast majority of the literature has approached the problem of trying to conceptualise the highly complex nature of L2 vocabulary knowledge, in two distinct ways. First through a 'developmental' approach, where vocabulary knowledge can be seen as a continuum of the development of word knowledge; breadth versus depth (Wesche & Paribakht, 1996), and more recently, through a 'dimensions' approach, quantifying lexical components of receptive and productive knowledge (González-Fernández, 2022). Nation (2006) suggests 6,000-7,000 word families are required for the effective L2 production of spoken forms. Schmitt et al (2017) estimate a rough lexical coverage rate of between 95-98% for 'acceptable' reading comprehension. A widely supported comprehensive model of word knowledge is Nation's (2013), word knowledge model, which measures both the receptive and productive aspects of L2 vocabulary (Cheng & Matthews, 2018; Li & Kirby, 2015; Webb, 2008). However, the issue of how to measure these lexical features has largely been a difficult endeavour for SLA research, and has sparked much debate (see Webb 2021). Meara (1996), points out, the problem with measuring L2 vocabulary is that often these tests fail to accurately measure what they intend to find out. Furthermore, L2 learners learn and acquire words in different ways than native speakers, which is hazardous for SLA researchers - who often used widely accepted - but very standardised vocabulary tests, failing to capture the true level of L2 vocabulary.

What does lexical richness tell us about an L2 vocabulary?

Lexical richness is important as it can be used to examine the relationship between vocabulary knowledge and vocabulary use (Laufer and Nation, 1995). In other words, it provides insight into the inner workings of the L2 lexicon and provides some clues as to how it is not only constructed, but also how it is accessed and used. So far studies have revealed that; L2 receptive vocabulary is generally much larger than L2 productive knowledge (Webb, 2008), that receptive vocabulary size is strongly associated with L2 reading, and interestingly as Schmitt et al (2017) found, also in written production. This shows the complex lexical network at work, showing a direct relationship between receptive and productive knowledge. Durrant and Schmitt (2009), found that adult EAP L2 learners reply more on low frequency collocations in written text, providing some evidence that L2 learning learn in similar formulaic ways as L1 children, providing support to the constructivist views by Ellis et. al. (2008), who argue, given

enough exposure to an L2, language develops similarly to native speakers. Similarly, Saito (2020) found that in terms of speaking, Japanese L2 speakers were significantly more understood by natives when using more low frequency collocations. The implications of this are significant, suggesting a richer knowledge of MWU can facilitate greater communication abilities. Indeed, the effect of learning word chunks has a snowball effect, allowing for deeper and more meaningful extrapolations to linguistic forms (Tomasello, 2003).

One particular carefully planned study by Sonbul and Schmitt (2009), showed that explicit classroom instruction in a ESP medical school, resulted in far greater gains in L2 learning compared to implicit leaning alone. This shows a key difference in the way that L2 learners acquire vocabulary compared to incidental learning, which is dominant within L1 literature (Nagy, 1997). Nation (2022), also notes that L2 implicit learning conditions are often much different than L1 conditions. Three lexical measurement tests of low frequency words were taken directly from the core textbook used on the ESP programme. Future studies should consider the L2 backgrounds of their subjects, and adapt their measurement techniques accordingly. Although the nature of a tailored test is extremely effective in measuring vocabulary knowledge for a particular group of L2 learners, it is also extremely inefficient, and thus, not always practical. Nevertheless, it is important to consider whether or not tests are appropriate for what they intend to measure (Meara, 1996; Webb, 2021). One extremely confounding variable which was glossed over in the study was that over 50% of participants restudied the target words between tests. A throwaway line was acknowledged in the study about affecting the results, but it was not discussed at length, or indeed if adjustments were made for the unexpected intervention.

To summarise, L2 vocabulary is highly complex, and is extremely hard to know exactly how words are stored, processed and produced with the lexical networks. The problem of operationalisation of words to date has created much debate within the literature. Nevertheless, scholarly work in the field has provided some excellent empirical evidence to explain more about L2 vocabulary. Studies presented in this short essay have shown a somewhat significant relationship between lexical richness and an L2 vocabulary knowledge. We saw in Schmitt et. al.'s (2017), study how the receptive and productive components are interwoven, and as larger L2 productive vocabulary is dependent on receptive vocabulary, we can thus infer greater lexical richness on the receptive side would naturally result in a larger (and richer) productive side. Conversely, MWU studies by Durrant and Schmitt (2009), and Saito (2020), show links that L2 learning occurs in similar ways to L1, but often lack lexical diversity and richness. The implications for pedagogy are that L2 may need further explicit instruction in MWU, just as Sonbul and Schmitt (2009), found in their ESP classroom study. It should be noted, that an overwhelming number of studies in the field are focussed on the English L2 learner, so we should be cautious in overgeneralising statements to other languages without empirical evidence. Moreover, current debate is still ongoing about appropriate measures of vocabulary knowledge so the validity of previous research must be taken into consideration when assessing previous outcomes of L2 vocabulary research.

ALSLA essay 2

This assignment is composed of three sections: an introductory section about the conceptualizations and operationalizations of lexical richness and vocabulary knowledge, followed by a tentative discussion on the relationship between lexical richness and learners' vocabulary knowledge, and then a conclusion.

1. Lexical richness and vocabulary knowledge

While lexical richness has been defined by different researchers (e.g. Laufer & Nation, 1995; Wimmer & Altmann, 1999; Daller, Turlik, & Weir, 2003; Bulté & Housen, 2012) in various ways, it refers to the quality of vocabulary in a written or spoken language sample in a broad sense. The concept of lexical richness has been operartionalized to characterize vocabulary use with a number of subordinate dimensions primarily including, but not limited to, lexical diversity and lexical sophistication (Crossley, Zhang, & Schache, 2011b; Read, 2000; Lu, 2012; Kyle, 2019). Lexical diversity concerns the range of unique words in a text (Daller et al., 2003; Yu, 2010; Zhang, 2014), while lexical sophistication measures the percentage of relatively low-frequency or advanced words (Laufer & Nation, 1995; Kim, Crossley, & Kyle, 2017). Indices of the lexical richness are useful for analyzing features of vocabulary use in written or spoken texts (Milton, 2009; Kyle, 2019). Vocabulary knowledge is recognized to be of significance for language use both in theoretical psycholinguistic dimensions underlying L2 production (Skehan, 2009) and empirical studies on L2 performance (de Jong et al., 2012; Dabbagh & Enayat, 2019). Before proceeding to discuss what lexical richness could tell us about learner's vocabulary knowledge, it is important to know the multiple dimensions involved in vocabulary knowledge. Researchers (Cronbach, 1942; Richards, 1976; Nation, 1990, 2001) have proposed various, yet complementary, taxonomies of vocabulary knowledge. Nation (2001) produces the most influential and comprehensive framework currently available (John, 2004; Daller, Milton, & Treffers-Daller, 2007; Schmitt, 2008; Milton, 2013). This framework categorizes vocabulary knowledge into three interrelated aspects: form, meaning and use. Each of them is subcategorized into three dimensions and additionally distinguished by both receptive and productive knowledge (see Table 1). The receptive-productive distinction is helpful for estimating learners' level of mastery of vocabulary at a receptive level (reading and listening), and at a productive level (writing and speaking) (Read, 2004; Schmitt, 2014).

2. The relationship between lexical richness and learners' vocabulary knowledge

As mentioned earlier, lexical diversity and lexical sophistication are two primary subordinate dimensions of lexical richness (Crossley et al., 2011b; Read, 2000; Lu, 2012; Kyle, 2019), and therefore let us discuss what each of them could tell us about learners' vocabulary knowledge. Lexical diversity is arguably an essential predictor of learners' productive vocabulary knowledge considering that it has been noted as being closely aligned with the quality of writing (e.g. Laufer and Nation 1995; Engber, 1995; Jarvis 2002) and speaking (e.g. O'Loughlin 1995; Daller et al., 2003; Malvern and Richards 2002) performances. However, inclusive findings were reported by Vermeer (2000) that employed a set of traditional measures to analyze the lexical diversity in spontaneous speech data of L1 and L2 Dutch children. The study found that the indices of lexical diversity obtained through different measures did not correlate with either the receptive or productive vocabulary test. To put it in another way, his research showed that there was not a clear correspondence between lexical diversity of spoken texts and students' vocabulary knowledge. It can be argued that Vermeer's (2000) conflicting findings may be attributed to his use of traditional lexical diversity measures such as the Type-Token Ratio (TTR). A number of researchers (e.g., Richards, 1987; Johnson, 1944; Malvern & Richards, 1997; McCarthy & Jarvis, 2010; Kyle, 2018) have strongly criticized that the TTR

is not a reliable measure because of its instability and extreme sensitivity to text length. In empirical studies (e.g., Broeder et, Extra, & Van Hou, 1993; Malvern, Richards, Chipere, & Durán, 2004), this measure has been found to produce widely varying values intrinsically skewed by the length of language samples despite the vocabulary knowledge of the same participants' remaining constant. Alternative measures such as D (Malvern & Richards, 1997; McCarthy & Jarvis, 2007) might present supporting evidence for the relationship between lexical diversity measures and learners' vocabulary knowledge. Taken together, lexical diversity obtained through valid and reliable measures is an indicator of students' productive vocabulary knowledge. However, it should be noted that it is far away from providing a complete picture of students' vocabulary knowledge. For instance, the measure does not consider how a word is used in the text, whether the word form is grammatically correct, or whether the meaning is correct in a particular situation (see Table 1 for Nation's (2001) vocabulary taxonomy).

Similar to lexical diversity, lexical sophistication also taps into learners' productive vocabulary knowledge because of the correspondence between sophisticated lexical production and writing quality (e.g. Laufer & Nation, 1995; McNamara, Crossley, & McCarthy, 2010; Crossley and McNamara, 2012, Guo, Crossley, & McNamara, 2013; Crossley & Kyle, 2018). However, the majority of the studies on the analysis of lexical sophistication heavily relied on frequency-based measures, which might underrepresent the multi-dimensional nature of students' vocabulary knowledge (Nation, 2001; Koizumi, 2012). Perhaps the best-known frequency-based measure is the Lexical Frequency Profile that was developed by Laufer and Nation (1995) to analyze the lexical sophistication of 65 L2 essays by classifying the words into four different frequency bands. They found that the LFP discriminated between students of different language proficiency levels, provided stable outcomes for two different compositions written by the same learners, and correlated highly with an independent measure of productive vocabulary knowledge. A number of subsequent empirical studies have supported the use of the LFP as an effective method to study or trace the lexical development of students (Lenko-Szymanska, 2002; Horst and Collins, 2006), predict learners' academic performance (Morris & Cobb, 2004), study the contribution of dictionary use to writing (East, 2005) and so forth. Researchers (e.g. Lin & Morrison, 2010; Catalán & Llach, 2017) have also offered support for using the LFP as a way for program designers and instructors to understand learners' vocabulary needs. Despite that Laufer & Nation (1995) provides a useful approach into obtaining the lexical richness of a text, the LFP is of inherent drawbacks and one could argue that it may not be able to provide us with a complete and detailed picture of learners' vocabulary use (Lamy and Jones, 2002; Higginbotham and Reid, 2019). The principal shortcoming of the LFP is that it removes word errors including incorrect derived forms from the analysis. The removal of lexical errors is likely to bias the lexical richness outcomes considering the relationship between the accuracy of lexical use and production performance (Astika 1993; Ellis, 1994; Engber, 1995; Wolfe-Quintero, Inagaki, & Kim, 1998). Moreover, the LFP does not take account of formulaic language such as collocations (Meara & Bell, 2001). A prototypical example provided by Muncie (2002), an empirical study using the LFP, is that instead of recognizing the expression put out as a collocation or a sophisticated use of vocabulary, the program would assign the two single words into the most frequent 1,000 range, respectively. This is surprising because that the use of formulaic language is an important part of lexical competence (Biber et al., 1999; Nation, 2001; Laufer & Waldman, 2011). In recognition of this problem, many researchers (e.g. Nation & Webb, 2011; Bulté & Housen, 2012; Bestgen & Granger, 2014) have suggested that the measure of lexical richness should include collocations to tap into the "use" dimension of Nation's (2001) vocabulary knowledge taxonomy (see Table 1). The inclusion of collocations in the measure of L2 lexical richness is

particularly warranted, as L2 learners tend to have limited ability to produce correct and idiomatic collocations (Nesselhauf, 2003; Webb & Kagimoto, 2009; Laufer & Waldman, 2011). Aware of the inherent shortcomings of frequency-based measures, researchers suggest that multiple values beyond frequency can be employed to analyze lexical sophistication (Crossley, Salsbury, & McNamara, 2009; Kyle, Crossley, & Berger, 2017).

3. Conclusion

Taken together, lexical richness obtained through reliable measures is an essential indicator of learners' productive vocabulary knowledge (Laufer & Nation, 1995). However, it is far away from providing us with a complete and detailed picture of students' vocabulary knowledge. Multiple reliable quantitative and qualitative measures can be used to tap into the multi-dimension nature of vocabulary knowledge (Nation, 2001).

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA (pp. 21–46). Amsterdam: John Benjamins.
Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. System, 41, 965–981.
Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical density in FL written production. Applied Linguistics, 16, 307–322.

ALSLA essay 3

The way the human mind acquires language is widely debated, but understanding how infants grasp their first language is only one part of the story. Second language acquisition is a complex process whereby many facets must be considered. An L2 learners' vocabulary knowledge is one window into understanding their proficiency level and lexical richness is a term which describes the depth of a learner's knowledge. Laufer and Nation (1995) examined lexical richness as how varied a learner's vocabulary is, as well as its size, among other features. We can examine this concept to understand how it may inform us of an L2 learner's vocabulary knowledge which in turn could help guide pedagogical approaches. This paper will address the question stated above by firstly considering Read's (2000) dimensions approach whilst taking a close look at collocations. Next, we will examine vocabulary size before finally probing creativity in learner's vocabulary knowledge and considering morpheme productivity and word associations.

Part 1

The dimensions approach as proposed by Read (2000) is one way to address the concept of vocabulary depth. Read regards vocabulary knowledge as a gradual process rather than something learners either have or don't have. He proposed that vocabulary knowledge is multidimensional, and each aspect of word knowledge can be ranked on a scale from 'no knowledge' to 'full mastery'. Various dimensions constituting word knowledge include spoken and written form, register and meaning. These components all contribute to lexical richness as it demonstrates the different layers to word knowledge. When an L2 learner scores the word highly (closer to full mastery), it indicates a more comprehensive vocabulary knowledge. The ability to use a word in many contexts (whether syntactic contexts e.g., the word in its figurative or literal sense, or physical contexts e.g. high register written piece in a journal or lower register conversation at the market) is relevant to lexical richness. This is because the learner demonstrates they can manipulate their vocabulary knowledge therefore demonstrating deeper lexical competence and knowledge.

We find the 'collocation' aspect of Read's dimensions approach particularly relevant when discussing L2 learners' lexical richness. Whilst it is unclear to what extent L2 learners use chunking as a method of acquisition likened to children learning their L1 (Wray 2002), it has been reported that collocations are necessary to make L2 speech sound more native like (Pawley & Syder, 1983). Saito (2020) states that collocations make a learner's speech more comprehensible. Some words can appear in many collocations, or multiword units, and understanding how the given word may change meaning or indeed serve varying functions in each collocation means the language user has a good grasp on that word's meaning. This indicates they can proficiently understand or use that word in multiple different contexts and in frequently occurring combinations of words. This mark of lexical richness as defined earlier therefore subsequently denotes the L2 learner's vocabulary knowledge as higher level and more native like as the lexical richness here achieved through engaging with collocations makes the language user more comprehensible and fluent. Returning to Laufer and Nation's (1995) definition, it demonstrates the variation in a learner's lexicon which is crucial for lexical richness. However, this approach has been criticised for being very difficult to measure reliably as learners subjectively self-report their own vocabulary knowledge.

Part 2

Whilst the dimensions approach focuses largely on the numerous usage contexts of a word, some theorists assign importance to determining the number of words needed to constitute a

sufficiently large vocabulary (particularly for receptive knowledge). The lexical richness of that speaker's lexicon will influence to what extent they can understand and make themselves understood. We can think of how many words are needed to constitute that 'richness'. Schmitt, Horst, Cobb & Schmitt (2017) refer to lexical coverage and vocabulary size. The former refers to the number of words in a stimulus one must know to arrive at the most relevant interpretation of it, and the latter refers to how many words that learner requires in their lexicon to meet said lexical coverage target. In order to understand 98% of a spoken discourse, Stæhr (2008) claims L2 learners require 5000-6000 word families, whilst Nation (2006) argues that 6000-7000 are needed. With written texts containing more low frequency words, to reach that 98% lexical coverage target, Hsueh-Chao & Nation (2000) argue that 8000-1000 word families are required.

The deliberations of quantifying vocabulary knowledge is directly relevant here, since the researchers referred to 'word families'. This means the individual must know the inflected forms of that word. As discussed in part three, morphology is one aspect of lexical richness, which in turn contributes to vocabulary knowledge. Therefore, even saying we need word families delves into the idea of lexical richness as it's not just about knowing a word or not but is about a sensitively interrelated and interdependent system whereby L2 learners' lexical richness directly improves their vocabulary knowledge.

An L2 learner's lexicon may reveal a lot about their motivation for learning. For example, if the learner has more jargon or specialist vocabulary, we may say their language is more lexically rich due to the higher number of low frequency words they can understand and produce. Their vocabulary knowledge subsequently informs us of their language needs e.g. having more of a focus on written medical journals, or a focus on spoken discourse in marketing.

Part 3

Using language creatively can be regarded as a sign of proficiency. Morpheme productivity refers to how an affix can create new word forms (Rastelli, 2020), in this paper it will also refer to the L2 learners' use and manipulation of morphemes to develop their proficiency. L2 learners gradually learn more complex morphemes to better manipulate morphosyntax to meet their communicative needs. Whilst European languages are marked with verb subject agreement and tense markers, morphology is crucial for meaning making in agglutinative languages. As the L2 learner acquires more morphemes and learns how to inflect more words they can communicate more complex ideas and broaden the contexts in which they can use words. By increasing their morpheme productivity and creativity, they therefore enrich their lexicon due to being able to deepen their vocabulary knowledge. For example, in Arabic the verb 'enjoy' is pronounced 'yastamta'. To transform it into the 3rd person masculine plural 'they enjoy', you must add the suffix 'eon' making the word 'yastamta-eon'. The near future tense is created by adding the prefix 'sa-'. This makes 'they will enjoy': 'sa-yastamta-eon'. This is but one small example of the variety of morphemes one agglutinative language contains, and vocabulary in Arabic is rich and varied facilitating great communicative opportunities for the motivated learner. Therefore, as the learner acquires more complex inflections increasing the morphosyntactic complexity of their utterances it suggests more ease and confidence of communicating in their target L2. This demonstrates how as the language user becomes more adept at harnessing the lexical richness of their mental lexicon in the form of morpheme productivity, they have access to a larger vocabulary knowledge and subsequently can communicate more ideas more fluently.

Whilst affixes allow us to elicit a huge amount of information and allow the language user to adapt their vocabulary knowledge to perform different grammatical functions, word association links to creativity in language production as learners vary their vocabulary usage to alter communication. Zareva (2007) found that L2 learners differ from L1 learners in syntagmatic relations but could engage with paradigmatic relations in a more native-like manner. This means L2 learners could draw on (high frequency) synonyms and antonyms without too much difficulty, however struggled with more thematic word associations. As having stronger syntagmatic relations is indicative of native-like speech, it could be argued that is a feature of lexical richness as it signals that the language user has a thorough understanding of the word and other related words (its word family). Therefore, we could infer that as the L2 learner gains proficiency they also gain awareness of syntagmatic relations thus enriching their lexis. This subsequently strengthens the links between words in their mental lexicon and enlarges their vocabulary knowledge again taking us back to Laufer and Nation's (1995) definition.

Conclusion

This paper has explored the link between lexical richness and L2 learners' vocabulary knowledge. We observed several variables to lexical richness in Read's dimensions approach and concluded that collocations are a key aspect of lexical richness to enlarge and cement vocabulary knowledge. The various figures proposed for vocabulary size to reach a lexical coverage of 98% highlighted the intrinsic link between the two components of lexical richness and vocabulary knowledge. Finally, we probed the significance of morpheme productivity for vocabulary knowledge and successful communication before considering the function of syntagmatic relationships to reinforce lexical richness in the L2 language user's learning journey. This points to the importance of taking a multifaceted approach to vocabulary acquisition as lexical richness, which heavily contributes to vocabulary knowledge, is comprised of many aspects. All the constituent parts (to lexical richness) build on one another, subsequently aiding the learner in increasing their proficiency.

References:

Hsueh-Chao, M. H., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. Reading in a Foreign Language, 13(1), 403-430.

Laufer, B., & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. Applied Linguistics, 16(3), 307–322.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? Canadian Modern Language Review, 63(1), 59-82.

Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory. In J. Richards & R. Schmidt (Eds.), Language and communication (pp. 191–226). Longman.

Saito, K. (2020). Multi- or Single-Word Units? The Role of Collocation Use in Comprehensible and Contextually Appropriate Second Language Speech. Language Learning, 70(2), 548-588.

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007). Language Teaching, 50(2), 212-226.

Staehr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. Language Learning Journal, 36(2), 139-152.

Rastelli, S. (2020). Contingency learning and perfective morpheme productivity in L2 Italian: A study on lexeme–morpheme associations with ΔP. Corpus Linguistics and Ling. Theory, 16(3), 459–486. https://doi.org/10.1515/cllt-2019-0071

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

Wray, A. (2002). Formulaic Language and the Lexicon. Cambridge: Cambridge University Press.

Zareva, A. (2007). Structure of the second language mental lexicon: How does it compare to native speakers' lexical organization? Second Language Research, 23(2), 123–153.

ALSLA essay 4

Nowadays, the quality of a second language learner's (L2) speech is associated with various characteristics, where vocabulary is argued to play a major role (Lewis, Nation, Schmitt etc). Indeed, one may not construct a sentence without any meaningful units of vocabulary that may be acquired in many possible ways. However, rather than being concerned with plethora of approaches and methods of teaching and acquiring vocabulary, this essay focuses on an equally important question: what measures are there of defining whether L2 learners' vocabulary is 'lexically rich', and how does it characterise their vocabulary and overall proficiency?

In search of appropriate estimative measures of lexical richness, many researchers came to the idea of lexical richness reflecting the quality of one's vocabulary in a language sample. It has been interpreted in different ways as some see it as variety of vocabulary units, whereas others apply a series of criteria to measure lexical richness. In 2000, Read, has introduced a model that suggests measuring lexical richness through lexical sophistication, lexical diversity, lexical density and the proportion of errors, having combined many previous works. While error rate may be deemed self-explanatory and is relatively easy to measure, only the first 3 dimensions are to be covered.

A common approach to defining lexical sophistication is that it equals the percentage of sophisticated or advanced words in a text (Laufer & Nation, 1995). As pointed out by Laufer and Nation, labelling words as advanced is highly dependent on the definition of the researcher while analysing their data. The most influential previous studies define advanced or difficult words as low-frequency words (Laufer & Nation, 1995; Read, 2000; Vermeer, Bogaards and Laufer 2004). This definition is based on the idea that the more frequent the word is, the easier it is for the learner to memorise it. The idea is supported by many studies and the actual teaching practice of educators in different contexts: it takes more cognitive effort to memorise and produce a vocabulary unit that a learner does not receive as their regular input. The concept of lexical sophistication correlates with the idea of language depth and breadth, representing both at the same time: being able to comprehend and produce an uncommon unit of vocabulary means that a learner has experienced both input and output opportunities that has helped them retain it.

Lexical diversity focuses on how unlike vocabulary units are, or the concept of "relative amount of repetitiveness" (Carroll, 1938). It may be used interchangeably with the terms lexical variability, lexical variation, and lexical variety (Engber, 1995). It is interesting to note that lexical diversity is sometimes used as a comprehensive term that is comprised of, but not limited to, lexical sophistication. This generalisation may seem confusing, although it can be argued that the concepts are closely intertwined. The key difference between them lies in what each aims at measuring, and in case of lexical diversity it is regarded as a separate notion due to its strong focus on the same token not being repeated in the language sample. It is believed that using totally diferent units that stem from different word families should aid in measuring one's proficiency, although it has been argued that the sample sixe may significantly impact the data when it comes to calculating the ratio of repeated or non-repeated units (Malvern, 2012).
From the point of its practicality, lexical diversity is an important component when estimating lexical richness. For researchers, it provides an idea of an appropriate correlation between the number of words acquired. For educators, it represents a standard to be aimed at: the more non-repeated words a learner is able to use in their sample, the higher their proficiency. The knowledge of vocabulary is quantified in this sense, making it accessible and relatively easy to

interpret and use in a classroom setting therefore giving a transparent idea of a learner's vocabulary knowledge.

Moving on to lexical density, it is a concept that suggests estimating the proportion of function words in relation to content words (Ure, 1971). The concept makes a causal claim: if content words prevail in a sample of a learner-produced text, whether written or spoken, their proficiency is higher than that of those with function words prevailing. Later, in 1985, Halliday proposed an alternative representation of the idea of lexical density: "lexical density is the number of lexical items as a ratio of the number of clauses". Some of the criticism aimed at measurement issues is broadly related to the problem of different approaches to define multi-word units as function or as content words (e.g. to take off may be considered a content word meaning to undress, or a combination of a content word and a function word). Nevertheless, the concept of lexical density says a lot about a learner's vocabulary knowledge: it emphasises the ability to use various vocabulary units in accordance with other parts of the sentence. Taking into account how limited the number of function words is, one may assume that prevalence of content words in a sample of text of learners with different levels of proficiency should differ to a significant extent.

Lexical density has been a controversial measure as it leaves a lot to a researcher's interpretation of one's sample, therefore, it may be worth attempting to see this measure applied to a receptive skill. Further, I am going to analyse an article that challenges lexical density in certain textbooks. In the paper "Lexical Density and Readability of Secondary Stage English Textbooks in Jordan" by Mohammad Ahmad Bani Amer (2021) the author aims at examining the lexical density and readability in English language textbooks in Jordan.

To begin with, the author presents a finding from a preparatory experiment, in which he randomly questioned students from schools about their opinion on the text from their textbooks. The author provides no data either about the schools that took part in the experiment or about the students. Thus, the preparatory stage leaves one questioning the scale and reliability of further research. As well as that, the author claims that the texts from the textbooks do not correspond to the level of students based on their answers, which cannot be considered a reliable source as there may be many unobserved factors leading to students' failure to comprehend the texts. The problem may be stated correctly, but it is poorly justified.
Further on, the author states that lexical density has a direct correlation with how well a text can be understood, which has been proved by many aforementioned researchers and therefore can be considered a valid claim. The objectives of the study, which are to determine the indices of lexical density in 2 textbooks "Action Pack11+12" for different years of study and to compare them accordingly, are relevant. It is stated multiple times that the textbooks are designated by the Ministry of Education, although there is no additional information on why these books are chosen for examination, nor is there any mention of funding.

The author provides a clear overview of the theory concerning lexical density and readability and chooses to use the Flesch's reading ease formula, which is a widely used formula for texts in English. When talking about his research design, the author claims to have chosen half the texts from each textbook, however, he does not mention on which basis it is done, which may significantly impact the results. Moreover, taking into consideration the second research question, which is to compare how the degree of lexical density and readability changes from Grade 12 to Grade 11, it can be argued that the change of the degree cannot be considered to change only between the two textbooks, but it should also be subject to development within each of the textbooks, which would be more representative of the actual language complication.

Finally, the author's interpretation of the results should not be considered accurate based on the impartiality concern mentioned before as the author argues that the data from the studied textbooks demonstrate a fluctuating pattern, but can nevertheless be deemed appropriately easy for students, which contradicts the problem investigated. The author also fails to mention what further research could be done.

Overall, when evaluating a learner's knowledge of vocabulary, the framework of lexical richness may be applied as it gives clear, easily quantifiable data that represents the knowledge in 4 different dimension. One should be aware of the replicability concern as there is no guarantee that different samples of texts should be able to provide the same results. To resolve it, one may want to conduct a longitudinal observation based on the aforementioned criteria, which will either enable to see the progress of a learner over time, of highlight potential issues in their acquisition process. The same framework may be applied to investigate whether certain texts are appropriate for various levels and what may be gained from presenting them to learners.

ALSLA essay 5

Vocabulary is a fundamental part of language learning. As an important index of vocabulary knowledge, lexical richness has gained some attention in L2 studies. This paper aims to analyse what lexical richness may and may not reveal about a learner's vocabulary knowledge when looking at lexical diversity, density, and sophistication. Since the past research majorly studied writing production in English, the analysis focuses on English learners' written vocabulary knowledge. The paper begins by defining lexical richness and vocabulary knowledge, followed by an analysis of literature concerning lexical diversity, density, and sophistication. An in-depth critique of Laufer and Nation's (1995) study is also included when scrutinising lexical sophistication.

Lexical richness and vocabulary knowledge are multidimensional constructs with no unanimity in terms of their definitions in L2 research. However, researchers mainly acknowledge that vocabulary knowledge includes productive and receptive knowledge, and both the depth of vocabulary knowledge (vocabulary size) and breadth of vocabulary knowledge (knowing aspects of words, and how well the words are known) need to be taken into account (Nation, 2022). In this paper, vocabulary knowledge covers all aspects of knowledge of and about words. Concerning lexical richness, the constructs within this umbrella term are more complex. Although the term was introduced by Yule (1944) as a parameter to characterise word usage distribution and word repetition rate, researchers later tended to use it to address a wider range of constructs. Some common ones are lexical diversity, lexical density and lexical sophistication (e.g. Laufer & Nation, 1995; Ure, 1971; McCarthy & Jarvis, 2007). Researchers have also developed various measures to evaluate these constructs, some of which are discussed in the following paragraphs. In addition, words within the lexical richness framework are operationally used as tokens, types, lemmas, and word families depending on different assumptions about learners' knowledge and research purposes (Kyle, 2020). Unless specified, all "words" in this paper are deemed as "tokens".

Studies regarding lexical richness have shed light upon L2 learners' vocabulary knowledge in different ways. A spotlight was put onto the written production of learners. The analysis starts by reviewing lexical diversity, which means the variety of individual words in a text (Kyle, 2020). Many empirical studies found that greater lexical diversity leads to better writing quality in an L2 (e.g., Engber, 1995; Kim, 2014; Qin & Wen, 2007). A particularly influential one by Engber (1995) analysed 66 essays composed by university English learners with different L1 backgrounds and varying proficiency levels. Engber's study highlighted a very strong correlation between L2 learners' lexical diversity and writing proficiency during time-restricted writing tasks. This could inform something about learners' receptive and productive knowledge: L2 learners who write more lexically diversely may possess higher lexical fluency (Webb, 2020) as they could access, process, and produce words more quickly than their counterparts within the same amount of time. In Engber's study, the evaluation of writing proficiency particularly accounts for the communicativeness of essays and whether word choices are appropriate in their contexts. That is, not only the breadth but also the depth of vocabulary knowledge was considered. It is therefore reasonable to infer that greater lexical diversity in writing signals a larger vocabulary size as well as better knowledge about the use of vocabulary. The reason is that ability to use vocabulary appropriately is built upon knowledge of form and meanings, and requires learners to understand collocations and constraints on the use of vocabulary (Nation, 2022). Few studies (e.g. Lu, 2012) investigating L2 oral production seem to echo the findings from the research in written production. However, these studies on written and oral production primarily collected data from adult learners and

targeted academic or formal texts. More studies investigating child learners' production and different genres are needed to further support the link between lexical diversity and lexical fluency as well as depth/breadth of vocabulary knowledge.

Another construct of lexical richness, namely, lexical density refers to the ratio of content and function words (Kyle, 2020). Unlike lexical diversity, studies generally found no significant relationship between lexical density and writing proficiency (Engber, 1995; Linnarud, 1986). This undermines the validity of studies using lexical density as a measure (e.g. Stegen, 2012) to monitor written proficiency. Limited research on oral production of L2 learners also suggests that lexical density could not effectively predict oral proficiency (Lu, 2012). However, lexical density is helpful when distinguishing between the modes and genres of language production (Halliday, 1985). Generally, written discourse has significantly more content words than spoken discourse due to differences in time limitations, efforts required, and people's belief that spoken discourse "lacks content" (McCarthy et al., 2020, p. 58). It is therefore deduced that lexical density could mirror the depth of vocabulary knowledge since an L2 learner needs sufficient knowledge concerning register and discourse aspects of words to produce an appropriate amount of content words in a text according to its discourses.

Lexical sophistication, the last construct discussed in this paper, refers to the number of difficult words (i.e. low-frequency words) used in a text (Kyle, 2020). The definition itself highlights that frequency is vital in quantifying lexical sophistication. An in-depth analysis of the study by Laufer and Nation (1995) is needed for its great impact in shaping the research on the lexical richness and its implications for different aspects of vocabulary knowledge. Laufer and Nation investigated the written production of 65 L2 learners with different English proficiency levels and tested the validity of their proposed measure, namely, Lexical Frequency Profile (LFP). According to Laufer and Nation (1995), LFP should reflect a learner's vocabulary size and act as a reliable and sensitive tool to measure productive knowledge. The LFP categorised words ("words" are word families in their research) used by learners based on their frequency, checked them against its corpus, and calculated the proportion of words of each category in a text. Words are categorised as follows: the first 1000 most frequent words, the second 1000 most frequent words, the academic words, and less frequent words beyond these categories. Laufer and Nation highlighted that it is better to calculate the first two categories and less frequent words for less frequent learners, whereas the latter three categories are for advanced learners. Nevertheless, their study did not provide clear criteria for "advanced learners", and the boundaries between different proficiency groups seemed questionable. The selected participants did not sit the same placement test. While some took a specifically designed test, others were allocated to a proficiency group only because they passed a departmental entrance exam at a different educational institute. How long they have passed such test remains unknown. Furthermore, the investigators asked all participants, including those lower-intermediate level learners to compose two texts, followed by a relatively long test to trigger their productive vocabulary knowledge. The lower-intermediate learners may find it struggling to complete all the tests. Consequently, their written proficiency may be affected by participant biases, especially by participant fatigue. Unfortunately, no further distinctions or sub-distinction were made for off-list words, meaning that some part of learners' depth of vocabulary knowledge may be neglected. These words with no corpus reference could be idiosyncratic words signalling lexical error, but they may also be words that show learners' encyclopaedic vocabulary knowledge—real-world knowledge developed for vocabulary over years (McCarthy, 1990). Studies on mental lexicon (e.g., Harrison, 2015) indicated that the longer the word is retained in one's mind, the richer the encyclopaedia for this word might be. A closer look at off-list words may offer more information regarding what aspects of words a

learner knows. Meara's (2005) computer simulation data further challenged the claim of LFP being a reliable tool for evaluating productive vocabulary knowledge. Meara found that LFP works effectively when learners' vocabulary size is disparate, but it fails to capture modest changes in vocabulary size and is not sensitive when learners' vocabulary sizes are similar. Despite the shortfalls, LFP enjoys its reputation of being easy to interpret and is one of the most widely used tools for measuring lexical sophistication. Subsequent studies using LFP or other measures generally found that lexical sophistication is also a powerful predictor of writing proficiency (Daller & Phelan, 2007; Ha, 2019) but not speaking proficiency (Lu, 2012). In fact, lexical sophistication is also strongly correlated with lexical diversity (Siskova, 2012). Altogether, if a learner uses more difficult words in writing, (s)he could probably use a wider range of vocabulary and possesses higher lexical fluency. (S)he may also know more about the use of vocabulary in addition to form and meaning. A note here is that the underlying assumption guides the perception of the word's difficulty in these studies: low-frequency words are normally learned later, and higher- frequency words are earlier. Nevertheless, learners may not always learn in such order. These studies tended to neglect learners' individual differences in vocabulary learning.

To conclude, lexical richness could reveal many aspects of vocabulary knowledge when looking at lexical diversity, density, and sophistication. Lexical diversity could mirror the breadth/depth of vocabulary and lexical frequency in writing, and probably speaking. Lexical density reflects learners' discourse and register knowledge. Lexical sophistication, especially when measured by LFP, could indicate vocabulary size in some conditions (i.e., when learners have disparate vocabulary sizes) and signal lexical fluency as well as knowledge concerning vocabulary use. Nevertheless, lexical richness is not convincingly robust when revealing factors such as learners' encyclopaedic aspects of knowledge or individual learning paths. Moreover, studies tended to engage with writing texts from adults. It is thus recommended that future research investigates young learners' lexical richness, and involves oral production as well as non-English L2 learners.

Daller, H., & Phelan, D. (2007). What is in a teacher mind? Teacher ratings of EFL essays and different aspects of lexical richness. In H. Daller, J. Milton & J. Treffers-Daller (Eds.), Modelling and assessing vocabulary knowledge (pp. 133-149). Cambridge University Press.
Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. Journal of Second Language Writing, 4, 138–155.
Ha, H. S. (2019). Lexical richness in EFL undergraduate students' academic writing. English Teaching, 74(3), 3-28.
Halliday, M. A . K. (1985). Spoken and written language. Deakin University Press.
Harrison, S. V. (2015). Word association: Exploring the L2 mental lexicon of Korean EFL learners. The Asian EFL Journal Quarterly, 461, 8-37.
Kim, J. (2014). Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study. English Teaching, 69(4), 27-51.
Kyle, K. (2020). Measuring lexical richness. In S. Webb (Ed.), The routledge handbook of vocabulary studies (pp. 454-476). Routledge.
Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: lexical richness in L2 written production. Applied Linguistics, 16(3), 307-332.
Linnarud, M. (1986). Lexis in composition: A performance analysis of Swedish learners' written English. CWK Gleerup.
Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. The Modern Language Journal, 96(2), 190-208.
McCarthy, P. M. (1990). Vocabulary. Oxford University Press.

McCarthy, P.M., Christian, M., & Slade, D. (2020). Discourse analysis. In N. Schmitt & M. P. H. Rodgers (Eds.), An introdution to applied linguistics(pp. 55-71). Routledge.
McCarthy, P. M. & Jarvis, S. (2010). Vocd: A theoretical and empirical evaluation. Language Testing, 24(4), 459-488.
Meara, P. (2005). Lexical frequency profiles: A Monte Carlo analysis. Applied linguistics, 26(1), 32-47.
Nation, I. S. P. (2022). Learning vocabulary in another language (3rd ed.). Cambridge University Press.
Qin X., & Wen, Q. (2007). Research on the development rules and characteristics of Chinese college students' written language writing ability. Social Science Press. Siskova, Z. (2012). Lexical richness in EFL students' narratives. Language Studies Working Papers, 4, 26-36.
Ure, J. (1971). Lexical density and register differentiation. Applications of Linguistics, 443-452.
Webb, S. (2020). The routledge handbook of vocabulary studies. Taylor & Francis Group.
Yule, G. (1944). The statistical study of literary vocabulary. Cambridge University Press.

ALSLA essay 6

1. Vocabulary knowledge and lexical richness
In this section, the definitions of key concepts related to "vocabulary knowledge" and "lexical richness" will be reviewed, referring to previous literatures.

1.1 Vocabulary Knowledge
Before moving to vocabulary knowledge, it is necessary to discuss the definition of vocabulary in the context of Second Language Acquisition (SLA). Several academic concepts are under the term "vocabulary", including "tokens" and "types" (concerned with word count), "functional word" and "content word" (based on grammatical category), "lemma" and "word family" which have practical indications for measuring vocabulary size (Read, 2000). A wealth of literatures in SLA have interpreted vocabulary knowledge from multiple perspectives (Meara, 1997; Laufer & Paribakht, 1998; Waring, 1998; Nation, 1990; Nation, 2001). Vocabulary knowledge is classified into receptive vocabulary knowledge (entails the ability for reading and listening) and productive vocabulary knowledge (for writing and speaking), and it is also suggested that the development from receptive to productive vocabulary knowledge mastery is more of a continuum process (Melka, 1997). Nation (2001) observed that vocabulary knowledge involves knowing the "form", "meaning" and "use" of the word (Nation, 2001). Each of the categories are divided into three dimensions, i.e., spoken, written and word part under "form"; form and meaning, concept and referents, associations under "meaning"; grammatical functions, collocations and constraints on use under "use". with all nine dimensions further divided into receptive and productive aspects.
The breadth (or size) and depth of L2 learner's vocabulary knowledge is distinguished (Anderson & Freebody, 1981; Schmitt, 2014); the former refers to the number of words for which L2 learners know at least some of the significant aspects of meaning, while the latter refers to the word which the L2 learners master all of the distinctions an ordinary adult would understand under normal circumstances (Anderson & Freebody, 1981). Henriksen (1999) proposed three dimensions of vocabulary knowledge, including the partial or precise knowledge, the depth of knowledge (the quality of L2 learners' vocabulary knowledge) and the receptive- productive use ability. Read (2000) developed the concept of vocabulary knowledge depth and divided it into precision of meaning, comprehensive word knowledge and network knowledge. Fluency, or how quickly L2 learners can retrieve the form or meaning of a given word from memory is later added as another dimension of vocabulary knowledge. (Daller et al., 2007; Šišková, 2012)

1.2 Lexical Richness
Previous studies have put forward different dimensions for measuring lexical richness. Linnarud (1986), Laufer and Nation (1995) summarized four measurements of productive lexical richness, i.e., Lexical Originality, Lexical Density, Lexical Sophistication and Lexical Variation. Based on that, researchers (Engber, 1995; continued to develop new concepts to complete the measuring framework, including lexical individuality, percentage of lexical errors, etc. Table 1 summarized the major measurements proposed by previous studies on L2 learner's lexical richness.

Table 1: Measures of Productive Lexical Richness
Researchers Measurements Formula Implication for vocabulary knowledge
Linnarud (1986);
Laufer and Nation (1995) Lexical Originality (LO)/ Lexical Individuality (LI) Number of tokens unique to one writer/Total number of tokens L2 Learner's performance compared to

other writers in the same group
Lexical Density (LD) Number of content tokens/Total number of tokens The percentage of words conveying information
Lexical Sophistication (LS) Number of advanced tokens/Total number of lexical tokens How advanced the L2 learners' vocabulary knowledge is
Lexical Variation (LV) Number of types/Number of tokens A productive strategy for L2 learners to express content (Engber, 1995)
Engber (1995) Percentage of Lexical Errors Sum of lexical errors per segment/ sum of lexical items per segment How accurate the learners' vocabulary knowledge mastery is
Source: Linnarud (1986); Laufer and Nation (1995); Engber (1995).
It should be noted that there are some controversies around certain measurements. For example, LD can get lower as a result of more occurrence of subordinate clauses and LS can only be effective with uniform standard being used.

2. Further discussion on the relations between lexical richness and L2 learner's vocabulary knowledge

Concerning the relationship of lexical richness and receptive/productive vocabulary knowledge, lexical richness is reflective of the productive vocabulary knowledge and more generally, the language proficiency of L2 learners, including both oral and written production. In terms of its relations with the breadth, depth and receptive-productive use ability of vocabulary knowledge, lexical richness seems to have more implications for breath of productive vocabulary knowledge. Moreover, lexical richness has direct relationship with the "use" of vocabulary knowledge, though "form" and "meaning" definitely serve as the basis of achieving greater lexical richness.
To some extent, all aspects of lexical richness measures have certain degrees of implications for L2 learners' vocabulary knowledge. Particularly, lexical sophistication and lexical variation have strong positive relationship with vocabulary knowledge, which has also been backed up by multiple previous empirical studies (Crossley et al., 2011; Vögelin et al., 2019). To illustrate, L2 learners' who have the ability to use more lower-frequency and advanced-level vocabulary are often seen as having better vocabulary knowledge, in other words, the higher Lexical Sophistication is a powerful indicator of vocabulary knowledge, both in terms of breadth and depth (Higginbotham & Reid, 2019).
Lexical richness can be used to quantify the diversity and breadth of L2 learners' vocabulary knowledge, and the measurements of lexical knowledge can be applied to study the relations between L2 learners' vocabulary knowledge and use (Real et al., 2020). Some scholars hold the view that vocabulary knowledge can be operationalized as lexical richness (Read, 2000; Lei & Yang, 2020).

3. In-depth critique on Lei & Yang's (2020) empirical study of L2 learners' lexical richness

Apart from theoretical exploration of lexical richness, there are also empirical studies that developed tools for measuring lexical richness (Halliday, 1985; Jarvis, 2002; Nation & Coxhead, 2015), examined how lexical richness impacted L2 learner's written (Dieter, 2014; Azodi et al., 2014; Real et al., 2020) and oral production (Lindqvist et al., 2013).
This section will present a critique on a corpus-based study conducted by Lei and Yang (2020), which compared the lexical richness of research articles written by Chinese PhD students (CPhD) with two other groups, i.e., the native beginner students' corpus (NBS) comprising unpublished research papers by native final-year undergraduate students and master students, and the native expert corpus (NE) comprising published research articles.
Lei & Yang's study aimed to address the two research questions, (1) to identify the different and similar patterns in terms of lexical richness among the three corpora; (2) to discuss how

“nativeness” and “academic expertise” impact the lexical richness of research articles. Based on such questions, the choice of the two corpora as references is adequately justified yet not without any limitations. On the one hand, the researchers managed to grasp the fundamental differences or decisive factor which separate the three groups. The discipline-specific feature of academic discourse was also given full consideration, and all three corpora used the same disciplines of “engineering, physics and biology”. On the other hand, the role of “academic expertise” in lexical richness of research articles was measured by comparing the NBS corpus and NE corpus. However, “academic expertise” might not be the only factor which led to variations in lexical richness the two groups; for example, since the native expert corpus only consisted of published research articles, peer reviews, teamwork, professional editing might be involved in the process, which could further increase lexical richness of the research articles in this corpus. To attribute the differences only to academic expertise might be over-simplification.

In terms of methodology, the definitions of terms (lexical diversity/lexical density/ lexical sophistication) were clear and consistent, and the choices of measuring tools (MATTR/ Lexical Density Calculator/AntwordProfiler) were appropriate as well. However, the researchers did not exhaust all the available dimensions of lexical richness discussed by previous literatures, such as lexical originality and percentage of lexical errors, which could also be effective measurements of lexical richness and L2 students’ vocabulary knowledge according to previous empirical studies (Engber, 1995). The sizes of three corpora were not balanced, with the CPhD corpus (565,722 word tokens) and NE corpus (672,094 word tokens) being much larger than the NBS corpus (188,493 word tokens). Though the researchers stated that they did so to exhaust all the available resources and the three measuring tools of lexical richness used are not relevant to the word tokens, it could be more ideal to balance the size of three corpora to avoid any potential distraction.
The researchers exercised caution in presenting the conclusion that in terms of the quality of RAs, “nativeness” had a limited role to play compared with “academic expertise”. They also related to this finding to other studies with similar conclusions. Since the sample was comparatively small (142 research articles for Chinese PhD corpus, and 71 and 128 for native beginner corpus and expert corpus respectively) and the non-native corpus only consisted of PhD students from China, therefore the generalizability of the conclusion to other countries where English is not the native language remains to be uncertain.

References

Anderson, R. C., & Freebody, P. (1981). Vocabulary knowledge. Comprehension and teaching: Research reviews, 77-117.
Azodi, N., Karimi, F., & Vaezi, R. (2014). Measuring the Lexical Richness of Productive Vocabulary in Iranian EFL University Students' Writing Performance. Theory & Practice in Language Studies, 4(9).
Batia Laufer, Paul Nation, Vocabulary Size and Use: Lexical Richness in L2 Written Production, Applied Linguistics, Volume 16, Issue 3, September 1995, Pages 307–322, https://doi.org/10.1093/applin/16.3.307
Coxhead, A., Nation, P., & Sim, D. (2015). Measuring the vocabulary size of native speakers of English in New Zealand secondary schools. New Zealand Journal of Educational Studies, 50(1), 121-135.
Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). What is lexical proficiency? Some answers from computational models of speech data. Tesol Quarterly, 45(1), 182-193.
Daller, H., Milton, J., & Treffers-Daller, J. (Eds.). (2007). Modelling and assessing vocabulary knowledge. Cambridge: Cambridge University Press.

Durán, Pilar, Malvern, David, Richards, Brian, & Chipere, Ngoni. (2004). Developmental Trends in Lexical Diversity. Applied Linguistics, 25(2), 220-242.
Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. Journal of second language writing, 4(2), 139-155.
Halliday, M. A. K. (1989). Spoken and written language. Oxford University Press, USA.
Henriksen, B. (1999). Three dimensions of vocabulary development. Studies in second language acquisition, 21(2), 303-317.
Higginbotham, G., & Reid, J. (2019). The lexical sophistication of second language learners' academic essays. Journal of English for Academic Purposes, 37, 127- 140.
Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. Language Testing, 19(1), 57-84.
Lei, S., & Yang, R. (2020). Lexical richness in research articles: Corpus-based comparative study among advanced Chinese learners of English, English native beginner students and experts. Journal of English for Academic Purposes, 47, 100894.
Lindqvist, C., Gudmundson, A., & Bardel, C. (2013). A new approach to measuring lexical sophistication in L2 oral production.
Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge, UK: Cambridge University Press.
Nation, I. S., & Webb, S. A. (2011). Researching and analyzing vocabulary. Boston, MA: Heinle, Cengage Learning.
Read, J. (2000). The nature of vocabulary. In Assessing Vocabulary (Cambridge Language Assessment, pp. 16-37). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511732942.003
Read, J. (2004). Vocabulary knowledge be defined. Vocabulary in a second language: Selection, acquisition, and testing, 10, 209.
Real, D. V. C., Cuizon, P. C., Necesito, R. D., Tuppal, C. P., Oducado, R. M. F., Lee, M. D., & Javenes, K. J. C. Lexical Richness of L2 Production Using Nation and Laufer's Lexical Frequency Profile. SDCA Asia-Pacific Multidisciplinary Research Journal Volume 2, December 2020, 32.
Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. Language learning, 64(4), 913-951
Šišková, Z. (2012). Lexical richness in EFL students' narratives. Language Studies Working Papers, 4, 26-36.
Vögelin, C., Jansen, T., Keller, S. D., Machts, N., & Möller, J. (2019). The influence of lexical features on teacher judgements of ESL argumentative essays. Assessing Writing, 39, 50-63.
Waldvogel, D. A. (2014). An analysis of Spanish L2 lexical richness. Academic Exchange Quarterly, 18(2), 8.

ALSLA essay 7

In vocabulary studies, the term lexical richness encompasses the ideas of lexical diversity, lexical density and lexical sophistication (Read, 2000, as cited in Chen, 2020; and Dawson et al., 2021), generally as exhibited in the linguistic production of an L2 learner. Lexical diversity is defined as the range and variety of the words produced by an L2 learner, while lexical density is the ratio of lexical words relative to the total amount of the words in the production (Chen, 2020; Dawson et al., 2021). Lexical sophistication, the last component of lexical richness, is measured by looking at the level of frequency of the words used by the L2 learner (Dawson et al, 2021). All of the elements of lexical richness paint a picture of an L2 learner's knowledge of the words that are in their vocabulary. This paper takes on the perspective of usage-based theories which state that language is learned through exposure in context (Ellis, 2008, Tomasello, 2003). It aims to explore how an L2 learner's lexical richness can shed light on the size and depth of their vocabulary and how it relates to the importance of input and exposure in L2 learning. An analysis of the Lex-30 vocabulary test (Meara & Fitzpatrick, 2000) will be given in the discussion about vocabulary size and examples of an L2 learner's linguistic production are taken from Lightbown and Spada (2013) to illustrate the points made in this paper.

Lexical richness and vocabulary size
The lexical richness exhibited in an L2 learner's linguistic production can indicate the size of their vocabulary. Through an L2 learner's lexical diversity, it is possible to measure the range of words in their vocabulary and thus give an estimate of the number of words that they know, as well as the internal makeup of their vocabulary. The lexical density in a text or a speech output gives us an idea of the proportion of lexical and function words in their vocabulary, while lexical sophistication allows us to know the frequency level of the words. Several tests have been devised with the purpose of measuring an L2 learner's vocabulary size, such as LexTale (Lemhöfer & Broersma, 2012), Lex-30 (Meara & Fitzpatrick, 2000) and the Vocabulary Size Test (Nation & Berglar, 2007, in Beglar, 2010). The Lex- 30 test, designed by Meara and Fitzpatrick (2000), seeks to measure an L2 learner's productive vocabulary size by asking the participant to give four words that are related to a word prompt. The participants are free to write down any word without a binding context for each prompt, with the intention of improving previous studies that used cloze tests to measure vocabulary. The design of the test is expected to elicit a wider variety of words from the participant to better represent their vocabulary, in which the test succeeds.

However, the words used as prompts in the test all fall under the category of lexical word, which skews the results into generating more lexical words than structure words. As a result, the Lex-30 test does not account much for the size of vocabulary that consists of structure words. The absence of context also removes another dimension of vocabulary from the test results: it is impossible to know if the participant knows how to use the word in context or if they really know the meaning of the word. With regards to the concepts of lexical richness and vocabulary size, the Lex-30 test provides information on an L2 learner's lexical diversity by assessing the range and variety of their vocabulary, as well as their lexical sophistication by seeing the frequency of the words elicited during the test. The test results then give an estimate of the learner's vocabulary size, despite not accounting for context and structure words.

Lexical richness and vocabulary depth
The lexical richness exhibited in an L2 learner's linguistic production can indicate the depth of their vocabulary. According to Nation (2022), knowing a word comprises knowing its form,

meaning and use. Norbert Schmitt elaborates on the concept of vocabulary depth by separating word knowledge into eight categories (spoken form, written form, meaning, grammar, collocation, register, frequency and association), according to which word knowledge lies on a continuum of no knowledge to full mastery (Schmitt, 2014). Instead of a dichotomous state of knowing and not knowing, word knowledge and vocabulary depth are described as a developmental process that spans across many dimensions, hence the term 'dimensions approach' to refer to this perspective (Schmitt, 2014). Through the lexical richness in a linguistic production, it is possible to see where an L2 learner's knowledge of a word stands in the spectrum of each dimension. The lexical density, diversity and sophistication in a linguistic output can be measured and then evaluated to generate an estimate of the learner's vocabulary depth.

In the example given in Lightbown and Spada (2013) of a text produced by an English learner with French as their L1, a spelling mistake of 'glasses' can be found in the form of the word '*glaces'. Not only does the mistake inform us of the level of lexical sophistication in the text by showing their use of a high frequency word, it also shows us the nuance in the learner's word knowledge.The learner is familiar with the concept of glasses and is able to map the concept to a word, but is less at ease with its form, as shown by the spelling mistake in the word form. Despite that, the learner clearly possesses the knowledge of how to use the word and the lexical phrases that can collocate with it, as can be seen from their attempt of building the phrase '*a king of glaces' (a kind of glasses) in the text. Their use of the word in the circumlocution around the word 'binoculars' also shows that they are able to associate the word 'glasses' to 'binoculars' due to their similar qualities.

The role of input in the process of building word knowledge

In the perspective of usage-based theories, input is the starting point of language learning and linguistic knowledge is built upon exposure to input in particular contexts, which is also the case for lexical development in second language learners (Tomasello, 2003). Exposure to input has been found to affect L2 learners' morphosyntactic development (Oller & Eilers, 2002) and their acquisition of lower frequency collocations (Durrant & Schmitt, 2009). Kawecki (2022) also found that in a corpus of texts produced by learners of L2 French with Indonesian as the L1, there is a strong reliance on the input received during the learning process, as words in the input are reproduced in the output text. The existence of a word in an L2 learner's vocabulary essentially indicates that the learner has encountered this word through some form of input, and an L2 learner's lexical richness can show the influence of the input that they are exposed to in the shaping of their word knowledge.

In the example given in Lightbown and Spada (2013), lexical richness gives an idea of the L2 learner's input through their use of collocations in their output text. The English L2 learner with Chinese as their L1 uses formulaic phrases such as 'one-horse open sleigh' and 'dashing through the town' that are recognizably taken from a popular Christmas carol ("Jingle Bells"). Another lexical phrase that appears in the same text is "This year Christmas comes soon!", which appears to be a variation of the popular formulaic expression "This year Christmas comes early!". This allows us some information into the L2 learner's input: s/he is familiar and has been exposed to high frequency input that are in popular use.

Conclusion

Lexical richness in an L2 learner's linguistic output can shed light on their word knowledge in terms of providing information about their vocabulary size and depth. By measuring lexical density, diversity and sophistication in a linguistic output produced by the student, the

frequency of the appearing words and how they are used in context can give us an estimate of the amount of words that the learner knows and how well s/he knows it. Analyzing an L2 learner's lexical richness also reveals the kind of input to which the learner has been exposed. Second language educators may take into consideration the lexical richness in the students' linguistic output to decide on the kind of input that is provided in class and how to modify it when necessary.

References

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. Language testing, 27(1), 101-118.

Chen, Yi-chen. (2020). Assessing the lexical richness of figurative expressions in Taiwanese EFL learners' writing. Assessing Writing, 43, 7-18.

Dawson, Nicola, Hsiao, Yaling, Banerji, Nilanjana, Tan, Alvin Wei Ming, & Nation, Katie. (2021). Features of lexical richness in children's books: Comparisons with child-directed speech. Language Development Research, Language Development Research, 2021.

Durrant, Philip, & Schmitt, Norbert. (2009). To what extent do native and non-native writers make use of collocations? International Review of Applied Linguistics in Language Teaching, IRAL, 47(2), 157-177.

Ellis, R. (2008). The study of second language acquisition (2nd ed.). Oxford, England: Oxford University Press.

Kawecki, Regis. (2022). In what ways are the Semarang University Indonesian learners of French developing an Interlanguage? European Journal of Applied Linguistics, 10(2), 227-235.

Lemhöfer, K.M, & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. Behavior Research Methods, 44(2), 325-343.

Lightbown, P. & Spada, N. (2013). How Languages are Learned. Oxford: OUP

Meara, Paul, & Fitzpatrick, Tess. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. System (Linköping), 28(1), 19-30.

McNeil, Levi. (2012). Using talk to scaffold referential questions for English language learners. Teaching and Teacher Education, 28(3), 396-404.

Nation, I. (2022). Learning vocabulary in another language (Third ed., Cambridge applied linguistics series). Cambridge.

Oller, D. K., & Eilers, R. E. (Eds.). (2002). Language and literacy in bilingual children(Vol. 2). Multilingual Matters.

Tomasello, M. (2003). Constructing a language: A usage-based theory of language acquisition. Cambridge, England: Harvard University Press.

Appendix B; Corpus of AI-generated essays

AI essay 1

What does 'lexical richness' tell us about an L2 learner's vocabulary knowledge?

Introduction

Lexical richness is a term used to describe the diversity and complexity of an individual's vocabulary. In the context of second language (L2) learning, understanding lexical richness can provide insight into an L2 learner's vocabulary knowledge, which is crucial in assessing their language proficiency and designing effective instructional strategies. This essay will provide an overview of what lexical richness is, how it is typically measured, and its relationship with vocabulary knowledge. We will also explore the factors that affect lexical richness in L2 learners, methods for assessing it, and the implications for L2 educators.

Definition of lexical richness

Lexical richness is a measure of the diversity and complexity of a learner's vocabulary, including the range of words used, their frequency, and their depth of knowledge. It is typically assessed through word lists, text analysis, or computational tools. High lexical richness indicates a greater variety of words used and a more sophisticated use of language.

The relationship between lexical richness and vocabulary knowledge

There is a strong correlation between vocabulary size and lexical richness. Learners with larger vocabularies tend to have more varied and sophisticated language use. However, word frequency also affects lexical richness. High-frequency words are often used more frequently but may not contribute as much to overall lexical diversity. Knowledge depth also impacts lexical richness, as learners with deeper conceptual understanding of words can use them in more nuanced ways.

Factors that affect lexical richness in L2 learners

Language proficiency level plays a significant role in determining lexical richness. Higher proficiency levels are associated with greater vocabulary diversity and sophistication. Language exposure also affects the development of lexical richness. Input quantity and quality are crucial for learners to acquire new words and use them appropriately. Motivation and interest in learning also play a role, as learners who are more engaged are likely to seek out new words and use them in meaningful ways.

Assessment of lexical richness in L2 learners

Assessing lexical richness can be done through different methods, such as using vocabulary tests, analyzing written or spoken texts, or employing computational tools. However, there are challenges and limitations to assessing lexical richness. It can be difficult to capture all aspects of lexical diversity, and different assessment methods may introduce potential biases.

Implications of lexical richness for L2 learners and educators

Measuring lexical richness can provide valuable information for L2 learners, such as identifying areas for improvement and tracking progress over time. It is important for L2 educators to incorporate vocabulary instruction in their classrooms, including strategies for promoting vocabulary acquisition and retention. Ongoing assessment and monitoring of vocabulary knowledge in L2 learners are crucial, as lexical richness is a dynamic construct that can change over time.

Conclusion

Lexical richness is a significant factor in understanding an L2 learner's vocabulary knowledge. It provides insight into the range and depth of a learner's vocabulary use, which is important in assessing their language proficiency and designing effective instructional strategies. Educators should consider lexical richness in evaluating L2 learners' language proficiency and designing appropriate instructional interventions. Future research should explore the relationship between lexical richness and other language skills and investigate effective instructional practices for promoting lexical diversity.

AI essay 2

The Significance of Lexical Richness in Assessing L2 Learner's Vocabulary Knowledge
Introduction
In the field of Second Language Acquisition (SLA), vocabulary knowledge is a crucial aspect of language proficiency. It is widely accepted that learners must acquire a certain number of words to be able to communicate effectively in a second language. However, simply knowing a large number of words is not enough; learners must also be able to use them accurately and appropriately in context. This is where the concept of lexical richness comes in. In this essay, we will define lexical richness, discuss its importance in assessing vocabulary knowledge in L2 learning, and answer the research question: What does 'lexical richness' tell us about an L2 learner's vocabulary knowledge?

Definition of Lexical Richness
Lexical richness refers to the diversity and complexity of the vocabulary used by a speaker or writer. It is often measured by examining the range and variety of words used, as well as the density and sophistication of the language. This can include factors such as the use of rare or unusual words, the ability to use synonyms and antonyms effectively, and the ability to use figurative language and idiomatic expressions.

Importance of Assessing Vocabulary Knowledge in L2 Learning
Assessing vocabulary knowledge is important in L2 learning for several reasons. First, vocabulary knowledge is closely linked to overall language proficiency. Learners who have a wide range of words at their disposal are better able to express themselves and understand others in a variety of contexts. Second, vocabulary knowledge is strongly related to reading comprehension. Without sufficient vocabulary knowledge, learners may struggle to understand written texts, which can negatively impact their academic success. Finally, vocabulary knowledge is a good predictor of long-term language learning success. Learners who have a strong foundation in vocabulary are more likely to continue learning and improving their language skills over time.

Research Question: What does 'lexical richness' tell us about an L2 learner's vocabulary knowledge?
The concept of lexical richness has become increasingly important in recent years as a way to assess vocabulary knowledge in L2 learning. By examining the diversity and complexity of a learner's vocabulary, we can gain insights into their overall language proficiency and potential for future success. In the following sections, we will review the literature on lexical richness and L2 learning, discuss our methodology for studying this concept, present our results, and discuss the implications for university examiners and language teachers.

Literature Review
Overview of Second Language Acquisition
Before we discuss lexical richness specifically, it is important to provide some context on second language acquisition. SLA is a complex and multifaceted process that involves many different factors such as motivation, age, input, and individual differences. One of the key aspects of SLA is vocabulary acquisition, which involves not only memorization of individual words but also the ability to use them in context.

Definition and Measurement of Vocabulary Knowledge

Vocabulary knowledge can be defined in several ways, including passive knowledge (the ability to recognize words) and active knowledge (the ability to produce words). It can also be measured in different ways, such as through tests of receptive vocabulary (such as multiple-choice tests) or productive vocabulary (such as writing or speaking tasks). However, these measures do not necessarily capture the full range of a learner's vocabulary knowledge.

Lexical Richness as a Measure of Vocabulary Knowledge

Lexical richness has emerged as an alternative measure of vocabulary knowledge that takes into account the complexity and diversity of a learner's vocabulary. This concept is closely related to collocation and phraseology, which refer to the way words are combined to create meaning. By examining the patterns of word use in a learner's speech or writing, we can gain insights into their overall language proficiency and potential for future success.

Previous Studies on Lexical Richness and L2 Learning

Several studies have examined the relationship between lexical richness and L2 learning. For example, Laufer and Nation (1995) found that lexical richness was a better predictor of reading comprehension than measures of vocabulary size alone. Similarly, McCarthy and Jarvis (2010) found that lexical richness was positively correlated with overall language proficiency in L2 learners. Other studies have examined the role of lexical richness in specific contexts, such as academic writing (Hyland and Tse, 2007) or spoken discourse (Simpson-Vlach and Ellis, 2010). Overall, these studies suggest that lexical richness is an important aspect of vocabulary knowledge that can provide valuable insights into a learner's overall language proficiency.

Methodology

Participants

To investigate the role of lexical richness in assessing vocabulary knowledge in L2 learners, we recruited a sample of 50 university students who were enrolled in an intermediate-level English course. These students came from a variety of linguistic backgrounds and had varying levels of English proficiency.

Data Collection

To measure lexical richness, we collected samples of each participant's writing using a standardized prompt. We then analyzed these samples using a computerized tool that calculated several measures of lexical richness, including type-token ratio (the number of unique words divided by the total number of words), frequency of rare words, and use of idiomatic expressions.

Data Analysis

We compared the results of our lexical richness analysis to other measures of vocabulary knowledge, such as scores on a standardized vocabulary test and self-reported vocabulary knowledge. We also conducted statistical analyses to examine the relationships between these different measures.

Results

Description of Participants' Lexical Richness Scores

Our analysis revealed significant variation in participants' lexical richness scores. Some participants showed high levels of lexical richness, while others had more limited vocabularies. Interestingly, there was not always a strong correlation between scores on the various measures of vocabulary knowledge. For example, some participants who scored highly on the vocabulary test had relatively low scores on measures of lexical richness.

Comparison of Lexical Richness Scores to Other Measures of Vocabulary Knowledge

We found that lexical richness was positively correlated with scores on the vocabulary test and self-reported vocabulary knowledge. However, it was also clear that lexical richness provided unique information about participants' vocabulary knowledge that was not captured by these other measures. For example, some participants who reported high levels of vocabulary knowledge still struggled with using words in context or understanding idiomatic expressions.

Implications of Results for L2 Learning and Assessment

Our results suggest that lexical richness is an important aspect of vocabulary knowledge that should be considered in L2 learning and assessment. While tests of vocabulary size can provide useful information, they may not capture the full range of a learner's vocabulary knowledge. Assessments that focus on lexical richness can provide more nuanced insights into a learner's overall language proficiency and potential for future success.

Discussion

Significance of Lexical Richness in Assessing Vocabulary Knowledge

The concept of lexical richness has become increasingly important in recent years as a way to assess vocabulary knowledge in L2 learning. By examining the diversity and complexity of a learner's vocabulary, we can gain insights into their overall language proficiency and potential for future success. Our study provides further evidence for the importance of this concept by demonstrating its relationship to other measures of vocabulary knowledge and its unique contributions to our understanding of learners' abilities.

Limitations and Future Directions of Research

Our study has several limitations that should be addressed in future research. First, our sample size was relatively small, and we did not have access to detailed information about participants' linguistic backgrounds or previous language learning experiences. Second, our analysis focused primarily on written language; future studies should also examine lexical richness in spoken language. Finally, our study did not explore the impact of lexical richness on specific language skills, such as reading comprehension or writing ability.

Implications for University Examiners and Language Teachers

The findings of our study have important implications for university examiners and language teachers. These professionals should be aware of the concept of lexical richness and its importance in assessing vocabulary knowledge in L2 learners. They should also consider incorporating assessments of lexical richness into their evaluation practices to gain a more complete understanding of learners' abilities. Finally, they should provide opportunities for learners to develop their lexical richness skills through activities such as reading, writing, and speaking practice.

Conclusion

Our study has demonstrated the importance of lexical richness in assessing vocabulary knowledge in L2 learners. By examining the diversity and complexity of a learner's vocabulary, we can gain insights into their overall language proficiency and potential for future success. This concept should be considered by university examiners and language teachers in their evaluation practices, and learners should be given opportunities to develop their lexical richness skills through a range of language learning activities. Ultimately, a strong foundation in vocabulary knowledge is essential for long-term language learning success.

AI essay 3

The Significance of Lexical Richness in L2 Learning

I. Introduction
Language proficiency is a key indicator of a learner's ability to communicate effectively in a second language. One important aspect of language proficiency is lexical richness, which refers to the variety and complexity of a learner's vocabulary. In this essay, we will explore the significance of lexical richness in L2 learning and what it tells us about a learner's vocabulary knowledge.

II. Understanding Lexical Richness
Lexical richness is typically measured by calculating the ratio of types (different words) to tokens (total occurrences of words) in a given text. This ratio gives an indication of the variety of words used, as well as the complexity of the learner's vocabulary. Different approaches to measuring lexical richness include qualitative measures, which involve examining the text for the presence of certain words and phrases, and quantitative measures, which involve counting the number of different words used in a text.

III. Implications of Lexical Richness in L2 Learning
A high level of lexical richness implies that the learner has a wide range of vocabulary knowledge, while a low level indicates that they lack variety in their word choice. Lexical richness has been shown to be closely linked with language proficiency, as higher levels of lexical richness indicate higher levels of language proficiency. Therefore, measuring lexical richness can provide valuable insight into a learner's level of language proficiency. Lexical richness plays an important role in both language production and comprehension, as it provides learners with the necessary tools to express their ideas clearly and accurately, as well as to understand what is being communicated to them.

IV. Factors affecting Lexical Richness in L2 Learning
Different learners may have different levels of lexical richness due to individual differences such as age, native language, and learning strategies. Learning contexts have a huge impact on lexical richness, as different contexts provide different kinds of input and output which can affect a learner's vocabulary knowledge.

V. Strategies for Enhancing Lexical Richness in L2 Learning
Effective vocabulary learning techniques such as spaced repetition and mnemonics can help learners acquire new vocabulary more quickly and effectively, thus increasing their lexical richness. Technology can be used to enhance lexical richness by providing learners with access to digital resources such as online dictionaries and flashcards for easy access to new vocabulary. Incorporating lexical richness into language assessment and instruction can help university examiners better evaluate learners' language proficiency, as well as guide teachers in designing effective lesson plans that focus on developing learners' vocabulary knowledge.

VI. Conclusion
In conclusion, lexical richness is an important indicator of a learner's language proficiency, as it provides insight into the depth and breadth of their vocabulary knowledge. Understanding the implications of lexical richness can help university examiners better assess learners' language proficiency, as well as guide teachers in designing effective lesson plans that focus on developing learners' vocabulary knowledge. Further research should be conducted on the

factors that affect lexical richness in L2 learning, as well as strategies for enhancing it. By prioritizing lexical richness in language instruction, we can help learners achieve greater success in their second language acquisition journey.

AI essay 4

What Does 'Lexical Richness' Tell Us About an L2 Learner's Vocabulary Knowledge?

Introduction
Lexical richness is the variety of words and word combinations used in a language. It is an important factor in evaluating L2 learners' vocabulary knowledge, as it helps to identify areas where they may need additional support. In this essay, we will explore the relationship between lexical richness and L2 vocabulary knowledge, and how it can be used to evaluate the effectiveness of L2 teaching methods.

Lexical Richness
Different ways can measure lexical richness, such as the number of different words used, frequency of word use, etc. Different types of lexical richness measures include the Type-Token Ratio (TTR), Measure of Textual Lexical Diversity (MTLD), and D (V) measure. These measures are used to evaluate vocabulary knowledge in L2 learning, as they indicate the level of vocabulary knowledge an L2 learner has achieved.

L2 Learning
L2 learning is the process of learning a language that is not one's native tongue. Factors that affect L2 learning include age of acquisition, motivation, and exposure to the language. Vocabulary knowledge is essential in L2 learning because it is essential to communication. Lack of vocabulary knowledge can lead to misunderstandings and miscommunications.

Vocabulary Knowledge
Vocabulary knowledge refers to the words and phrases that a person knows and understands in a language. There are two types of vocabulary knowledge: receptive vocabulary (words a person understands when they hear or read them) and productive vocabulary (words a person can use in speaking or writing). Vocabulary knowledge is essential for effective communication and can help L2 learners adapt to different social situations.

Second Language Acquisition
Second language acquisition is the process by which a person learns a second language. The stages of second language acquisition include preproduction, early production, speech emergence, intermediate fluency, and advanced fluency. Vocabulary knowledge is essential to communication at all stages of L2 learning.

Conclusion
Lexical richness is an important factor in evaluating L2 learners' vocabulary knowledge. Vocabulary knowledge is essential to effective communication in L2 learning. The implications of lexical richness in evaluating L2 learner's vocabulary knowledge can identify areas where L2 learners may need additional support. Future research directions include further exploration of the relationship between lexical richness and L2 vocabulary knowledge. As university examiners, it is important to understand the significance of lexical richness in evaluating an L2 learner's vocabulary knowledge and to consider this factor when evaluating their language proficiency.

AI essay 5

Introduction
Lexical richness is an important aspect of second language (L2) learning that reflects a learner's vocabulary knowledge and language proficiency. It refers to the diversity and variety of words used by an L2 learner, including factors such as word frequency, complexity, and appropriateness. Vocabulary knowledge plays a significant role in L2 acquisition as words are the building blocks of language. This essay will explore what lexical richness tells us about an L2 learner's vocabulary knowledge and its relevance to L2 learning.

Lexical Richness and Vocabulary Knowledge
Lexical richness is a strong indicator of a learner's vocabulary knowledge. A learner with a high level of lexical richness has a larger vocabulary and is able to use words in a variety of contexts. In contrast, a learner with a low level of lexical richness may struggle to express themselves accurately and fluently. Therefore, lexical richness is a useful tool for measuring L2 proficiency as it reflects the learner's ability to use language accurately and appropriately.

The relationship between lexical richness and vocabulary knowledge is complex. It is not simply a matter of knowing more words, but also being able to use them effectively in communication. Learners with high levels of lexical richness are able to use words in context, which enhances their communicative competence. In contrast, learners with low levels of lexical richness may struggle to use words accurately and appropriately in different contexts.

The role of lexical richness in measuring L2 proficiency is significant. It is often used in conjunction with other measures such as grammatical accuracy, fluency, and comprehension to provide a comprehensive assessment of language proficiency. For example, a learner may have good grammatical accuracy but limited vocabulary knowledge, leading to restricted communication abilities.

Factors Affecting Lexical Richness
Several factors affect the development of lexical richness in L2 learners. These include:
Environment and exposure to L2: Learners who are exposed to L2 in naturalistic settings or immersion programs tend to have higher levels of lexical richness. Exposure to authentic materials such as books, movies, and TV shows can also contribute to the development of lexical richness.

Learning strategies: Learners who use effective vocabulary learning strategies such as spaced repetition, contextual learning, and active recall tend to have higher levels of lexical richness. These strategies help learners retain and use new vocabulary more effectively.
Cognitive factors: Some learners may have cognitive factors such as working memory capacity or processing speed that affect their ability to acquire and use new vocabulary.

Research Studies on Lexical Richness
Numerous studies have investigated the relationship between lexical richness and L2 proficiency. These studies have used a variety of methods to measure lexical richness including word frequency lists, text analysis, and self-report measures. Studies have consistently found a strong positive correlation between lexical richness and L2 proficiency. Learners with higher levels of lexical richness are able to produce more accurate and fluent language. The findings suggest that vocabulary knowledge is a key factor in L2 acquisition and should be a focus of L2 instruction.

Assessment of Lexical Richness

Quantitative measures include word frequency lists, type-token ratios, and other statistical measures. Qualitative measures involve analyzing the context and appropriateness of words used by learners. The validity and reliability of measures vary depending on the method used. It is important to choose appropriate measures that accurately reflect the learner's vocabulary knowledge and language proficiency.

Conclusion

In conclusion, lexical richness is an important aspect of L2 learning that reflects a learner's vocabulary knowledge and language proficiency. Factors such as environment, learning strategies, and cognitive factors can affect the development of lexical richness. Research studies have consistently found a positive correlation between lexical richness and L2 proficiency. L2 teachers should focus on developing learners' vocabulary knowledge through effective instruction and exposure to authentic materials. Learners should use effective vocabulary learning strategies to improve their lexical richness and overall language proficiency. Future research could investigate the effectiveness of different vocabulary learning strategies and the role of cognitive factors in the development of lexical richness.

AI essay 6

The Importance of Lexical Richness in Assessing L2 Learners' Vocabulary Knowledge

As university examiners, it is our responsibility to assess the language proficiency of L2 learners accurately. One crucial component of language learning is vocabulary knowledge, which enables learners to communicate effectively in the target language. However, assessing vocabulary knowledge can be challenging, as it involves evaluating not only the number of words known but also their complexity and breadth. This is where the concept of lexical richness comes into play. In this essay, we will explore what lexical richness tells us about an L2 learner's vocabulary knowledge, including its theoretical background, methods of assessment, empirical studies, and implications for examiners.

I. Introduction

A. Definition of lexical richness

Lexical richness is a measure of the variety of words and expressions used in a text. It is used to assess the complexity and breadth of an individual's vocabulary knowledge and language proficiency. Lexical richness can be measured using various indices, such as word frequency counts, type-token ratios, lexical diversity indices, and semantic density measures.

B. Importance of assessing vocabulary knowledge in L2 learning

Vocabulary knowledge is one of the most important components of language learning. It is essential for learners to understand and use words and expressions in context to communicate effectively in the target language. Assessing vocabulary knowledge can help identify areas that need improvement and inform teaching practices.

C. Significance of lexical richness in evaluating L2 learners' vocabulary knowledge

Lexical richness provides a useful insight into an L2 learner's vocabulary knowledge and can be used to evaluate their language proficiency. A high level of lexical richness indicates a broad vocabulary range and a better understanding of the nuances of the target language.

II. Theoretical Background

A. Second language acquisition theories

Lexical richness has been studied within the context of various second language acquisition theories, such as the Input Hypothesis and the Output Hypothesis. These theories suggest that exposure to a variety of language input and opportunities for language production are essential for vocabulary acquisition.

B. Vocabulary acquisition in L2 learning

Vocabulary acquisition is a complex process that involves both conscious and unconscious learning strategies, as well as multiple contexts for use. Learners use various techniques to acquire new words, such as memorization, guessing from context, and learning word families.

C. Role of lexical richness in L2 vocabulary development

Lexical richness is believed to play an important role in L2 vocabulary development, as it provides learners with exposure to a greater variety of words and expressions in the target language. A broader vocabulary range can also help learners understand the nuances of the language and communicate more effectively.

III. Methods of Assessing Lexical Richness

A. Types of lexical richness measures

There are several methods for assessing lexical richness, such as word frequency counts, type-token ratios, lexical diversity indices, and semantic density measures. Each method has its own advantages and disadvantages, which should be carefully considered when selecting a measure for a particular purpose.

B. Advantages and disadvantages of different measures
Word frequency counts are easy to calculate but may not accurately reflect the complexity of the text. Type-token ratios provide information on the variety of words used but do not account for the frequency of those words. Lexical diversity indices consider both variety and frequency but may be time-consuming to calculate. Semantic density measures evaluate the number of different meanings expressed in a text but may not capture the full range of vocabulary knowledge.
C. Practical considerations in selecting lexical richness measures
Practical considerations such as time constraints and cost should also be taken into account when deciding which measure to use. For example, word frequency counts may be more suitable for large-scale studies, while lexical diversity indices may be more appropriate for smaller-scale studies.

IV. Empirical Studies on Lexical Richness and L2 Vocabulary Knowledge
A. Findings on the relationship between lexical richness and L2 vocabulary knowledge
Studies have found that there is a positive correlation between lexical richness and L2 vocabulary knowledge, with higher levels of lexical richness indicating greater levels of vocabulary knowledge. For example, a study conducted by Meara (1996) found that the type-token ratio was a good predictor of vocabulary size in L2 learners.
B. Factors affecting the relationship between lexical richness and L2 vocabulary knowledge
Several factors can affect this relationship, such as the type of text used for analysis, the language proficiency level of the learner, and the language context in which the text is used. For example, a study conducted by Laufer and Nation (1995) found that the relationship between type-token ratio and vocabulary size was stronger for high-frequency words than low-frequency words.
C. Implications of research findings for L2 vocabulary assessment and instruction
The findings from these empirical studies have important implications for both assessment and instruction, as they suggest that lexical richness can be used to accurately evaluate learners' vocabulary knowledge and inform teaching practices. For example, teachers can focus on providing learners with exposure to a variety of language input and opportunities for language production to enhance their lexical richness.

V. Conclusion
A. Summary of key points
This essay has explored the importance of lexical richness in assessing L2 learners' vocabulary knowledge, including its theoretical background, methods of assessment, empirical studies, and implications for examiners.
B. Implications for examiners in assessing L2 learners' vocabulary knowledge
Examining lexical richness provides examiners with valuable insights into an individual's level of L2 vocabulary knowledge, which can help to accurately assess their progress and inform their instruction.
C. Recommendations for future research on lexical richness and L2 learning
Further research is needed to better understand how lexical richness can be used to assess and teach L2 learners more effectively. For example, future studies could investigate the relationship between lexical richness and language proficiency in different language contexts and examine the effectiveness of different instructional methods for enhancing lexical richness.
In conclusion, lexical richness is a valuable tool for assessing L2 learners' vocabulary knowledge. It provides insights into the complexity and breadth of an individual's vocabulary range, which can inform teaching practices and accurately evaluate learners' progress. As

examiners, we should consider using lexical richness measures in our assessments to ensure we assess L2 learners' vocabulary knowledge accurately.

AI Essay 7

What does ‘lexical richness’ tell us about an L2 learner’s vocabulary knowledge?
Language is a mode of communication that is essential for human beings to interact with one another. As such, learning a second language is becoming increasingly important in today's globalized world. Second language acquisition refers to the process of acquiring a new language after already acquiring a first language. Vocabulary knowledge is crucial for effective communication in any language, and it plays a critical role in second language (L2) learning. Lexical richness refers to the variety and complexity of vocabulary used by a language learner. It is a significant indicator of a learner's proficiency level and their ability to effectively communicate in the target language. This essay aims to explore the significance of lexical richness in L2 learning, its relationship with vocabulary knowledge, and the factors that affect it.

Lexical richness is the extent to which a language learner uses a wide range of vocabulary items and can effectively express themselves using them. It serves as an essential indicator of a learner's proficiency level and overall language skills. Measuring lexical richness can be done through various tools such as vocabulary tests, text analysis, and computer software programs. Lexical richness is closely related to vocabulary knowledge, and learners with higher levels of lexical richness tend to have a more extensive and diverse vocabulary.

Several studies have examined the relationship between lexical richness and vocabulary knowledge in L2 learners, providing valuable insights into the significance of lexical richness in language learning. For example, a study by Laufer and Nation (1995) found that learners with higher levels of lexical richness had better comprehension and recall of texts they read. Another study by Crossley et al. (2011) found that lexical richness was a significant predictor of essay quality in L2 learners.

Individual differences can significantly affect learners' lexical richness and overall language proficiency. Age, motivation, and aptitude are some individual factors that can influence learners' lexical richness. Younger learners tend to have higher levels of lexical richness than older learners, as they are more receptive to new vocabulary. Motivation and aptitude also play a critical role in learners' lexical richness, as motivated and talented learners tend to acquire new vocabulary more effectively.

Exposure to the target language and effective instructional methods can enhance learners' lexical richness and improve their overall language skills. Contextual factors such as the amount of exposure to L2 and instructional methods can influence learners' lexical richness. For example, learners who are exposed to L2 more frequently tend to have higher levels of lexical richness. Effective instructional methods such as extensive reading, vocabulary games, and explicit instruction can also help learners enhance their lexical richness.

Measuring lexical richness is an effective way to assess a learner's vocabulary knowledge, which is critical for tracking their progress and identifying areas for improvement. Teachers should focus on enhancing learners' lexical richness through effective teaching strategies and interventions. Effective teaching strategies such as extensive reading, vocabulary games, and explicit instruction can help learners enhance their lexical richness. Several studies have investigated the effectiveness of various instructional methods and interventions for improving learners' lexical richness.

In conclusion, lexical richness plays a crucial role in L2 learning and serves as an essential indicator of a learner's proficiency level and vocabulary knowledge. Measuring lexical richness can help teachers assess learners' vocabulary knowledge, track their progress, and identify areas for improvement. Teachers should focus on enhancing learners' lexical richness through effective teaching strategies and interventions. Further research is needed to explore the relationship between lexical richness and other factors that influence language learning. Ultimately, understanding the significance of lexical richness in L2 learning can help learners acquire new vocabulary more effectively and communicate more fluently in the target language.

Appendix C; MSc ALSLA formative assignment instruction sheet

UNIVERSITY OF OXFORD
M.SC. APPLIED LINGUISTICS AND SECOND LANGUAGE ACQUISITION

Formative Assignment
Michaelmas Term 2021

In all assignments, use your own words wherever possible or appropriate, but also try to demonstrate your mastery of the topic's *metalanguage.*

Write no more than 1,500 words (excluding references but including any footnotes/endnotes).

Submit this assignment as a Word file directly to your supervisor via email no later than noon on Monday 14th November 2022 (i.e. by 12h00 on Monday of 6th Week).

Formative assignment question:

- What does 'lexical richness' tell us about an L2 learner's vocabulary knowledge? In your answer provide an overview of the relevant research and an in-depth critique of at least one study.

Please remember!

- This is a purely formative assignment. It does not contribute to your overall mark on the course. You do not need to worry about it at all – it is simply a chance to learn and improve.
- You will receive feedback from your own supervisor highlighting strengths and areas for improvement.
- Your supervisor will not give you a mark for this assignment. The examined assignments will be significantly longer (2,500 words) and so we are not able to apply the same mark scheme to this shorter piece of work.
- Nonetheless, your supervisor may indicate in broad terms where the assignment might be expected to sit in relation to the mark scheme, were it to be scaled up appropriately for a longer, examined assignment. For example, your supervisor might say that your work has some features which are consistent with a Low Pass / High Pass / Merit / Distinction.

Appendix D; List of linguistic features

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Feature	H1	H2	H3	H4	H5	H6	H7	AI1	AI2	AI3	AI4	AI5	AI6	AI7
2	Establishes common applied linguistics abbreviation to use later	1	1	0	1	1	1	0	1	1	0	1	1	0	1
3	First person plural pronoun	1	1	1	0	0	0	1	1	1	1	1	1	1	0
4	Error in choice of article	1	1	1	0	0	0	0	0	0	0	0	0	0	0
5	Gives essay a third person neutral singular pronoun / "one's"	1	0	0	0	0	0	0	0	0	0	1	0	0	0
6	Establishes dates of literature used	1	0	0	0	0	0	0	0	0	0	0	0	0	0
7	Erorr in use of comma	1	0	0	0	0	0	0	0	0	0	0	0	0	0
8	Error in citation	1	0	0	0	0	0	0	0	0	0	0	0	0	0
9	Run-on sentence	1	0	1	1	0	1	1	0	0	0	0	0	0	0
10	Only the first word capitalized in subheading	1	1	0	0	0	1	1	1	0	0	0	0	0	0
11	British English	1	0	1	1	1	1	0	0	0	0	0	0	0	0
12	American English	0	1	1	0	0	1	0	0	1	1	0	0	1	0
13	Missing comma	1	0	1	1	0	0	1	0	0	0	0	0	0	0
14	Incomplete clause	1	0	0	0	0	0	0	0	0	0	0	0	0	0
15	APA in-text citations	1	1	1	0	0	0	0	0	1	0	0	0	0	0
16	Error writing "et al."	1	0	0	0	0	0	1	0	0	0	0	0	0	0
17	Error in use of dash	1	0	0	0	1	0	0	0	0	0	0	0	0	0
18	Error in choice of tense	1	1	0	1	0	0	0	0	0	0	0	0	0	0
19	Question as a subheading	1	0	0	0	0	0	0	0	1	0	0	0	0	0
20	Use of 'and' and '&' in in-text citation alternatively	1	0	0	0	0	0	0	0	0	0	0	0	0	0
21	Informal register	1	0	0	0	0	0	0	0	0	0	0	1	0	0
22	Starting two clauses with "that"	1	0	0	0	0	0	0	0	0	0	0	0	0	0
23	Missing noun after "this"	1	0	1	0	0	0	0	0	1	0	0	0	0	0
24	Use of abbreviation but not specified what the full form is	1	1	1	0	0	1	1	0	1	1	1	0	0	0
25	Spelling error	1	0	0	1	0	0	0	0	0	0	0	0	0	0
26	Semantic error (e.g. learning instead of learner)	1	1	1	0	1	1	0	0	1	0	0	0	0	0
27	Use of passive sentences	1	0	0	0	0	0	0	0	0	0	0	0	0	0
28	Mentions a claim but does not analyze it	1	0	0	0	0	0	0	0	0	0	0	0	0	0
29	Use of suggestion for future research	1	0	0	0	0	0	0	0	0	0	0	0	0	0
30	Use of the same adverb twice in one sentence	1	0	0	0	0	0	0	0	0	0	0	0	0	0
31	No subheading for introductory or concluding paragraph	1	0	0	0	0	0	0	0	0	0	0	0	0	0
32	Missing Oxford comma	1	1	1	1	0	1	0	0	0	0	0	0	0	0
33	Confusing referent due to lengthier clause	1	0	0	0	0	0	0	0	0	0	0	0	0	0
34	Signposts structure of essay	1	0	0	0	0	0	0	1	1	0	0	0	0	0
35	Numbering subheadings and paragraphs	0	0	0	0	0	1	0	0	0	0	0	0	0	0
36	Use of Oxford comma within in-text citation	0	0	0	0	0	0	0	0	0	0	0	0	0	0
37	Use of tables	0	0	0	0	0	1	0	0	0	0	0	0	0	0
38	Missing comma around "therefore", "but", "thus", "as well as"	0	1	1	0	1	1	0	0	0	0	0	0	0	0
39	Use of "e.g." in longer in-text citations while citing multiple sources	0	1	0	0	0	0	0	0	0	0	0	0	0	0
40	Mixing e.g. and e.g., in one essay (AE vs. BE)	0	1	1	0	0	0	0	0	0	0	0	0	0	0
41	Erorr with apostrophe	0	0	1	0	0	1	0	0	0	0	0	0	0	0
42	"(In) This paper/essay"	0	0	1	1	1	0	1	1	1	1	1	1	1	1
43	Dividing subheadings using "Part 1, Part 2"	0	0	1	0	0	0	0	0	0	0	0	0	0	0
44	Use of contraptions	0	0	1	0	0	0	0	0	0	0	0	0	0	0
45	Error in number	0	0	1	0	0	0	0	0	0	0	0	0	0	0
46	Extremely long parenthesis	0	0	1	0	0	0	0	0	0	0	0	0	0	0
47	Incorrect in-text citations	1	0	1	1	0	0	1	0	0	0	0	0	0	0
48	Missing preposition	0	0	1	0	0	0	0	0	0	0	0	0	0	0
49	Missing comma before "as"	0	0	1	0	1	0	0	0	0	0	0	0	0	0
50	Adverbs before and after verbs	0	0	1	0	0	0	0	0	0	0	0	0	0	0
51	Use of "&" instead of "and"	0	0	1	0	0	0	0	0	0	0	0	0	0	0
52	Error in writing a number (e.g. 1000 instead of 10,000)	0	0	1	0	0	0	0	0	0	0	0	0	0	0
53	Writing numbers in their word formats	0	0	1	0	0	0	0	0	0	0	0	0	1	0
54	Use of second person singular pronoun	0	0	1	0	0	0	0	0	0	0	0	0	0	0
55	Missing pronoun	0	0	1	0	0	0	0	0	0	0	0	0	0	0
56	Use of "one" for fomal register and gender-neutral language	0	0	0	1	0	0	0	0	0	0	0	0	0	0
57	Use of colon to separate clauses	0	0	0	1	1	0	1	0	1	0	1	0	0	0
58	Use of "In [date], [author]"	0	0	0	1	0	0	0	0	0	0	0	0	0	0
59	Error in citing quotation	0	0	0	1	0	1	0	0	0	0	0	0	0	0
60	Missing quotation marks	0	0	0	1	0	0	0	0	0	0	0	0	0	0
61	First person singular pronoun	0	0	0	1	0	0	0	0	0	0	0	0	0	0
62	Error in choice of preposition	0	0	0	1	0	0	0	0	0	0	0	0	0	0
63	Error in paragraph break	0	0	0	1	0	0	0	0	0	0	0	0	0	0
64	Using number in the number format	0	0	0	1	0	0	0	0	0	0	0	0	0	0
65	Error in conjunction	0	0	0	1	0	0	0	0	0	0	0	0	0	0
66	Incorrect/unnecessary capitalization of nouns	0	0	0	1	0	1	0	0	1	0	1	0	1	0
67	Use of "unless specified"	0	0	0	0	1	0	0	0	0	0	0	0	0	0
68	Missing article	0	0	0	0	1	0	0	0	0	0	0	0	0	0
69	"(s)he"	0	0	0	0	1	0	0	0	0	0	0	0	0	0
70	i.e. with comma (AE)	0	0	0	0	1	0	0	0	0	0	0	0	0	0
71	Use of 1.1., 1.2.	0	0	0	0	0	1	0	0	0	0	0	0	0	0
72	More than one word capitalized in subheading	0	0	0	0	0	1	0	0	1	1	1	1	1	0
73	Establishing uncommon, personalized abbreviations	0	0	0	0	0	1	0	0	0	0	0	0	0	0
74	Missing "and"	0	0	0	0	0	1	0	0	0	0	0	0	0	0
75	Literatures (with an s)	0	0	0	0	0	1	0	0	0	0	0	0	0	0
76	Dividing clauses with semi-colons	0	0	0	0	0	1	0	0	1	0	0	0	0	0
77	Error in punctuation (use of period instead of comma, forgot to close parenthesis, mis	0	0	0	0	0	1	1	0	0	0	0	0	0	0
78	Comma before etc.	0	0	0	0	0	1	0	0	0	0	1	0	0	0
79	Use of (1)	0	0	0	0	0	1	0	0	0	0	0	0	0	0
80	Conclusion paragraph	1	1	1	1	1	0	0	0	0	0	0	0	0	0
81	Use of secondary sources	0	0	0	0	0	0	1	0	0	0	0	0	0	0
82	Use of first name for authors in some sentences	0	0	0	0	0	0	1	0	0	0	0	0	0	0
83	"s/he"	0	0	0	0	0	0	1	0	0	0	0	0	0	0
84	No indicators of British or American English	0	0	0	0	0	0	1	1	0	0	1	0	0	1
85	Parenthesis for definitions	0	0	0	0	0	0	0	0	1	0	1	0	0	0
86	Research paper format	0	0	0	0	0	0	0	0	1	0	0	0	0	0
87	Use of Roman numerals to separate subheadings and sections	0	0	0	0	0	0	0	0	0	1	0	0	1	0
88	Breaking a section off with a colon and using a colon for the sub-sub-sections	0	0	0	0	0	0	0	0	0	0	0	1	0	0
89	Untitled preface before the introduction	0	0	0	0	0	0	0	0	0	0	0	0	1	0
90	Separation of sub-sub-sections by lettering (A, B, C)	0	0	0	0	0	0	0	0	0	0	0	0	1	0
91	Sub-sub-headings with first word capitalized only	0	0	0	0	0	0	0	0	0	0	0	0	1	0
92	No subheadings at all	0	0	0	0	0	0	0	0	0	0	0	0	0	1
93	Total # of features	34	14	29	21	13	23	14	6	16	6	11	6	10	4

Common features:

#	Feature	Example(s) in human-written essays	Example(s) in AI-generated essays
1	Establishes common applied linguistics abbreviation to use later	"second language (**L2**)" (human essay 1) "Type-Token Ratio (**TTR**)"	"second language (**L2**)" (AI essay 1) "Second Language

	in the essay.	(human essay 2)	Acquisition (**SLA**)" (AI essay 2)
2	Uses the first person plural pronoun.	"Next, **we** will examine…" (human essay 3) "...allows **us** to know the frequency…" (human essay 7)	"...**we** will explore…" (AI essay 3) "...tells **us** about an L2 learner's…"(AI essay 5)
3	Uses third person neutral singular pronoun and/or subject.	"Firstly, **it** will briefly explain…" (human essay 1). Here, 'it' is used to refer to the essay.	"...not **one's** native tongue." (AI essay 4)
4	Capitalizes only the first letter of the first word in a subheading, with the rest of the words uncapitalized.	"**Vocabulary** knowledge and lexical richness" (human essay 6) "**Lexical** richness and vocabulary size" (human essay 7)	"**Definition** of lexical richness", "**Factors** that affect lexical richness in L2 learners" (AI essay 1)
5	Uses American English.	"conceptualizations and operationalizations" (human essay 2) "summarized" (human essay 6)	"memorization" (AI essay 2) "prioritizing" (AI essay 3)
6	Uses APA in-text citations.	"(Wesche & Paribakht, 1996)" (human essay 1) "(Pawley & Syder, 1983)" (human essay 3)	"(Hyland and Tse, 2007)" (AI essay 2)
7	Uses a question as a subheading.	"What does lexical richness tell us about an L2 vocabulary?" (human essay 1)	"Research Question: What does 'lexical richness' tell us about an L2 learner's vocabulary knowledge?" (AI essay 2)
8	Uses informal register at times.	"**So far** studies have revealed…" (human essay 1)	"It is not **simply a matter** of…" (AI essay 5)
9	Is missing a noun after "this".	"This indicates…" (human essay 3)	"This can include…" (AI essay 2)
10	Uses an abbreviation but does not specify what the full form is.	"**L2** learners…" (human essay 6) "...with French as their **L1**…" (human essay 7)	"...**L2** learners'..." (AI essay 4)
11	Has a semantic error.	"...L2 **learning** learn…" instead of learner(s) (human	"...discuss our **methodology…**" instead of

		essay 1)	methods (AI essay 2)
12	Signposts the upcoming structure of the essay.	“First, it will briefly explain… / Next, this paper will explain… / It will attempt to critically assess…” (human essay 1)	“This essay will provide… / We will also explore…” (AI essay 2)
13	Uses “This paper/essay”.	“...**this essay** focuses on…” (human essay 4) “**This paper** aims to…” (human essay 5)	“**This essay** will explore…” (AI essay 5) “**This essay** aims to explore…” (AI essay 7)
14	Writes numbers in their word format.	“As discussed in part **three…**” (human essay 3)	“**One** crucial component…” (AI essay 6)
15	Uses a colon to separate clauses.	“...representing both at the same time**:** being able to comprehend…” (human essay 4)	“There are two types of vocabulary knowledge**:** receptive vocabulary…” (AI essay 4)
16	Uses a semicolon to separate clauses.	“...word part under “form”**;** form and meaning…” (human essay 6)	“...on written language**;** future studies…” (AI essay 2)
17	Has unnecessary capitalization of certain nouns.	“...from **Grade** 12 to **Grade** 11…” (human essay 4)	“...**Type-Token Ratio**…” (AI essay 4)
18	Has more than one word capitalized in the subheading.	“1.1 Vocabulary Knowledge” (human essay 6)	“II. Understanding Lexical Richness” (AI essay 3)
19	Uses a comma before ‘etc.’.	“...percentage of lexical errors**, etc.**” (human essay 6)	“...frequency of word use**, etc.**” (AI essay 4)
20	Has no indicator of British or American English.	In human essay 7, there are no words that indicate whether the student has used British or American English, as they have only used words found in both dialects of English throughout.	In AI essays 1, 4, and 7, there are no words that indicate whether the AI has used British or American English, as it has only used words found in both dialects of English throughout.

Features related to linguistic errors:

#	Feature (error)	Example(s)
1	Error in choice of article.	"...what we can learn about **an** learners L2 vocabulary." (human essay 1) [instead of "a learner's"]
2	Error in use of comma.	"... (SLA) research seems to suggest**,** that one's superiority, or sophistication of lexical ability, allows…" (human essay 1) [instead of "to suggest that"].
3	Error in citation.	"Schmitt et al (2017)" (human essay 1) [instead of "Schmitt et al. (2017)"].
4	Missing comma.	"An L2 learners' vocabulary knowledge is one window into understanding their proficiency level and lexical richness is a term which describes the depth of a learner's knowledge." (human essay 3) [missing comma before "and"].
5	Missing Oxford comma.	" i.e., Lexical Originality, Lexical Density, Lexical Sophistication and Lexical Variation." (human essay 6).
6	Error in writing 'et al.'.	"... as Schmitt **et al** (2017) found,..." (human essay 1) "...views by Ellis **et. al.** (2008), who argue…" (human essay 1) [instead of "et al."].
7	Error in use of dash or hyphen.	"...which is hazardous for SLA researchers - who often used widely accepted - but very standardised.." (human essay 1).
8	Error in choice of tense.	In human essay 2, they start a section with the present simple tense, and have the present perfect continuous tense within the same sentence. However, they switch to past continuous tense, breaking the flow of the paragraph. See paragraph 2 for ALSLA essay 2 in appendix [X].
9	Spelling error.	"reply" instead of "rely" (human essay 1). "leaning" instead of "learning" (human essay 1). "sixe" instead of "size" (human essay 4).
10	Confusing referent of lengthier clause	"...vocabulary size is strongly associated with L2 reading, and interestingly as Schmitt et al (2017) found, also in written production." (human essay 1).
11	Missing comma around conjunctions.	"...they also gain awareness of syntagmatic relations **thus** enriching their lexis." (human essay 3). "This mark of lexical richness as defined earlier **therefore subsequently** denotes the…" (human essay 3).

12	Error in use of apostrophe.	"...an L2 learners**'** vocabulary knowledge." (human essay 1) [instead of "an L2 learner's vocabulary knowledge].
13	Use of contractions (in an academic essay).	"...something learners either have or **don't** have." (human essay 3).
14	Error in number.	"**These** components all contribute to lexical richness as **it** demonstrates the different layers…" (human essay 3) ["these" = plural, "it" = singular].
15	Incorrect in-text citations.	"(Wray 2002)" instead of "(Wray, 2002)" (human essay 3).
16	Missing preposition.	"...L2 learner's vocabulary knowledge as higher level and more native like" (human essay 3) [missing preposition after "as", e.g., as at a higher level].
17	Missing comma before "as".	"...and more native **like as** the lexical richness here achieved through engaging with collocations makes the language user more comprehensible and fluent." (human essay 3).
18	Use of "&" instead of "and" in an academic paper.	"Schmitt, Horst, Cobb **&** Schmitt (2017) refer to lexical coverage…" (human essay 3). "...Hsueh-Chao & Nation (2000) argue that…" (human essay 3)
19	Error in writing a number.	"...argue that 8000-**1000** word families are required." (human essay 3) [instead of 10,000]
20	Missing pronoun.	"This means L2 learners could draw on (high frequency) synonyms and antonyms without too much difficulty, **however struggled** with more thematic word associations." (human essay 3) [missing pronoun after "however"].
21	Error in citing quotation.	"Later, in 1985, Halliday proposed an alternative representation of the idea of lexical density: "lexical density is the number of lexical items as a ratio of the number of clauses"." (human essay 4).
22	Missing quotation marks.	"...(e.g. **to take off** may be considered a content word…)" (human essay 4) ["to take off" should be in quotation marks].
23	Error in choice of preposition.	"...Mohammad Ahmad Bani Amer (2021) the author aims **at** examining…" (human essay 4) [preposition should be "to"].
24	Error in paragraph break.	In ALSLA essay #4, there is a paragraph break between two paragraphs which are dealing with the same concept

		with one paragraph being significantly shorter than the other (see Appendix A).
25	Error in choice of conjunction.	"This generalisation may seem confusing, **although** it can be argued that the concepts are closely intertwined." (human essay 4) [instead of "though"].
26	Missing article.	"The reason is **that ability** to use vocabulary appropriately is built upon…" (human essay 5) [missing article before "ability"].
27	Missing "and".	"...since the native expert corpus only consisted of published research articles, **peer reviews, teamwork**, professional editing might be involved.." (human essay 6) [missing "and" before "teamwork"].
28	Incorrect plural form.	"...referred to previous **literatures**." (human essay 6) [instead of literature]
29	Error in punctuation.	"Based on that, researchers (Engber, **1995;** continued to develop new concepts…" (human essay 6) [parenthesis not closed].
30	Missing comma before "etc.".	"...vocabulary is argued to play a major role (Lewis, Nation, Schmitt **etc**.)" (human essay 4)
31	Incomplete clause.	"...in two distinct ways. **First through a** 'developmental' approach, where…" (human essay 1) (missing relative clause before "first" which would link the previous sentence with this one).

Personalization features:

#	Feature	Example(s)
1	Establishes dates of literature used.	"Next, this paper will explain [...] by reviewing the relevant literature over the past 30 years to help understand…" (human essay 1).
2	Use of tables.	Present in human essay #6.
3	Use of "unless specified".	"Unless specified, all "words" in this paper are deemed as "tokens"". (human essay 5)
4	Establishes uncommon, personalized abbreviations to use later.	"Chinese PhD students (CPhD)" (human essay 5).

Length features:

#	Feature	Example(s)

1	Has run-on sentences.	“Some words can appear in many collocations, or multiword units, and understanding how the given word may change meaning or indeed serve varying functions in each collocation means the language user has a good graph on that word’s meaning.” (human essay 3).
2	Use of “e.g.” in longer in-text citations while citing multiple sources.	“[...] (e.g. Laufer & Nation, 1995; Ure, 1971; McCarthy & Jarvis, 2007).” (human essay 5)
3	Extremely long clauses in parentheses.	“[...] (whether syntactic contexts, e.g., the word in its figurative or literal sense, or physical contexts e.g. high register written piece in a journal of lower register conversation at the market) [...]”. (human essay 3).

Structure features:

#	Feature	Example(s)
1	Has a conclusion paragraph.	All ALSLA essays have a concluding paragraph except essay #6.

Pronoun features:

#	Feature	Example(s)
1	Uses the second person singular pronoun.	“To transform it into the 3rd person masculine plural ‘they enjoy’, **you** must add the suffix ‘eon’ making the word ‘yastamta-eon’.” (human essay 3).
2	Uses the first-person singular pronoun.	“Further, **I** am going to analyse an article that challenges lexical density in certain textbooks.” (human essay 4).

Style features:

#	Feature	Example(s)
1	Uses British English.	“Conceptualisations” (human essay 1). “criticised” (human essay 3). “characterise” “labelling” (human essay 4).
2	Uses “and” and “&” in in-text citation alternatively.	“(Laufer **and** Nation, 1995)” “(Cheng **&** Matthews, 2018; Li **&** Kirby, 2015; Webb, 2008)” (human essay 1).
3	Starts two clauses with “that”.	“So far studies have revealed **that**; L2 receptive vocabulary is generally much larger than L2 productive knowledge (Webb, 2008), **that** receptive vocabulary size is strongly associated with L2 reading, and interestingly as Schmitt et al (2017) found, also in written production.” (human essay 1).

4	Uses passive sentences.	"Three lexical measurement tests of low frequency words were taken directly from the core textbook used on the ESP programme." (human essay 1).
5	Uses the same adverb twice in one sentence.	"Although the nature of a tailored test is **extremely** effective in measuring vocabulary knowledge for a particular group of L2 learners, it is also **extremely** inefficient, and thus, not always practical." (human essay 1).
6	Uses Oxford comma within an in-text citation.	"(Astika 1993; Ellis, 1994; Engber, 1995; Wolfe-Quintero, **Inagaki, & Kim**, 1998)." (human essay 2).
7	Mixes British and American English in one essay	"...(whether syntactic contexts **e.g.,** the word in its figurative or literal sense, or physical contexts **e.g.** high register written piece in a journal of lower…)" (human essay 3).
8	Uses adverbs before and after verbs.	"...they can **proficiently** understand…" "...focuses **largely** on…" (human essay 3).
9	Uses numbers in the number format.	"...first **3** dimensions…" (human essay 4).
10	Use of "(s)he".	"**(S)he** may also know…" (human essay 5)
11	Uses "i.e."/ "e.g." with a comma.	"...e.g.,..." (human essay 3) "...i.e.,..." (human essay 5)
12	Use of "(1) (2) etc.".	"...two research question, **(1)** to identify…" (human essay 6)
13	"Use of "s/he".	"...**s/he** is familiar…" (human essay 7)

#	Feature	Example(s)
1	Uses parentheses for definitions.	"...passive knowledge **(the ability to recognize words).**." (AI essay 2)

Content Features:

#	Feature	Example(s)
1	Mentions a claim but does not analyze it.	In the concluding paragraph, ALSLA student 1 introduces a claim about the ongoing debate regarding measures of vocabulary knowledge, without elaborating or linking it back to the previous arguments. See ALSLA essay 1 in Appendix A.
2	Has suggestions for/acknowledges future	"Future studies should consider the L2 backgrounds of their subjects, and adapt their measurement techniques

	research with respect to an article.	accordingly." (human essay 1). "The author also fails to mention what further research could be done." (human essay 4).

#	Feature	Example(s)
1	Has an untitled preface before the introduction.	Present in AI essay #6.

Sources and references features:

#	Feature	Example(s)
1	Use of "In [date], [author]"	"Later, **in 1985, Halliday** proposed an alternative …" (human essay 4).
2	Use of secondary sources.	"(**Read, 2000, as cited in Chen, 2020**; and Dawson et al., 2021)" (human essay 7).
3	Use of first name for authors in some sentences.	"**Norbert Schmitt** elaborates on the concept…" (human essay 7).

Subheading features:

#	Feature	Example(s)
1	Has no subheading for introductory or concluding paragraphs.	ALSLA essay 1 does not have a subheading for its introductory or concluding paragraph, though it does have subheadings for the body paragraphs.
2	Numbers subheadings and paragraphs.	ALSLA essay 6 numbers its subheadings as such: "**1.** Vocabulary knowledge and lexical richness"
3	Divides subheadings using "Part 1, Part 2 etc.".	ALSLA essay 3 labels its subheadings as "**Part 1 // Part 2**".
4	Uses "1.1, 1.2 etc.".	ALSLA essay 6 labels sub-subheadings with decimals of numbers such as; "**1.1** Vocabulary Knowledge".

#	Feature	Example(s)
1	Uses Roman numerals to separate subheadings and sections.	"**I**. Introduction / **II.** Understanding Lexical Richness // **III.** Implications of Lexical Richness in L2 Learning …" (AI essay 3)
2	Breaks off a section with a colon and uses a colon for the sub-sub-section.	"Several factors affect the development of lexical richness in L2 learners. **These include:** **Environment and exposure to L2:** Learners who are

		exposed to L2 in naturalistic settings or immersion programs tend to have higher levels of lexical richness. Exposure to authentic materials such as books, movies, and TV shows can also contribute to the development of lexical richness. **Learning strategies:** Learners who use effective vocabulary learning strategies such as spaced repetition, contextual learning, and active recall tend to have higher levels of lexical richness. These strategies help learners retain and use new vocabulary more effectively." (AI essay 5)
3	Separates sub-sub-sections by lettering.	"II. Theoretical Background // **A.** Second language acquisition theories // **B.** Vocabulary acquisition in L2 learning" (AI essay 6).
4	Capitalizes the first letter of the first word of the sub-sub-headings.	"III. Methods of Assessing Lexical Richness // A. **T**ypes of lexical richness measures" (AI essay 6).
5	Does not have subheadings.	AI essay #7 does not have subheadings.
6	Research paper format.	Out of the 7 AI essays, only essay #2 has the format of a research paper with sections such as methodology, participants, data collection etc.

Appendix E; List of n-grams

ALSLA essay 5-grams

ALSLA essay #	5-gram	Freq
1	relationship between lexical richness and	2
2	a complete and detailed picture	2
	complete and detailed picture of	2
	is far away from providing	2
	it is far away from	2
	lexical diversity and lexical sophistication	2
	lexical richness and vocabulary knowledge	2
	of learners' productive vocabulary	
	knowledge	2
	picture of students' vocabulary knowledge	2
	us with a complete and	2
	with a complete and detailed	2
3	of lexical richness as it	2
	the language user has a	2
4	can be argued that the	2
	it can be argued that	2
5	lexical diversity density and sophistication	3
	the depth of vocabulary knowledge	3
	about the use of vocabulary	2
	and possesses higher lexical fluency	2
	at lexical diversity density and	2
	knowledge when looking at lexical	2
	lexical fluency as well as	2
	lexical richness and vocabulary knowledge	2
	looking at lexical diversity density	2
	of vocabulary and possesses higher	2
	range of vocabulary and possesses	2
	to the quality of esl	2

	vocabulary and possesses higher lexical	2
	vocabulary knowledge when looking at	2
	when looking at lexical diversity	2
6	an estimate of the learner	2
	and sophistication in a linguistic	2
	density diversity and sophistication in	2
	diversity and sophistication in a	2
	example given in lightbown and	2
	give an estimate of the	2
	given in lightbown and spada	2
	how to use the word	2
	in the example given in	2
	lexical density diversity and sophistication	2
	lexical richness and vocabulary size	2
	lexical richness exhibited in an	2
	linguistic production can indicate the	2
7	lexical richness of research articles	3
	learners' vocabulary knowledge and the	2
	lexical sophistication and lexical variation	2
	the lexical richness of research	2
	vocabulary knowledge and lexical richness	2

Grouped n-grams:

ALSLA Essay	5-grams	Grouping into larger n-grams
2	"a complete and detailed picture" "complete and detailed picture of"	"a complete and detailed picture of"
	"is far away from providing" "it is far away from"	"it is far away from providing"
	"us with a complete and" "with a complete and detailed"	"us with a complete and detailed"
4	"it can be argued that" "can be argued that the"	"it can be argued that the"
5	"lexical diversity, density, and sophistication" "looking at lexical diversity, density" "at lexical diversity, density, and" "when looking at lexical diversity"	"vocabulary knowledge when looking at lexical diversity, density, and sophistication"

	“vocabulary knowledge when looking at”	
6	“lexical richness of research articles” “the lexical richness of research”	“the lexical richness of research articles”
7	“and sophistication in a linguistic” “density, diversity and sophistication in” “diversity and sophistication in a” “lexical density, diversity and sophistication”	“lexical density, diversity and sophistication in a linguistic”
	“example given in Lightbown and Spada” “in the example given in” “given in Lightboawn and Spada”	“In the example given in Lightbown and Spada”
	“lexical richness exhibited in an” “linguistic production can indicate the”	“lexical richness exhibited in an L2 learner’s linguistic production can indicate the”

AI essay 5-grams

AI Essay #	n-gram	Freq
1	and designing effective instructional strategies	2
	assessing their language proficiency and	2
	factors that affect lexical richness	2
	in assessing their language proficiency	2
	language proficiency and designing effective	2
	new words and use them	2
	proficiency and designing effective instructional	2
	relationship between lexical richness and	2
	that affect lexical richness in	2
	the diversity and complexity of	2
	the relationship between lexical richness	2
	their language proficiency and designing	2
2	and potential for future success	5
	language proficiency and potential for	5
	overall language proficiency and potential	5
	proficiency and potential for future	5
	can gain insights into their	4
	gain insights into their overall	4
	in assessing vocabulary knowledge in	4
	insights into their overall language	4
	into their overall language proficiency	4
	of lexical richness in assessing	4
	the concept of lexical richness	4
	the diversity and complexity of	4
	their overall language proficiency and	4
	university examiners and language teachers	4
	we can gain insights into	4
	and complexity of a learner	3
	by examining the diversity and	3
	diversity and complexity of a	3
	examining the diversity and complexity	3

for university examiners and language	3
found that lexical richness was	3
implications for university examiners and	3
lexical richness in assessing vocabulary	3
other measures of vocabulary knowledge	3
richness in assessing vocabulary knowledge	3
to other measures of vocabulary	3
vocabulary we can gain insights	3
a strong foundation in vocabulary	2
a way to assess vocabulary	2
an important aspect of vocabulary	2
and self reported vocabulary knowledge	2
as a way to assess	2
aspect of vocabulary knowledge that	2
become increasingly important in recent	2
capture the full range of	2
concept of lexical richness has	2
develop their lexical richness skills	2
does lexical richness tell us	2
for example some participants who	2
full range of a learner	2
has become increasingly important in	2
importance in assessing vocabulary knowledge	2
important aspect of vocabulary knowledge	2
important in recent years as	2
in recent years as a	2
increasingly important in recent years	2
is an important aspect of	2
its importance in assessing vocabulary	2
knowledge the concept of lexical	2
learning by examining the diversity	2
lexical richness has become increasingly	2
lexical richness is an important	2

	lexical richness tell us about	2
	lexical richness was positively correlated	2
	long term language learning success	2
	of lexical richness has become	2
	question what does lexical richness	2
	recent years as a way	2
	research question what does lexical	2
	richness has become increasingly important	2
	richness is an important aspect	2
	richness tell us about an	2
	richness was positively correlated with	2
	role of lexical richness in	2
	significance of lexical richness in	2
	suggest that lexical richness is	2
	test and self reported vocabulary	2
	that lexical richness is an	2
	that lexical richness was positively	2
	the full range of a	2
	the role of lexical richness	2
	their lexical richness skills through	2
	to assess vocabulary knowledge in	2
	to develop their lexical richness	2
	vocabulary knowledge the concept of	2
	vocabulary test and self reported	2
	way to assess vocabulary knowledge	2
	what does lexical richness tell	2
	years as a way to	2
3	as guide teachers in designing	2
	as well as guide teachers	2
	can help university examiners better	2
	designing effective lesson plans that	2
	effective lesson plans that focus	2
	focus on developing learners vocabulary	2

	guide teachers in designing effective	2
	in designing effective lesson plans	2
	language proficiency as well as	2
	learners language proficiency as well	2
	lesson plans that focus on	2
	on developing learners vocabulary knowledge	2
	plans that focus on developing	2
	proficiency as well as guide	2
	teachers in designing effective lesson	2
	that focus on developing learners	2
	well as guide teachers in	2
4	an important factor in evaluating	2
	is an important factor in	2
	of lexical richness in evaluating	2
	relationship between lexical richness and	2
	the relationship between lexical richness	2
	vocabulary knowledge is essential to	2
5	the development of lexical richness	4
	higher levels of lexical richness	3
	vocabulary knowledge and language proficiency	3
	able to use words in	2
	affect the development of lexical	2
	correlation between lexical richness and	2
	have higher levels of lexical	2
	is an important aspect of	2
	learning that reflects a learner	2
	levels of lexical richness are	2
	lexical richness and vocabulary knowledge	2
	lexical richness are able to	2
	lexical richness is an important	2
	lexical richness may struggle to	2
	of lexical richness are able	2
	of lexical richness may struggle	2

	positive correlation between lexical richness	2
	relationship between lexical richness and	2
	richness is an important aspect	2
	studies have consistently found a	2
	tend to have higher levels	2
	the relationship between lexical richness	2
	to have higher levels of	2
	use effective vocabulary learning strategies	2
6	relationship between lexical richness and	3
	the relationship between lexical richness	3
	a variety of language input	2
	and breadth of an individual	2
	and opportunities for language production	2
	as word frequency counts type	2
	assessment empirical studies and implications	2
	background methods of assessment empirical	2
	communicate effectively in the target	2
	complexity and breadth of an	2
	counts type token ratios lexical	2
	diversity indices and semantic density	2
	effectively in the target language	2
	empirical studies and implications for	2
	example a study conducted by	2
	exposure to a variety of	2
	for example a study conducted	2
	frequency counts type token ratios	2
	importance of lexical richness in	2
	including its theoretical background methods	2
	indices and semantic density measures	2
	input and opportunities for language	2
	its theoretical background methods of	2
	knowledge including its theoretical background	2
	language input and opportunities for	2

	learners with exposure to a	2
	lexical diversity indices and semantic	2
	lexical richness can be used	2
	methods of assessment empirical studies	2
	of assessment empirical studies and	2
	of language input and opportunities	2
	of lexical richness in assessing	2
	ratios lexical diversity indices and	2
	richness can be used to	2
	studies and implications for examiners	2
	such as word frequency counts	2
	the complexity and breadth of	2
	the importance of lexical richness	2
	theoretical background methods of assessment	2
	to a variety of language	2
	to communicate effectively in the	2
	token ratios lexical diversity indices	2
	type token ratios lexical diversity	2
	variety of language input and	2
	variety of words and expressions	2
	vocabulary knowledge including its theoretical	2
	word frequency counts type token	2
7	higher levels of lexical richness	4
	significance of lexical richness in	3
	the significance of lexical richness	3
	acquire new vocabulary more effectively	2
	an essential indicator of a	2
	areas for improvement teachers should	2
	as an essential indicator of	2
	as extensive reading vocabulary games	2
	can influence learners lexical richness	2
	effective teaching strategies and interventions	2
	enhancing learners lexical richness through	2

essential indicator of a learner	2
extensive reading vocabulary games and	2
focus on enhancing learners lexical	2
for improvement teachers should focus	2
games and explicit instruction can	2
have higher levels of lexical	2
help learners enhance their lexical	2
improvement teachers should focus on	2
learners enhance their lexical richness	2
learners lexical richness through effective	2
learners with higher levels of	2
lexical richness through effective teaching	2
on enhancing learners lexical richness	2
reading vocabulary games and explicit	2
relationship between lexical richness and	2
richness through effective teaching strategies	2
serves as an essential indicator	2
should focus on enhancing learners	2
such as extensive reading vocabulary	2
teachers should focus on enhancing	2
tend to have higher levels	2
the relationship between lexical richness	2
through effective teaching strategies and	2
to have higher levels of	2
vocabulary games and explicit instruction	2
with higher levels of lexical	2

Appendix F; MSc ALSLA grading rubric for summative assessments

&70+ (Distinction level)

At the upper end of this range (80+), assignments will generally have the followingcharacteristics:

Evidence of all the criteria at 70+ and in addition:

1. Clear evidence of individual searching and extensive reading of original sources.
2. Excellent balance between breadth of research/issues discussed, and depth of importantpapers/evidence examined.
3. Persuasive analysis/critique is both at the 'macro level' (the overall research output in thefield) and at the 'micro level' (limitations of individual studies).
4. Originality of insight.

At the lower level of this range 70-79 assignments will generally have the followingcharacteristics:

1. Addresses the question thoroughly and explicitly
2. Excellent balance of breadth and depth, appropriate to the question addressed. Sourcesare addressed in an appropriate level of detail.
3. Extensive reading, which goes beyond the routine sources and shows evidence ofindividual searching.
4. Excellent understanding of theories and issues for teaching, learning, and/or theory (asappropriate), with no key issues or arguments being omitted.
5. Able to grapple thoughtfully and insightfully with the implications of theories and issues,including in relation to teaching and learning (where appropriate).
6. Literature and evidence are subjected to persuasive critical evaluation at both micro andmacro levels of discussion.
7. Evidence of penetrating and/or original insight and analysis.
8. Strong and clear lines of argumentation; tight and logical structure.
9. Writes clearly, logically and concisely, with flair.
10. Excellent presentational standards.
11. Excellent adherence to academic conventions.
12. Study design (where chosen) is extremely well presented and shows excellentunderstanding of research methods.

65-69 (Merit Level)

Assignments will generally have the following characteristics:

1. Addresses the question clearly and thoroughly
2. Good balance of breadth and depth. Sources, including empirical studies, are addressedin appropriate detail.
3. Extensive reading, with some evidence of individual searching.
4. Sound understanding theories and issues, with most of the main issues and argumentsbeing covered.

5. Thoughtful consideration of implications of theories and issues for teaching, learning,and/or theory (as appropriate).
6. Some strong and convincing critical evaluation of literature and evidence
7. Evidence of individuality of insight
8. Clear lines of argument throughout. Clear structure.
9. Clear, logical and concise writing.
10. Very high presentational standards
11. Good adherence to academic conventions
12. Study design (where chosen) is well presented and shows very good understanding ofresearch methods well beyond the basic level. There are no major design flaws.

50-64 (Pass level)

At the upper end of this range assignments will generally have the following characteristics:

1. Addresses the question reasonably clearly and thoroughly.
2. Reasonable balance of breadth and depth. Sources, including empirical studies, arediscussed in sufficient detail.
3. Extensive reading, but sticks mostly to reading lists provided.
4. Clear evidence of understanding theories and issues.
5. Draws appropriate implications for teaching, learning, and/or theory (where appropriate).
6. Some evidence of critical reflection, though not always sufficient or sufficientlypersuasive.
7. Beginning to show evidence of individuality of insight.
8. Generally well structured. Generally logical lines of argument throughout.
9. Generally good clear writing.
10. High presentational standards
11. Generally adheres to academic conventions.
12. Study design (where chosen) is clear and shows good understanding of research methods and their implications, though there may be some minor design flaws.

At the lower end of this range 50-59 assignments will generally have the followingcharacteristics:

1. Reader sometimes has to infer the way in which the candidate has addressed the question.
2. Balance of breadth and depth is not optimal: there may be too much breadth and not enough depth, or vice versa. Sources are not always adequately explained.
3. Evidence of having read a reasonable amount of literature, including some key papers.There is too much reliance on secondary sources.
4. There is evidence of understanding theories and issues, but there may be some gaps inunderstanding and/or omissions in the way they are explained. There are some unsubstantiated assertions.
5. There is some attempt to identify implications for teaching, learning, and/or theory (asappropriate), though these are not always convincing or adequately founded based onthe evidence that has been presented.
6. There are attempts at critical reflection, but sometimes weak, tokenistic or unconvincing.

7. The lines of argument tend to follow well-worn paths; there is little or no evidence ofindividuality of insight.
8. Reasonably well structured with some attempt at pulling themes together. Lines ofargument may not always flow logically.
9. Writing quality is acceptable but may not always be clear, and more proofreading wouldbe warranted.
10. Reasonable presentational standards.
11. Some gaps in adherence to academic conventions.
12. Study design (where chosen) is sufficiently clear and there are few serious design flaws.

49 or lower (at fail level)

Assignments will generally have the following characteristics:

1. Difficult to determine how the question is addressed.
2. Shows little evidence of either breadth or depth. Issues/studies are not discussed in sufficient detail. Rarely sufficient discussion of the issues.
3. Little evidence of reading, and far too many secondary sources.
4. Inadequate understanding of theories and issues. Too many expressions lifted from the literature without demonstrating understanding.
5. No, or unconvincing, exploration of implications for teaching, learning, and/or theory (asappropriate).
6. Superficial reporting of evidence, no critical discussion.
7. No evidence of insight.
8. Little or no cohesion between ideas.
9. Poor writing quality; sentences are sometimes tortuous and incomprehensible.
10. Poor presentational standards; little or no evidence of proofreading.
11. Poor adherence to academic conventions.
12. Study design (where chosen) is unclear and/or seriously flawed, showing lack ofunderstanding of research methods, even at a basic level.

Appendix G; CUREC form

SECTION A: Filter for CUREC 2 application		
This section determines whether the application for ethics review should be made using this form (CUREC 1A) or the CUREC 2 form (for research with more complex ethical issues).		
Please indicate with an 'X'.	**Yes**	**No**
Does the research involve the deception of participants?	☐	☒
Are the research participants vulnerable in the context of the research, or classed as people whose ability to give free and informed consent is in question? For example, Participants aged 16 or under (also answer question A5); Participants aged 16 – 18 (refer to competent youths for guidance); adults at risk; Note the University's Safeguarding Guidance and Code of Practice and its implications for researchers involving young people or adults at risk.	☐	☒
By taking part in the research, will participants be at risk of criminal prosecution or significant harm?	☐	☒
Does your research raise issues relevant to the Counter-Terrorism and Security Act (the Prevent Duty), which seeks to prevent people from being drawn into terrorism? Best Practice Guidance 07 on the Prevent Duty provides further guidance.	☐	☒
If you answered **'No'** to all the questions above, go to Section B. If you answered **'Yes'** to any question above, continue to question 5 below.		
Is your project covered by a CUREC Approved Procedure?	☐	☐
If yes, list the CUREC Approved Procedure(s) you will follow		
If you have answered '**No**' to all questions 1-4, go on to **Section B**. If you answered '**Yes**' to ANY of questions 1-4, and answered '**No**' to question 5, **stop** completing this form and do not submit it for ethical review. You will instead need to submit a CUREC 2 application form. If you answered '**Yes**' to **any** of questions 1-4, and your project is covered by an Approved Procedure, **go on to Section B**. If more than one Approved Procedure applies, contact the SSH IDREC or your DREC for advice on whether a CUREC 2 form should be submitted instead.		

SECTION B: Researchers	
Name of Principal Investigator or student's supervisor	Hamish Chalmers
Department or Institute	Education
University of Oxford email address	hamish.chalmers@education.ox.ac.uk
Copy and paste the following six rows as necessary to complete for each additional researcher who will be involved in this study, including student(s) and those external to the University.	
Name of researcher or student	Noorin Iqbal
Department or Institute	Education

University of Oxford email address	noorin.iqbal@education.ox.ac.uk
Role in research	Student
Degree programme, if student research	MSc Applied Linguistics and Second Language Acquisition
The whole research team	
Have the researchers undertaken research ethics and integrity training?	**Yes** ☒ **No** ☐
Please provide details of any research ethics and integrity training undertaken, including the dates of the training. Alternatively state relevant research experience.	'Research Integrity Core Course', University of Oxford, completed January 2023. 'Foundations of Educational Research 1' module, completed Michaelmas 2022. 'Foundations of Educational Research 2' module, Hilary 2023 (ongoing). 'Research Methods' module, Michaelmas 2022 and Hilary 2023. The PI has extensive experience of conducting research with human participants and was a member of the DREC for three years.
State any conflicts of interest and explain how these will be addressed.	None.

SECTION C: The research project	
Title of the research project	
A comparison of the linguistic encoding of artificial-intelligence-generated academic essays and academic essays written by MSc ALSLA 2022-23 students.	
Anticipated start date of the aspect of the research project involving human participants and/ or personal data (**dd/mm/yy**).	Start when ethics approval is obtained.
Anticipated research end date (**dd/mm/yy**).	10/08/23
Provide a brief lay summary of the aims and objectives of the research. This should cover the questions it will answer and any potential benefits. (max 300 words)	
This research project aims to examine and compare the linguistic features present in essays written by master's students on a specific linguistics topic and essays generated by artificial intelligence software on the same topic. These linguistic features include, but are not limited to, vocabulary tokens, n-grams (sequences) of words as phrases, grammatical errors and typos, transition sentences, and paragraph-specific phrases (e.g., "in conclusion" as a frequent opening sentence for conclusions). I aim to create a rubric that outlines these features with respect to both types of essay, in order to view the similarities and differences between the two text types (if any). The primary potential benefit of this research project is that it will provide a concise list and explanation of the linguistic features present in AI-generated text. As AI-generated text is becoming increasingly popular with students having free access to software such as ChatGPT, its misuse amongst students for their assignments is also becoming a potential threat to academic integrity and honesty. Thus, it is important to	

understand how humans write as opposed to how AI "writes", for academic staff to distinguish between the two, especially if they are unfamiliar with their students' writing styles.	
Please indicate the methods to be used (indicate with an '**X**'):	
Analysis of existing records	☐
Snowball sampling (recruiting through contacts of existing participants)	☐
Use of casual or local workers e.g. interpreters (refer to guidance in BPG 01: Researcher safety)	☐
Participant observation	☐
Covert observation	☐
Observation of specific organisational practices	☐
Participant completes questionnaire in hard copy	☐
Participant completes online questionnaire or other online task (refer to guidance in BPG 06: Internet-mediated research)	☐
Using social media to recruit or interact with participants (refer to guidance in BPG 06: Internet-mediated research)	☐
Participant performs paper and pencil task	☐
Participant performs verbal or aural task (e.g. for linguistic study)	☐
Focus group	☐
Interview (refer to guidance in BPG 10: Conducting research interviews)	☐
Audio recording of participant (you will generally need specific consent from participants for this)	☐
Video recording of participant (you will generally need specific consent from participants for this)	☐
Photography of participant (you will generally need specific consent from participants for this)	☐
Others (please specify below) Collecting formative essays written by the MSc ALSLA 2022-23 cohort.	☒
Provide a brief summary of the research design and methods. What will research participants be asked to do? (max 300 words) Please also submit a copy of the questions participants will be asked, if applicable, or some information about the sorts of topics that will be covered.	

For the data collection stage, the MSc ALSLA 2022-23 cohort will be asked to submit their formative assignment from Michaelmas 2022 (if they consent to participate in this research project), prior to the feedback they received from their supervisors. This was the first assignment the ALSLA students had to submit after their enrolment into the course. It was a formative assignment, and therefore did not form part of their formally assessed work, and was not credit bearing. The participants will be asked to copy and paste their essays into a textbox in an anonymous form link sent to their email, without any information that can lead to potential identification (e.g., no names, candidate numbers, names of supervisors etc.). Once this has been done, I will copy and paste each essay into a Word Document and upload it into Nexus 365 OneDrive for data protection. I will then ask AI software (such as Moonbeam) to generate the same number of essays as I got from the cohort. For the data analysis stage, I will extract the top 10-20 most frequent vocabulary tokens from both samples via AntConc. I will also extract the most frequent 3, 4, and 5 n-gram sequences of words from both samples via AntConc. I will then analyze the vocabulary used in both types of essays (form, meaning, register, tense, presence of collocations etc.). I will extract transition sentences and opening sentences of different paragraphs (such as conclusions) from both text types. I will look for the presence of any grammatical errors and typos. Based on these features, I will create a matrix of all features with binary labels (0 if a feature is absent and 1 if a feature is present). With respect to statistical analysis, I will use the Jaccard's coefficient to see how similar or dissimilar the texts are. Lastly, I will present the findings, and discuss pedagogical implications and suggestions for further research.			
List the location(s) where the research will be conducted, including any other countries.	Oxford, United Kingdom.		
Clarify which parts of the research will be conducted in-person and which will take place remotely, e.g. online.	Collecting the students' essays will be conducted remotely, through an anonymous Microsoft Forms link. Asking AI software to generate essays will also be done remotely.		
If your research involves fieldwork or travel and your department requires a travel risk assessment, will you have completed and returned a risk assessment form beforehand? Please indicate with an '**X**'. (This must be approved by your department before you travel. If you are travelling overseas, you are advised to take out University travel insurance.) Refer to guidance available from your Department, the Safety Office, the Social Sciences Division, and the Humanities Division, and on travel for University business.	Yes	☐	
	No	☐	
	Not required in this instance	☒	
In the case of international or collaborative research, explain how you will address any ethical issues specific to the local context. Please provide details of the local review, approval or permission obtained or required. Refer to the BPG 16: Social science research conducted outside the UK and the Code of Conduct for Ethical Fieldwork. If there will be no local review, explain why not. Please mention any stakeholder or community engagement that has been/ will be undertaken in relation to the research. Please also address any physical or psychological risks for Oxford researchers and local fieldworkers in Section G.			
Not applicable.			

Name of departmental/ peer reviewer (if applicable)	Not applicable.
External organisation funding the research and grant reference (if applicable)	Not applicable.
Please refer to the CUREC Best Practice Guidance and list any that have been used to develop your research.	BPG 09 Data collection, protection and management.

SECTION D: Recruitment of research participants		
Number of participants	23	
How was the number of participants decided?	I aim to examine the linguistic encoding of the formative assignments submitted by the ALSLA cohort. Since I aim to anonymize the data, I will not be using my own essay. Hence, out of the 24 students in the cohort, there will be potentially 23 participants.	
Age range of participants	20 and over.	
Inclusion criteria	MSc ALSLA 2022-23 students	
Exclusion criteria	I cannot include my own work in the data, as I aim to analyze the documents without knowing who has written it for the purpose of this project and comfort of my colleagues.	
Indicate with an '**X**' all intended recruitment methods **Please submit copies of the recruitment material that will be used, e.g. advertisement text, introductory email text.**	Poster advert	☐
	Flyer	☐
	Email circulation	☒
	Social media (e.g. Twitter, Facebook)	☐
	Website	☐
	In-person approach	☒
	Snowball sampling	☐
	Recruitment sites (e.g. Mechanical Turk)	☐
	Existing contacts or volunteer database	☐
	Other (please specify):	☐
How will potential participants be identified and approached?	The participants' emails will be taken from the email directory at the University of Oxford. They will then be emailed with the details of the project, consent form, and an anonymized link to submit their essays.	
Will informed consent be obtained from the research participants or their parents/ guardians? If not, please explain why not.	No, as all participants are adults over the age of 18, their parental/guardian consent is not	

	required, as they are able to give informed consent themselves.
For each activity or group of participants, explain how informed consent will be obtained from the participants themselves and/ or their parents/ guardians, if applicable. How will their consent be recorded? **Please submit copies of all participant-facing materials for review.** E.g.: Recruitment material (e.g. emails, posters) Information for participants to read (or hear) before they agree to take part (e.g. written information or, if applicable, an outline oral information script). A document to record informed consent. Further guidance and templates.	I first intend to verbally explain my project to the MSc ALSLA cohort in class, listing the points of consent. Then, I will attach the relevant consent points to the email and form I will be sending out as part of the document submission. I will give my cohort the opportunity to ask me any concerns/questions at the end of my talk and via email. Their consent will be recorded through the Microsoft Form link where they will be asked to read through the consent and click a button if they agree, whilst submitting their assignment.
Provide details of any payments and incentives and the rationale for providing these. Further guidance in Best Practice Guidance: 05 Payments and incentives in research.	None.
Describe how participants may withdraw from the study may withdraw any personal information they have provided from the study State any limits to withdrawal, for example once the data has been anonymised or at some other specified stage prior to publication. Make sure participants are aware of any withdrawal limits.	The participants may withdraw from the study at any point of this research project, even after they have sent in their documents. However, to do so, they would have to reach out to me and tell me which essay was theirs, since I will be anonymizing all of them (e.g., by telling me information about the content). Here, there can be a drawback, as I will then know which essay was their work. Their document will then be removed from my sample.

SECTION E: Research data
All information provided by participants is considered research data for the purpose of this form. Any research data from which participants can be identified is known as personal data; any personal data which is sensitive is considered special category data. Management of personal data, either directly or via a third party, must comply with the requirements of the UK General Data Protection Regulation (UK GDPR) and the Data Protection Act 2018, as set out in the University's Guidance on Data Protection and Research. In answering the questions below, please also consider the points raised in the Data Protection Checklist and Data Protection Screening Assessment and whether, for higher-risk data processing, a separate Data Protection Impact Assessment may also be required for the research. Advice on research data management and security is available from Research Data Oxford and your local IT department. Advice on data protection is available from the Information Compliance team.

For guidance on conducting internet-mediated research, refer to CUREC's Best Practice Guidance 06: Internet-mediated research.			
What data will be collected? (Indicate with an '**X**')			
Screening documents	☐	Task results (e.g. questionnaires, diaries)	☐
Consent records (e.g., written consent forms, audio-recorded consent, assent forms)	☒	IP addresses (refer to Best Practice Guidance 09: Data collection, protection and management for guidance)	☐
Contact details for the purpose of this research only	☐	Field notes	☐
Contact details for future use (guidance)	☐	Photographs	☐
Opt-out forms	☐	Information about the health of the participant (including mental health)	☐
Audio recordings	☐	Previously collected (secondary) data	☐
Video recordings	☐	Data already in the public domain. Specify the source of the data:	☐
Transcript of audio/ video recordings	☐	Other, please specify: Formative assignments written by MSc ALSLA in Michaelmas 2022 (anonymized).	☒

During the course of the research, where will **each type of** research data be stored?	After copying and pasting the essays from the submission forms into a Word Document, I will upload the file to Nexus 365 OneDrive with my University of Oxford email address, in order to ensure data protection. The file will then be deleted from my laptop hard disk drive. With respect to the AI-generated essays, I will also upload the document to Nexus 365 OneDrive with my University of Oxford email address, and delete it off my laptop's hard disk drive.		
Who will have access to the research data during the project?	The data will only be available to the student researcher during the project and the supervisor. Responsible members of the university with responsibility for data control may be given access to the data on request. The individuals will not be identifiable from the data as the form will be completely anonymous and they will be asked to remove any personal information prior to submission.		
Please complete this section if your research involves the	**Please indicated with an 'X'.**	**Yes**	**No**
	Are data access agreements in place for access to and use of this secondary data? (If so, please attach these.)	☐	☐

use of secondary (i.e. previously collected) data.	Did the individuals agree that their data could be used for this purpose?	☐	☐
	Could anyone (including members of the research team) link the data back to an individual or individuals? If this is a possibility, please explain how the associated ethical issues will be addressed:	☐	☐
How do you intend to share the research data at the end of the project?	Depositing in a specialist data centre or archive	☒	
	Submitting to a journal to support a publication	☒	
	Depositing in an institutional repository	☒	
	Dissemination via a project or institutional website	☒	
	No plans to share the data	☐	
	Other (please specify):	☐	
How do you intend to report and disseminate the results of the research? (Indicate with an 'X')	Thesis publication	☒	
	Publication in a peer reviewed journal	☒	
	Publicly available report	☒	
	Conference presentation	☒	
	Publication on a website	☒	
	Pre-registration	☒	
	Report to a research funder	☒	
	Providing participants with a lay summary of the results	☒	
	Submission for academic assessment	☒	
	Other (please specify):	☐	

Explain what will happen to the data at the end of the research project. This question must be answered for each type of data, including completed consent forms.
The raw data will be destroyed in accordance with the UK General Data Protection Regulation and the Data Protection Act. The documents will be permanently deleted from the OneDrive. The processed data and consent forms will be retained on OneDrive.

SECTION F: Protection of research participants and their personal data		
How identifiable will the participants be from the **research**	Directly identifiable from the information included	☐
	Pseudonymised/ indirectly identifiable	☐
	Not identifiable – data is anonymous	☒

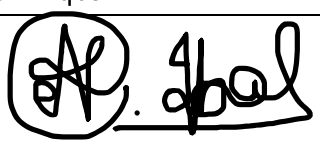
outputs? (Indicate with an 'X')	Other, please specify:	☐
To what extent will the **data** be de-identified? How identifiable will any individuals be from the research data? Describe any measures you will take towards assuring confidentiality, potential risks to confidentiality.	Since the nature of the formative assignments did not require the MSc ALSLA cohort to include any personal details about them, no aspect of the collected data can be identifiable once asked to remove information on the cover page. I will be copying and pasting the text from the form to a document on my own laptop, hence, there will be no meta-data to de-identify through Word document features. Personal names and candidate numbers will be strictly prohibited to mention anywhere on the form by the participants upon submission, assuring confidentiality, and any potential risk to confidentiality.	
How will you ensure that third parties (e.g., interpreters and transcribers) are aware of and adhere to the measures described in this form?	Not applicable.	

SECTION G: Risks and benefits of the research
Will the research involve topics that could be considered sensitive? If so: Please provide more detail or supporting information (such as the interview questions) to show the range of questions; Explain what steps will be taken to reduce risk of distress; Consider seeking advice from within your Department or from the ethics committee including whether the application might benefit from additional ethics review (e.g., via a CUREC 2 application).
No potential sensitive topics.
Describe any additional burden or risks to the participants or others, including the potential for any indirect negative consequences. Explain the steps you will take to address these.
No potential additional burden or risks to the participants or others.
Describe any physical or psychological risks to the researcher(s) (including local fieldworkers or research assistants) and the steps you will take to address these.
Not applicable.

Describe any benefits of the research, both to participants and to others. Outline the processes put in place to enable equitable research (see BPG 16 Social science research conducted outside the UK for further guidance).
The participants will benefit in terms of understanding how they write and how AI writes. Though their work will be anonymized, the research findings will present to them features of their writing, in comparison with AI writing. Additionally, professors can benefit from this research if they are made aware of how AI writes, in order to help them in case there are suspicions of ghost AI writing.
Comment on the societal impact.
As AI-generated essays are becoming more prominent, with students having free access to such resources, there is a rising threat to academic honesty and integrity on both school and higher education levels. With this research project, I aim to examine the linguistic encoding of both types of essays (i.e. written by students and by AI), which can lead to a better understanding of how these types of essays are similar and distinct. Thus, I can potentially propose solutions (such as a rubric of features) that can help teachers and professors detect or question when such academic violations are being committed.
Give details of any other ethical issues or relevant information.
Not applicable.

SECTION H: Professional guidelines		
Please indicate with an '**X**' at least one set of professional guidelines you will follow.		
Research specialism/ methodology	**Association and guidance**	
Anthropology	Association of Social Anthropologists of the UK	☐
Computer Science	ACM Code of Ethics and Professional Conduct	☐
Criminology	British Society of Criminology Statement of Ethics	☐
Education	British Educational Research Association Ethical Guidelines for Educational Research	☒
Geography	American Association of Geographers Statement on Professional Ethics	☐
History	Oral History Society of the UK Ethical Guidelines	☐
Internet-mediated research	Association of Internet Researchers Ethical Guidelines British Psychological Society: Ethics Guidelines for internet-mediated research Association for Computing Machinery Code of Ethics and Professional Conduct	☐
Management	Academy of Management Code of Ethics	☐
Political Science	American Political Science Association (APSA) Guide to Professional Ethics in Political Science	☐

Politics	Political Studies Association. Guidelines for Good Professional Conduct	☐
Psychology	British Psychological Society Code of Ethics and Conduct	☐
Social research	Social Research Association: Ethical Guidelines	☐
Socio-legal studies	Socio-Legal Studies Association: Statement of Principles of Ethical Research Practice	☐
Sociology	The British Sociological Association: Statement of Ethical Practice	☐
Visual research	ESRC National Centre for Research Methods Review Paper: Visual Ethics: Ethical Issues in Visual Research	☐
Other professional guidelines		☐

SECTION I: Endorsements and signatures	
Please ensure this form is endorsed by the Principal Investigator (or student's supervisor), the Head of Department (or nominee) and, if student research, by the student themselves. **The SSH IDREC Secretariat accepts either option below. If you have a DREC, check which signature option it prefers.** **Option 1: direct email endorsements** Each of the signatories should submit an email from a University of Oxford email address, indicating their acceptance of the responsibilities listed below. **Option 2: signatures** Please scan the signed form and email it to us as a PDF. Pasted images of signatures cannot be accepted.	
Endorsement by the Principal Investigator/ student supervisor and student, if applicable	
I/ we the researchers understand my/ our responsibilities as Principal Investigator (and student, if applicable) as outlined in the guidance on the CUREC website. I/ we declare that the answers above accurately describe the research as presently designed, and that the ethics committee will be informed of any changes to the project which affect the answers to this form. I/ we will inform the relevant IDREC if the Principal Investigator changes.	
Name of Principal Investigator	
Principal Investigator's signature	
Date	
Name of student (if applicable)	Noorin Iqbal
Student's signature	
Date	06 February 2023

Departmental endorsement – from the Head of Department or nominee (Another senior member of the department may sign where the head of department is the Principal Investigator, or where the Head of Department has appointed a nominee. Example nominees include Deputy Head of Department, Director of Research, or Director of Graduate/ Undergraduate Studies.)	
On the basis of the information available to me, I confirm that: I am aware of the research proposed and have read this application; To the best of my knowledge, the proposed design and scientific methodology do not raise ethical concerns; I support this research in principle, subject to ethical and other necessary reviews.	
Signature	
Name	
Role	
Date	

Consent Form via Email

Dear MSc ALSLA 2022-23 students,

My dissertation focuses on comparing the linguistic encoding of students' essays and AI-generated essays. For my study, I have chosen to analyze the MSc ALSLA cohort's formative assignments prior to submission. This is due to this assignment being the first one submitted the cohort had to submit to the University of Oxford, and reflects the cohort's writing style prior to receiving any feedback from the supervisors.

Please read the statements below. If you are happy with all of the statements, please press the button on the form below confirming you agree to participate in this study. This will be considered to constitute giving your consent to participate in the study. Next, please copy and paste your essays (with citations, endnotes, and bibliography) into the textbox below. Please remember to **REMOVE** the cover page of your assignment and any information that can be used to identify you, such as your name, candidate number, submission date, and name of supervisor. If you have used any anecdotal information with regards to your personal life or experiences in the essay, please remove it as well.

The essays will be stored on Nexus 365 OneDrive for data protection. They will not be shared with anyone except the researcher during the entirety of the research project. When using the data in the analysis, no identifiable information will be mentioned.

To withdraw from this study, please email me at noorin.iqbal@education.ox.ac.uk. You will be asked to identify your work, in order for me to remove your data.

Please submit your work by 17/03/23. After this date, the form will no longer accept submissions.

If you have any questions about the research or the statements below, please do not hesitate to contact me at noorin.iqbal@education.ox.ac.uk or in-person.

I confirm that I have read and understand the information sheet for the study, "A comparison of the linguistic encoding of artificial-intelligence-generated academic essays and academic essays written by MSc ALSLA 2022-23 students" for Noorin Iqbal's MSc ALSLA dissertation. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.

I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason, and without any adverse consequences or penalty.

I understand what will happen to my data.

I give the researcher(s) permission to use my formative assignment anonymously.

I give permission for the researcher(s) to quote me directly using a participant number (e.g., 1 participant wrote, "[x]").

I give permission for the researcher(s) to re-contact me to clarify information.

I agree for anonymized research data collected in this study to be used in other research studies.

I am happy to take part in the research.

[] I have read and understood the consent form above and I agree to share my formative assignment with Noorin Iqbal for her dissertation.

Thank you very much for your time and interest to participate in this study.

Best regards,

Noorin Iqbal

Appendix H; CUREC Form Approval

UNIVERSITY OF OXFORD DEPARTMENT OF EDUCATION 15 Norham Gardens, Oxford OX2 6PY

student.curec@education.ox.ac.uk; staff.curec@education.ox.ac.uk **DREC Chair: Professor Liam Gearon**

CONFIDENTIAL

Noorin Iqbal noorin.iqbal@education.ox.ac.uk 6 March 2023

Dear Noorin,

Research Ethics Approval – CUREC 1a Ethics Approval Reference: EDUC_C1A_23_073

Study Title: A comparison of the linguistic encoding of artificial-intelligence-generated academic essays and academic essays written by MSc ALSLA 2022-23 students

The above application has been considered on behalf of the Departmental Research Ethics Committee (DREC) in accordance with the procedures laid down by the University for Ethical Approval of all research involving human participants.

I am pleased to inform you that, on the basis of the information provided to the DREC, the proposed research has been judged as meeting appropriate ethical standards, and accordingly, approval has been granted.

Please continue to follow *COVID-19: CUREC guidance on research involving human participants* https://researchsupport.admin.ox.ac.uk/governance/ethics/coronavirus

If relevant, please also check the CUREC website for their **best practice guidance**: https://researchsupport.admin.ox.ac.uk/governance/ethics/resources/bpg

Amendments

Should there be any subsequent changes to the study, you should submit details to the DREC for consideration and approval. Details of changes must be listed on an amendment form.

Please do not hesitate to contact me if you have any queries. 1

Good luck with your research study.

Yours sincerely, Aliya Khalid

Lecturer in Comparative and International Education Member of DREC
Department of Education
University of Oxford

