Verb + noun collocations in L1 and L2 English spoken language examinations:

Introducing the Trinity Lancaster Corpus of L1 Spoken English to investigate formulaic language

Lorrae Fox

This thesis is submitted for the degree of

Doctor of Philosophy

Lancaster University

Department of Linguistics and English Language

January 2024

Abstract

This thesis investigates spoken verb + noun collocations in L1 and L2 English candidates undertaking a language exam. It further introduces the new Trinity Lancaster Corpus of L1 Spoken English (TLC-L1), a corpus developed to align in context to the TLC-L2 (Gablasova et al., 2019). By considering differing proficiency levels of L2 speakers, the research takes a pseudolongitudinal approach to investigate collocational development with core findings including the influence of topic and register on the use of collocations in a language exam, which are also reflected in the TLC-L1 corpus, as well as evidence of a nonlinear developmental trajectory of English language learners and their use of verb + noun collocations. Finally, the thesis brings three major contributions and implications to the field: (1) methodological with the development and application of a new corpus, (2) theoretical through the analysis of collocations in the under-investigated mode of speech and (3) pedagogical with suggestions and examples of corpus-informed language teaching materials.

Table of Contents

List of Tables

List of Figures

Acknowledgements

Firstly, I would like to thank Professor Vaclav Brezina and Dr. Dana Gablasova, my supervisors, for their guidance during my PhD. I would not be the researcher I am today without their expertise and enthusiasm, for which I am very grateful. I would also like to thank my panel members during my time at Lancaster for their helpful discussions during key stages in my research progression.

This thesis would not have been possible without the NWSSDTP CASE studentship I was awarded; I am incredibly grateful to the funding bodies involved in this project for giving me the opportunity to pursue my PhD research. Many thanks also go to every participant and Trinity College London examiner who contributed their time and voice to the TLC-L1, and to Ruth for transcribing each word.

I would like to give appreciation for all my friends and colleagues, old and new, that have celebrated my achievements with me and offered kindness during the challenges. Special thanks to Linda and Mel; your good cop/bad cop dynamic was so needed way back in that soulless Starbucks and every pub since. Mary-Ann, Joanne and Claire, thank you for your belief in me while also showing me how to lead with empathy. Will and Luke, thank you for your illuminating words of wisdom, encouragement, and emoji-titled playlists. Always. I wouldn't have survived without you both. And Katie, I cannot express how incredibly grateful I am for you. Thank you so much for your unwavering support; I promise I will come visit you more now.

Finally, I would like to express my heartfelt gratitude to my parents, Rachel and Dave, who this thesis is dedicated to. Thank you, both; Mum for always being on the end of the phone and Dad for the reminders to work smoothly.

Declaration

I declare that this thesis has been composed solely by myself and represents the result of my own original research. It has not been submitted, in whole or in part, for the award of a higher degree in this University or elsewhere.

<signature>

Lorrae Fox

Chapter 1: Introduction

1.1. Overview

Language is inherently formulaic as it is acquired, produced and processed in word combinations (Wray, 1999). Spoken language relies heavily on this phrasal usage for ease of comprehension and production due to the transient nature of the mode (Biber et al., 1999). These multi-word sequences or expressions can be further broken down into differing types of formulaic language, including collocations; these are word combinations that have a higher-than-chance co-occurrence (Brezina, 2018). Mastery of collocations, alongside other phrasal chunks, contributes to a speaker's phraseological competence, which is a core aspect of overall fluency in language (Wray, 2002; Paquot, 2018) and perceived spoken fluency (Cobb, 2003).

This research focuses on the use of collocations within a spoken learner corpus and a native speaker corpus, aiming to describe the nature of one type of collocation, verb + noun, within a language testing context of these two groups of speakers. The thesis also serves to introduce a new corpus of language testing data: the Trinity Lancaster Corpus of L1 Spoken English (TLC-L1).

This introductory chapter provides a rationale for the study (Section 1.2) and introduces the thesis's aims and scope (Section 1.3). It further establishes key concepts that will be explored in depth within the Literature Review in Chapter 2 (Section 1.4) before mapping an outline of the thesis in Section 1.5.

1.2. Rationale

Formulaic language is essential for communicative fluency and therefore important for language learners to develop. A general motivation for this study is to investigate formulaic language in L1 and L2 spoken language to advance our knowledge of this core feature of communication, specifically focusing on how collocations are used.

Collocations are one type of formulaic language that research suggests language learners find particularly challenging (e.g. Altenburg & Granger, 2001; Laufer & Waldman, 2011, see Section 2.4 for further discussion) especially when compared to native speaker use of these phraseological chunks (e.g. Vedder & Benigno, 2016, see Section 2.4.1.2 for further discussion). Furthermore, there are mixed results from investigations into how proficiency impacts the use of collocations in language learners and how this is related to their overall collocation development (e.g. Thewissen, 2015 and Vedder & Benigno, 2016

who found conflicting results regarding the connection between collocation knowledge and language proficiency by looking at learner errors, see Section 2.4.1.1 for further discussion). Alongside these points, a general preference for using written language in corpus research, in part due to availability of data, has further motivated this study to investigate spoken formulaic language to further extend our understanding of collocation use.

As well as these theoretical motivations, the thesis also has methodological motivations with the introduction of the TLC-L1, which was in part developed in this project, as well as further study of the Trinity Lancaster Corpus (TLC-L2). This latter dataset is the largest corpus of spoken learner English currently available, with 4.2 million words of interactive language (Gablasova et al., 2019). The creation of this corpus was a much-needed addition to the field of learner language research due to the necessity for more spoken data to be available for study. With this addition came a demand to create a new corpus primarily as a comparison for the TLC-L2 but also to use as its own dataset to investigate how L1 speakers engage in an interactive, examination context. This is because it can be valuable to compare learner language to a large informal corpus such as the British National Corpus 2014 (Brezina et al., 2021; BNC2014); however, there is also an understanding that informal conversations are a very different context to formal language examinations, which will likely impact the language used. To mitigate this impact, using corpora within the same context is meaningful for research. The TLC-L2 and the new TLC-L1 both use the Trinity College London Graded Examination in Spoken English (GESE) as the examination context. The TLC-L1 corpus is unique as there are no major corpora of L1 speakers undertaking this specific examination or any other English language examination. As previously mentioned, there is also a lack of spoken language research using corpora in general due to the logistical and financial challenges involved in creating spoken language corpora. Therefore, this was a further rationale for the thesis study to contribute to the currently limited conversation on collocation use in spoken language in native speakers and English language learners with supporting the development of a new corpus.

The elements set out in this rationale uncover a gap in the current conversation which this thesis aims to fill.

1.3. Aims and scope of the thesis

Overall, this thesis aims to contribute to theoretical and methodological discussions and presents pedagogical suggestions. Theoretically, the study adds to the current conversation on formulaic language use by providing an up-to-date overview of phraseology and collocation research. More specifically, the thesis focuses on learner corpus research findings and considers these previous results to situate this study and guide methodological decisions. Furthermore, the thesis works to fill the gap in current research investigating spoken language as it empirically investigates the use of verb + noun collocations in L1 and L2 speech, which is noted to be under-investigated in corpus research (see Section 2.4). This is despite the generalised acknowledgment of how crucial formulaic language is during speaking for perceived fluency and increased listener comprehension (Wray, 2002) and that this is likely due to faster retrieval of the semantic units (Cobb, 2018).

The thesis takes a blended phraseological and frequency-based approach to the definition and extraction of collocations, an approach that is suggested by Granger (2018) which is comparatively rare in the field but gaining momentum (Gablasova et al. 2017; Lee, 2019). Furthermore, the thesis reflects on the advantages and limitations of the corpus approach in the study of phraseology in general, and more specifically for collocations. It also applies this reflection to corpus linguistics as a methodology to investigate the fields of language teaching and language testing and the thesis furthers the current discussion of the idea of the L1 norm in language learner research in Sections 2.4.1.2 and 7.2.1.2.

The thesis undertakes a descriptive investigation of collocations in learner language, contributing to the complexity of current research findings about formulaic language development based on proficiency (see Section 2.4.1.1) while also investigating collocations in a comparable L1 corpus. This study goes further in that this is the first instance of verb + noun collocations being studied within the TLC-L1 and the TLC-L2 corpora. The thesis findings also contribute to the understanding of the importance of context for the interpretation of the use of collocations, focusing on the influence of register and topic within language examinations.

Regarding methodological contributions, the study introduces the TLC-L1 dataset having been in part compiled during this PhD project. The thesis describes the corpus design, nature of interaction and data collection context as well as discussing some differences

between the L1 and L2 Trinity Lancaster Corpora. This adds to the discussion on corpus creation with exploration of core considerations during the process in order to help others with their decision making when compiling corpora (see Section 3.2). The thesis also aims to introduce some pedagogical implications, outlining potential pathways to pedagogical uses of the findings and giving an example of a corpus-informed teaching activity.

Regarding scope, the present study focuses on one type of collocation – verb + noun – to thoroughly explore this in the two corpora. For the TLC-L2, speaker proficiency has been chosen as the main variable, looking at three proficiency levels based on the Common European Framework of Reference (CEFR; Council of Europe, 2001) and these levels are B1, B2 and C1/C2 combined. L1 background is acknowledged to potentially factor into collocation production (see Section 2.4.1.3) with an in-depth investigation of this noted to be beyond the capacity of this project but an exciting avenue for further research. Finally, the study takes the opportunity to begin the descriptive account of the new TLC-L1 and find commonalities in collocation usage with the TLC-L2, rather than taking a strictly contrastive approach between the L1 and L2 datasets. This descriptive approach opens the discussion for further research with the novel dataset.

Overall, this project is significant as it brings new evidence to light about the use of verb + noun collocations in L1 and L2 spoken language based on the TLC-L1 and TLC-L2 leading to theoretical and methodological contributions and pedagogical implications.

1.4. Key concepts in the thesis

1.4.1. Phraseological and frequency-based approaches

The phraseological approach to investigating collocations involves a focus on the semantic relationship between the words in the construction. By assessing usage based on the lexical elements, semantic bonds and degree of fixedness within the collocation, this approach acknowledges the grammatical rules of language from a more subjective perspective, with Cowie (1998) being an early proponent of this approach to the study of collocation.

In contrast, the frequency-based approach focuses on word co-occurrence based on quantitative evidence from an objective perspective. This evidence is determined using simple measures such as frequency of co-occurrence and more complex statistical

measures such as association measures which assess the strength of the relationship between words based on different parameters of that association (Brezina, 2018).

This thesis will take a blended approach to the identification of collocations as this has been suggested to be effective for learner corpus research by Granger (2018) and undertaken by Gablasova et al. (2017) and Lee (2019). This approach has been decided because both the phraseological and frequency-based approaches have strengths and weaknesses which are discussed in further detail in Section 2.1.A combination of the two approaches means collocations can be extracted based on their phraseological structure, ensuring the results are linguistically similar in this way, which is essential for learner language research as different types of language present different challenges for L2 speakers (Granger, 2018). Then, investigating these linguistic items in terms of measuring the frequency of co-occurrence and dispersion helps operationalise collocation quantitatively which ensures potential for replicability and using these thresholds helps to minimise the fuzziness of the phraseological approach (Granger, 2018). Finally, further support from qualitative concordance analysis can consider the context of these collocations in use The research will also view the resulting collocations from a phraseological perspective considering fixedness and commutability where appropriate in the discussion. Further exploration of the benefits and limitations of these approaches is detailed in Section 2.1 of the literature review while a definition for collocation in this thesis is presented in Section 1.4.3.

1.4.2. Corpus-based approach

Corpus linguistics is a methodology that uses a large database of authentic language production to study linguistic patterns, and this is "a finite-sized body of machine-readable text, sampled to be maximally representative of the language variety under consideration" (McEnery & Wilson, 2001, p. 32). A corpus-based approach involves the study of language using corpus linguistics as a methodology; within this, there are further methods of analysis. This thesis will employ some of those core methods, such as concordancing, to analyse the language in use in the authentic setting. The analysis will take a mixed methods approach, first detailing quantitative findings that will then guide the research to where deeper qualitative analysis will be most meaningful based on the proposed research questions. Corpus linguistics is particularly relevant to the study of learner language due to its ease in application when studying specific linguistic features that are of interest for language learning and teaching research, such as syntagmatic

language features like lexical bundles and collocations (Granger, 2020). A recent scoping review from Tan and Azmi (2021) also notes the current prevalence of using corpus linguistics in language learning research. This evidence means a corpus-based approach is an appropriate and effective methodology for this thesis.

1.4.3. Definition of collocation for the thesis

It is important to clearly define terms within research, especially when there are differing approaches to a concept such as collocation, to ensure clarity in interpreting findings. As the thesis takes a blended approach to the operationalisation of collocation for this learner corpora research following recommendations from Granger (2018), for the purposes of this thesis, collocations are those verb + noun word combinations that adhere to the following parameters:

(1) Automatically extracted from the corpora using a restricted CQL query that has been tested for high precision and high recall.
(2) Manually checked to ensure there is a syntagmatic relationship.
(3) Set frequency and dispersion thresholds to establish the collocational status.

Further details regarding the query, procedure and other methodological decisions can be found in Section 3.3. and 3.4.

1.5. Map of the thesis

Chapter 1 has introduced a brief overview of this thesis, detailing the rationale behind the research, presenting aims, defining scope, and explaining key concepts, including phraseological, frequency- and corpus-based approaches taken for this work. Chapter 2 presents the literature review, beginning with how researchers have defined collocation according to differing perspectives before further discussing how collocations are beneficial for both understanding and producing language. This chapter also includes findings from corpus linguistics research into collocations, narrowing this focus to learner corpus research with critical reviews from the perspective of this research. It then highlights a summary of the current state of the art regarding what is known about the intersection of collocations, corpus linguistics and language learning. Chapter 2 also introduces the research questions to be interrogated by the thesis based on the extensive literature review. Chapter 3 presents the methodology section which describes the components of the two corpora used in the research and the data collection procedure for creating the TLC-L1 corpus. Corpus design and metadata breakdown are also presented

here as well as detailing the methodological decisions taken for the data analysis for each corpus. To begin to answer the research questions, Chapter 4 presents the results of the TLC-L2 analysis based on three areas: shared verb + noun collocations, frequent verb types within verb + noun collocations and patterns in high frequency delexical verb + noun collocations. It approaches this analysis considering the variable of language proficiency. Chapter 5 presents the results of the TLC-L1 analysis, following a similar pattern to Chapter 4 by looking at frequent verb + noun collocations as well as unique combinations, frequent verb types with verb + noun collocations and finally patterns in high frequency delexical verb + noun collocations. To further explore the research questions of the thesis, Chapter 6 brings the research together to critically consider the results from both analyses and link findings here to the previous literature, highlighting major themes and considerations for theoretical contributions to the field. Finally, Chapter 7 reviews the main findings to the research questions, explores methodological contributions and proposes pedagogical implications. The chapter finishes by acknowledging limitations and discussing how to use these to fuel further research. Finally, closing remarks brings the thesis to a conclusion.

Chapter 2: Literature Review

This chapter provides an overview of the literature offering a theoretical underpinning of the thesis as well as the basis for methodological decisions taken in the study. In Section 2.1 it first explains how collocations can be defined, exploring both the phraseological and frequency-based approaches, leading to the operationalisation of collocation for this research to combine the two. It also introduces research supporting how collocations are beneficial to the comprehension and production of language. Next, Section 2.3 introduces the discussion around corpus linguistics as an effective methodology for investigating collocations and outlines some core considerations that need to be addressed when creating corpora. In Section 2.4, the main section of the literature review, the discussion moves to how language learning research has thus far investigated the phenomenon of collocations focusing on four aspects: the speakers, the types of collocations, the context of using collocations and the applications of collocation research. Within this major section, 2.4.1 details how learners use formulaic language, including the comprehension and production of collocations based on speaker proficiency, exploring research comparing L1 and L2 use of formulaic language before outlining how L1 language and cultural background can also impact it. After this, Section 2.4.2 explores research into the

specific type of collocation under review in this thesis: verb + noun collocations, and gives justifications for this choice to exemplify evidence of phraseological competence in spoken English. Then, Section 2.4.3 considers two contextual variables: topic and register, while the last section considers how collocational research can have an applied impact within two fields: language teaching and language testing. Finally, a summary of the literature highlights the core areas this thesis aims to explore and expand on before the four research questions are stated in Section 2.6.

2.1. How do we define collocation?

Since the term collocation was first introduced by Firth (1957), there have been different approaches to defining it. Firth's now famous quote within formulaic language research, "you shall know a word by the company it keeps" (Firth, 1957, p. 179), still forms the basis of the majority of the definitions of collocations, identifying the two essential characteristics of collocations: i) the recurrence of word combinations and ii) the frequency of their co-occurrence. In addition to these fundamental properties of collocations, researchers have offered more refined definitions and operationalisations of the term, focusing on different key characteristics of the construct (Granger, 2018). As Gries (2013) noted, given the different approaches to defining what is a collocation, the construct is best seen as "a radial category whose different senses are related to each other and grouped around one or more somewhat central senses, but whose senses can also be related to each other only rather indirectly" (p. 138). Nevertheless, more than 60 years on from Firth's original quote, researchers are still finding the concept challenging to pin down, with Saito (2020) recently confirming that "a precise definition has been elusive" (p. 550).

The differing definitions primarily draw on distinct traditions to the study of collocations (Granger & Paquot, 2008): the phraseological and frequency-based approaches. The former approach specifies the co-occurrence of certain lexical elements, semantic bonds, and degree of noncompositionality of meaning as a basis for defining collocation. The approach is based on the Russian tradition of phraseology led by Mel'čuk and later adopted by Cowie (1998). Alternatively, the frequency-based approach, stemming from Firth (1957) and later Halliday (1966), operationalises collocation using statistical association measures that take into account the frequency of occurrence of the combinations, looking beyond the semantics of the words and drawing on textual structure and syntactic relationships (Gablasova et al., 2017). To further complexify this

challenge in defining collocation, specific combinations can be defined differently depending on the approach taken. For example, Webb et al. (2013, p. 110) present the case of *pull strings*. From a phraseological perspective, this would likely be classified as an idiom due to the two words' underlying meaning and semantic opacity when combined. In contrast, the frequency-based approach would consider it a collocation. Although defining collocation is a complex issue that has yet to be truly resolved, there are some similarities that both approaches favour regarding the definition. Firstly, they consider collocations to be word combinations appearing together more frequently than by chance. Further to this characteristic, collocations are usually described as consisting of two components in terms of the number of elements (or collocates) that make up the collocation. These can either be adjacent (for example, bigrams such as *according to*) but can also regularly co-occur within a certain distance from each other (e.g., *make [a] decision*). As Evert (2008) notes, despite a well-established intuition that words tend to co-occur, there is still some disagreement within linguistics as to what precisely constitutes a collocation, with each approach highlighting the importance of different properties. These properties of collocations that have been traditionally recognised in the literature (Evert, 2008; Gries, 2013) are: i) degree of fixedness, ii) semantic unity, iii) frequency of occurrence and iv) number of collocates. Boers and Webb (2018) further review defining collocations according to the phraseological vs frequency-based approaches which presents a thorough examination of the issues. For a more concise discussion, the following section outlines the collocational properties particularly highlighted in each approach and those especially important to this thesis.

2.1.1. The phraseological approach to defining collocation

The phraseological approach places focus on the semantic relationship between words. This relationship is typically based on a subjective judgement (usually by native speakers of the language) which considers the grammatical rules of the language. Two of the key properties of this approach are the degree of fixedness and semantic unity, which will be discussed in more detail below.

'Fixedness' or 'restrictedness' of word combinations refers to the lexical and syntactic flexibility in the combination of two words and the interchangeability of the elements with which each of the constituent words can combine. The degree of fixedness is a

critical concept in phraseological research on formulaic language, with word combinations evaluated from the perspective of a continuum that goes from free to fixed, and a certain degree of fixedness is required for each combination to be considered as 'formulaic'(Gyllstad & Wolter, 2015; Laufer & Waldman, 2011). Different approaches have been adopted to evaluate and describe the degree of fixedness of two words, highlighting different features of the relationship between two words. Opacity and commutability have been two primary criteria in determining the degree of fixedness of word combinations (Cowie, 1994; Howarth, 1998). Opacity refers to how literal the meaning of the word combination is, with combinations such as *take a train* and *under the table* being considered as more literal than combinations such as *take the fall* and *under the weather* which are considered more opaque. Commutability refers to whether the elements within the combination can be substituted, e.g., in verb + noun structures. For example, in the combination *take a train*, *train* could be replaced with another form of transport, or the verb could be changed to *catch, get, miss* etc. *Take the plunge,* instead, due to its opaque meaning of undertaking something, has a lower degree of commutability – substituting *plunge* would then change the meaning of the phrase. Drawing on work by Cowie (1994) and Howarth (1998), Nesselhauf (2005, p.30) used the criterion of commutability and identified the following types of combinations on the free to fixed continuum, referring to verb + noun combinations: 1. Verb combinable with (virtually) every noun (*want a pen*), 2. Verb combinable with a large group of nouns (*kill a man*), 3. Verb combinable with a small but well-delimitable semantic group of nouns (*drink water*), 4. Verb combinable with a sizable group of nouns, but there are exceptions (*commit a crime*), 5. Verb combinable with a small set of nouns (*shake one's head*).

However, there are also some issues with the applicability of such categorisations to word combinations. For example, Moon (1998) noted that fixedness is a key criterion for formulaic language but pointed out that this can be variable; therefore, a question arises: How fixed does a phrase need to be in order to be considered formulaic? To answer this, a continuum should be considered regarding the terminology rather than attempting to form discrete categories with set boundaries. The reasoning for this is linked to the degrees of fixedness, frequency, and non-compositionality of such 'FEIs' – Moon's grouped term for formulaic expressions and idioms. By creating a database of these FEIs, she found that fixedness was a key feature of these types of phrases but also noted their unstableness, finding that 40% varied lexically. She states that this results in "doubt on

the viability of the notion of the canonical form" (1998, p. 121), as the most restricted words also seem to be the most infrequent.

A second property that is especially relevant to the phraseological approach is that of semantic unity. This term describes the extent to which a phrase functions as one semantic unit with a specific global meaning that is different to the sum of its parts (Granger & Paquot, 2008). This property of word combinations is also described in terms of their (non)compositionality. One factor of semantic unity is that there are restrictions on what phrases are typically used due to semantic coherence; for example, 'pregnant' is frequently used to describe a particular state and often includes a female (Allerton, 1984). Beyond these semantic coherence restrictions, words that co-occur because they are favoured by speakers, i.e., they prefer to choose one option over another, can be deemed to be collocations because frequent usage plays a role. However, frequency alone is insufficient for a phrase to be considered a collocation. For example, phrases like 'much of the' may frequently occur in language. However, there is a lack of semantic unity as there is no specific global meaning of the phrase, and, therefore, it would not be considered a collocation.

Overall, when considering fixedness and semantic unity, this can lead to collocations being further classified according to these characteristics; Nesselhauf (2005, p.22-23) proposes that collocations can be classified in three ways: by their (1) syntactic characteristics which are "classified according to the word classes in which their elements appear" e.g., lexical collocations (such as verb + noun) and grammatical collocations (such as noun + preposition), (2) semantic characteristics (of the collocator – e.g., figurative (*deliver [a] speech*), delexical (*make + recommendations*) and technical (*try [a] case*) verbs and (3) commutability of elements *(foot [a] bill* where the meaning of *foot* here means it can only be combined with *bill*). These three characteristics are frequently discussed when research into collocations comes from a phraseological perspective and will be highlighted throughout this thesis, such as delexical verb + noun collocations that are highly frequent.

Although there are clear benefits to investigating collocational use using the phraseological approach, there are also some limitations that need to be considered when moving forward with research. Firstly, there is a need to make "binary choices" (Granger, 2018, p.231) between collocations and other combinations and this means there is no acknowledgment of the progressive nature of understanding and using formulaic

language. This then leads to general fuzziness of defining collocations within the phraseological approach (Granger, 2018) as well as an understanding there are also fuzzy borders with other fields of linguistic inquiry (Granger & Paquot, 2008, p. 28-29). This means that researchers investigating collocations need to be transparent with what their operationalisation of collocation is as "the delimitations between different types of word combinations are not necessarily identical" (Nesselhauf, 2005, p. 17). Some choose to adopt the frequency-based approach to ensure they have an objective way to define collocation based on frequency counts (Wolter & Gyllstad, 2013). However, when going beyond frequency counts in research, it can be beneficial to combine the approaches by considering the phraseological approach primarily and then incorporating the frequency-based approach using frequency as a defining criterion to create a mixed approach to the study of collocations (Nesselhauf, 2005).

2.1.2. The frequency-based approach to defining collocations

In contrast to the phraseological approach, the frequency-based approach focuses on the quantitative evidence of word co-occurrence. This means the judgement as to whether a phrase is deemed a collocation is more objective as the approach seeks to determine whether the co-occurrence is due to chance. This judgement can be measured based on simple measures such as frequency of co-occurrence but is often done utilising more mathematically complex association measures. Association measures are "statistical measures that calculate the strength of association between words based on different aspects of the co-occurrence relationship" (Brezina, 2018, p. 67). Several different association measures can produce varying lists of collocations from the same data as they each highlight distinct parts of the collocational relationship based on two main dimensions: frequency and exclusivity.

The first aspect of the collocational relationship highlighted in the frequency-based approach is frequency which "refers to the number of instances in which a node and collocate occur together in a corpus" (Brezina, 2018, p. 71). Node refers to the word under study while collocate is the word co-occurring with the node). The second aspect is exclusivity which "refers to a specific aspect of the collocation relationship where words occur only or predominantly in each other's company" (Brezina, 2018, p. 71). These two aspects can be applied in diverse ways in the frequency-based approach, which results in different combinations being foregrounded in the same corpus data.

According to Brezina (2018), deciding what constitutes a collocational relationship can be done in different ways. One such way is basing the judgement on the measure of frequency of co-occurrence. This is done by producing a rank-ordered list, based on frequency, of the node and their collocates. In turn, this means there is no baseline (reference point) to see if the collocation is occurring more often than by chance. This baseline is created by comparing the observed frequencies (the number of times the words occur in the corpus) with the expected frequencies (the number of times the words would be expected to appear in the corpus by chance).

Instead, another possibility for defining collocation, according to Brezina (2018), can be done involving applying more complex association measures. One such way is by using a random co-occurrence baseline (the 'shake the box' model), which compares two types of frequency to determine if the combination occurs more often than expected by chance alone. These frequencies are (1) the observed frequency of collocation and (2) the frequency of the node (the word of interest) and collocate outside of the collocation window. The strength of collocation is calculated by association measures, of which there are many that each highlight various aspects of the relationship between the node and collocate based on the two dimensions of frequency and exclusivity of the combination. Other collocation measures, such as Cohen's d, consider other dimensions of collocations, such as dispersion which refers to the distribution of the collocates in the corpus, and directionality (Delta P), which can indicate the direction of attraction between the collocate to the node.

However, there are also some limitations to the frequency-based approach. Using no baseline (reference point) and only ordering according to the frequency of co-occurrence of the two words in the combination means that function words will typically show as the most frequent collocates due to their frequency in language in general (Brezina, 2018). Certain association measures can assist with this issue by including other dimensions of the collocational relationship, such as exclusivity. However, although there are multiple association measures available, there is no one 'best' measure to use as it depends on the individual research questions (Gablasova et al., 2017). Instead, there are often trends to use some association measures over others, such as Mutual Information in L2 research (Gablasova et al., 2017), and the most used measure within the research area may not be the best fit for the research questions. Therefore, it is crucial to assess what is needed to be measured by defining what collocations are of interest.

2.2. How are collocations beneficial for understanding and producing language?

Pawley and Syder (1983) were among the first to consider how English speakers know to select the most idiomatic and nativelike constructions with limited processing capacity and attributed this to a knowledge of 'sentence stems' that are lexicalised, resulting in fluent language – a feature of overall phraseological competence. According to the two researchers, "fluent and idiomatic control of a language rests to a considerable extent on knowledge of a body of sentence stems which are institutionalised or lexicalised" (p. 191).

Generally, it has now been well-established that collocations are a core feature of language, with estimates of 50% (De Cock et al., 1998; Erman & Warren, 2000) to 70% (Hill, 2000) of written and spoken English categorised as some form of formulaic language. Furthermore, Biber et al. (1999) state how formulaic language is considered to be a core characteristic of spoken language with further suggestions that more instances are present in spoken than in written language (Brazil, 1995; Leech, 2000). Evidence also comes from early corpus studies such as Jackendoff (1997), who found formulaic language was used as often as individual words within a TV show (p. 156). Ellis et al. (2008) further explain that this prevalence of formulaic language within spoken utterances is because "speech is constructed in real-time and imposes greater working memory demands than writing" (p. 376), meaning that there is more reliance on preconstructed forms due to these demands. This idea is further supported by research from Tavakoli and Uchihara (2020), extending that multi-word sequences also "enhance fluency because they facilitate access and retrieval of lexical units and free up attentional resources that are needed to deal with other aspects of speech performance" (p.511). There are also differences in how spoken and written language is used; Shin and Nation (2008) found that when looking at the most frequent collocations in both spoken and written English only 15 collocations occur in both top 50s. The here-and-now nature of spoken language is reflected in items like *this morning*, and its personal and interactional nature is reflected in items like *thank you*. Therefore, formulaic language, such as collocations, is fundamental in spoken language for increased fluency and often differ in type from written collocations; consequently, learning context and mode specific collocations is something of particular value for language learners developing their speaking skills.

As well as increasing fluency, collocations are also advantageous for the processing of language as they tend to be understood more readily than other less formulaic

combinations of words (Conklin & Schmitt, 2008; Yamashita, 2018). Studies have shown this is because they are processed differently than single words with increased ease of retrieval. There are differing views of how this processing advantage occurs. Myles and Cordier (2017) state that a given speaker has formulaic language "stored whole in their lexicon or because it is highly automatised" (p. 10). The distinction is made between storage and automatisation as it is challenging to prove holistic storage; therefore, the definition from Myles and Cordier (2017) instead places emphasis on the processing advantage rather than claiming the sequences are "stored and retrieved whole from memory" (Wray, 2002, p. 9).

Previous research from Ellis (1996) and further investigation from Cobb (2018) claim that the processing occurs within working memory, as formulaic language is 'chunked'; the combinations are stored as one semantic unit. Ellis (1996, p. 107) states that this chunking is "the development of permanent sets of associative connection in long-term memory" and increases language fluency. Cobb (2018) expands on this, noting that there is "low-cost handling of formulaic patterns qua chunked, single items, with working memory thereby left free to handle a relatively small number of truly novel, unpredictable constructions" (p. 196). Therefore, when formulaic sequences such as collocations are used rather than single words, this frees up working memory space due to faster and easier retrieval of the set semantic units, and this then leads to increased language fluency as the speaker is then able to deal with more complex language.

Regardless of whether the language is stored or highly automatised, the consensus is that formulaic language is pervasive and important for fluency. This is especially relevant for spoken language as Wood (2015) proposes that using formulaic sequences is beneficial for both the speaker's perceived fluency and for increasing comprehension for listeners. Further, Lin (2018) notes that studies from Dechert (1983) and Raupach (1984) have found that it is the language learners who use formulaic language that seem to speak with a "distinctive fluency". More recently, Martinez and Schmitt (2012) note that several researchers found that producing formulaic language can positively impact the overall impression of learners' language ability (Boers et al., 2006; Ellis & Sinclair, 1996; Lewis, 2008; Ohlrogge, 2009). Considering the importance of formulaic language, such as collocations, in increasing fluency, it is noteworthy that many studies have found that L2 learners find this aspect of language challenging to engage in. For example, Spöttl and McCarthy (2004, p. 191) state that formulaic language is "the last and most challenging

hurdle in attaining near nativelike fluency", as even the most advanced learners still struggle with this aspect. This struggle may have to do with differences in language processing between native speakers and language learners. Myles and Cordier (2017) explain that the advantages for native speakers includes the ability to process highly idiomatic phrases. This is compared to language learners, where only the most transparent or frequent formulaic phrases have this advantage in processing. The advantage also differs individually with Schmitt et al. (2004) noting that not all frequent combinations are holistically processed for everyone. Finally, when compared to single words, Durrant (2008) also states that "collocations are in general much rarer, much more diverse, and much more strongly tied to specific areas of discourse than are individual words" (p. 3), so learners need to learn not only the appropriate formulaic language but also when to use this based on the appropriate situation. Semantic transparency is also a factor to consider regarding the processing of formulaic language, with Gyllstad and Wolter (2016) finding that as well as frequency impacting processing, collocations being more semantically transparent combinations also meant the processing was increased when compared to less semantically transparent constructions such as idioms. Finally, both L1 and L2 speakers process collocations subconsciously quicker and more accurately than novel phrases (Ellis et al., 2008; Sonbul, 2015).

Overall, evidence from psycholinguistic research into the processing of collocations adds to the picture as to why collocations are a meaningful aspect of language for further study as there is evidence that not only does it reflect patterns of usage, but collocations could also tell us more about how speakers, and specifically language learners, store, retrieve and understand language.

It has been established how important formulaic language and specifically, collocations are within spoken language to aid in processing and fluency, which is why it is also crucial to explore why it is challenging for language learners to acquire. Altenburg and Granger (2001) and Laufer and Waldman (2011) are just two of a number of studies that have found even the most advanced learners still demonstrate collocational errors, with Erman et al. (2015) finding an under-representation of collocation frequency within L2 speakers when compared to a native speaker group. Language learners and collocation research will be explored more thoroughly in Section 2.4; here the importance of understanding and using collocations can be seen, and that it is also challenging for learners to master.

2.3. Corpus linguistics and collocations

2.3.1. Why is corpus linguistics as a methodology valuable for the study of collocations, particularly in learner language?

Since corpus linguistics first developed as a methodology for studying authentic language use, it has been used to study formulaic language and specifically collocations. Granger (2020) notes that these large databases of language samples are particularly fruitful for studying any syntagmatic language features such as collocations and lexical bundles which are two distinct types of formulaic language. The former are co-occurring lexical words occurring more frequently than expected by chance while the latter "can be regarded as extended collocations: bundles of words that show statistical tendency to co-occur" (Biber et al. 1999, p. 989). Such language features benefiting particularly from corpus investigation is due to corpus software tools being easy to use when identifying and extracting these language features automatically, meaning analysis of large amounts of data is more time efficient and straightforward to do, which is only improving further with more innovative methods (Liu, 2021). Also, corpus linguistics can work to develop descriptions of these formulaic units as documented in language corpora (e.g., Evert, 2005; Gries, 2008; Sinclair, 1991). Corpus methods also go hand in hand with the frequency-based approach to defining collocations due to the need to quantitatively measure the frequency and exclusivity of word combinations under this approach.

Furthermore, collocations can also be visualised using corpus software tools such as #LancsBox (Brezina et al., 2015) which is valuable because patterns can be found in a different way using such analysis. This corpus-driven approach is advantageous as although it is limited in scope in the fact that only one feature of phraseological competence is being investigated, Chen and Baker (2016) state that the corpus approach allows for "a more systematic and thorough examination of learner language" (p. 878). As well as the ease of extracting collocations, corpus linguistics is also beneficial because it is inherently effective at helping the researcher find patterns in language, so the association combinations of collocations lend themselves to this kind of methodology.

Within the field of language learner research, corpus linguistics has become well-established as an effective methodology for exploring data. Learner corpora research generally takes one of two approaches: text-internal and text-external, the latter using a native speaker reference corpus to compare the learner corpus to (Tavakoli & Uchihara, 2020). There can also be pseudolongitudinal studies such as Granger and Bestgen (2014),

where different speakers at different proficiency levels at one point in time can be studied in order to look at developmental patterns in second language acquisition. This is also true of studying collocations, as Bestgen and Granger (2018) note that learner corpus research is beneficial because it can focus on multiword units rather than on single words for measures of lexical richness. This is supported by Tan and Azmi (2021) who conducted a scoping review on language learning research that used collocational competence to measure overall language competency and found that nine of the 21 articles used corpus methodology, therefore, demonstrating that it is a frequently used and well-established, appropriate methodology for the study of collocations. Corpus linguistics can also be used to complement other areas of language learning research that involve the study of collocations, such as second language acquisition (SLA) studies. Granger (2020) notes that using large corpora can be especially conducive to supporting the typically much smaller sample sizes of more experimental research within the field of SLA (p. 7). Finally, corpus linguistics methods can not only be used to research collocations to understand how people use them but also to help within language teaching, such as creating corpus-informed textbooks to ensure authentic language is being taught and even bringing tools into the classroom for data-driving learning (DDL) activities (Pérez-Paredes & Mark, 2021).

2.3.2. What are some core considerations for corpus design and compilation for language learning research?

In general, corpora design and creation need to follow specific guidelines to ensure that the data collected will be useful for the research questions to be explored. To do this, Egbert (2017) states that researchers "should be careful to learn about the sampling design and practices of existing corpora before using them for research and follow sound principles of corpus design and sampling when constructing a new corpus" (p. 560). These principles should also be followed for learner corpora. As well as general considerations for corpora construction, additional factors need to be taken into account for learner corpora.

The first factor regarding the design of the corpus is establishing when the data has been sampled. Most native speaker corpora include samples of language from one point in time, e.g., the Spoken BNC 2014 (Love et al., 2017) which contains speakers from a set period of time. Longitudinal native speaker corpora are rarer but typically include the

same speakers sampled multiple times during first language acquisition, i.e., as children, such as the CHILDES English Manchester Corpus (Theakston et al., 2001).

For language learning research, there is often an emphasis on finding patterns in language development; therefore, one snapshot in time may not be sufficient to explore this variable. Longitudinal corpora, where data has been taken from one speaker over multiple periods, seems like the most logical way to document any language proficiency development. Siyanova-Chanturia (2015) is one example of a longitudinal learner corpus, and the author claims this is a more effective way of measuring the development of collocational use as each speaker can be tracked over time, thus taking into account individual differences which are known to have an impact on language production. Schmitt et al. (2004) state these individual differences are likely to influence the acquisition of formulaic language, too, with Halim and Kuiper (2018) also finding evidence of this. However, despite the obvious benefits of longitudinal data collection to investigate language development, obtaining this data is resource intensive and, therefore, difficult to gather. Instead, many learner corpora opt for a cross-sectional design where data is taken from speakers at differing proficiency levels and so any analysis can take a pseudolongitudinal approach to examine these points in time across language development. One cross-sectional learner corpus is the TLC-L2, where speakers are different learners at different proficiency levels at one singular point in time, so analysis regarding development is seen as "a slice (…) to piece together actual development" (Gass et al., 2013, p. 36). More learner corpora are cross-sectional than longitudinal due to the challenges involved with collecting data of the latter type based on the time and financial costs that are incurred, as well as the likelihood for attrition during the data collection process. Another consideration of particular importance when designing learner corpora is the definition of proficiency used when talking about the speaker's language competence. This is crucial to, firstly, consider thoroughly and, secondly, to be explicit about within the corpus description to ensure accurate comparisons can be made between research studies (Carlsen, 2012).

This degree of comparability is vital for language learning research when a L1/L2 comparison is being made. If there are differences between the two corpora that are not just related to the variable that is being observed, then this results in challenges when make a meaningful comparison (Gablasova, 2020). Therefore, the creation of comparable corpora, for example, to ensure the language context is the same, is essential and valuable

within language learning research. This level of comparability is demonstrated with corpora such as the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin et al., 2010) and its comparison corpus the Louvain Corpus of Native English Conversation (LOCNEC; De Cock, 2004). This has led to valuable learner corpus research into linguistic features such as evaluative adjectives (De Cock & Frankenberg-Garcia, 2011) and pragmatic markers (Aijmer, 2011; Buysse, 2017) and demonstrates the importance of having comparable corpora. Furthermore, it should be noted that not all spoken learner corpora are comparable. For example, the LINDSEI and LOCNEC corpora may seem to be similar to the TLC-L2 and the new TLC-L1 in terms of design as they are spoken corpora of L1 and L2 English. However, there are distinct differences that have impact on the generalisability of findings between the two sets of corpora. A major difference is that the TLC-L2 and TLC-L1 speakers are engaging in a spoken examination context (Gablasova et al., 2019) rather than an informal interview (Gilquin et al., 2010). This means the tasks undertaken vary considerably i.e. a picture description task in the LINDSEI/LOCNEC compared to the Discussion and Conversation tasks in the TLC-L1/L2 where speakers are engaging in Trinity College London's Graded Examination in Spoken English and are required to maintain the dialogic interaction (Trinity College London, 2021). As stated by Caines and Buttery (2017), task plays a significant role in the use of language. In addition, the LOCNEC is compiled of 117,427 words (De Cock, 2004) compared to the 523,205 of candidate language in the TLC-L1 (see Section 3.2.2. for the full breakdown of corpus size); therefore, the former is also less than a quarter of the size of the latter, allowing for less opportunity of use when investigating linguistic choices (Caines & Buttery, 2017). Consequently, developing new comparable corpora is of value to expand the research possibilities.

Granger (2020) acknowledges the challenges of designing corpora of language learners stating that "the perfect learner corpus does not exist and will never exist, as no single corpus can answer all the research questions that L2 researchers aim to answer" (p. 254). Therefore, the goal should not be striving to achieve perfection when it is impossible. Instead, it is important to consider the purpose of the corpus creation and how it has been designed, and essentially, be explicit in this when describing characteristics of the corpus to ensure that conclusions drawn from the data by others can be reliable (Carlsen, 2012). Furthermore, Durrant and Schmitt (2009, p. 162) note that "identifying native texts that are equivalent in type to non-native writing is, as other researchers have noted, highly

problematic (Granger et al., 2002, p. 40; Lorenz, 1999, p. 14)". Overall, there is a need to design a corpus with a specific purpose in mind to ensure that it will be accurate to use to investigate the research questions set out to be researched. Furthermore, as it is compiled, the choices and characteristics of the corpus also need to be made very explicit so that (1) other researchers using the corpus can do so with confidence that they have chosen an appropriate corpus for their study and (2) when using results, the comparison to other studies can be explained accurately.

2.4. Collocations and language learning research

The following sections of the literature review will consider how collocations and language learning research have come together in several ways. Firstly, this section will focus on speakers learning an additional language covering how their receptive and productive knowledge of collocations is linked to their L2 proficiency, how their L1 background influences the use of collocations and how learners compare to native speakers in their collocational competence. Then, the literature review moves focus to a specific type of lexical collocation: verb + noun collocations, as these have been frequently studied in the context of language learning research. Next, context will be considered, namely how topic and register can influence the production of collocations in learners. Finally, the literature will give an overview of how collocations have been investigated within language teaching and language testing and how results from research can be applied to these fields. Bestgen and Granger (2018) note the difficulty in attempting to synthesise results from studies investigating L2 phraseology and attribute this in part to the criteria researchers used to identify phraseological units, whether taking a quantitative or qualitative approach.

Gyllstad (2007), later supported by Nizonkiza (2017), notes that research into collocations and language learners typically falls into two categories. The first is experimental research involving receptive and productive collocational knowledge tests. In contrast, the second involves using learner corpora (and sometimes reference corpora) to analyse authentic language production on a large scale. The literature review will continue to make this distinction between the two research approaches while also aiming to bring them together to create a fuller picture of the complexity of collocations.

2.4.1. The Speakers: L1 and L2

2.4.1.1. What does the research show about L2 proficiency and use of collocations?

The relationship between L2 proficiency and learners' use of collocations is a complex one that many researchers have tried to unpick; Paquot and Granger (2012) note that it is difficult to establish what influence proficiency can have on L2 production of collocations in part due to differing definitions of proficiency used by researchers. Although Bachman's (1990, p. 16) definition is frequently applied, this being "the knowledge, competence, or ability in the use of a language, irrespective of how, where, or under what conditions it has been acquired" the issue is how this can then be broken down into differing proficiency levels in a way that development can be described and measured. Carlsen (2012) supports this point by explaining that proficiency levels within learner corpora are frequently ill-defined. Gablasova (2020) further highlights this point in that there is a need to decide on how representative a corpus is for any SLA or language testing research project, and in particular, researchers need to know how proficiency was established for findings to be meaningful. This is further complicated due to changing conceptual and theoretical ideas surrounding language proficiency. For example, Leung (2022) explores the history of what it has meant to be proficient in a language and how this has developed over time with an overall move away from the native speaker being the "universal reference" (p. 74) that learners should aspire to be like. They further explain the complexity of the concept, saying that "some aspects of language proficiency may be beyond static description" (p. 76).

The importance of defining proficiency has been frequently raised in the last ten years. Consequently, more explicit descriptions of the characteristics of corpora are included when researchers describe their data. Small, specialised corpora such as the NOn-native Spanish corpus of English (NOSE; Díaz-Negrillo, 2012) can have fuzzy boundaries of proficiency levels, whereas others, usually larger corpora, benefit from explicit alignment to a proficiency framework. One example of this is the TLC-L2 corpus (Gablasova et al., 2019) which includes data from the Trinity College London Graded Examinations in Spoken English (GESE). The GESE as an exam underwent an extensive alignment process to a well-established and frequently used proficiency framework (Papageorgiou, 2007). This means that, although the candidates progress through multiple grades in the GESE with small and manageable increases in difficulty, each of these grades can also be linked to a corresponding CEFR level. Accordingly, findings can be more accurately

compared to other research that uses the CEFR framework when investigating proficiency, rather than using less well-defined terms such as ‘upper intermediate’ or ‘beginner’. González Fernández and Schmitt (2015) acknowledge this too, by making links between their research and Laufer and Waldman’s (2011) study but noting that “it is impossible to know how the participant proficiency levels of the two studies compare” (p. 112). Having a more explicit description of how proficiency is defined is of benefit to learner corpora research as a whole.

To complicate matters further, research into language proficiency involves assessing learners’ knowledge at different stages and what is meant by knowledge can be further broken down. There are two ways that learners’ collocational knowledge is measured: productive, by focusing on the use of collocations and receptive, by testing understanding of collocations. Lee (2021) claims that productive knowledge is more challenging for language learners to develop than the receptive understanding of collocations. As corpus linguistics can only speak to what is actually produced in language, it is important to consider research into learners’ receptive collocational knowledge and how this adds to the picture of an individual’s overall productive phraseological competence as the two are obviously linked in some way; speakers need to have receptive knowledge before they can produce collocations, although there is some evidence that one type of knowledge cannot necessarily predict the other (e.g., Zareva et al., 2005). There is evidence of a difference between receptive and productive knowledge of collocations in language learners; Kamarudin et al. (2020) found that the learners in their study had a significantly higher mean score for the receptive collocation knowledge test than for the productive collocation knowledge test. When considering results from different studies, it is important to be aware of whether the researchers are measuring receptive or productive knowledge of formulaic language, such as collocations.

This next part of the section will explore these two aspects of collocational knowledge in L2 speakers and discuss how findings have been linked to proficiency. The connection is complex, and the following research exemplifies this.

Regarding a potential link between receptive knowledge of collocations and proficiency, some studies have found evidence of a connection. Investigating receptive knowledge typically involves experimental study designs with tests of collocational awareness set for participants. This differs from spontaneous communication, which can assess how

speakers produce collocations in an authentic language exchange. One example of testing for receptive collocational knowledge comes from Uchihara et al. (2021), who used a word association task to elicit responses from 40 first-year Japanese undergraduate students learning English. These participants ranged from B1 to C1 proficiency in English based on the CEFR scale. The authors were looking at oral proficiency and measured this in two ways: subjective human judgements and objective measures of fluency and lexical richness, e.g., pace of speech. Overall, the researchers found collocational knowledge was a predictor for oral proficiency as they discovered that the speakers who used more low-frequency collocations, measured by t-scores, were observed to speak at a faster pace and used less silent pauses; they were also subjectively judged to be more proficient.

Further to this finding of low t-score collocation usage, those speakers using high MI score collocations were also considered more lexically proficient. Although Uchihara et al. (2021) acknowledge the small participant group, the findings indicate this would be worthwhile investigating further. Lee (2021) measured the collocational knowledge of Korean and Mandarin L1 learners of English through a phrase acceptability judgement task. The study found proficiency, measured using a cloze test, to indicate success in the acceptability judgement of what L2 collocations were grammatical but non-nativelike in their construction; however, proficiency was not linked to success in identifying L2 collocations. This ability to discriminate between nativelike and non-nativelike collocations was deemed an essential step in the "developmental path" of L2 collocational receptive knowledge (Lee, 2021, p. 205). Another study to use a cloze test but for measuring collocational competence of lexical and grammatical collocations rather than proficiency comes from Keshavarz and Salimi (2007), who found a statistically significant relationship between TOEFL test scores and collocational competence in their 100 Iranian EFL learners. However, a limitation comes from the fact that the cloze test used had multiple options to choose from, which may have impacted the results.

There has been more research into productive knowledge of collocations, either using elicitation tasks or looking at corpus data. The relationship becomes even more complex with no real definitive claims to be stated other than the consensus that collocation learning is a "highly complex and multifaceted process" (Keshavarz & Salimi, 2007, p. 452). Those studies that have found evidence of a connection between proficiency and collocational (or more general phraseological) competence can be organised into different categories based on three parameters; (1) frequency: the number of collocations differs

based on proficiency level, (2) types measured by association measures (AMs): different proficiency levels use different types of collocations on the frequency/exclusivity scale and errors and appropriacy: (3) proficiency levels vary in how they accurately produce collocations based on the L1 norm and the context. The following section will discuss these three parameters in more detail before introducing alternative evidence to complicate the connections further.

The first parameter within investigations into phraseological competence, and possibly the most researched so far, is frequency. This is looking at how many collocations (or other formulaic sequences) learners use and whether this frequency changes depending on proficiency. Namvar (2012) found a strong positive correlation between learners' general language proficiency and their knowledge of collocations; however, their general proficiency level was measured by a writing test about an unforgettable experience which was then holistically marked to measure proficiency by Linguistics and Education PhD students. As previously discussed, how proficiency has been measured may have impacted the findings and how generalisable they are to other research. Similar issues with results were observed by Nizonkiza (2012), who also found proficiency increase aligned with learners' development of productive collocational knowledge. They used TOEFL scoring to assess proficiency but then grouped the results into five levels based on the number of learners in each group, aiming to have at least 30 students for statistical reasons. This makes it challenging to assess what stage of language development the students were at as they were grouped in comparison to each other rather than on a more objective scale. Also looking at L2 writing, both Laufer and Waldman (2011) and Paquot and Granger (2012) found that higher proficiency learners used more collocations than those at a less advanced developmental stage. However, Paquot and Granger (2012) further explain that high proficiency L2 speakers may increase the number of collocations they use compared to low proficiency users; though, the quality is impacted during this development. So, there may be a more frequent occurrence of collocations, but these are not used accurately. This could be due to what is suggested by Thewissen (2008) as increased phraseological richness as high-level users become more confident and creative in their attempts to use formulaic language, the usage increases. However, a similar number of errors is maintained as those with lower language proficiency.

Moreover, using lexical collocations has also been correlated with writing fluency, an important aspect of proficiency (Hsu, 2007). Two more studies have found evidence of a

link between proficiency and phraseological knowledge based on frequency measures but have also posed that development is slow and nonlinear. Firstly, Forsberg and Bartning (2010) approached their research using a pseudolongitudinal study of L2 French learners and found significant differences in written lexical formulaic sequences as these increased at higher CEFR levels. However, this difference was not found in adjacent levels, such as A2/B1 and B2/C1 but between the A2/B2/C2 levels. This suggests development is slow, and one level increase on the CEFR scale may not be enough to capture formulaic language development. Finally, Nizonkiza's (2017) findings also support this notion of slow development and suggest this may also be nonlinear in growth. Using TOEFL scores to assess proficiency as in their earlier 2012 study, Nizonkiza also found collocation knowledge was particularly slow to develop at the lower levels of proficiency but then seems to gain momentum as the learner enters the intermediate stage of language learning before this then stabilises at advanced proficiency and potentially even levelling off at this stage. Tavakoli and Uchihara (2020) found that there was a difference in the frequency of use between the participants in their low B1 group and the C1 group. The lower proficiency speakers engaged in more repetitive and redundant task-related multi-word sequences (MWSs) than the higher proficiency C1 group, who were found to use the same n-grams. However, these were used more creatively, such as changing the structure of the MWSs or using synonyms. Therefore, this study showed a qualitative difference in using oral formulaic language within this language testing environment. Finally, moving to n-grams, another type of formulaic language, Kyle and Crossley (2015) conducted a cross-sectional study with L2 speakers and found human rated oral proficiency scores positively correlated with the frequency of n-grams overall. This meant that those speakers who were more proficient were also using a greater amount of L1 speaker target-like formulaic language features. Therefore, speech was rated as more proficient based on the frequency of formulaic language and how L1 speaker-like it was.

Another aspect of collocational competence and its relationship to proficiency comes from research that uses the frequency-based approach to measure the association strength of collocations. Different measures have different results because they highlight different aspects of a collocation relationship, i.e., frequency and exclusivity. Bestgen and Granger (2014) found a significant decrease in the use of collocations with average t-scores (highlighting frequency) over six months with no change in average mutual information (MI) scores (highlighting exclusivity) in their longitudinal study of L2 writing. As the

writers developed, they used fewer high-frequency collocations, but their highly exclusive collocation use stayed the same. In a further study measuring collgrams (constructions with features of both n-grams and collocations), Bestgen and Granger (2018) again found evidence of a connection between proficiency and collocation use; here, they found an increase in the sophisticated use of collgrams across three years, and this was seen by an increase in low t-score bigrams, a decrease in high t-score bigrams and finally an increase in high MI score bigrams. This means that over time writers were using formulaic language that was more strongly associated based on exclusivity than frequency. This finding has also been supported by research from Eguchi and Kyle (2020), where an increase in proficiency level also meant an increase in the use of more strongly association (MI score) collocations, this time in spontaneous speech. Finally, González Fernández and Schmitt (2015) set out to measure productive knowledge of collocations in a group of 108 Spanish L1 speakers learning English to answer the question of how many collocations L2 learners use and how well they do this, as learner output does not tell the whole story; there may be collocations that learners know but do not have the opportunity to use. Contrary to other research, they claim that collocations are relatively easy for L2 learners. Their results show that many are typically known, with the productive collocation knowledge test results averaging over 56%. However, the researchers acknowledge that at 13.67 mean years of study in English, this could have impacted the findings. Learners could be demonstrating more high-frequency collocations (i.e., higher t-score collocations) because they encounter them during their day-to-day life rather than more mutually exclusive collocations that are more salient for native speakers but only applicable in specific situations. Therefore, it is not only the fact they are encountering them more but the fact that there are situations that are setting them up to encounter them more, i.e., daily life rather than a specific context. The researchers posit "whether frequency, as derived from specialised corpora better representing learner usage, might be a superior way to predict collocation knowledge" (González Fernández & Schmitt. 2015, p. 114). Overall, there does seem to be some consensus that strongly associated combinations, as measured by t-score and MI score, differ in use depending on a learner's proficiency level.

The third parameter that has been found to intersect with proficiency and formulaic language use is errors and appropriacy. Thewissen (2015) found that the production frequency of collocations did not seem to be linked to proficiency, but instead, the errors

produced were different in type. Bestgen and Granger (2018) also found that even at relatively advanced levels of language proficiency, learners are still engaging with a noticeably high rate of errors within their production of collocations. In addition, Chen and Baker (2016) found lower-level proficiency learners were using conversational lexical bundles in writing compared to high level learners, demonstrating that proficiency could be linked to appropriate contextual use of formulaic language.

Conversely, some research into errors and accuracy suggests no significant connection between collocational knowledge and proficiency. One such study from Vedder and Benigno (2016) was unable to find a relationship between collocation proficiency and overall L2 proficiency when considering both the frequency and the accuracy of the use of collocations by the learners. The authors acknowledge this could be due to the study only using low-intermediate and intermediate learners and that there may be a more salient relationship at higher proficiency levels, i.e., B2 upwards on the CEFR scale. This links with the idea of the 'slow development' of collocations, as findings from Forsberg and Bartning (2010) and Nizonkiza (2017) have suggested. The researchers note that "collocation development has also been shown to follow a u-shape curve and to be influenced by the syntactic complexity of the collocation types" (Vedder & Benigno, 2016, p. 26), which ties in with the notion of nonlinearity in the development of collocational knowledge. Siyanova-Chanturia and Spina (2020) also found that L2 phrasal production may actually "get worse as a function of time before it can slowly and gradually get better" (p. 452) based on their study of noun + adjective combinations in written L2 Italian. Regarding inaccuracy of collocations, Nesselhauf (2005) found that L2 use of verb + noun collocations maintained a high level of inaccuracy (around 1/3 of all used) regardless of how long the learner had been studying. Finally, Laufer and Waldman (2011) posit there is relationship between collocational development and proficiency and also found that errors were similar in both more and less proficient EFL learners, thus demonstrating that the picture is unclear.

The above literature has focused on text-internal measures of collocational competence. Further to this, there has been research using text-external measures, allowing L2 production data, typically from corpora, to be compared to L1 data to gain a sense of proficiency in a different way, i.e., how do learners use collocations in comparison to how a native speaker uses them. The current research points to there being mixed results that may depend on the learners that have been studied and the methods employed in the

research, but with some consensus that collocational use does develop, albeit in a nonlinear way, and most likely in the sense of diverse types of collocations being used. The importance of researching types of collocations used is further supported by Cobb (2003), who notes that the use of collocations may increase the perceived fluency of a learner. However, overuse of collocations can occur, leading to inappropriate usage. This section has demonstrated the challenge of uncovering a clear relationship between proficiency and collocational competence.

Another aspect to keep in mind when researching the relationship between collocations and language proficiency is the individual variation of a language learner's developmental journey (Lowie & Verspoor, 2015). De Bot et al. (2017) explains that one conceptual framework that focuses on the individual within language development is Complex Dynamic Systems Theory (CDST) which combines Complexity/Chaos Theory (CT) and Dynamic Systems Theory (DST). According to CDST language development is a complex system with different components that are interacting, changing and adapting over the course of time. The dynamics within language development therefore mean that one component influences others (Larsen-Freeman, 2012). Stages of development are not linear in growth but constantly ebb and flow with variation based on the individual learning the language. Consequently, L2 development is not necessarily a ladder (Larsen-Freeman, 2006) but a constantly changing interactive system. In particular, researchers working within the framework of CDST note the importance of formulaic sequences as these linguistic features in learner language are frequently found to be in flux (Larsen-Freeman, 2006 p.3). This notion of a non-linear trajectory in language development has been supported by previous research explored in this section regarding the u-shaped curve (Forsberg & Bartning, 2010; Nizonkia 2012; 2017; Siyanova-Chanturia & Spina, 2020; Vedder & Benigno, 2016) and has also been found in more recent research from Brezina and Fox (2021) using the TLC-L2 where phraseological development was found to be slow with large individual differences found between learners. When considering research into L2 development, particularly formulaic sequences such as collocations, there should be an awareness that the nature of development is "full of progress and regress" (Duan & Shi, 2021) and that this may be reflected within any investigative results.2.4.1.2. What does the research show when comparing L1 and L2 use of collocations?

Four recurrent themes emerge from the literature when considering how language learners use collocations in relation to native speaker usage. Firstly, diversity of collocations. De Cock et al. (1998) proposed the diversity of formulaic language as a significant difference between the two groups, as they found that L2 writers use fewer formulaic sequences than L1 writers. Ädel and Erman (2012) also found this to be the case with lexical bundles where native speakers used a wider range of this formulaic language type – 130 – compared to 60 lexical bundles occurring in the learners' written texts. Finally, Granger (2018) notes that current findings indicate that both the quality and quantity of collocations develop as learners become more proficient in a language, but there is still often the case of overuse of a small number of types for L2 users; this demonstrates a lack of collocational diversity when compared to L1 speakers.

The second recurrent theme follows from Granger's (2018) point regarding overuse. A common analytical approach to comparing learner and native speaker language concerning formulaic language has been considering the overuse and underuse of phrases. An earlier study from Granger (1998) was one of the first to begin to explore the overuse/underuse of collocations. She found that learners underused native-like expressions and tended more towards using atypical expressions instead. This core result has been echoed elsewhere, such as by Erman et al. (2015) and Bestgen and Granger (2018). Li and Schmitt (2009) also found overuse of specific lexical phrases in their longitudinal case study, and Chen and Baker (2010) found this in an academic writing context too – overuse of informal lexical bundles in academic writing and underuse in the typical academic writing formulaic sequences – showing the phenomena occurring within a specific context. Finally, as well as collocation use in general, overuse and underuse seem to be linked with erroneous formulaic language. Vedder and Benigno (2016) found that L2 speakers also produced a higher number of erroneous collocations when compared to the L1 speakers in their study; in fact, the latter group were noted to barely produce errors at all, though the overall number of incorrect collocations from L2 speakers was lower than expected.

Overuse and underuse also link to the third recurrent theme of L1 and L2 comparative research into formulaic language: investigations into the frequency and exclusivity of collocations. Durrant and Schmitt's (2009) findings demonstrated a difference in the use of high MI score collocations, with non-native speakers significantly underusing them when compared to native speakers. They also found a statistically significant difference

between how frequently rare combinations are used in the long texts, with a higher proportion found in the native speaker group. Finally, the researchers also noted a significant overuse of high t-score collocations by non-native speakers compared to native speakers when considering collocation tokens rather than types. These findings from Durrant and Schmitt (2009) were among the first to explore the frequency and exclusivity of collocations, from a frequency-based perspective, in native speaker and learners' language. This study was supported by Ellis et al. (2008), who noted that there was psycholinguistic evidence for why highly frequent collocations are used more by non-native speakers while native speakers favour highly mutually exclusive collocations. More recent research has found much of the same evidence, with González Fernández and Schmitt (2015) noting that high MI score collocations are "likely to be especially salient" (p. 98) for native speakers and it is because of this that when non-native speakers do not utilise them in writing, it can be especially noticeable. This is also the case in speech, with Saito (2020) finding that using more infrequent formulaic language significantly impacted perceived L2 oral proficiency. Low frequency combinations containing infrequent, abstract and complex words were found to be strong determiners of L1 raters' scores on comprehensibility and lexical appropriateness of the L2 speakers. This factor of frequency and exclusivity also impacts comprehensibility; for example, Saito and Liu (2022) discovered that when collocations were operationalised using MI scores (exclusivity) rather than t-scores (frequency), L2 speech that was distinctive was evaluated to be more comprehensible. Overall, there is evidence that collocations are used differently by L1 and L2 speakers based on the frequency and exclusivity of these combinations, and the differences between these collocations can impact perceived L2 oral proficiency and comprehensibility.

The fourth and final recurrent theme in the research into how L1 and L2 speakers use formulaic language is appropriacy. One study from Hyland (2012) found that competent and appropriate use of collocations can distinguish between novice and expert use in a range of genres within an EAP context, while Siyanova and Schmitt (2008) focused their research on 810 adjective-noun collocations produced by Russian learners of English in the International Corpus of Learner English (Granger et al., 2002; ICLE) and found that, based on MI scores and frequency measures, 45% were used appropriately. Interestingly, this was then found to be not hugely different from native speakers. Overall, using collocations that are appropriate to the context seems to be important for distinguishing

novice and expert speakers though appropriacy is challenging to measure based just on the collocations alone; more qualitative analysis can look to see the collocations in context and, using corpus linguistics, within concordance lines.

2.4.1.3. What does the research show about L1 background and the use of collocations?
Another variable that has been found to influence learners' use of collocations is L1 cultural and language background. In contrast to the varying results investigating language proficiency, there is some agreement regarding the impact of L1 background influence in L2 collocation use (Granger, 2018). So much so that some studies, such as González Fernández and Schmitt (2015), highlight the need to control for transfer effects by only selecting one L1 background in their research. Others have posited that L1 can influence through negative transfer, positive transfer or L1 avoidance. Regarding these transfer effects, research from Lee (2019) and Choi (2019) has found evidence of negative transfer from L1 that led to collocational errors due to this interlingual factor. An explanation for this difference between L1 and L2 collocation use comes from Wolter and Gyllstad (2011), who acknowledged that an "L1 may have considerable influence on the development of L2 collocational knowledge" (p. 430). In this study, the finding was attributed to more effective processing from the L2 speakers when presented with L1-L2 collocations (translation equivalent collocations) than L2-only collocations (items only acceptable in English, not Swedish). The researchers noted that priming might factor in this difference, incorporating a psycholinguistic aspect to collocation usage. This concept of variance in processing L1 and L2 collocations is further supported by Yamashita and Jiang (2010), indicating that differences in the production of collocations are likely to have a psycholinguistic basis. Granger (2018) also explains that the deep entrenchment of L1 collocations is a factor in the late development of L2 collocations. Finally, there may not be just a linguistic difference between the L1 and L2 but also impact from speakers' cultural background. Namvar (2012) found evidence of both positive and negative L1 transfer on the use of collocations as well as noting a cultural factor to the challenges met by the participants in this study.

This notion of L1 transfer becomes more complex if attempting to separate receptive and productive knowledge of collocations. One study from Lee (2021) investigated L1 transfer effects on L2 collocations in low-intermediate to advanced learners focusing on recognising unacceptable word combinations. Using a framework from Jarvis (2000), Lee

compared the L1 target group to two other typologically distinct languages, focusing on English language learners from Korean and Mandarin backgrounds as well as native speakers of English to investigate the complex relationship of L1 influence on L2 collocations. The study found no evidence of L1 transfer when participants were tasked with recognition of unacceptable L2 collocations, and this was the case even at the lowest level of proficiency. This is counterevidence to findings often reported in L2 writing where word-for-word translation occurs from the speaker's L1, suggesting that L1 transfer occurs at the productive level but not at the receptive level. Learners are aware of non-nativelike L2 collocations but may still produce them regardless of this receptive knowledge. This adds to the support that collocational competence is highly complex to achieve.

As well as L1 transfer, Bestgen and Granger (2014) note that another phenomenon, L1 avoidance, can also be a factor in the collocational errors of learners. Finally, Bahns (1993) entered this discussion early from a language teaching perspective, noting that this influence could be beneficial to language learners as a necessity to teach and learn every collocation would be overwhelming. Instead, the focus should be placed on teaching collocations with no equivalent in the L1 and relying on positive L1 transfer, using phraseological constructions from the first language in the additional language as an acquisition strategy.

Bestgen and Granger (2018) state that a potential reason for a noticeably high rate of errors within even advanced learners' production of collocations is due to the L1 background influence, which seems to override the proficiency level. This could be due to topic complexity. Kreyer (2021) found an increase in L1 interference at specific points of learners' language development. The author proposes this may be due to the increased complexity of the topics the learners needed to communicate. Therefore, they were more reliant on translating from their German L1 rather than using collocations that would be more typical for English writing. Furthermore, even at high proficiency levels, collocations can be influenced by the learners' L1. Cao and Badger (2021) found that Vietnamese learners of English used unconventional collocations at an occurrence of only around 7%. However, 40% of these were said to be directly influenced in some way by their L1. So, even when minimal inaccuracies exist, these are still heavily influenced by L1.

This difference in perception could also contribute to collocational usage errors due to the distance between L1 and L2 backgrounds. Wang and Shaw (2008) found this to be the case with an impact on error rates in usage. This impact was thought to be caused by the 'risk' felt by learners when producing language. The study found that a closer first language to English, in this case, Swedish, resulted in more risks and errors, whereas Chinese speakers were more conservative in usage and thus produced fewer errors. This finding suggests the importance of the speakers' choices; production is influenced by comprehension and language background, and potentially, confidence in usage increases with proficiency, leading to more inappropriate collocations. Therefore, when considering overall phraseological competence within a language test, neither a linear increase in the number of collocations found in speech nor a linear decrease in inappropriate use of collocations can be a simple indicator of increased language proficiency. Considering the diversity of collocations used and the risk-taking of speakers is also essential.

Types of inappropriate collocations used also seem to differ depending on the L1 background. Thewissen (2008) investigated how learners' L1 may influence types of error concerning levels of English language proficiency and found there may be evidence of L1 differences when considering grammaticality and acceptability errors in both frequency of occurrence overall and type of error. Furthermore, Shih (2000) notes that Taiwanese learners utilised specific lexical simplification strategies based on their L1 background. Cross and Papp (2008) found differences between L1 Chinese, Greek and German-English learners. The Chinese-English learners were found to use the collocations more frequently but with a higher error rate. There was also a difference in the types of error produced depending on the learners' L1 background, giving further evidence to the importance of this variable in collocation research. Again, this supports the research interest into how L1 background may influence phraseological competence regarding types of L2 collocations.

Overall, these studies have implications for future research; pure frequency of collocations may not be the sole focus of interest. Instead, there could be a difference in the diversity of what formulaic language is used. Considering the language testing field, it would be of value to investigate if speakers are using more of the standard, expected collocations but also producing more errors and whether these learners could be said to

be at a higher proficiency of language than those who use fewer collocations in their speech but are more accurate.

These studies demonstrate a need for the understanding that collocations can be influenced by L1 background when conducting research into this feature of formulaic language. There is ample opportunity for further research into how L1 background may affect collocational usage in language learners, with L1 transfer, both positive and negative, and L1 avoidance possible reasons for learners' differences in using collocations. Using corpora with rich metadata such as the Trinity Lancaster Corpus (Gablasova et al., 2019), investigations can also extend beyond language and into the cultural background, adding to the research conducted by Namvar (2012), for example, considering differences between L1 Spanish speakers from Mexico and Spain.

2.4.2. The Type: Verb + Noun Collocations

2.4.2.1. What has research found so far regarding how verb + noun collocations are used?

Recent research has investigated several different types of phraseological collocations, such as adjective + noun (Zhang & Chen, 2006; Takač & Lukač, 2013; Granger & Bestgen, 2014), verb + adverb (Paquot, 2019), verb + preposition (Kamarudin et al., 2020) and adverb + adjective (Granger, 1998). The most common type of phraseological collocation researched is verb + noun (e.g., Laufer & Waldman, 2011). This may be because it is especially prevalent in language; for example, a search within the Trinity Lancaster Corpus of L1 Spoken English (TLC-L1) shows this as both the most frequent form of collocation and the most widely dispersed across candidate speakers. Within verb + noun collocations, these can be broken down into smaller categories of type based on the constituents of the collocation; the word under study (node) and the word co-occurring (collocate). This thesis will focus on lexical collocations, specifically verb + noun collocations, and highlight select high frequency delexical collocations for case studies to engage in a more in-depth qualitative analysis.

Some researchers, such as Choi (2019), investigated learner use of delexical collocations. Here, "delexical" is understood to indicate that the meaning of the verb depends on the meaning of the noun, or another definition being "verbs with little meaning" (Sinclair, 1990, p. 147); overall, the speaker needs to know what verb to use with what noun based on knowledge of the phrase as a fixed expression rather than the individual vocabulary items. If learners can use these appropriately, it could be argued that this is one indicator

of phraseological competence. Allerton (1984, p. 33) noted that the choice of verbs within delexical collocations is mostly an arbitrary one and that they are "semantically unmotivated"; therefore, if the learners are not choosing them based on what meaning they bring, they are choosing to use verbs based on what they perceive to be accurate – potentially through what they have learned as collocates. This is further supported by Chi et al. (1994, p. 162) "as such verbs carry no significant meaning, it is likely that students will choose the wrong verb-noun collocate unless they have previously learned it as a chunk".

High frequency verbs such as *make*, *take*, *have*, *do*, and *get* are also especially worthwhile to investigate as a node component of verb + noun collocations as these verbs tend to have neutral connotations in use and therefore, there is minimal topic and register bias (Chi et al., 1994). This is especially important considering the data being used in this study because it is not natural conversation but language examination data. The results can take a broader view beyond this test by minimising the potential effects of topic and register bias as much as possible, such as from specific topics set by the exam. Section 2.4.3.1 explores the importance of topic within such analysis in more detail. High frequency verb collocations are also of value to investigate as there have been calls to include these more explicitly in language pedagogy (Nguyen & Webb, 2017); knowing more about how learners and L1 speakers use these collocations is of value for teaching.

Further to this minimisation of topic bias, Zinkgräf (2008) points out that the use of delexical verbs shows students understand that the language has restricted collocations while Yan (2010) notes that learners find them difficult to use because the verb adds little meaning to the whole phrase with verb + noun collocations accounting for 50% of all lexical collocation errors in their study. Therefore, it could be argued that not using high frequency verb + noun collocations appropriately indicates a lack of phraseological competence. This is an opportunity for production data, such as a corpus, to add to the understanding of learners' collocation use and compare this to L1 speakers. Further rationale for considering high frequency verbs, specifically in delexical verb + noun collocations, comes from Altenberg and Granger (2001, p. 174), as they are said to have: basic meanings, different semantic fields, high frequency equivalents in other languages, polysemy and overall seem to be problematic for language learners. These points raised are also related to the variable of language proficiency being investigated in this thesis.

There have been numerous studies into verb + noun collocation use in various contexts. Some consider all verb + noun collocations, whereas others highlight key high frequency, delexical verbs. When considering English collocations, by far the most common verb node is *make,* which was both a sole focus for some (Babanoğlu, 2014; Gilquin, 2007; Kim, 2002; Lee & Na, 2015; Lin & Lin, 2019; Sawaguchi & Mizumoto, 2022) while others included it within a more extensive analysis of multiple verbs, frequently also involving *take,* e.g. *make* and *take* (Du et al., 2022), *make, take* and *get* (Ma & Kim, 2013) and *make, take, give* and *do* (Kreyer, 2021).

Babanoğlu (2014) focused on *make*'s lexical and grammatical use as a verb in argumentative writing. The research looked at how Turkish EFL learners used this verb appropriately and also considered potential L1 transfer effects by comparing three corpora TICLE (Turkish L1), JPICLE (Japanese L1) and LOCNESS (English L1). Appropriateness was measured by assessing the overuse, underuse and misuse of the *make* verb within the TICLE compared to the native speakers in the LOCNESS. The research found that the overall frequency of *make* was different between the Turkish and Japanese learner corpora when compared with the L1 corpus. However, this underuse of *make* by the L2 learners was not statistically significant. There was some evidence of L1 transfer in the Turkish learners' use of *make* when conducting error analysis; some used *give [a] decision* rather than *make [a] decision.* Although based on writing rather than speech, this research indicates that it would be valuable to investigate the use of specific high frequency verbs that occur within verb + noun collocations.

Similar to Babanoğlu (2014), Kim (2002) focused on the verb *make* and how Korean EFL learners use this along with typical collocates within their writing. Kim found that the learners either underused or misused both the delexical as well as the idiomatic use of *make*; this was in comparison to overusing the causative function of the verb. The study claims that both proficiency and L1 transfer contribute to the findings. Again, this shows that delexical verbs, in particular, are challenging for learners to produce. Finally, Ma and Kim (2013) looked at the use of high frequency delexical verbs within seven Korean EFL textbooks as well as the collocational knowledge of 209 Korean high school students. The researchers found that, of the high frequency delexical verbs, *make, take* and *get* were the most frequently used in the books. Furthermore, they also found that, although the participants felt their knowledge of verbs was sufficient when undergoing a collocation test, the correct answer rate was only 38%, demonstrating that this type of collocation is

challenging for English language learners. The authors call for a need to teach these delexical verb collocations explicitly. This research also indicates the value of focusing on specific high frequency delexical verb collocations; here, knowledge was assessed using a collocation test to elicit responses. Further research into how learners produce and use these collocations within a spoken examination would be valuable to investigate to contribute to the growing literature. Finally, using the ICNALE, a corpus of written learner English, Lin and Lin (2019) looked at the difference between native English speakers and Asian learners' use of the verb *make*, considering both the lexical and grammatical features. Specifically, they link overuse and underuse to overall language proficiency. Much like Kim (2002), Lin and Lin found significant underuse of *make* by English learners when using the verb in its delexical use compared to native English speakers. They mention a limitation that not all of these high frequency verbs are likely to be equally problematic to learn for L2 speakers and that further research should investigate "multiple high-frequency verbs and make cross comparisons" (p. 13).

Sawaguchi and Mizumoto (2022) used a bilingual corpus (Kansai University Bilingual Essay Corpus; Yamanishi et al., 2013) to investigate the use of *make* + noun collocations in writing. This corpus had learners complete a written essay on the same topic in their L1 (Japanese) and their L2 (English). The researchers focused on the L1 influence on the use of the L2 collocations and explored any effects from proficiency level too. Based on the essays, they found no development in the learners' productive collocational knowledge. However, there was evidence of L1 influence on collocation use changing according to proficiency level – this means that proficiency alone might not impact collocation use, but L1 can influence it in combination with proficiency. However, there were proficiency-specific uses of delexical *make*. The researchers suggest extending the study into other high frequency verb + noun collocations like *take*.

Some studies have taken the overuse/underuse approach to describe how learners use collocations compared to L1 speakers. For example, Suzuki (2015) aimed to answer two research questions using corpus-based methods. Firstly, the author was interested in whether Japanese learners of English tended to overuse or underuse the high frequency verb of *get*. The second research question involved a comparison of the L2 speakers with native speakers of English. Using written essay data from the ICNALE (Ishikawa, 2013), Suzuki found that the L2 speakers typically overused *get*, and this was claimed to be due to a reliance on the *get* + noun construction, as *get* was also significantly underused in other types of construction with the verb. The author found that the learners engaged in

what they term 'atypical combinations' such as using *get* with the nouns *money, friend* and *thing*, attributing this to the possibility that the learners lack "collocational knowledge of the verb *get* and tend to rely on the open-choice principle when they reproduce English sentences" (Suzuki, 2015, p. 15). This concept of the open choice principle and the idiom principle was explored by Erman and Warren (2000) following on from Sinclair's (1991) initial work on how speakers construct phrases. The two concepts are two concepts related to how speakers use and store lexical knowledge. The open choice principle involves word-for-word combinations while the idiom principle involves multi-word combinations as preconstructions (Erman & Warren, 2000, p. 29). Therefore, Suzuki (2015) argues that, based on the atypical collocations found with *get*, the learners are creating the collocations by putting words together rather than understanding them to be typical preconstructed phrases.

Another study by Lee and Na (2015) suggests that L1 influence may be a factor in the overuse and underuse of a specific verb. The study focused on using *make* as a verb in writing, comparing two corpora of Korean EFL learners and native speakers. A major finding was that the L1 Korean speakers overused *make* overall when compared to the native speaker corpus and that this was evident even with the most advanced learners. However, the learners also underused *make* when it was being used delexically; this supports the finding from Kim (2002) and demonstrates the difficulty of this type of verb usage for English language learners as Lee and Na (2015) note that incorrect choices for collocates were also evident in the corpus. They note that there may have been some L1 influence regarding these differences due to the "overlapping meaning between the English verb *make* and the Korean verb *mandeulda*" (p. 22).

Some studies suggest one influence on L2 verb + noun collocation production could be that of language proficiency. Laufer and Waldman (2011) looked at these phraseological combinations in the writing of L1 Hebrew speakers across three proficiency levels. Much like the results detailed previously in Section 2.4.1.1, they noted seemingly conflicting results. More advanced students were found to produce more errors, rather than less, as would be assumed with language development. The authors attribute this to a rise in confidence in taking risks with language as part of learners' development. L2 learners underproduce verb + noun collocations in their written discourse when compared to NSs of a similar age. Conflictingly, there is also evidence that learners may minimise risk-taking with verb + noun collocation use, as found in another study from Gilquin (2007).

Here, *make* collocations were the focus and the research used three forms of data; elicitation data of fill-in exercises and judgement tests compared with a corpus of 'free production.' The research found a 7% error rate and underuse of these collocations within the corpus data compared to 51% and 43% in the fill-in exercise and judgement test data, respectively. The advanced French-speaking learners were more reluctant to take chances within their free-writing compositions, which led to a relatively low error rate and overall underuse of the collocations.

Previous research into the nature of verb + noun collocations has found that L1 background can be a factor in L2 language production. For example, Zinkgräf's (2008) investigation into verb + noun 'miscollocations' in L1 Spanish found that university students' writing exhibited notable errors when using these collocations and attributed the cause to negative transfer from L1. The author considers this to be due to direct translation from the L1 or a considerable amount of overlap in meaning between the two languages. She also finds a discrepancy between language proficiency and competence in using collocations, as well as recurrent patterns in the miscollocations used. Others have found that learners can rely on patterns from their L1. Luzón-Marco (2011) noted that learners used atypical verb + noun combinations in technical writing due to a reliance on using patterns models from their L1. This was fuelled by having issues with the phraseology of the discourse generally, echoing that language context matters, as well as difficulty with sub-technical language and high frequency verbs. Finally, Juknevičienė (2008) investigated Lithuanian learners of English and L1 speakers using the LICLE and LOCNESS corpora and found that delexical collocations, such as those using high frequency verbs, are less likely to match or have some equivalent in L1, so there are more errors due to this L1 influence. There were also set patterns in collocation use due to the L1, such as the confusion of *make* and *do* in English, as these have a single-word equivalent in Lithuanian. She also found that learners did not produce collocations with abstract nouns in the same way that was evident in the LOCNESS. This links to Saito and Liu (2022), who found that "L2 learners use of content words become more diverse, abstract, infrequent and complex in nature with conversational experience" (p.20). This also links in with a move away from the 'lexical teddy bears' of Hasselgren (1994) which the author defines as structures that learners feel most comfortable with and thus tend to overuse. Previous findings have also suggested error type rather than frequency of use may be more directly linked to language proficiency (Thewissen, 2015) and raises further

questions with regards to how best to measure phraseological competence, as attempts at linguistic innovation may be present in higher proficiency speakers but also considered inaccurate when compared to the L1 norm established in research. For example, Garner (2020) looked at three groups of L2 writers categorised as on low, medium and high language proficiency. By analysing the collocation frequency, diversity and association strength using covarying collexeme analysis, Garner found the higher proficiency writers used more diverse, less frequent and more strongly associated verb + noun collocations. Finally, Paquot (2019) studied verb + noun structures in B2, C1 and C2 EFL writers and found that these structures could distinguish the C2 speakers from the other groups based on average MI scores; only the most proficient speakers were using these collocations accurately.

As well as accuracy in collocation use, research into verb + noun collocations suggests nonlinear progress in learners' acquisition of these structures. Kreyer (2021, p. 99) notes the "uneven and slow development of collocational competence". His study took a longitudinal approach and examined 83 German L1, English L2 speakers across four years. This data came from the Marburg Corpus of Intermediate Learner English (MILE; Kreyer, 2015) which contains written learner English from grades 9 (14-15 years old and around A2/B1 proficiency) to 12 (17-18 years old expected to be achieved B2 level). The texts included cover a variety of written registers and types and the author's purpose was to record a substantial number of learners across a long period of time. However, the author claims this dataset is still too small in some ways, namely for the purpose of tracking the use of one verb + noun collocation over several years. Subsequently, even the most frequent collocations were not used frequently enough for this kind of analysis. Therefore, there is tension with what data we can collect, and the research methods must consider this. Kreyer (2021) found that the normalised collocate frequencies for the delexical verbs *do, give, make* and *take* decreased to less than half as time passed, from 4.9 to 2.1 per 1,000. He notes this is in direct contention with Laufer and Waldman's (2011) findings of increased collocation use and posits that a reason may be that the students are expanding their vocabulary and, therefore, less reliant on using verbs that have multiple meanings. Instead, they are diversifying their collocation use.

A study by Du et al. (2022) examined how A1, A2, B1, B2, C1, and C2 EFL students used *make/take* + nouns in their writing based on their proficiency level. The researchers used the EF-Cambridge Open Language Database (EFCAMDAT; Geertzen et al., 2013)

and extracted 3,600 scripts for analysis. Using the BNC as a reference corpus, Du and colleagues calculated the t-score of each *make/take* + noun combination found in the EFCAMDAT based on the observed frequency in the BNC. 61% of these combinations had a t-score higher than 3.9 and were thus categorised as collocations. Then, the researchers undertook three approaches to analysis, firstly categorising the collocations based on their semantic fields using the UCREL Semantic Analysis System (USAS; Rayson et al., 2004), secondly annotating nouns with difficulty levels using the English Vocabulary Profile (EVP; Kurtes & Saville, 2008). Finally, the researchers calculated the lengths of noun elements, again using EVP. Regarding semantic fields of the noun elements, Du et al. (2022) found that beginners tended to use collocations that contained concrete nouns and everyday activities compared to the advanced learners who used more nouns within abstract semantic fields like social or political topics. The EVP analysis also found that beginners tended to use less complicated and shorter noun elements within the collocations compared to the advanced group; however, this was only found between the extremes of the proficiency levels, as there was no significant difference between A1 and A2 speakers for this. The authors conclude that this provides evidence for collocation development as proficiency increases and suggests these results could be used to teach learners specific semantic field collocations appropriate to their level of proficiency. It is interesting to note that the study chose the t-score measure to assign the category of collocation to the combinations. As seen in previous research within this literature review, choosing a different association measure, such as MI score, may have led to very different results. However, this provides evidence of collocational development within the frequency dimension and would be of value to explore further.

Considering the literature explored in this section, verb + noun collocations are especially interesting to investigate as a measure of phraseological competence, with case studies focusing on high frequency delexical verb collocations particularly useful to highlight. By looking at the frequency and distribution of verb + noun collocations across language proficiency levels, we can better understand phraseological competence in spoken learner English.

2.4.3. The Context: Topic and Register

2.4.3.1. How does topic impact learners use of language?

It has been established that topic can be a variable within language research and that this needs to be controlled to ensure the validity of results (Paquot, 2020). This is very much the case for learner corpus research as well, particularly at the point of data collection. Alexopoulou et al. (2017) note that "the topics used to elicit the L2 samples shape the language that is represented in the corpus" (p. 181). Topic influence can affect language learners in different ways, particularly when undergoing assessments. Firstly, background knowledge and topic familiarity about a specific subject has been found to impact language assessment performance. One example from Huang et al. (2016) specifically looked at L2 speaking assessments. They compared integrated speaking test tasks (where input was provided for candidates to use to aid the creation of their replies) with independent speaking test performance and how they were each associated with knowledge of topic. The study found a significant impact on performance based on topical knowledge in both conditions (integrated and independent speaking tasks), which was also found to be topic dependent. From this, it can be said that topic influences speaking test performance in general, and this needs to be considered when creating language exams. Later research from Khabbazbashi (2017) and Yoon (2021) have also indicated similar findings regarding the impact of topical knowledge on performance.

A second way topic can affect learners' language is the influence it can have on elicited specific linguistic features, with Suzuki (2015) finding that "variations of words in the data tend to depend on the topics of the essays" (p. 7). However, not only vocabulary seems to be influenced by topic, but also formulaic language. L2 candidate control of topic has been found to be positively connected to stronger certainty adverbials in spoken language exams (Gablasova & Brezina, 2015), while Cortes (2004) found lexical bundles used by learners writing within different disciplines were influenced in this way. A later study also on lexical bundles from Paquot (2014) specifically noted that it was not just content word frequency and use influenced by topic but also found an influence in tense preferences, including the overuse of lexical bundles that included the modal verb *will*. This also supports findings from Hinkel (2009), who studied modal verbs in L1 and L2 writing and found there was a significant effect on the frequency of obligation and necessity modals in L2 writing which were found to occur mainly during topics such as parental roles and responsibilities as well as family duties. Hinkel suggests that the high

frequency of these types of modals may be linked to the L1 cultural background of the Chinese, Japanese and Korean speakers within the study. Paquot (2014) had anticipated some topic influence on the language and had factored this in as an explanatory reason for lexical bundles' presence in the learners' writing. Despite this factoring, she notes that there were still topic-influenced lexical bundles, which were found to be mainly functioning as referential expressions in contrast to stance markers and discourse organisers that were found to be overused due to L1 influence. If these findings are occurring for lexical bundle analysis, it would be interesting to see if there is topic influence in some way regarding a different kind of formulaic language type, such as collocations.

It is important to research topic when expanding our understanding of phraseological collocations because such research has previously been conducted from a frequency-based approach (Gablasova et al., 2017) and issues were found with only considering association measures which are inherent in this approach. Gablasova et al. (2017) found differences in MI-score when looking at collocations from two different speakers who discussed different topics within the TLC-L2 (Gablasova et al., 2019). In this case, using MI-scores derived from the BNC meant that those speakers discussing social conditioning had lower MI-scores than those talking about global warming due to the technical terms that were necessary to include in their collocational selection. The researchers also note the need to be mindful of topic and task when interpreting collocational patterns (Ellis et al., 2015; Forsberg & Fant, 2010). Therefore, it is of value to analyse topic in order to shape our understanding of phraseological collocations.

Overall, topic is an important variable to look out for in the analysis, and it is also essential to control for topic as much as possible within the analysis. In this thesis, this will be done by using two tasks led by different interlocutors. The Discussion task is candidate-led as they have chosen and prepared a topic to introduce within the exam, thereby engaging in schema activation ahead of the assessment and easing the cognitive load of language retrieval (Leblanc & Fujieda, 2012). The Conversation task is examiner-led; the topics have been created by trained item writers and selected by the assessment board, with the examiner choosing on the day which to introduce or having the candidate pick randomly from two unseen lists.

2.4.3.2. How does register impact learners use of formulaic language?

Staples et al. (2015) state "registers are language varieties associated with a particular configuration of situational characteristics and purposes" and "can be defined at any level of generality" (p. 505). One example of a register frequently studied in language learning research is that of academic writing. The research link between academic writing as a register and formulaic language, especially lexical bundles, is strong. Hyland (2008) states that clusters (of formulaic language) can actually help to define registers. More evidence of this link comes from the fact that most studies that have set out to consider the influence of register on formulaic language production have been looking at language within academic writing (e.g., Durrant & Mathews-Aydinli, 2011) or within other genres of writing such as email (Fritz et al., 2022) and personal descriptions (Burgos, 2018). This focus on written production has left a gap for research into register within spoken language. There is evidence that formulaic language differs depending on the mode (written vs spoken language e.g. from Ellis et al., 2008), so what has been found in the many studies within (academic) writing regarding register-influenced formulaic language choices may not be directly seen in spoken language. However, there is also evidence that language learners utilise similar linguistic choices in academic writing as in their speech due to a lack of genre awareness of the discourse situation (Gablasova et al., 2017). Considering this, what we do know is that there is indeed an influence of register on language use, and this should be considered in research. For example, Biber et al. (2004) noted different functioning lexical bundles between speech (stance and discourse organising) when compared with writing (referential). They attributed this difference to the processing involved in the two modes. This was later confirmed by Staples (2015), adding to Biber et al. (2004) that "variation in the frequency, function, and fixedness of lexical bundle use suggests meaningful differences in the way that formulaic language is processed in speech as well as the functions for which it is used" (p. 276).

Considering how language learners' proficiency factors into the relationship between register and formulaic language use, Chen and Baker (2016) undertook a study of lexical bundles within L2 expository and argumentative essays. These learners were divided into three groups based on their level of proficiency according to the CEFR scale: B1, B2 and C1 speakers. As the essays were rated, they found evidence that B2 was a crucial point in transitioning from informal to more formal academic writing based on the more appropriate lexical bundles. The learners were able to produce more appropriate language

for the specific writing context; it will be of interest to study how this may be reflected within a spoken language examination setting.

Regarding register-influenced collocations, Sinclair (1991, p. 109) suggests the relationship between register and collocation by stating that when a register choice is made, "all the slot-by-slot choices are massively reduced in scope or even, in some cases, pre-empted." Neshkovsha (2019) later confirmed this idea of a strong connection between collocations and registers and establishes two types of collocations based on this relationship: common collocations, which are frequent in general conversation and register-specific collocations that occur within a specific context, such as *dummy object* occurring within the field of IT. The context limits the opportunities for language choices of a speaker, and this is very much the case within the language examination register and context too.

One significant influence on language choices within the examination register is the presence of the examiner; this is especially true in oral proficiency language exams, where the examiner is frequently also the interlocutor that the candidate needs to engage with to complete the assessment. This significantly impacts the conversation's power dynamics and overall speaker dominance within the conversation (Young & Milanovic, 1992). Furthermore, one aspect of the exam register is how the examiner supports the candidate during the assessment. One of the first studies to describe what language occurs during an oral proficiency assessment comes from Lazaraton (1996), who set out to describe the linguistic and interactional support from the native speaker examiner to the candidate. She found the interaction involved (1) priming topics to scaffold speech, (2) engaging in collaboration to help candidates complete acceptable responses, (3) providing evaluative responses after a candidate answers, (4) echoing and/or correcting responses, (5) repeating questions while changing speech pace, including pauses and overarticulation, (6) stating questions as statements that need yes/no confirmation, (7) drawing conclusions for the candidates and (8) rephrasing questions to check or facilitate answers. These can be said to be typical features of the spoken language exam register. This study shows the value of description as a first point – we need to know what is happening within a register to be able to analyse it further. This also led to a further focus on discourse analysis within language testing (Lazaraton, 2012).

The language exam is a register whereby a variety of language is elicited not only based on the apparent task differences but due to the ethos of the language testing board developing these assessments. For example, Trinity College London (2023a) state publicly that their exams offer "conditions that support 'bias for best' where candidates are encouraged to demonstrate what they can do with language and aren't marked down for what they can't do", which follows on from Merrill Swain's original idea to bias for best in language assessment in the early 1980s (Fox, 2004). Furthermore, certain language is produced in order to fulfil the functions of the exam, which is distinct from language that is influenced by the topic of conversation. Seeing how this may be reflected within the linguistic choices of the candidates would be beneficial to explore further.

The limited research thus far into the combination of collocations, exam register, and spoken language has been acknowledged with calls for expansion to investigate how register can affect association strength between words, taking a frequency-based perspective to the study (Gablasova et al., 2017) while a phraseological approach to the study of register and collocations would also be complementary to the research so far. As the focus has also been on register influence on academic writing, expanding this to continue adding to the description of a spoken examination context would also be valuable to the field.

2.4.4. The Applications: Language Teaching and Language Testing

2.4.4.1. How has research into collocations been used within language teaching?

Bahns and Eldaw (1993) were some of the first researchers to propose explicitly teaching collocations to EFL students. Since then, the call for interventions to teach formulaic language has only increased as many believe that there is still an insufficient emphasis on this crucial linguistic feature in the language classroom (Granger & Bestgen, 2014; Tan & Azmi, 2021; Nesselhauf, 2005; Martinez & Schmitt, 2012; Siyanova-Chanturia, 2015). As previously discussed in this literature review, this focus comes from the evidence that mastery of formulaic language is an essential aspect of overall language fluency. However, the research has shown the difficulties many learners face, even those most advanced proficiency speakers, with some evidence that formulaic language needs more target exposure when teaching than other language forms (Forsberg & Fant, 2012) and so it is clear to see why the calls to action are continuing.

One such way to teach formulaic language more explicitly that has gained traction in recent years has been to engage with corpus linguistics to inform pedagogy; data-driven learning (DDL) is an approach to teaching first proposed by Johns (1991). It has been defined more recently by Liu (2021) as "the direct use of authentic corpus data to conduct student-centred discovery learning activities" (p. 180). There are three main ways in which corpora can be applied to language teaching; these are using corpus-influenced materials, corpus-cited texts and corpus-designed activities (Bennett, 2010).

Success in teaching and learning collocations has been found using DDL methods. Daskalovska (2015) looked into using corpus-based activities for learning verb + adverb collocations and compared the use of these activities to more traditional and typical activities found in course books. Those participants who used the concordance lines in the corpus-based activities were found to do better on all parts of the subsequent test than those using traditional activities. The author proposes this was due to the exposure to many authentic language examples in the concordance lines, attention focusing activities to the grammatical patterning of the collocations and summarising activities, which further helped the participants to memorise the collocations. Other studies have also involved using concordance lines, such as Rezaee et al. (2015), where using a concordancer was found to have a statistically significant effect on both learners' receptive and productive knowledge of collocations, further providing evidence for the benefits of DDL.

Using corpus tools can also facilitate learners' acquisition of collocational knowledge. Liu (2021) undertook research comparing learners' knowledge development from the use of a collocational dictionary to those engaging with #LancsBox (Version 2) (Brezina et al., 2015). This corpus linguistics tool goes beyond concordance lines as it also includes a feature called GraphColl which can visualise collocational relationships within a corpus. Liu found a slight improvement in overall knowledge of collocations, with more training cited as being potentially beneficial to this group and future cohorts, as many wanted to continue using the tool to help with their language learning. As well as more extensive interventions to help support language teaching, Tan and Azmi's (2021) scoping review has shown that even an indirect corpus approach can help with language teaching, which "may be the most suitable and practical approach that can cater to almost all levels of proficiency whilst consuming a limited amount of resources" (p. 115). Finally, a recent meta-analysis from Boulton and Cobb (2017) details the positive effects of this "corpus

revolution" (p. 388) in both applied linguistics and in language teaching. Future directions are plentiful and include more focus on how learners participate with DDL and promote more diversity in the learners engaging with it, and in the learning environments it is situated (Pérez-Paredes, 2022).

There are some core considerations with teaching formulaic language, such as collocations, as Nguyen and Webb (2017) explore, noting that care needs to be taken when deciding what to focus on in class as time constraints make it impossible to cover every possibility. Shin and Nation (2008), in their research uncovering the most frequent collocations in spoken English, note that although frequency is an important criterion for deciding what to teach, it should not be the sole focus; difficulty and range of use are two other criteria to consider. Antle (2013) echoes this reminder that "it is important to consider our students' needs, level and motivation" when developing activities (p. 353). Research can have an impact on helping this decision process; a study from Durrant and Schmitt (2009) led to the recommendation to support learners in acquiring high MI score collocations to develop their overall phraseological competence. More research into which collocations prove most challenging for learners and which collocations work to provide the level of perceived fluency needed for successful communication will help teachers decide what to draw their students' attention to within the classroom while using corpus linguistic methods to support this instruction.

2.4.4.2. How has research into collocations been used within language testing?

As there is now a consensus that teaching formulaic language is an essential part of teaching a language, the developing knowledge around how we use and teach formulaic language has also piqued an interest in how this can be applied to language testing – once exposed to the linguistic constructions, how do language students produce formulaic language and collocations under test conditions? Alderson (1996) was one of the first researchers to identify the potential of using corpora within language testing. Consequently, language testing as a field has leaned on corpora for a variety of purposes, including the design and validation of tests (Deshors et al., 2016; Hawkey & Barker, 2004) and connecting linguistic items, such as error types, with proficiency levels (Granger & Thewissen, 2005) whilst also validating proficiency scales (Carlsen, 2012). L1 speaker corpora can be used to establish the lexicogrammatical characteristics of the target language to test, while learner corpora can give insights into usage according to many variables, such as L1 background or age.

Regarding collocations within language tests, Saito and Liu (2022) researched the relationship between the ease of understanding of L2 speakers by L1 and L2 raters and their use of collocations measured by t-score and MI score. They found that conversational experience in the L2 speakers influenced mutually exclusive word combinations and coherency. Overall, they found that "raters rely substantially on collocations while making intuitive judgements, particularly when the lexical context of speech is relatively limited and predictable" (p. 19). Although the picture description task was found to rely on raters' collocational awareness more than the oral proficiency interview, it is still emerging evidence that the use of collocation does indeed impact judgements of language proficiency and, therefore, of value to explore further in language testing corpora such as the TLC-L1 and the TLC-L2.

As well as considering how collocations are used within language tests, there have also been attempts to assess learners' productive knowledge of phraseology. These tests of collocational knowledge frequently included translation tasks alongside cloze tests (Bahns & Eldaw, 1993; Biskup, 1992; Farghal & Obiedat, 1995; Gitsaki, 1999); however, Gyllstad (2007) notes that reliability measures for these tests are either problematic or simply not reported within the research.

Further tests have been created to assess written formulaic language proficiency, as discussed in Gyllstad and Schmitt (2018), such as CONTRIX (Revier, 2009); however, the authors note the difficulty in creating a standard measure for phraseological competence. This is due to several factors, including the various subcategories of formulaic language, such as collocations and idioms and due to the challenges surrounding the identification of formulaic language, whether considered from a frequency-based or phraseological approach. Regarding the former approach, another challenge for measuring phraseological competence comes from the association measures frequently chosen. MI score has been found to be effective as a measure of phraseological sophistication in academic writing (Paquot, 2019; Paquot et al., 2021), but depending on the corpus used, the variables included may mean that MI scores as well as means and medians as measures of central tendency are not able to capture learners' phraseological development or overall competence (Paquot et al., 2022). Paquot et al. (2021) also acknowledge that "mean MI is a very crude measure of a learner's phraseological competence" (p.143). Despite this, it was still considered to be the most appropriate

option for the analysis, which shows the challenge involved in measuring phraseological competence.

Instead of testing solely for phraseological competence, including criterial features within existing rating scales that exemplify proficiency in phraseological use may be more beneficial. Römer (2017, p. 486) notes that the TOEFL spoken test does include reference to formulaic language, but this includes criteria such as "[s]ome low-level responses may rely heavily on practiced or formulaic expressions". An issue arises here as formulaic expressions are seen as a 'low level' response. Although research has shown there can be overuse of collocations at lower levels, this criterial descriptor does not capture the nuance of phraseological competence, considering variables such as proficiency level or L1 background. Römer (2017) also comments on the need to operationalise phraseology or lexicogrammar within such scales and the importance of doing so by bridging the gap between corpus linguistics and language testing.

However, there are arguments against changing descriptors, such as that from Xi (2017, p. 572), who contends that neither holistic human scoring nor analytic rubrics should include "descriptors of small, frequency/count-based linguistic elements, even if they have some relevance to the construct" such as specific collocations. Even though collocations can be extracted from a corpus, it may not be fitting to include them in criterial rubrics.

Overall, there are significant benefits to using corpus methods within language testing research, firstly in the development and application of learner corpora as Park (2014) remarks that learner corpora can effectively describe a specific variety of language. This in turn avoids the issue of learners striving to achieve 'nativelike' proficiency when only corpora of native speakers are used to establish language assessment criteria. There is a need to acknowledge the presence of different varieties of English, and 'expert speaker' could be a more appropriate goal for learners. There is much variation within native English speakers, so should this variation within English language learners be attributed as a deficit or a varietal difference? Instead, learner corpora could be used to develop tests for assessing how a learner is developing a specific variety of a language rather than only comparing how similar they are to a native speaker.

Learner corpus research can achieve the above because a large amount of data from speakers can be analysed; this can remedy the representativeness issue typically arising

from experimental data in SLA research (Granger, 2018). However, there also needs to be a consideration that corpora can only show production and not the perception or comprehension of language. This is particularly vital to consider when investigating learner data and considering language competence, as the speakers are only producing a certain level of fluency and accuracy of a language at that moment in time. Corpora are language samples, so researchers need to be cautious about generalising outside these specific contexts and speakers. Further, language tests are carefully crafted to produce real-life usage of specific linguistic features from the candidates. Corpora of examination language, therefore, may not be indicative of spontaneously occurring informal speech but are a valid context for study. As Laufer and Waldman (2011) state, learner corpora are invaluable at providing language performance data for research.

As well as benefits, there are core considerations when using corpus methods in language testing research. When focusing on seemingly minor aspects of lexicogrammatical usage, we must be careful in claiming how crucial these are in an overall language assessment. This review has established that L2 speakers may not be proficient in noticing nuanced linguistic elements (Durrant, 2014) and, overall, human raters are also not especially skilled in noticing frequency-based linguistic elements either. We need to be mindful as to how relevant considering specific lexicogrammatical features are in assessing language; this is of benefit when considering how best to apply seemingly abstract research findings to inform more concrete practices within language testing, such as developing assessment criteria. Furthermore, there may be hesitation in researching scripted language such as that in the IELTS language test and therefore it is of value to ensure language examinations that elicit spontaneous language exchanges, such as the GESE (Trinity College London, 2021), are also studied. The latter arguably provides a clearer view of the language learners can produce in real-life spontaneous interactions.

In addition, corpus linguistics, language testing and phraseological research can also combine to be beneficial; for example, Bestgen and Granger (2014) considered one way to investigate collocational production in assessment through an L1/L2 comparison. The study aimed to analyse how phraseological competence can affect L2 writing proficiency and text quality assessment. They did this by looking at the quantity and quality of bigrams. It was noteworthy that they did not only calculate formulaicity based on 'text-internal' measures; these are just calculated based on learner text. They also used 'text-external measures', which are calculated based on an external resource – a large corpus

of texts covering a "broad spectrum of native language use" (p.38). They note their reference corpus choice as a limitation due to the fit between this and their overall aims of the study; further evidence that context is vital.

However, an L1/L2 comparison can potentially set up the 'native speaker' as the goal in language acquisition which can be problematic. This is an issue in measuring phraseological competence. Competent speakers as often seen as those who can also successfully deviate from the norms in certain language situations or for certain communicative functions, such as humour, when appropriate. Considering that high proficiency speakers may still produce the same number of errors as low proficiency speakers, Wray (2002) notes that several researchers have found L2 learners to rely on creativity resulting in the overuse of words, which they believe to be synonyms when producing collocations. This erroneous belief of collocational equivalence would be a reason for L2 speakers to produce atypical collocations in the target language as they do not realise the 'fixedness' of the particular language chunk, which could mean they are instead adhering to the open choice rather than idiom principle needed for successful use of collocations (Wang, 2016). However, without knowing the intention, we cannot attribute creativity or lack of fixedness to a collocational error. Howarth (1998) also comments on this notion of creativity in word combinations as L1 speakers producing atypical collocations might be considered creative, especially outside academic prose. However, within learner data, a concern is what we can decide is creativity and what is simply incorrect. Considering these deviations from the norm, there are also differences between errors (knowledge-related) and mistakes (performance-related) in production. Assigning a deviation as an error or mistake within corpus data is difficult as we are unaware of the process behind the production. Therefore, an L2 speaker who produces a nonstandard collocation may do so because of creativity, mistake, or error or because they demonstrate an emerging variety of that language. Deshors et al. (2016) discuss how corpora can be used to find innovative linguistic structures in L2 English. However, identifying these is tricky as frequency alone cannot be relied upon. Labelling a construct as an innovation relies on the analyst as, again, we cannot be sure of the linguistic intention. Durrant and Schmitt (2009, p.168) suggest classifying word combinations "across a scale of collocational strength" rather than assigning correct/incorrect judgements. Finally, Gablasova et al. (2017) also warn against interpreting L2 speaker

choices as language proficiency inadequacies. Instead, learner corpora such as the TLC-L2 (Gablasova et al., 2019) should be used to explore the development of competencies.

2.5. Summary of the literature

This review of the current literature covers substantial ground as the intersection between phraseology, corpus linguistics and language learning is vast. The research in this thesis draws on the theoretical and practical considerations outlined above; the following points represent the core theoretical ideas pulled from the review that have influenced the methodological and analytical decisions of this research going forward.

- Corpus linguistics is an appropriate method to investigate the phenomena of collocations.
- There are differing ways to defining collocations depending on what approach is taken.
- Defining collocation is a complex issue.
- There is a lack of research into spoken English language use due to the difficulty in obtaining data.
- Using collocations are challenging for language learners at all proficiency levels, even those who are advanced.
- There is a lack of consensus as to how proficiency impacts collocation usage.
- One conceptual framework to consider individual differences within L2 language development is that of Complex Dynamic Systems Theory.
- There is a difference in the use of collocations between L1 and L2 English speakers.
- Looking at collocation use can help inform language teaching and testing materials.

Considering this summary, the empirical research in this thesis adopts a combined approach to the study of collocations using corpus methods, first identifying collocations of interest according to their phraseological properties using a specific Corpus Query Language (CQL) query to capture verb + noun collocations based on their syntactic structure. The collocations will then be analysed according to this phrasal relationship as well as considering the statistical collocational status of the collocations of interest. This blended phraseological and frequency-based approach has been suggested to be used by Granger (2018) with further support coming from Szudarski (2023) and has been implemented recently by researchers such as Gablasova et al. (2017) and Lee (2019).

2.6. Research Questions

Based on the exploration of the literature, the following research questions emerged:

RQ1: To what extent are there differences in the (1) frequency and (2) distribution of use of verb + noun collocations amongst L2 English speakers at B1, B2 and C1/C2 level?

RQ2: To what extent is there evidence of topic influence on speaker choice of verb + noun collocations in the TLC-L1 and the TLC-L2 corpora?

RQ3: To what extent is there evidence of register influence on speaker choice of verb + noun collocations in the TLC-L1 and the TLC-L2 corpora?

RQ4: How do TLC-L1 and TLC-L2 speakers use high frequency delexical verb + noun collocations in spoken examination language?

These research questions are addressed in Chapter 4 which explores the TLC-L2 and Chapter 5 which explores the TLC-L1, while Chapter 6 brings together a general discussion of the two corpora. Descriptions of the Trinity Lancaster Corpora and details of the methodological decisions taken in the data analysis for the two will now be introduced.

Chapter 3: Methodology

This chapter introduces the methodological decisions for the study. First, it gives descriptions of the two corpora under investigation in this thesis, beginning by giving a succinct overview of the TLC-L2 including corpus design and speaker-related characteristics. The section then moves to provide the rationale for choosing this corpus for this study (Section 3.1.5). The TLC-L1 is introduced in Section (3.2) and begins with the rationale for its development before exploring key parts to the design and data collection. Section 3.3 and 3.4 detail the procedure taken for the data analysis in each of the corpora before concluding the chapter in Section 3.5.

3.1. Description of the Trinity Lancaster Corpus of L2 spoken English (TLC-L2)

The TLC-L2 is a 4.2-million-word corpus of spoken interactional English (Gablasova et al., 2019). It is compiled from Trinity College London's Graded Examinations in Spoken English (GESE; Trinity College London, 2016) and contains candidate (L2) and examiner (L1) language engaging in an interactive language proficiency interview. The main objective of the examination is to measure speaking ability in a range of tasks designed to generate different types of spoken production (both monologic and dialogic). A trained examiner conducts the exams as an L1 speaker of English, who acts as both the

interlocutor and examiner. The data in the TLC-L2 corpus comes from 2,053 L2 speakers of English from three different proficiency levels (B1 – threshold, B2 – intermediate, and C1/C2 – advanced) and a range of linguistic backgrounds. More specifically, the GESE follows a graded system with Grade 1-12 exams available for learners to take. The TLC-L2 cor–us contains examinations from Grades 7 and above. Table 1 below shows the mapping of the GESE grades onto the CEFR scale as calibrated by Trinity College London (Papageorgiou, 2007). The TLC-L2 is available in its entirety for analysis through Sketch Engine (Kilgarriff et al., 2014). In this study, a subset of the corpus, consisting of the candidate language in the Conversation and Discussion tasks, was used (see Section 3.3. below for a rationale).

Table 1 GESE grades and their CEFR alignment

GESE Grade	1	2	3	4	5	6	7	8	9	10	11	12
CEFR level	A1			A2			B1		B2		C1/C2	

3.1.1. Nature of interaction

The corpus includes language elicited as part of an examination, between one candidate and one examiner, under timed conditions within an institutional setting. The exams are designed to elicit communicative language speaking skills in a semi-formal setting and examiners are tasked with encouraging the candidate's best performance rather than strictly following a script. Therefore, the language is more prepared than informal conversation but is not scripted or prepared in the same sense as other examinations such as the International English Language Testing System (IELTS) examination (Read, 2022) and can be considered spontaneous speech. Candidates engage in the GESE exams for numerous purposes, including work, study, immigration, employment or leisure. In total, four core speaking tasks are included in the corpus, along with a listening task and the introductory section (in which the interlocutors introduce themselves). The four speaking tasks include Presentation, Discussion, Interactive, and Conversation. Each of these tasks involves different types of communication, and the tasks undertaken depend on the examination grade being taken. Table 2 provides an overview of the four core tasks, and Table 3 shows the size of each sub-corpus (in terms of the number of words) according to each speaking task and by speaker role.

Table 2 Overview to the four GESE speaking tasks adapted from Gablasova et al. (2019)

Task	**Topic familiarity**	**Interlocutor roles**	**Type of interaction**	**Grades**
Presentation	pre-selected topic	candidate-led	monologic	10-12
Discussion	pre-selected topic	jointly led	dialogic	4-12
Interactive	general topic	candidate-led	dialogic	6-12
Conversation	general topic	jointly led	dialogic	1-12

Table 3 Number of tokens per speaker role in each speaking task

	Presentation	**Discussion**	**Interactive**	**Con-versation**	**Total**
Candidate	200,562	981,400	374,371	954,102	2,510,435
Examiner	25,642	586,668	364,029	690,446	1,666,785
Total	226,204	1,568,068	738,400	1,644,548	4,177,220

The current study is based on a subset of the TLC-L2 corpus and, in particular, on the data from two of the speaking tasks, the Discussion and Conversation tasks. The two tasks were selected as they are available at each of the three proficiency levels, allowing us to observe any developmental patterns that may be related to the proficiency level of the speakers. Regarding the topic of the interaction, the Discussion task topic is chosen by the candidates based on their interests and expertise. In contrast, the Conversation task topics are introduced by the examiner. As a result, the language elicited in two tasks represent communication on a broad range of topics, thus reducing the possibility of a topic bias, i.e., the effect that a particular topic may have on the lexical and grammatical choices of the speakers. Regarding the format and structure of the speaking tasks, at the C (C1 and C2) level, the Discussion task is preceded by a Presentation task, in which the L2 speaker talks for five to six minutes on a topic of their own choice. The Discussion task then involves a conversation between the candidate and examiner related to the topic of the Presentation task. At the two lower proficiency levels, B1 and B2, the candidates briefly introduce their topic first, and then the discussion continues from this point for

around 5 minutes. The Conversation task takes five minutes and involves the examiner inviting the candidate to engage in two topics of general interest. A list of possible topics is available to (B1, B2 and C1) candidates before the exam, so there is some familiarity with topics chosen to suit either younger or older candidates. C2 candidates engage in any topic the examiner considers most suitable for the exam.

3.1.2. Corpus design

3.1.3. Speaker-related characteristics

The TLC-L2 corpus contains several speaker-related variables. Relevant to this thesis is the information about L2 users' proficiency level.

Table 4 shows the breakdown of the number of speakers at each proficiency level. The largest group of speakers are the threshold B1 speakers with 933 candidates. This is decreased for B2 (intermediate) with the smallest group being the advanced speakers of C1/C2. Table 5 lists the most frequent linguistic backgrounds represented in the corpus. Although linguistic background will not be considered as a variable in the analysis, it is important to acknowledge the wide variety of language backgrounds due to the impact this can have on collocation use (see Section 2.4.1.3. for further exploration).

Table 4 Number of speakers at each proficiency level

L2 CEFR Band	**No. of speakers**
B1	933
B2	805
C1/C2	315

Table 5 Number of speakers from major linguistic backgrounds

Country (L1 background)	**No. of speakers**
Argentina (Spanish)	196
China (Mandarin, Cantonese)	290
India (Bengali, Gujarati, Hindi, Kannada, Konkani, Malayalam, Marathi, Marwari, Sindhi, Tamil)	248
Italy (Italian)	346
Mexico (Spanish)	312
Spain (Spanish)	347

In addition to the major linguistic backgrounds in Table 5, the TLC-2 contains data from the following language backgrounds: Albanian, Arabic, Bulgarian, Czech, Danish, French, German, Japanese, Korean, Lithuanian, Marwari, Persian, Polish, Portuguese, Romanian, Russian, Sinhala, Slovak, Telugu, Turkish and Ukrainian.

The thesis will be using a subset of the corpus, namely the candidates' Discussion and Conversation task data only. From this point, this subcorpus will be described.

3.1.3.1. Proficiency in English

Within SLA research, proficiency is one of the core variables under study as it can give insight into language development at different stages of knowledge and, from the perspective of practical application of the research, have implications for improved language teaching. However, this has been a problematic variable to investigate in learner corpus research, due to different ways in which proficiency has been established in learner corpora. For example, rather than direct rating of proficiency, proxy variables such as length of study or grade obtained have been used to estimate learners' proficiency (Paquot & Granger, 2012). One advantage of the TLC-L2 lies in the fact that it contains direct measures of candidates' proficiency. Further, each task is rated A-D (with A being the highest mark) with a final overall exam rating of Distinction/Merit/Pass/Fail. Only the exams that received a Pass were included in the TLC-L2, so it can be assumed the candidates have achieved the proficiency set by the exam. Table 6 provides an overview of the number of speakers and number of words at each level of proficiency as well as the mean number of words per speaker in the TLC-L2 Discussion task and Conversation task subcorpora.

Table 6 Overview of the number of speakers and the size of the corpus in terms of the number of tokens and mean individual contribution per proficiency level

L2 CEFR Band	**No. of speakers**	**No. of tokens in subcorpora**	**Mean no. of tokens per speaker**	**Standard deviation**
B1	933	742,854	796.20	197.41
B2	805	815,479	1,013.02	223.28
C1/C2	315	377,169	1,197.36	265.68

As can be seen from Table 6, the majority of the corpus used in this study is comprised of B1 and B2 level English speakers undertaking the GESE with 1,558,333 tokens in total. It can also be seen there is a distinct increase in the average number of tokens per speaker, and the subsequent standard deviation, as the proficiency increases. It appears that as speakers' proficiency increases, they are likely to be able to talk at a faster rate and contribute, on average, more words in each of the tasks; however, it appears that there is considerable individual variation in the length of L2 speakers' contributions. Given the smaller number of speakers at the C1 and C2 levels of proficiency, in the thesis these have been combined and treated as speakers at C-level of proficiency. Therefore, the proficiency levels featured in this study can also be considered as Threshold (B1) – Intermediate (B2) – Advanced (C1/C2), as suggested by Gablasova et al. (2019).

3.1.3.2. Linguistic and cultural background

Table 7 shows the breakdown of speakers according to their linguistic and cultural background, with the major groups being represented here.

Table 7 Overview of speaker tokens, means and standard deviation by major linguistic and cultural background subcorpora

Country (L1 background)	**No. of speakers**	**No. of tokens**	**Mean no. of tokens per speaker**	**Standard deviation**
Argentina (Spanish)	196	179,975	1,046.37	407.81
China (Mandarin, Cantonese)	290	325,360	1,121.93	521.91
India (Bengali, Gujarati, Hindi, Kannada, Konkani, Malayalam, Marathi, Marwari, Sindhi, Tamil)	248	300,346	1,211.07	528.86
Italy (Italian)	346	446,308	1,289.91	620.11
Mexico (Spanish)	312	378,251	1,212.34	631.91
Spain (Spanish)	347	428,463	1,234.76	547.26

The largest linguistic group for number of tokens is Italy (Italian); this group also has the highest mean number of tokens per speaker. This could be due in part to the level of exams these candidates are taking. Higher level GESE exam takers will have a higher mean number of tokens due to their increased English language proficiency and/or because they are more proficient and taking higher grade exams, are engaging in more speaking tasks and therefore having more opportunity to speak.

3.1.4. Additional metadata

The TLC-L2 contains additional information about the speakers included in the corpus. This metadata includes, age, gender, completed education, age of exposure to English and learning history for the candidates among others, as well as examiner age and length of time examining. The full details of this metadata can be found in the paper from Gablasova et al. (2019).

3.1.5. Rationale for selecting the TLC-L2

This study used a 1,900,000-word candidate-only subset of the corpus containing two tasks within the GESE: the Discussion and Conversation. The Discussion task involves the candidate pre-selecting their own topic while the examiner comments and asks questions. The Conversation task is based on general topics with the examiner choosing from a specific list of proficiency level-appropriate options. Both tasks are designed to be jointly led interactions and dialogic in nature and occur at all proficiency levels within the corpus. The TLC-L2 was selected as suitable for this thesis due to three major factors. Firstly, it is currently the largest spoken corpus of its kind, providing a sufficient amount of linguistic data for the meaningful analysis of collocations in spoken L2 production. Secondly, the corpus represents interactive communication, a type of communication which is very common, but which has so far not received much attention in studies on formulaic language. Finally, the corpus contains speakers at different proficiency levels and includes direct measures of L2 proficiency. This information is not always available in learner corpus research, making it difficult to assess the impact of proficiency on the collocational patterns in L2 production.

Table 8 Mean number of words and standard deviation per speaker in each task

	Presentation		**Discussion**		**Interactive**		**Conversation**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD

Candidate	636.70	164.04	478.03	130.27	334.26	109.10	464.74	174.23
Examiner	82.19	49.13	285.76	103.84	325.03	96.32	336.31	108.44

Table 8 demonstrates the monologic and dialogic nature of the tasks. The mean for the Presentation task and candidates is far higher than for the other tasks, as this is five-minute formal presentation on a discursive topic. Likewise, this is much lower for the examiners. Interestingly, the means for both the candidates and examiners are most similar in the Interactive task – this is intended to be the most fully co-constructed tasks of the GESE, and this is demonstrated when considering the average number of tokens per speaker.

3.2. Description of the Trinity Lancaster Corpus-L1 (TLC-L1)

The following sections introduce the TLC-L1 as a new corpus of L1 spoken English. As this is the first study to use the TLC-L1, particular attention is given to the corpus design including speaker-related characters (Section 3.2.3.1) and the nature of the interaction (Section 3.2.4) as well as the training of the participants (Section 3.2.5) and the data collection context (Section 3.2.6) to gain a sense of the corpus as a whole and thus to better understand the results from this research project. Section 3.2.6 provides an account of decisions made in relation to the data collection, which was a significant component of this project. Finally, Section 3.2.7 presents similarities and differences between the TLC-L1 and TLC-L2 from the examiners' perspective which helps to demonstrate the comparability of the two corpora.

3.2.1. Rationale for the TLC-L1 development

The Trinity Lancaster Corpus of L1 spoken English interaction (TLC-L1) was developed as a reference corpus for the Trinity Lancaster Corpus of L2 spoken English (TLC-L2; Gablasova et al, 2019; see also Section 3.1. for a description). The corpus was developed as part of the project led by Dana Gablasova (PI) and supported by funding from the following organisations: Trinity College London; the Department of Linguistics and English Language (Lancaster University); the Centre for Corpus Approaches to Social Science (CASS) (Lancaster University); the Faculty of Arts and Social Sciences (Lancaster University), and by the North West Social Science Doctoral Training Partnership (NWSSDTP). Trinity College London are an examination board that tests for performance not only in their language testing suite of exams such as the Graded

Examinations in Spoken English (GESE) and Integrated Skills in English (ISE) but also within music, drama and teaching, focusing on communicative and performance skills in these examinations. The corpus consists of 968,877 words from 203 conversations between two L1 British English speakers: one in the candidate role of which there are 203 unique speakers; and one in the examiner of which there are six unique speakers. Each of the candidates represents different combinations of social groups in terms of gender, age, region, educational and socio-economic backgrounds. These L1 users took part in the same speaking tasks (following the GESE; Trinity College London, 2016) as the L2 speakers in the TLC-L2, making the two corpora directly comparable in terms of the nature (genre, register and mode) of communication. To ensure comparability with the TLC-L2, the same transcription guidelines and transcriber were used for this corpus. This comparability is also why British English was chosen as the variety of English language to be collected. The TLC-L2 candidates are learners in countries where British English is typically taught and they undertook the GESE from Trinity College London, which is a British examination board.

There are several corpora available to date that represent L1 spoken British English. The British National Corpus 1994 (BNC; Aston & Burnard, 1998) and the BNC2014 (Brezina et al., 2021; Love et al., 2017) are two of the most widely used corpora which include balanced samples of British English. These two corpora are comprised mainly of written language with 10 percent of the corpus data – around 10 million words – representing spoken language. The BNC has been credited to be one of the most widely accessible corpora of its kind and has led to the creation of research-driven tools such as *Word Frequencies in Written and Spoken English* (Leech et al., 2001), a dictionary of written and spoken English. The BNC2014 has been released as an updated version of the BNC to be more representative of present-day English (i.e., British English in the period of 2012 to 2016). The corpus includes approximately 10 million words of spoken interaction, representing informal spontaneous conversation (Love et al., 2017). In addition to general corpora of British English, specialised corpora representing spoken British English in specific domains are available for analysis as well. An example of a genre-specific dataset is the British Academic Spoken English Corpus (BASE) (Thompson & Nesi, 2001). This corpus includes 1.75 million words of spoken academic language recorded at a UK university, spanning 4 academic divisions in 160 lectures and 38 seminars.

The corpora mentioned above are all effective for investigating L1 spoken British English in differing contexts and genres such as informal conversation (e.g., the BNC and BNC2014) or academic discourse (BASE). As such, they have been used as a reference corpus in a large number of studies that primarily focused on investigating L2 English use, acting as a reference point or a benchmark against which the L2 production is compared and interpreted. For example, Römer and Garner (2019) used the BNC as a "proxy for L1 usage" (p. 211) when studying verb-argument constructions (VACs) in L2 spoken English in the TLC-L2, finding that language proficiency impacted the productivity of VAC usage in L2 speakers with more advanced learners showing more similar patterns to the BNC L1 usage. More recently, the BNC2014 was used in Brezina and Fox (2021) which investigated adjective + noun collocations in spoken L2 English in a balanced sample of the TLC-L2. In the study, the researchers noted that frequent adjective + noun collocations found in the BNC2014 (such as *bloody hell*) were missing entirely from the L2 use in the TLC-L2. The researchers argued that this difference was likely due to the difference in the language represented in the two corpora and that the collocations were missing due to the more formal nature of the GESE exam context when compared to the informal conversations in the BNC2014.

However, while these corpora have been successfully used as reference points in previous studies, it has been argued that there may be issues related to the validity of the results based on the comparison of corpora that may not be fully comparable in terms of the communicative contexts (e.g., genre, register and task) in which the language was produced. Several studies demonstrated that impact of specific spoken genre and task on the language produced by the L1 as well as L2 speakers. For example, looking at collocational patterns and strengths within a corpus, Gablasova et al. (2017) note the importance of being mindful of genre and task when comparing L2 to L1 language, otherwise the explanations and interpretations of the results may be due to the corpus composition rather than driven by the speakers in the corpus. Several studies have investigated the effect of task on L2 speaking production including Wei (2011) who found that Chinese learners of English used different types of discourse markers (e.g., *and, but, also, so*) across the speaking tasks representing description, narration, comparison and apology. In another study, Neary-Sundquist (2013) noted that task structure affected frequency of pragmatic markers in both L1 and L2 English speakers. Finally, Marín

Cervantes (2019), using the TLC-L2, found that multi-word verbs in L2 spoken English were also affected by task type while noting large inter-speaker variation.

Therefore, to increase the validity of the conclusions of a study based on an L1/L2 comparison, it is essential to have a reference corpus that is comparable with the L2 corpus as fully as possible. As the speaking tasks used in the TLC-L2 were considered to be quite specific (even though they represent the broader genre of interactive, semi-formal spoken production) it was considered crucial to use a reference corpus in which L1 speakers would be engaged in the same type of interaction. While this type of discourse shares some characteristics of informal conversation (e.g., high degree of interactiveness) such as represented in the BNC1994 and BNC2014, it is also different in several other respects (e.g., more formal setting, specific roles in the speaking tasks) making the use of these two corpora as reference points less appropriate. The linguistic setting and the language represented in the TLC-L2 is perhaps more similar to a more formal academic context and the academic language produced therein; however, comparison to the BASE corpus would not be suitable as much of the data comes from university lectures which are typically much less interactive than the dialogic conversations in the GESE language exam and therefore the TLC-L2. In addition, the context of a high-stakes exam as well as the power relationship between the candidate and examiner further contribute to making the GESE language exam, and the language represented in the TLC-L2, a unique communicative setting that necessitates a new reference corpus: the TLC-L1.

3.2.2. Corpus size

The data for the TLC-L1 was collected over the period of 3 years, from February 2018 to June 2021. Overall, the contributions from 206 L2 speakers was collected, 203 of which were included in the final corpus. Three interviews were not included as the recordings were damaged. The corpus used in this study contains 833,878 tokens across four speaking tasks (the whole corpus contains additional parts, such as Greeting, which are not included in the analysis in this study). Table 9 below provides an overview of the number of tokens produced by speakers in each category (test candidates and examiners) across the four tasks as well as the overall number of tokens in the corpus. Table 10 presents the mean number of words and standard deviation per speaker in each task when considering speaker role.

Table 9 Frequency of tokens per speaker role in each speaking task

	Presentation	**Discussion**	**Interactive**	**Conversation**	**Total**
Candidate	167,056	120,612	86,556	148,981	523,205
Examiner	12,600	94,987	95,643	107,443	310,673
Total	256,424	215,599	182,199	179,656	833,878

Table 10 Mean number of words and standard deviation per speaker in each task

	Presentation		**Discussion**		**Interactive**		**Conversation**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Candidate	822.94	231.91	594.15	153.65	426.38	134.24	733.90	167.45
Examiner	67.74	53.20	467.92	133.67	471.15	143.88	529.28	134.76

During the data collection, six different examiners (L1 speakers) were used as interlocutors for the speaking tasks. The primary selection criteria for these speakers included: i) Comparability with the TLC-L2 and ii) Feasibility. With respect to the first criterion, examiners that were included as interlocutors in the TLC-L2 were selected to increase the (direct) comparability between the two corpora. Inclusion of the same set of examiners in the same speaking tasks with L2 and L1 speakers respectively allows for a direct comparison of L1 and L2 production, not available in any other corpora at the time of writing. Second, examiners' availability at the time and location of the data collection had to be taken into consideration. In particular, one examiner took part both in the face-to-face and the online data collection (for the description of the Language setting in which the interviews took place see Section 3.2.4 below) in order to balance and minimise the impact of having two formats of interviews (in person and online) on the language produced in the speaking tasks.

3.2.3. Corpus design: Structure and variables

A long tradition of research has demonstrated that speakers vary in how they use language depending on their social characteristics (e.g., age, gender, region, level of education and socio-economic background; Labov, 1966; Lakoff, 1975; Tannen, 2000) as well as depending on the linguistic setting in which the production takes place due to the

changing communicative purpose within these settings e.g., register effects from spoken, written or online communication (Biber, 2012; Goulart et al., 2019). There are general corpora that seek to represent English use across a broad range of language users, such as the BNC (Aston & Burnard, 1998) and the BNC2014 (Brezina et al., 2021; Love et al., 2017); however, they usually do not control the type of linguistic setting/communicative event more closely. On the other hand, more specialised corpora such as the LOCNEC (De Cock, 2004) follow a more defined design regarding the inclusion of specific speaking tasks but usually represent a rather narrow population of speakers in terms of social characteristics such as age and level of education. For example, LOCNEC, which serves as the reference corpus for the LINDSEI (Gilquin et al., 2010), includes L1 British English speakers at university. The TLC-L1 seeks to represent both i.e., L1 speakers in a specific type of speaking tasks, as well as speakers that represent a broader range of language users in terms of social characteristics (as far as possible). The following sections offer a more detailed description of TLC-L1 in terms of the speaker-related characteristics as well as the linguistic setting in which the language included in the corpus was produced. The sampling followed a combination of stratified and convenience sampling which will be further explained in Section 3.2.6.

3.2.3.1. Speaker-related characteristics

The TLC-L1 corpus recruitment process worked to ensure an adequate balance of a range of different social variables within speakers being represented in the corpus. The following sections provide a description of the corpus in terms of different sociolinguistic characteristics of the speakers. In particular, they focus on age, gender, education, occupation/social grade and region.

3.2.3.1.1. Age

The corpus contains speakers in the age range from 11 years old to 72 years old, closely reflecting the age distribution in the TLC-L2 with the age range of 8 years old to 72 years old. To ensure comparability in terms of age categories, the TLC-L1 was categorised using the same age groups as the TLC-L2. Table 11 provides an overview of the number of speakers and number of tokens in each age group.

Table 11 Overview of speaker tokens, means and standard deviation by age group

Age group	**No. of speakers**	**No. of tokens**	**Mean no. of tokens per speaker**	**Standard deviation**
Young (8-15)	13	22,646	1,742.00	372.81
Adolescent (16-19)	31	75,852	2,446.84	500.81
Young adult (20-35)	89	237,497	2,668.51	393.56
Middle adult (36-50)	28	73,470	2,623.93	370.77
Older adult (51 and older)	38	102,502	2,697.42	451.23
Unknown	4	11,238	2,809.50	208.06

In terms of the structure of the corpus (and amount of the data) according to the age of the speakers, almost half of the data comes from a specific age band, the young adult speakers aged 20-35 with 237,497 tokens. Older adults (51 and older) represent the second largest group of speakers with 102,502 tokens, with adolescent and middle adult speakers representing the next largest group at 75,852 token and the young speakers having the smallest representation at 22,646. Some participants declined to give their age at the time of data collection and this accounts for the Unknown category with 11,238 token from 4 speakers (see Table 11). This distribution is reflected in terms of the data (number of tokens) produced by each of these groups. In terms of the average contribution by individual speakers, the individual samples appear relatively similar across age categories with the exception of the 8–15-year-olds, who – with 1,742 tokens produced on average by the speaker – contributed considerably fewer words than speakers in the other age groups. This could be in part due to the level of literacy development of these speakers and also possibly related to their confidence to engage in the tasks (i.e., to engage in a conversation with an unfamiliar adult). Apart from this age group, when considering individual contributions, it can be noted that there is a less than 300 token difference between the group with the highest mean (older adults) and the second lowest mean (adolescents).

3.2.3.1.2. Gender

In terms of the gender of the speakers in the corpus, the dataset contains 130 female speakers and 73 male speakers with the former contributing 329,966 tokens to the corpus, while the latter contributed just under 200k words, as can be seen in Table 12. The mean number of tokens per speaker shows that male and female speakers on average contributed a very similar number of words (with the difference on average being just over 100 tokens).

Table 12 Overview of speaker tokens, means and standard deviation by gender

Gender	**No. of speakers**	**No. of tokens**	**Mean no. of tokens per speaker**	**Standard deviation**
Female	130	329,966	2,538.20	444.78
Male	73	193,239	2,647.11	517.57

3.2.3.1.3. Education

Participants' level of education is another social variable with an impact on the language produced both in native language use (e.g., Schneider & Barron, 2008) and in foreign language learning (Magogwe & Oliver, 2007). To collect the information, participants were asked to state their level of completed education. The distribution of the data in the corpus according to this variable is shown in Table 13 below. As can be seen from the table, the majority of the speakers included in the corpus have completed tertiary education (Bachelors, Masters and Doctoral combined: 59.61%), though the largest singular group of speakers is those who have completed Secondary education (33%). Overall, the mean tokens per speaker generally increases as the completed level of education increases, with Masters decreasing slightly before an increase at Doctoral level. Between the fewest and most years spent in education, there is a difference of 838.72 mean average of tokens per speaker; this is likely influenced from age and general level of literacy as well as years spent in education.

Table 13 Overview of speaker tokens, means and standard deviation by highest level of completed education

Completed education	No. of speakers	No. of tokens	Mean no. of tokens per speaker	Standard deviation
Primary	15	28,825	1,921.67	419.81
Secondary	67	181,698	2,711.91	472.10
Bachelors	59	162,068	2,746.92	441.48
Masters	44	119,571	2,717.52	435.43
Doctoral	18	49,687	2,760.39	85.05

It is interesting to note that the level of education may have played a role in speakers' willingness to participate in the research (and in the interviews). Those who have completed tertiary education are likely to have had some experience of research, either reading about it or conducting it themselves, while other participants may have been less familiar with research and therefore more apprehensive to take part. This was taken into consideration during recruitment by targeted recruitment and ensuring research materials were presented in a way that was accessible also to people without previous experience of or familiarity with research.

3.2.3.1.4. Socio-economic background: Occupation and social grade

It is widely acknowledged that social class is important to include as a variable in language research, though there are challenges with defining and operationalising this (Ash, 2004). For the TLC-L1 it was decided to follow the lead of the Spoken BNC2014 (Love et al., 2017). Firstly, to categorise the speakers according to their socio-economic background, their occupation was used as the primary variable. When establishing the relationship between the occupation and the social grade (and grouping the occupations according to the social grade), a convention used in the development of the Spoken BNC2014 (Love et al., 2017) was followed. The procedure first involved using the National Statistics Socio-economic Classification (NS-SEC), which generates an NS-SEC code based on the occupation title and is used by the UK government for census data collection. This was accessed via an online interactive website which involves entering a job title and then classifies it into a larger category representing different social grades (Office for National Statistics, 2010). The results were then mapped onto the Social Grade used in the creation of the Spoken BNC2014, following Table 14 below. It should be noted that the classification used for TLC-L1 differs in one respect from that used in the

Spoken BNC2014, namely those in education have been further categorised as ‘Pupil’ (denoting pupils in primary and secondary education) and ‘Student’ (denoting students in higher education) as it was argued that these two categories are not sufficiently distinguished in the original NS-SEC classification, in which they are subsumed under the ‘unemployed’ social group.

Table 14 NS-SEC classification standards mapped on to Social Grades (adapted from Love et al., 2017, p. 332)

NS-SEC	**Description**		**Social Grade**	**Description**
1	Higher managerial, administrative and professional occupations		A	Higher managerial, administrative and professional
1.1	Large employers and higher managerial and administrative occupations			
1.2	Higher professional occupations			
2	Lower managerial, administrative occupations		B	Intermediate managerial, administrative and professional
3	Intermediate occupations		C1	Supervisory, clerical and junior managerial, administrative and professional
4	Small employers and own account workers	**MAPS ON TO …**		
5	Lower supervisory and technical occupations		C2	Skilled manual workers

6	Semi-routine occupations	D	Semi-skilled and unskilled manual workers
7	Routine occupations		
8	Never worked and long-term unemployed	E	State pensioners, casual and lowest grade workers, unemployed with state benefits only
*	Students/unclassifiable		

Looking at the breakdown of the speakers by social grade in Table 15, it can be seen that the student group comprises the largest amount of data in the corpus with 207,320 words produced by 80 speakers. This group includes a range of speakers at different levels of higher education e.g., both first year undergraduate students as well as people finishing their doctoral degrees are included. Next, A and B, containing managerial, administrative and professional staff (e.g., chief executive officers, barristers, and nurses) both represent a large category in the corpus, with over 70,000 and 80,000 tokens respectively. Pupils as well as category C1 and E speakers represent the next three groups of speakers in terms of the size of their contribution (approx. 30-47k). The smallest groups are D with 5 speakers and C2 with 3 speakers with 16,000 and just under 9,000 tokens respectively. It is interesting to note that the Pupil group has the fewest mean number of tokens per speaker at 1,975.2; this is likely due to the young age of the speakers in this group and level of linguistic development. This also demonstrates the validity of separating the Pupils and Students from the E social group category as there is considerable variability in both the mean number of tokens and the standard deviation for each of these groups.

C1, C2 and D have the fewest speakers which is likely due to the data collection process. As discussed above, in relation to speakers' educational level and their willingness to participate in research projects, it is likely that those who are unfamiliar with what research entails may be more hesitant to engage with it. Further, the data collection mostly took place during typical working hours during the week due to the availability of the examiners involved. To counter each of these issues during the data collection process, the research call for participants was disseminated with a non-academic audience in mind and through channels such as social media and word of mouth. In addition, some evening

and weekend hours were also scheduled to ensure as much accessibility as possible to those working. It should be noted that while the number of speakers (and the size of the evidence) for individual social grades as defined in Table 15 may not be fully balanced by each grade, it is possible to further meaningfully categorise the speakers according to their profession/level of education (e.g., categories A, B and C1 all contain managerial, administrative and professional staff, while categories C2 and D both contain manual workers).

Table 15 Overview of speaker tokens, means and standard deviation by social grade

Social grade	**No. of speakers**	**No. of tokens**	**Mean no. of tokens per speaker**	**Standard deviation**
A	31	86,420	2,787.74	325.22
B	27	72,158	2,672.52	358.98
C1	13	37,123	2,855.62	438.45
C2	3	8,966	2,988.67	498.84
D	5	16,053	3,210.60	378.54
E	20	47,761	2,388.05	448.39
Pupil	24	47,404	1,975.17	484.10
Student	80	207,320	2,591.50	384.48

3.2.3.1.5. Regional distribution

While acknowledging the effect of region on speakers' language use, for example, Wardhaugh and Fuller's (2015, p.142) introduction to regional variation as well as Culpeper and Gillings (2018) who investigated politeness variation in British English and found a general tendency in use of formal, polite expressions across a North/South divide, region has not been included among the primary variables in the sampling frames for the corpus. This was partly due to the complexity of this variable; the planned size of the corpus (e.g., approx. 200 speakers) would not allow for a balanced sampling from across different regions of the UK while also ensuing a balanced distribution of other key social variables (e.g., age). However, while data collection took place primarily in Lancaster, speakers representing different regions in the British English use were included in the corpus. This was due to factors such as natural mobility of speakers across the UK which was further enhanced as part of the data collection took place in a university context with

students and staff moving for study/work from different locations. Also, as previously mentioned, the move from the face-to-face, in person exam recordings to online came with the advantage of a wider sample of regions included in the corpus due to geographical availability.

The speakers in the corpus were regionally categorised according to a self-reported response to the question "where have you spent most of your life?" From this, the participants' answers were further categorised following classification used in the Spoken BNC2014. For example, someone self-reporting 'Lancaster' would be assigned to Level 1 – UK; Level 2 – English; Level 3 – North. As noted by Love et al. (2017), this ensured maximum specificity from the self-reported responses while still allowing speakers to categorise themselves with flexibility. In particular, the Level 3 dialect categories describing the supra-region of the self-reported response were especially of interest as this level was found to be the most specific and most comprehensive with the data given from the speakers – i.e., speakers often did not include specific towns to provide the Level 4 category of 'region'. Table 16 provides the overview of the number of speakers and the size of the evidence according to the regions.

Table 16 Overview of speaker tokens, means and standard deviation by supra-region

Supra-region	**No. of speakers**	**No. of tokens**	**Mean no. of tokens per speaker**	**Standard deviation**
Midlands	24	64,717	2,696.54	483.45
Non-UK	4	11,124	2,781	202.23
North	113	300,154	2,653.23	499.29
Northern Ireland	1	2,354	2,354	0
Scotland	5	12,777	2,555.4	776.47
South	46	125,877	2,736.46	489.21
Unknown	8	19,816	2,4777	274.26
Wales	2	5,030	2,515	181.02

Around half of the speakers represented in the corpus are from the North (55.67%) with speakers from other locations accounting for the other half. This is to be expected as the data collection was initially in-person and took place in North-West England; therefore,

this variable was in part determined by the availability of the data and willingness of participants to be involved. The group with the highest mean number of tokens per speaker is Non-UK while the lowest is Northern Ireland; however, only one participant accounts for the Northern Ireland data which may be the reason why it is lower. One of the requirements to take part in the data collection was for the speaker to identify as speaking British English as their first language rather than placing restrictions on where they were born; this accounts for the Non-UK group. The participants in this group include speakers born in Oman, United Arab Emirates and the Isle of Man who identify as L1 British English speakers.

3.2.3.1.6. Language learning experience

As well as gathering metadata on the social characteristics of the speakers, information on their language learning experience was also collected. This included data on languages other than English learned and frequently used by the speakers, previous experience of spoken language examinations, and their current use of academic language. The purpose of these questions was to gain a deeper insight into the background of the speakers in terms of their experience with language learning and assessments as this could influence their language use within the GESE.

3.2.3.1.7. Additional languages learned and spoken

To better understand speakers' experience with learning and knowledge of other languages, the speakers were asked to state i) if they have learned a language other than English and ii) if they commonly use an additional language other than English. Over three quarters of the group reported they had studied another language at some point in their life (see Table 17) and many of these speakers had learned more than one (see Table 18), to varying degrees of proficiency. The learning experience varied; some mentioned they had been exposed to learning many years ago within a formal educational context while others were currently learning the language more casually.

Table 17 Additional languages learned by the speakers

Response	**No. of speakers**	**Percentage of speakers**
Yes	157	77.34
No	39	19.21
N/A	7	3.45

Table 18 Overview of additional languages learned

Language	**No. of speakers**
French	112
German	59
Spanish	53
Italian	14
Japanese	11
Mandarin	10
Latin	6
Russian	6
Polish	5
Unknown	5
Arabic	4
Dutch	4
British Sign Language; Catalan; Korean; Welsh	3 each
Greek; Portuguese; Swedish; Thai	2 each
Amharic; Bulgarian; Cantonese; Czech; Hebrew; Hindi; Kannada; Malay, Old English, Old French, Old Icelandic, Serbian; Urdu	1 each

Next, with respect to the use of an additional language, as shown in Table 19, most speakers did not report using an additional language; given that the 'non-applicable' response is likely to mean 'no' this gives a total of 137 out of 203 speakers (67.49%) not using a language other than English. Table 20 below provides a list of the additional languages reported by the speakers in the corpus. French, German and Italian represent the three most frequent additional languages used by 44 of the 66 speakers (66.67%) who use an additional language. This is likely due to the popularity of these languages being taught within compulsory education in the UK and the proximity of the countries in which these languages are commonly spoken to the UK i.e., France, Germany and Italy as they are popular holiday destinations, confirmed by many of the speakers reporting they used their additional language while on holiday.

Table 19 Additional languages used by the speakers

Response	**No. of speakers**	**Percentage of speakers**
Yes	66	32.51
No	127	62.56
N/A	10	4.93

Table 20 Overview of additional languages used

Language	**No. of speakers**
French	23
German	13
Italian	8
Mandarin	6
Spanish	4
Arabic	3
Dutch	3
Polish	3
British Sign Language	2
Japanese	2
Cantonese; Catalan Chinese; Kannada; Korean; Malay; Portuguese; Punjabi; Sign Supported English	1 each

3.2.3.1.8. Academic language use

As part of information on their language use experience, speakers were asked to report whether they currently used language in an academic setting. Experience with academic language use was considered to be a potential factor in speakers' performance in the context of a language examination which involves (semi-)formal conversations taking place in an institutional setting. Further, one of the tasks in the GESE (the Presentation task) involves giving an oral presentation, a task that may be particularly common in an academic setting, in which the ability to use academic language and conventions (e.g., structuring of information) may be of advantage. As can be seen from Table 21, a significant proportion of the speakers in the corpus reported that they currently used language in an academic setting. When asked to clarify their response, speakers reported

various professional settings they use English in including education, recruitment, healthcare, digital literacy, and administration and for more general purposes such as writing, reading, communication and presenting. Although these are not strictly academic contexts, it is interesting to note the interpretation of this question by most of the participants was whether they used academic *language* in a setting.

Table 21 Number of speakers that currently use language in an academic setting

Response	**No. of speakers**	**Percentage of speakers**
Yes	159	78.33
No	44	21.67

3.2.3.2. Language proficiency

Since L1 speakers, like L2 speakers, differ in their mastery and knowledge of the target language (Hulstijn, 2015), language proficiency of the speakers in the corpus was measured using two different methods: a) a test of vocabulary and b) assessment of speakers' performance in the GESE.

3.2.3.2.1. Vocabulary knowledge

The first measure of language proficiency for the TLC-L1 participants involved a test of vocabulary knowledge. The speakers were assessed through the use of a c-test to give us an indication of their general language proficiency. The c-test was first introduced by Raatz and Klein-Braley (1982) as an alternative to the cloze test. The c-test consists of four short texts (see Appendix 7). In each of these, the first and last sentences are given in full, with the task involving filling in parts of missing words in the rest of these texts, ensuring the added words follow grammatical and lexical rules, and semantically fit into the text. Despite the fact that the test is based on filling in lexical items, it is considered to measure a more general ability to use language (Raatz & Klein-Braley, 1982). The c-test is constructed as a graded assessment, with the language in the paragraphs progressively becoming more complex. Out of the total 25 points that could be awarded on the test, the average score among the participants was 90.41 with a standard deviation of 10.56. This lower average score for the younger speakers is to be expected due to their overall level of literacy.

3.2.3.2.2. Speaker performance

The second measure of language proficiency used for this corpus was assessment of the L1 speakers' performance in the GESE. Each participant's performance was graded by the Trinity College London examiner in the same way as is typically done for L2 speakers engaging in the GESE. For each participant, the examiner graded each of the tasks either A, B, C, or D with a final result as a P (pass) or F (fail). There was an assumption that all native speakers involved in the study would be competent enough in spoken English to pass the Grade 12, the highest level attainable for the GESE. Although most did pass, there were variable results for individual tasks which indicates this second measure also serves to highlight how well the participant engaged with the communicative purpose of the task.

3.2.4. Nature of interaction

3.2.4.1. Linguistic setting: Speaking tasks

The language for the corpus was elicited in the same way as for the TLC-L2: the participants took part in the GESE interviews with a trained examiner who was also an L1 English speaker. Unlike in TLC-L2, all the interviews followed the guidelines for Grade 12 of GESE – this is the highest level of speaking examination offered by Trinity College London and maps as a C2 (advanced) level of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). The interview consists of five speaking tasks; a brief overview to four of these tasks (which were included in the current analysis) in terms of the familiarity with the topic by the candidate, interlocutor roles and the type of interaction can be seen in Table 22. A more detailed description of each of these speaking tasks can be found in Section 3.1.1 as well as Gablasova et al. (2019) and Trinity College London GESE levels and resources (2023b).

Table 22 Overview to the four GESE speaking tasks adapted from Gablasova et al. (2019)

Task	**Topic familiarity**	**Interlocutor roles**	**Type of interaction**
Presentation	pre-selected topic	candidate-led	monologic
Discussion	pre-selected topic	jointly led	dialogic
Interactive task	general topic	candidate-led	dialogic
Conversation	general topic	jointly led	dialogic

3.2.4.2. Medium of communication

The format of communication involved in collecting the data for the corpus included both face-to-face in person and online interviews. While the interviews during the period from February 2018 to July 2019 were conducted in the face-to-face format, due to the COVID-19 pandemic, all Trinity College London exams were moved to an online interface in May 2020. As a result, all data collection for the TLC-L1 was moved to an online format as well. This gave the project the opportunity to include both face-to-face in person and online recordings of the exams. The online data collection began six months after the examination board's transition to online examinations; this delay was to ensure that the examiners had had enough experience with the new way of delivery before taking part in the data collection for the TLC-L1. Overall, approximately a quarter of the corpus data was collected using the online interface (58 interviews), with the rest conducted in the face-to-face in person format (145 interviews).

3.2.5. Training

A major consideration when collecting data for this corpus concerned preparation and training of the participants to take part in the GESE examination. It is likely that many of the L2 candidates in the GESE would have received guidance and training during English language classes and/or would have carefully studied the guidelines and materials (e.g., videos) available on the Trinity College London website (2023b) in preparation for taking the high-stakes exam. To ensure that the L1 speakers understood the format of the exam (e.g., the purpose and expectations in the individual tasks), guidelines and training was provided to them as well (Appendix 6). For the first round of data collection in February 2018, this training involved a face-to-face group session prior (e.g., a week in advance) to the exam interview taking place. This was eventually replaced by the procedure in which written guidelines were sent to the participants before the exam which included the description of the speaking tasks and the roles of the examiners and test-takers in each of the tasks. In addition to the written guidelines, all participants were invited to contact one of the researchers for further information or clarification at any point before taking part in the interviews. Shortly before taking the exam, the researcher repeated the key information about the exam/speaking tasks and checked participants' understanding of the GESE format and the requirements of the speaking tasks. Following the change to the online data collection, the guidelines and information about the exam was updated to reflect the new format of the interviews.

3.2.6. Data collection context

Data collection for TLC-L1 took place between February 2018 and March 2021. Data collection initially took place face-to-face with up to 15 examinations per day. Before the day, each participant (and their legal guardian if under-18 years old) read the participant information sheets (Appendices 3, 4 and 5). Consent was collected before the day for the younger participants (Appendix 2) while consent was collected before entering the examination room for those who were over-18 (Appendix 1). During the time before the exam, participants were also required to fill out the background questionnaire (Appendix 8). Once finished with the examiner, the participant returned to the researcher and finished the c-test while supervised. The researcher did not stay in the room (either physical or online) during the data collection to ensure comparability with the TLC-L2 setting where the candidate and examiner were alone for the dialogue.

In March 2020, the COVID-19 pandemic resulted in data collection moving from face-to-face in person examinations to online interactions. This move came with benefits and limitations to be acknowledged. A major benefit for this move was the ability to involve a wider audience geographically for the data collection. Previously, only those who could attend a specific location were able to be recorded for the corpus. This meant there was difficulty recruiting from beyond the university population which would impact the reliability of the corpus being somewhat representative of L1 speakers. This can be seen in the demographics of the corpus coming from a large proportion of university students and staff, based on educational background. Online data collocation meant that the speakers were able to participate from beyond only those with access to an out-of-town campus in one city. Conversely, the online format meant that recruitment of participants relied on a certain level of digital literacy, which was a limiting factor. However, the COVID-19 pandemic had led many people to engage with popular video calling software, so this resulted in being less of a challenge than anticipated.

Sampling involved a mixture of convenience sampling initially with recruitment based around the local area due to the need to travel for face-to-face data collection with the examiners. Gift vouchers were offered to thank participants for their time. After the move to online data collection, sampling took on a stratified approach and the online mode helped to facilitate this. The current demographic data of participants was analysed and from this, underrepresented groups were invited to take part. Due to the make-up of the TLC-L2 including under-18s, younger participants were especially encouraged to

contribute with one round of data collection being held in a local high school for this reason. Overall, the sampling for the corpus aimed to make the data as representative as possible for British English L1 speakers while also maximising the amount of data collected.

A difference noted between the in person and online formats was the fact that the online construct of the GESE may allow for those who are nervous taking the examination to 'hide' behind the screen which may have encouraged a different group of people to volunteer to take part in the data collocation. However, the online examiners both noted they felt more difficulty in putting people at ease in this format compared to in person. Regarding validity for the test, it was believed that some of the L1 participants had scripts with them for the topic presentation, despite being told not to do this. Clearly, this was impossible to control for if they were off the screen as the L1 participants were within their own homes. Although this is in an issue in the authenticity of the task fulfilment, the examiners mentioned that L2 candidates frequently presented topics that were clearly memorised – not dissimilar to reading from a script. Interestingly, this is a difference between the two groups of speakers that ends up with a similar outcome – a highly rehearsed speech.

3.2.7. Differences between the speakers in TLC-L1 and TLC-L2

3.2.7.1. Overall

Following each data collection, an interview with the examiner(s) was held to reflect on the process of conducting GESE examinations with L1 speakers as opposed to L2 speakers. Six interviews were conducted overall with four examiners, with approximate duration of 3 hours (30 minutes each on average). The guiding questions of the interview were:

a) Overall, how did the examinations go, in your opinion?
b) Were there any differences between the L1 and L2 speakers taking the Grade 12 exam?
c) Was there anything that you felt the L1 speakers struggled with in the exam?
d) Did you focus on any specific linguistic features within the exam?

Following the bottom-up coding of the interviews, the following two major themes emerged in examiners' reflection regarding the similarities and differences in their experience with examining L1 and L2 speakers: i) motivation of the speakers for taking

the exam and ii) linguistic competence of the two groups of speakers. This feedback helps to explore the comparability of the two corpora based on the perspectives from the examiners who were involved in the data collection for both the TLC-L1 and TLC-L2. This also further demonstrates the suitability of the TLC-L1 for the purposes of this research as a second dataset to use alongside the TLC-L2 corpus.

3.2.7.2. Motivation of the speakers

The GESE examinations are frequently taken by the L2 speakers as a high-stakes exam with real-life implications (e.g., for the success of visa applications or acceptance in a university course) and this can influence the preparation and performance of the speakers during the exam. For example, some of the L2 speakers may take courses to prepare for the exam, or they might have taken this or a similar exam before. The L1 speakers, on the other hand, had different motivations to take part. Speakers who completed the research study were paid for their participation; this could have led to a feeling of obligation to do well with a positive impact on the level of preparation for their exam. Further, it was noted that many of the L1 speakers were nervous, in part due to the unfamiliarity of the experience. Similarly, L2 speakers would likely have a level of anxiety during their exams but due to the high-stake implications instead. Motivation may have been a factor in how appropriately the speakers prepared for the exam too. It was noted by five of the examiners that the L1 speakers performed less successfully on the Presentation task in part due to the topic they had selected. The task calls for an argumentative topic, which then leads into further conversation between the examiner and candidate in the Discussion task. Many of the L1 speakers prepared more narrative topics instead such as discussing the impact of Brexit rather than taking a stance and presenting an opinion that could then be challenged by the examiner in the Discussion task. Further, the structure of the Presentation needed to include an introduction, main points, and conclusion, which was frequently missing, while many of the L1 speakers went over the five-minute time limit. All these factors indicate that the speakers were likely prepared to engage in the study but did not fully understand the purpose of the Presentation task, despite the emphasis on producing a discursive presentation that was put forth within the recruitment and training.

3.2.7.3. Linguistic competence

The second area of difference when interviewing L1 and L2 speakers reported by the examiners concerns the linguistic competence of the speakers. First, there is the difference between the speakers related to the mastery of the language in terms of grammar, lexicon

and pragmatics. All the examiners reported that the L1 speakers did not make grammatical errors in the same way as the L2 speakers do. They noted that although there were times when the most appropriate word was not chosen, there was a distinct lack of fossilised errors that they typically encounter in examinations with advanced L2 speakers of English. One examiner also reported that L1 speakers had a tendency to wait to ensure they were using the word they were intending, whereas L2 speakers used less of this kind of hesitation, possibly because of their focus on demonstrating fluency. Regarding pragmatics, one examiner reported that the shared background in terms of language and culture impacted the use of language; it was easier to use certain phrases and talk about current events because there was some assumption of shared knowledge. Second, the examiners reported that the difference in proficiency had implications for the type of interaction in the examinations. During L2 examinations, the examiners adopt a 'bias for best' approach to the interaction (Fox, 2004) and always adapt their language to encourage the candidate to achieve their highest possible level of proficiency. With the L1 speakers, most of the examiners noted that they did not need to do this. Further, and linked to the shared cultural understanding, the examiners felt it easier to build rapport with the L1 speakers. This is also in part due to engaging in Grade 12 examinations, where the level of interaction should be more advanced too. However, while the L1 speakers were expected to be proficient in the use of their native language, the examiners observed issues with their ability to demonstrate the communicative strategies required by individual speaking tasks (e.g., being proactive when being required to ask the examiners questions). This could in part be due to the issues with the amount of preparation for the exam noted above. Examiners noted it could also be due to a lack of confidence, especially from younger candidates and understanding of what the appropriate language/communicative strategies for the given context is. Finally, it is interesting to note the differences mentioned by examiners between the face-to-face and online context. Although the online examinations sometimes had time lag and lacked the interpersonal connection that sharing a space would give, both examiners who conducted the data collection online observed that they believed that some L1 and L2 speakers felt more confident when taking part in the interview in this format (than in a shared physical space), and in the case of the L1 speakers, in a home environment rather than an unknown setting.

3.2.8. Summary

This section provided the rationale for building a new corpus, the TLC-L1, as well as a description of its composition. The corpus represents an important addition to the L1 corpora of spoken British English constructed to date: First, it represents a reference corpus for the TLC-L2 corpus, making the two corpora directly comparable in terms of the speaking tasks. As such, it offers a crucial reference point for interpretation of patterns in the TLC-L2. Second, the TLC-L1 is important as a corpus of L1 in its own right, as a new dataset available for studying spoken L1 English, since – compared to corpora representing written English – there are still just a limited number of corpora representing spoken language. It is especially unique among L1 corpora of spoken English in that it represents language from individual speakers performing across a number of tasks, making it possible to study variation both within and between L1 speakers. Finally, the TLC-L1 also has the potential to be used to investigate language produced in two different settings, as both face-to-face and online conversations are included in the corpus. Therefore, there was a clear need for building this new corpus as the TLC-L1 is an essential resource both for this thesis and for wider research opportunities.

3.3. TLC-L2 analysis procedure

The TLC-L2 data analysis was undertaken through a combination of quantitative and qualitative methods and programs. Firstly, the TLC-L2 dataset is hosted in Sketch Engine (Kilgarriff et al., 2014) which was used to search the corpus data for the target collocations. The query uses a complex combination of restrictions using the CQL conventions (https://www.sketchengine.eu/documentation/corpus-querying/). The query searched for verb ([tag="VV.*"]) + noun combinations ([tag="NN.*"&word!="lot"]), including any potential intervening elements such as hesitations or repetition ([tag="(D|A|U|PPH1|PPHO1|PPHO2|PPIO1|PPIO2|PPY).*"]{0,3}([tag="(R|U|MC|MD).*"]{1,3}[tag="(J|U|CC).*"]{1,2}|[tag="(J|U|CC).*"]{0,3})[tag="(NN|U).*"&word!="lot"]{0,3}), and further ensured the combinations occurred within the same speaker turn (within <u/>). The query also includes results with nouns as both direct and indirect objects to give a fuller picture. The query is as follows:

[tag="VV.*"][tag="(D|A|U|PPH1|PPHO1|PPHO2|PPIO1|PPIO2|PPY).*"]{0,3}([tag="(R|U|MC|MD).*"]{1,3}[tag="(J|U|CC).*"]{1,2}|[tag="(J|U|CC).*"]{0,3})[tag="(NN|U).*"&word!="lot"]{0,3}(meet[tag="NN.*"&word!="lot"] [tag!="NN.*"]0 1)within <u/>

After testing, it was found that the query has a high precision and high recall. The precision was 95% with non-accurate hits involving inclusion of proper nouns, such as *English* and *French*, and instances where the speaker has used an incorrect form that has led to mistagged words e.g., *towers attacks* has been tagged as verb + noun. When consulting the concordance lines to infer meaning, the speaker likely intended *towers attacks* as two nouns and part of a larger noun phrase:

(1) Candidate 5_6_AR_28: my father er used to tell me a lot about this this about the twin **towers attacks**

Recall was tested on a sample of six texts with three male and three female speakers across four language backgrounds and different proficiency levels; this was done to ensure there was an even spread of data to be checked. These texts were manually annotated for instances of verb + noun collocations and compared with the results of the automated procedure. The recall was 96.09%. Missed hits included an idiom "feather his own nest", an instance of 'unclear text' where the transcriber has made an informed decision on the utterance and a compound noun exclusion "get a driving license". The obtained precision and recall levels were deemed high enough for query to be used for the present study.

As the focus of this part of the research is on L2 speech, only the results from the candidates were included. To further refine the results, the query was restricted to the Discussion task and the Conversation task only. The decision for this was justified as these two tasks are present in each of the major proficiency levels under investigation (B1, B2, C1/C2) and there is a balance between candidate-led (Discussion) and examiner-led (Conversation) interactive tasks. This split between the two speaker groups in the dialogic tasks means there is not a bias towards either. This choice also means the data included is comprehensive and relevant to the research questions for the study. From the above query and in the specific tasks and speakers, 43,644 verb + noun combinations were extracted from the TLC-L2. This is used as the core dataset for the research with further cleaning taking place for some of the analysis which is detailed in this section with justifications. The 43,644 instances were used so an overall grouped frequency of occurrence could be found and compared across the groups with the understanding that further analysis could then consider the internal modifications and use of the collocations in context.

The first major research questions of this thesis is to investigate the frequency and distribution of use of verb + noun collocations amongst L2 English speakers at B1, B2 and C1/C2 level. It was decided that one approach to studying this would be to focus on shared collocations. This was to ensure there was some evidence of the collocation distributed across all the groups and was defined as collocations that appeared within each of the three levels. For example, if the collocation was present in B1 and C1/C2 but not B2, it would not be considered a shared collocation and therefore not included in the dataset for this strand of the analysis. The processing of the results to get to this stage involved combining the lemmas for both the node (verb) and the collocate (noun) and removing internal modifications such as adjectives and hesitations; these internal modifications will be later considered in the concordance analysis where relevant. Overall, this processing resulted in collocations being combined by type. For example, *taking time, took time, take a long time* are three occurrences of the verb + noun collocation *take + time*. Further processing considered each concordance line of the verb + noun collocations and whether these adhered to the phraseological principles of a verb + noun collocation that is necessary for further analysis. Some instances were then removed, for example, the *thank* + noun construction that occurred frequently such as *thank you bye, thank you sir* and *thank you goodbye.* As will be explored within the analysis sections of this thesis, these could be considered register-influenced collocations in that they are expected within a certain conversational context; however, in the majority of the instances, they occur at the very end of the examination as the candidates are leaving. This means that, although categorised as appearing in the Conversation task, the collocations are not part of the dialogue of this task and subsequently should be omitted from further analysis. Finally, specific collocations were also removed from further analysis as they did not fit the required phraseological pattern the research is focused on. These are *think + people, know + people* as these tended to not fit the verb + noun construction grammatical pattern. Further, *learn + English, speak + English, study + English, know + English* all included a proper noun. This was due to a tagging inconsistency that meant the query was not able to filter these. After this data cleaning, there were 9,674 shared collocations ready for further investigation. From these, after excluding internal modifications and grouping by lemma, 201 verb + noun collocations were found to be shared by the B1, B2 and C1/C2 groups within the Discussion and Conversation tasks.

The remaining analysis of the 201 shared verb + noun collocations considers:

1. frequency and dispersion of the verb + noun combined
2. frequent verb types
3. frequent nouns that collocate with the frequent verbs

The above will be looked at quantitatively based on occurrences per proficiency level and task, and by percentage of speaker use where appropriate. Further analysis is undertaken qualitatively using concordance lines to see how the verb + noun collocations are being used contextually. This analysis will work to partly answer Research Questions 2 and 3 regarding the use of topic-influenced and register-influenced verb + noun collocations in the TLC-L2.

Finally, an investigation into high frequency delexical verb + noun collocations will take place. In particular, three verbs have been chosen for further study based on two factors: their high frequency of occurrence found in the TLC-L2 and previous research has found these especially fruitful to study (e.g., Gilquin, 2007; Kim, 2002; Ma & Kim, 2013). These verbs are *get*, *make* and *take*. Once frequencies have been found, the analysis will focus on one collocation per verb, per proficiency level for further in-depth concordance and qualitative analysis. This decision was made in part based on the most frequent collocation per level; however, this will also take comparative frequency into account between levels too. After exploration of the concordance lines to look for patterns in the high frequency delexical verb use, helping to in part answer Research Question 4 based on the TLC-L2 data "are there patterns in how these speakers use high frequency delexical verb + noun collocations in spoken examination language?", the analysis will take a final step to consider the semantic categories of nouns speakers use with these high frequency delexical verbs. To do this, the decision was made to extend the analysis back out to the original 43,644 dataset and find the combinations that are unique to each proficiency level. Doing this analysis with the larger dataset and not the 201 shared collocation types means that there is additional breadth to the analysis which is needed before going into further depth regarding semantic categories. Furthermore, the previous analysis has considered what is present across all proficiency levels while this step considers what is unique to each of the three groups, adding a new layer of analysis to the research. Once the unique collocations were found using Microsoft Excel, they were grouped by lemma removing internal modifications. Finally, the decision was made to partly replicate the

procedure from Du et al. (2022) by tagging the unique noun lemmas for each of the three verbs (*get*, *make*, *take*) across the three proficiency levels (B1, B2, C1/C2) according to semantic categories as defined by the UCREL Semantic Analysis System (USAS) tagger (Rayson et al., 2004). From this, the percentage of each noun semantic category for the verbs was found and will be further explored in the analysis to see how the patterns of high frequency delexical verbs change across proficiency levels, adding further depth to answering RQ4 of this thesis.

3.4. TLC-L1 analysis procedure

In broad terms, the procedure to analysis the TLC-L1 data was based on the same procedure for the TLC-L2. This was decided so as to ensure comparability within the results and in order to answer the research questions posed for exploration in the thesis, while also acknowledging the differences between the data – the TLC-L1 containing one proficiency group while the TLC-L2 includes three. The same query as detailed in Section 3.3 was used to extract the verb + noun combinations for the TLC-L1 corpus. This was decided as it had been tested for precision and recall and found to be accurate enough to proceed with for the TLC-L2 analysis. Although a different dataset, the language was considered to be similar enough due to context (an interactive dialogic language examination) to continue to use this query. From this query, 6,312 instances of verb + noun combinations were found. It was decided to further process these combinations before data analysis took place to ensure the final dataset could be classified as collocations according to their frequency and dispersion. The reasoning behind this choice of an extra step in the processing when compared to the TLC-L2 analysis was to be more stringent on the parameters of what is a collocation for the L1 speakers than the L2 speech as the latter may have more creativity within the combinations. If this was done with the L2 corpus data, there was potential to remove instances of verb + noun collocations that would be of benefit to acknowledge and analyse further as L1 and L2 speakers use language differently which is further explored in Section 2.4.1 of the literature review. Another reason for this additional step was to approach the categorisation from a frequency-based perspective by considering frequency and dispersion of the collocations as well as from a phraseological approach with the query created. This blended approach has been discussed further and justified in Section 2.1.1 and Section 2.1.2.

This additional step of processing involved comparing the initial 6,312 instances found in the TLC-L1 to the BNC2014 based on the frequency and dispersion of occurrences in both corpora. To do this, the collocations found were lemmatised with internal modifications removed. Then, each collocation was searched for within the BNC2014 using #LancsBox X (Brezina & Platt, 2023). The CQL query for this changed depending on the collocation; the following is an example search for *make* + *decision* [hw="make" pos="V.*"] []{0,3} [hw="decision" pos="N.*"]. The frequency and dispersion of each collocation was found. Based on Nesselhauf (2005), it was decided to implement a cut-off of 50 instances for dispersion to ensure the results were both meaningful and manageable. This means that moving forward, unless otherwise stated, the verb + noun collocations under investigation from the TLC-L1 also occur at least 50 times in the BNC2014 and across 50 different texts in the latter corpus. From this process, 2,150 verb + noun collocation tokens (1,181 types) were selected for further analysis in Chapter 5. As a major part of this analysis is looking at frequency data, it was decided that ranked (raw) frequency counts were appropriate to use as comparison were not being made, unlike the proficiency level comparisons in the TLC-L2 data.

The process of investigating frequent collocations alone ran a risk of overlooking the distinctive collocations that L1 speakers may use. Therefore, as well as considering the widely dispersed collocations, the analysis also took time to focus on unique collocations to the L1 dataset to see if there was anything notable. This again leads to a more holistic overview of the dataset by allowing space to see creativity in the language as well as looking at the most frequently occurring collocations in the L1 speakers.

To uncover the TLC-L1 unique collocations, the original dataset of 6,312 verb + noun collocations with frequency and dispersion information from the BNC2014 was revisited. From this, it was found that 316 of the collocations only occurred in the TLC-L1 corpus and not the much larger BNC2014. After concordance analysis, it was found that all instances were grammatically sound verb + noun combinations according to their phraseological construction. There was nothing of note regarding their grammar; all were grammatically correct and usual. From this, 32 of the combinations (around 10%) were randomly selected using a random number generator. These were then searched for in the EnTenTen20 (Jakubíček et al., 2013) using the same CQL query as the BNC2014 searches; this additional reference corpus is compiled from web texts and was one of the largest corpora at the time of the analysis with over 43 billion tokens. After searching the

EnTenTen20, collocations were categorised as yes/no if appearing in both corpora. If no, the concordance lines in the TLC-L1 were consulted and explored in Section 5.2 as unique verb + noun collocations.

3.5.Summary

This chapter gives an overview of the TLC-L2 corpus design and rationale for using for this study. It also introduced the new TLC-L1 corpus, detailing some of the methodological decisions involved in the data collection process and the rationale for its creation. Finally, the chapter outlines the data analysis procedure and provides justifications for the decisions involved. Overall, the TLC-L2 and TLC-L1 analysis take slightly different approaches due to the varying nature of the data. As the TLC-L2 analysis has the added variable of language proficiency level that is being investigated in this thesis, the analysis takes the approach of looking at shared collocations (collocations that appear in all three proficiency levels) as well as unique collocations (collocations that only appear in one proficiency level and not the remaining two). The purpose of this approach is to investigate what is mutually occurring to and to see what is exclusive in order to see how formulaic language use changes and develops across proficiency levels. Furthermore, the TLC-L2 data was not compared to the BNC2014 in this way to avoid placing the BNC2014 as the norm for the analysis due to the issues with this explored in Section 2.4.4. It was decided to investigate the learner language from a descriptive standpoint initially, so the data was not cleaned by filtering it to the standards or expectations or comparison to what is present in the BNC2014. However, this was done with the TLC-L1 and the BNC2014 as both are native speaker corpora. Overall, this chapter introduces the new TLC-L1 corpus, further explores the TLC-L2 corpus and outlines the decision making involved in the analysis of these corpora for this thesis.

Chapter 4: TLC-L2 Results and Discussion

This chapter presents the results from the TLC-L2 analysis and begins the discussion of these results in the context of current literature to answer the proposed research questions. First, Section 4.1 explores all instances of verb + noun collocations before detailing occurrences of shared verb + noun collocations in Section 4.2. Next, Section 4.3 investigates frequent verb types within verb + noun collocations with collocation patterns in high frequency delexical verb + noun collocations explored through three verb case studies in Section 4.4. Finally, Section 4.5 brings the chapter to a close.

4.1. All verb + noun collocations

The analysis will first consider all verb + noun collocations that have been extracted based on the query developed in Chapter 3. This will give an overview to the nature of verb + noun collocations as a whole in the TLC-L2, describing frequency and dispersion of these collocations so that decisions can be made to clean the data for further analysis. Overall, the TLC-L2 contains 43,644 verb + noun collocations based on the original query.

Table 23 Verb + noun collocations: frequency in proficiency group and means per speaker in the TLC-L2

Proficiency	**Absolute Frequency**	**Tokens**	**Relative Frequency (per 1k)**	**Mean (SD)**	**Range**
B1	18332	742,908	24.68	19.65 (8.24)	55
B2	17320	815,523	21.24	21.52 (9.60)	58
C1/C2	7992	377,178	21.19	25.37 (10.54)	65

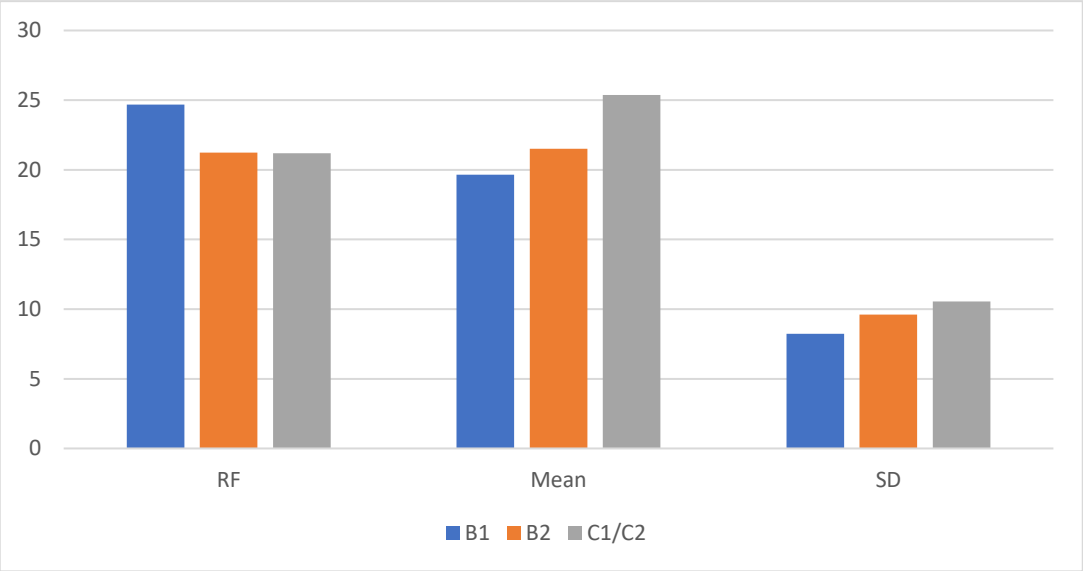

Figure 1 Frequency breakdown of all verb + noun collocations across proficiency groups in TLC-L2

Table 23 shows the breakdown of mean tokens per speaker for each proficiency level and illustrates an increase in the mean number of tokens per speaker in these tasks as the proficiency level increases. Further data visualisation is shown in Figure 1. This means that the advanced levels (C1/C2) are using more tokens in the same tasks than the lower proficiency groups. The standard deviation demonstrates that the individual variability of the number of tokens per speaker increases as the proficiency level increases; in other words, there is a wider range of values meaning that there is more individual variation of the frequency of use between speakers. Language use is varied and so is language development; therefore, these results further support the idea that individual variation is important to keep in mind when studying phraseological development echoing recent findings from Omidian et al. (2021).

All speakers in this corpus used at least one verb + noun combination as defined by the query. As the collocation type is present in all speakers, it can be stated that this is an appropriate collocation type choice to use to investigate phraseological competence in these speakers, and as well as in the L1 corpus in Chapter 5. This is because corpus data can tell us something about language use that is present in the sample, but if the evidence is not there, it is difficult to claim why this is the case. This is particularly important when considering learner language and investigating individual collocations (Kreyer, 2021). For example, if a certain language feature is not present, it could be that the speaker did not have the opportunity to use it but equally could be that they are unaware of the feature or that they do not have the language competence to use it adequately at the time of the

data collection. There is a decrease in the relative frequency of collocations in C1/C2 group overall data, but an increase in the average number of collocations used per speaker within the group; this could mean that certain speakers are using more of this type of collocation but that this usage is occurring alongside a variety of other language choices thus leading to increased diversity in the texts. This links to Granger's (2018) observations that quality and quantity of collocations increases as learners develop their language proficiency, though L2 speakers seem to use fewer instances of formulaic language when compared to L1 speakers, even at more advanced proficiency levels (Ädel & Erman, 2012). Saito and Liu (2022) have also noted this increase in diversity of language with developed conversational experience.

As shown in Table 23, there could be an effect of text length as the mean tokens per speaker is higher for the C1/C2 group and this could be impacting the average number of collocations used per speaker i.e., they are using more collocations simply because they have more opportunity to as they are using more language overall. To investigate this, a one-way ANOVA was used to see if there was a difference between the groups outside of text length and found a statistically significant effect of proficiency on the use of verb + noun collocations*:* $F(2, 2050) = 48.06$; $p < 0.001$. The size of the effect is small, $\omega = 0.209$. To understand where the difference lies between the groups, post-hoc tests (Bonferroni) found the significant difference is between B1 (threshold) and C1/C2 (advanced) groups ($p < 0.001$). This supports previous findings from Forsberg and Bartning (2010) that there are differences in collocational usage, but these are not typically in adjacent levels i.e., B2-C1/C2. These findings give credence to going ahead with further analysis.

The above analysis considers every possible instance of a verb + noun collocation based on the original query. The reason for including this section is to demonstrate that there are collocations occurring more frequently at the C1/C2 advanced proficiency level and that narrowing to the subset of collocations will be of value to show the development of the collocations that are actually there. The research will now consider a subset of these collocations after data cleaning has taken place; this is further explained in Chapter 3. These shared collocations have been chosen to undergo more in-depth qualitative analysis for two reasons: 1) they are present or 'shared' across the 3 speaker groups and 2) this means their usage can be compared between the groups.

4.2. Shared collocations

4.2.1. Overview

The TLC-L2 contains 9,674 verb + noun collocations that have been extracted based on the query developed within the methodology in the previous chapter; these collocations occur in every speaker group so that they can be said to be shared collocations. Within these collocations, there are 3,700 types; this includes all internal modifiers such as *write a fantasy book* and *write a good book* and these are classed as two different types.

Table 24 Descriptive statistics of shared verb + noun collocations in TLC-L2

Proficiency	**Absolute Frequency**	**Tokens**	**Relative Frequency (per 1k)**	**Mean (SD)**	**Range**
B1	4390	742,908	5.91	5.05 (3.35)	33
B2	3518	815,523	4.31	4.70 (3.26)	21
C1/C2	1766	377,178	4.68	5.77 (3.75)	25

Table 24 shows the highest relative frequency usage of these verb + noun collocations is at the lowest level of proficiency in the corpus – B1 speakers – this is before there is a dip in the relative frequency use of these collocations at the B2 level and increasing again at the advanced level. Interestingly, the mean per speaker is higher at C1/C2 level than B1 level, but there is again a dip in this number at the B2 level.

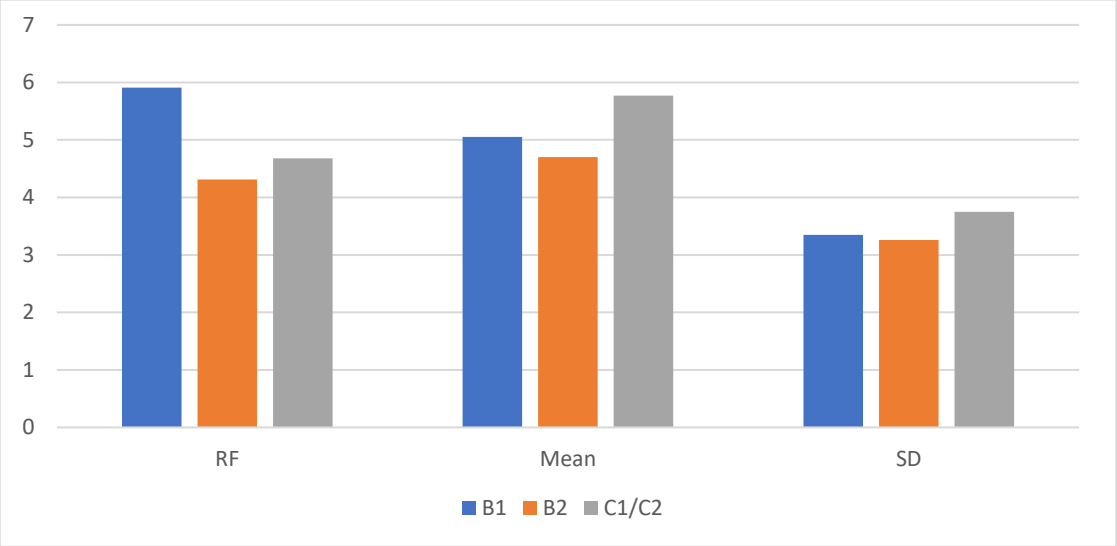

Figure 2 TLC-L2 speaker averages for using verb + noun collocations

Figure 2 shows the relative frequencies, means and standard deviations of the use of verb + noun collocations across the three language proficiency groups in the TLC-L2.

Comparing Table 24 with Table 23, it can be seen there is not a linear development of collocation usage based on frequency alone. In fact, in both the entire dataset and the shared collocation dataset, B1 group has the highest relative frequency of the groups. The mean usage increases steadily in the 'all collocations' data set but there is a noticeable decrease at the B2 level for mean and range. These descriptive statistics show there is not a set pattern of development when viewing verb + noun collocations in this way echoing results from Vedder and Benigno (2016) who noted that this development can follow a u-shape curve. This dip in the B2 usage of collocations in L2 learners is further supported by Siyanova-Chantura and Spina (2020) who noted language development could in fact get worse before it improved over time.

The research will now only consider the verb and the noun of these collocations. Removing the internal modifications and grouping by lemma, there are 201 types of verb + noun collocations shared by the 3 groups. The forthcoming analysis will consider frequency and dispersion of these collocations. Table 25 shows the 30 most frequent verb + noun collocation types in the TLC-L2 based on raw frequency and shown per speaker proficiency group.

Table 25 30 most frequent verb + noun collocation types in the TLC-L2 by raw frequency

	B1	**B2**	**C1/C2**	**Total**
read + book	189	117	27	333
learn + language	254	25	18	297
take + care	74	77	104	255
spend + money	132	28	36	196
play + game	91	65	38	194
spend + time	58	88	42	188
earn + money	66	80	36	182
give + money	132	34	15	181
play + football	113	42	7	162
save + money	131	17	4	152
help + people	56	66	27	149

get + job	52	65	32	149
get + money	75	39	30	144
find + job	42	67	25	134
listen + music	71	49	9	129
need + money	93	24	10	127
buy + clothes	84	23	15	122
buy + thing	73	23	25	121
watch + TV	70	45	6	121
see + people	39	43	32	114
use + phone	75	21	14	110
pay + attention	44	44	16	104
take + photo	46	55	1	102
make + people	20	57	23	100
play + tennis	70	25	3	98
know + thing	32	29	36	97
think + thing	27	37	31	95
like + music	45	38	11	94
watch + movie	45	33	16	94
make + feel	34	47	12	93

Looking at Table 25, it can be seen that the vast majority of the most frequent verb + noun collocations across all proficiency levels can be said to be related to the topic of hobbies or general interests. Reading, learning, and playing as well as watching television are all frequently mentioned in this initial, overall view of the most commonly used collocations of this type. Some notable exceptions to this theme are *take + care*, *help + people*, *see + people*, *pay + attention*, *make + people* and *make + feel* though again, they are in the minority in this top 30. It should be noted that combining all the instances in this way does not account for relative frequency and as the B1 group is the largest speaker-wise and the B2 group the largest token-wise, further analysis is needed. On the surface however, this could be indicating there are topic-influenced collocations being used highly frequently, namely related to hobbies and interest, and this choice of topic has previously been found to be one influence on speakers' overall language choices (Khabbazbashi, 2017; Suzuki, 2015; Yoon, 2021).

Table 26 Top 10 verb + noun collocations – ranked by raw frequency across proficiency levels and overall

B1	**B2**	**C1/C2**	**Overall**
learn + language	read + book	take + care	read + book
read + book	spend + time	use + internet	learn + language
give + money	earn + money	spend + time	take + care
spend + money	take + care	play + game	spend + money
save + money	find + job	spend + money	play + game
play + football	help + people	earn + money	spend + time
need + money	play + game	know + thing	earn + money
play + game	get + job	get + job	give + money
buy + clothes	think + government	see + people	play + football
get + money	make + people	think + thing	save + money

Table 26 shows the 10 top frequent collocations that occur ranked across the speaker groups. This again demonstrates the common occurrence of hobbies or general personal interests within the data. This is expected based on the nature of the GESE and language tests in general as speaking about oneself is a core developmental stage in learning languages, and this is usually occurring at and beyond the threshold of the B1 examination level when considering the CEFR descriptors (Council of Europe, 2001). For example, one of the Conversation task topics as specified by Trinity College London for Grade 5 (CEFR – B1) is "recent personal experiences" (Trinity College London, 2021, p. 27).

Table 27 Relative frequencies of the top 10 most frequent verb + noun collocations per 1,000 tokens

	B1	**B2**	**C1/C2**	**Total**
read + book	0.25	0.14	0.07	0.17
learn + language	0.34	0.03	0.05	0.15
take + care	0.10	0.09	0.28	0.13
spend + money	0.18	0.03	0.10	0.10

play + game	0.12	0.08	0.10	0.10
spend + time	0.08	0.11	0.11	0.10
earn + money	0.09	0.10	0.10	0.09
give + money	0.18	0.04	0.04	0.09
play + football	0.15	0.05	0.02	0.08
save + money	0.18	0.02	0.01	0.08

Table 27 uses different shades of highlighting to show which collocations are occurring most frequently in which proficiency groups; the darkest highlight is the most frequently occurring. This shows that the top 10 verb + noun collocations overall are driven by their frequent occurrence in the B1 group as 7 out of the 10 top collocations are all most frequent in the B1 group (70%). This could be due to the effect of exam topic on the language used and therefore classed as 'topic-influenced collocations'. Further discussion of this can be found in Section 4.3.2 of this thesis. There is also evidence of nonlinear usage looking at the frequencies, supporting evidence from Duan and Shi (2021) that formulaic language, such as collocations, have a tendency to develop in this way . The B1 group has 3/10 (30%) of collocations occurring the least when considering relative frequency compared to the other two groups while the B2 group has a mixed split between most and least frequently occurring collocations. The C1/C2 group have the most relatively frequent occurrences of *take + care* which can be considered to be the most abstract collocation out of the 10; the others are more concrete actions related to hobbies, sports and jobs. This supports previous findings from Du et al. (2022), who note that beginners tend to use more concrete nouns within collocations compared to the more advanced speakers using nouns with more abstract semantic fields such as feelings. Further exploration of semantic categories is done in Section 4.4.

Finally, the analysis considers individual speakers to gain more of a sense of the distribution of the collocations, as a high frequency could be demonstrating very few speakers using the collocation repeatedly. Both frequency and dispersion need to be analysed when understanding the nature of collocation usage as two dimensions of formulaicity (Saito & Lu, 2022).

Table 28 Percentage (and raw count) of speakers using verb + noun collocations in each group

	B1	**B2**	**C1/C2**
read + book	9.97 (93)	7.33 (59)	5.71 (18)
learn + language	18.11 (169)	2.48 (20)	3.81 (12)
take + care	4.93 (46)	6.58 (53)	17.46 (33)
spend + money	10.08 (94)	2.98 (24)	8.89 (28)
play + game	5.57 (52)	5.71 (46)	7.30 (23)
spend + time	4.93 (46)	8.32 (67)	10.48 (33)
earn + money	5.36 (50)	6.96 (56)	8.25 (26)
give + money	9.97 (93)	3.23 (26)	3.81 (12)
play + football	6.86 (64)	3.73 (30)	1.27 (4)
save + money	9.43 (88)	1.61 (13)	1.27 (4)

In Table 28, the darkest highlight shows the group with the largest percentage of overall speakers using this collocation. The B1 proficiency group has 5 collocations as the highest percentage of speakers using the collocations, so there are more of the speakers using the collocation. Overall, more individual B1 speakers are using the same top 10 collocations at this lower proficiency level (distribution), coupled with the previous table highlighting that the B1 speakers as a group use these top 10 collocations the most frequently. This could give an indication of the overall collocation diversity in that it could be the B1 speakers are sticking with the so-called lexical teddy bears (Hasselgren, 1994); these are the collocations they know and rely on when speaking and possibly engage in overreliance on these, similar to findings from Suzuki (2015).

The above establishes the impact that one proficiency group can have on the data if we consider it all as 'L2 speaker data'. This means it is necessary to look further at breaking down what is happening within and between the language proficiency groups as patterns could be due to different topics under discussion at each level and therefore differing opportunities to use language. This speaker opportunity of use is further explored in Buttery and Caines (2012). Alternatively, it could be because there is a developmental difference in the language proficiency. Table 29 shows the top 10 collocations per proficiency group and demonstrates that *learn + language* ranks highly in the B1 group data but does not appear in the top 10 for either the B2 group or the C1/C2 group. *Read*

+ book is the most frequent collocation in the B2 group with 0.14 occurrences; however, this is not as frequent as in the B1 group, with 0.25 occurrences. Despite this, it is still worthy of further investigation to see if topic has an influence as *read + book* does not occur in the C1/C2 top 10.

Table 29 Top 10 collocations per proficiency group ranked by relative frequency

B1		**B2**		**C1/C2**	
learn + language	0.34	read + book	0.14	take + care	0.28
read + book	0.25	spend + time	0.11	use + internet	0.13
give + money	0.18	earn + money	0.10	spend + time	0.11
spend + money	0.18	take + care	0.09	play + game	0.10
save + money	0.18	find + job	0.08	spend + money	0.10
play + football	0.15	help + people	0.08	earn + money	0.10
need + money	0.13	play + game	0.08	know + thing	0.10
play + game	0.12	get + job	0.08	get + job	0.08
buy + clothes	0.11	think + government	0.07	see + people	0.08
get + money	0.10	make + people	0.07	think + thing	0.08

4.2.2. Topic-influenced collocations

The previous sections have established that topic may have had an influence on the types of verb + noun collocations used in the corpus. This section focuses on four collocations that may be influenced by the topic within the exam. This would make sense in relation to the fact that each GESE grade comes with specific topics that the examiner will introduce in tasks e.g., Grade 5 (CEFR- B1) topics for the Conversation task can be selected from the following areas: festivals, means of transport, special occasions, entertainment, music and recent personal experiences (Trinity College London, 2021, p. 27). These will be cross-checked with the Trinity College London Exam Information: GESE Specifications (2021) for further investigation. Choices have been based on (1) their ranked relative frequency within the speaker proficiency group and (2) how this compares to the collocation ranking within other groups. One collocation has been chosen for each proficiency group; B1 (*learn + language*) and B2 (*read + book*) while two collocations have been chosen for C1/C2 proficiency group (*use + internet* and *take + care*). This is because there are no subject specific topics for Grade 12, the highest of the

GESE grades taken by the most advanced speakers. The topics for Grades 10 and 11 are also more complex and elaborately presented to candidates in that they must pick from a list from two options given by the examiner who will then choose a subject (Trinity College London, 2012). Therefore, two collocations were chosen in order to look at this in more depth.

4.2.2.1. *Learn + language (B1)*

Table 30 Occurrences of learn + language collocation by task across proficiency groups

	Frequency	**Speakers**
B1	**254**	**171**
Conversation	237	163
Discussion	17	8
B2	**25**	**20**
Conversation	14	12
Discussion	11	8
C1/C2	**18**	**12**
Conversation	7	6
Discussion	11	6
Total	**297**	**203**

Beginning with *learn + language,* Table 30 shows that there are more occurrences in the Conversation task than the Discussion task overall when considering the proficiency levels combined. However, the split between the two tasks is fairly similar for B2 groups at 14 occurrences in the Conversation and 11 in the Discussion task and likewise for the C1/C2 group with occurrences of 7 (Conversation) and 11 (Discussion). By far the most notable aspect of Table 30 and the *learn + language* collocation is that the B1 group has the most instances at 254/297 (85.52%) and within this group the task that contained the majority of these occurrences is the Conversation with 237/254 (93.31%). This indicates that it is likely the verb + noun collocation has been influenced by a specific topic choice as the Conversation task is a dialogue with topics chosen by examiners. These topics are aligned to specific subjects based on the GESE grades the candidates are sitting (Trinity College London, 2021). For example, B2 speakers are undergoing exams at grades 7-9. Topics for Grade 7 candidates include: education, national customs, village and city life,

national and local produce and products, early memories and pollution and recycling (Trinity College London, 2021, p. 35). For B1 candidates undertaking Grades 5 & 6 their topics include: travel, money, fashion, rules and regulations, health and fitness, and, crucially, learning a foreign language (Trinity College London, 2021, p.29).

Within the Conversation task and the B1 speaker group, there are 237 instances of the *learn + language* collocation with 163 different speakers using it. Further evidence of the fact that this is a topic-influenced collocation can be seen by the presence of one specific adjective when looking at the concordance lines. Many of the instances of *learn + language* are actually some variation of the larger phrase *learn + **foreign** language;* in fact, 116/237 (48.95%) of the instances include this:

Conversation task; B1 speaker

(1) Examiner: good erm and er what do you think is the best way to **learn a foreign language** in the classroom or outside the classroom ?
Candidate 2_6_IN_42: I think we should **learn a foreign language** outside the classroom because we can explore our world outside the cla= outside the classroom inside the classroom with the four walls we cannot do anything

Example (1) shows the candidate using the same phrasing as introduced by the examiner in the previous conversational turn.

As a comparison, Example (2) is offered below. This example is found in the Discussion task where the speaker has chosen a specific topic to present before the examiner asks them questions on this. The candidate is also an advanced speaker which is highlighted by the increased complexity of the conversational exchange, the examiner taking a more challenging stance in their questioning of the candidate and the candidate expressing complex ideas about the nature of language. This increase in complexity follows research from Saito and Liu (2022) who found conversational experience increases learners use of complex and abstract content words.

Discussion task; C1/C2 speaker

(2) Candidate CH_18: I think it's is kind of cultural thing
Examiner: yeah I accept that position but er surely ev-every thought can be expressed in every language can't it ?

Candidate CH_18: mm but don't you think that you know er when you when you er **learn a foreign language** y-you sometimes you just wanna know why and and then you just I I resist the idea of erm changing my personality when I **learn a foreign language**

Examiner: and what makes you think so ?

Candidate CH_18: well the cos I uhu erm I want to use the language as a tool not as a n= er I don't want the language to control me I want to control what I say

This speaker, Candidate CH_18 also uses the collocation 3 times within this task, accounting for some of the 11 instances of the collocation in the advanced group and Discussion task but only occurring from 6 speakers overall. This demonstrates the rarity of this collocation in the B2 and C1/C2 speaker groups, shown not just in lack of frequency but dispersion too.

A final indicator that *learn + language* is a topic-influenced, as well as B1 speaker specific, is the relative frequencies of this across the speaker groups. At 0.34 per 1,000 words, it is by far the most frequent collocation in the B1 subcorpus and this is compared to 0.03 per 1,000 words in B2. Interestingly, there is a slight increase in the use of *learn + language* at C1/C2 level to 0.05 per 1,000 words but still is not close to the frequency within B1.

4.2.2.2. *Read + book (B2)*

Table 31 Occurrences of read + book collocation by task across proficiency groups

	Frequency	**Speakers**
B1	**189**	**99**
Conversation	60	42
Discussion	129	57
B2	**117**	**60**
Conversation	27	21
Discussion	90	39
C1/C2	**27**	**18**
Conversation	11	9
Discussion	16	9

Total	**333**	**177**

Table 31 shows that most instances of the collocation *read + book* occur in the Discussion task (70.6% overall), and this is the case across all three proficiency groups. The Discussion task is candidate-led so the L2 speaker will have chosen what topic to focus on in this instance. Although the relative frequency is lower than the B1 group occurrences at 0.25 per 1,000 words, this collocation has the most relatively frequent occurrence in the B2 group with 0.14 per 1,000 words. Focusing on the B2 group, it can be seen that 21 speakers use the collocation 27 times for the Conversation task (77.78%), while 39 speakers use it 90 times in the Discussion (43.33%). This is a notable difference between the two tasks when considering distribution; there are fewer speakers in the Discussion task using the *read + book* collocation more times when compared to the examiner-led Conversation task. The candidates have chosen their own topic and potentially, this has led to an overreliance on using this collocation based on the repetition that is evidenced here (which would support findings from Hasselgren, 1994 and Sukuzi, 2015). Looking at the concordance lines, further evidence of this can be seen. One speaker, Candidate 2_7_CH_4, uses the collocation 9 times throughout their Discussion task with the first instance used to introduce their topic for the dialogue in Example (3) before they continue on the interaction, as seen in Example (4). It can also be seen that there is a slight variance in tense and number i.e., *reading books* compared to *read the book*. This supports findings from Paquot (2014) who has also found there to be an influence of topic on tense preferences.

Discussion task; B2 speaker

(3) Candidate 2_7_CH_4: well my topic is **reading books**

(4) Candidate 2_7_CH_4: I love reading book so I began to like **read those books**

The second most frequent usage from one speaker comes from Candidate 2_7_IN_41 who uses *read + book* 7 times in their Discussion task. Example (5) shows the first mention of the topic; however, this is slightly different to Candidate 2_7_CH_4 as they do not use the collocation in the topic introduction but do then continue on to using it within their dialogue as seen in Examples (6) and (7).

Discussion task; B2 speaker

(5) Examiner: so we'll start with the topic what what are we going to talk about?
Candidate 2_7_IN_41: my favourite book

(6) Candidate 2_7_IN_41: but this book I think everybody should **read this book**

(7) Candidate 2_7_IN_41: because er I just er **read the book** when I feel it is interesting

Overall, this is further evidence that the TLC-L2 contains some verb + noun collocations that are influenced by topics chosen, in this case the candidate subject of choice, supporting claims from Paquot (2020).

4.2.2.3. *Take + care (C1/C2)*

The first of two collocations recurrent in the C1/C2 group that will be under further investigation here is *take + care.* Table 32 shows the breakdown of occurrences of this verb + noun collocation across the two tasks and three groups of speakers. It can be seen that the majority of the occurrences are found within the C1/C2 speaker group. It can also be noted this is based on raw frequency; in fact, the relative frequencies further show how more common the collocation is in this advanced group of speakers compared to the others. With 0.28 occurrences per 1,000 words, *take + care* is found far more frequently in the C1/C2 group than in the B1 group, with 0.10 occurrences, and in the B2 group with just 0.09 occurrences per 1,000 words. Considering the task type as well, these instances for the advanced speakers occur more frequently in the Conversation task – the examiner-led interaction – than the Discussion task. This is the same, although to a lesser extent, in the B2 group, and inverted for the B1 group. This shows that the collocation is occurring more frequently as the proficiency increases, with a greater percentage of occurrences occurring in the examiner-led task as this proficiency increases too, supporting findings from Laufer and Waldman (2011) and Paquot and Granger (2012) regarding a link between proficiency and frequency of collocation use. This could be evidence that the collocation is influenced by the topics set by the exam item writers. Further investigation is needed by considering the concordance lines.

Table 32 Occurrences of take + care collocation by task across proficiency groups

	Frequency	**Speakers**
B1	**74**	**48**

Conversation	33	23
Discussion	41	25
B2	**77**	**54**
Conversation	44	34
Discussion	33	20
C1/C2	**104**	**59**
Conversation	78	45
Discussion	26	14
Total	**255**	**161**

Before going further into the concordance analysis, it is important to note the Trinity College London topics for Grades 10, 11 and 12 as these are the graded exams that comprise the C1/C2 level group of speakers. According to Trinity College London (2021), the GESE topics for Grade 10 Conversation task comes in two lists: List A comprises roles in the family, communication, the school curriculum, youth behaviour, use of the internet, designer goods; and List B includes international events, equal opportunities, social issues, the future of the planet, scientific developments, and stress management (p. 47). For Grade 11, the two lists include independence, ambitions, stereotypes, role models, competitiveness, young people's rights, the media, advertising, lifestyles, the arts, the rights of the individual and economic issues (p. 49). There is no set list for the Conversation task topics for Grade 12 (the highest grade) as "candidates are expected to be able to enter into discussion on any subject that the examiner deems appropriate for the individual candidate" (p. 51), though the age is considered when an examiner chooses a topic for the specific candidate.

Considering only the C1/C2 concordance lines of the Conversation task, Table 33 shows the breakdown of the 78 instances of *take + care* based on topics and number of speakers to consider the distribution of these topics.

Table 33 Breakdown of the topics introduced in the Conversation task for C1/C2 speakers

Topic	**Occurrences**	**Speakers**
Roles in the family	57	27
Equal opportunities	8	5

Social issues	2	2
Use of the internet	2	2
Capital punishment	1	1
Stress management	1	1
Space exploration	1	1
The internet	1	1
International aid	1	1
Independence	1	1
Politics	1	1
Lifestyles	1	1
International events	1	1
Total	**78**	**45**

Looking at the breakdown of the topics, the influence of the topic 'roles in the family' can be seen in the overall results. Many of the instances including the construction *take + care + of <family member>* or some variation on this. This is exemplified below in Examples (8) and (9) where the examiners use the specific phrase "roles in the family" to introduce the topic and then a few turns later, the candidate uses the collocation *take + care* with the additional prepositional phrase *of <family member>*.

Conversation task; C1/C2 speakers

(8) Examiner: so let's now do the conversation and we're going to choose from list A so list A **roles in the family**
Candidate 2_CH_2: and if have a single child she will now have to **take care** of them

(9) Examiner: right and er finally the conversation okay er so I'd like to talk about **roles in the family** er so erm are family roles erm traditional?
Candidate 2_IN_15: most of the time the woman does a part-time job so that she can actually **take care** of the husband or the children

There are some other instances of *take + care* used in other Conversation task topics such as social issues (Example (10) and space exploration (Example (11).

(10) Candidate 2_ME_11: because you have to **take care** of your body you don't have to eat you don't have to eating salt and I love salt and I would I don't wanna say I love it but I like it

(11) Candidate 2_IN_2: l-lots of other things you know like like space exploration doesn't only mean making told you about they can be er pollution they can be lot of other er controlled matter which can be **taken care** of

The collocation is still being used, but it can be seen how much of an influence topic can be on the language used within an examination and the value of analysing concordance lines to investigate potential reasons for collocations occurring more frequently that perhaps typically expected (McEnery & Hardie, 2012). *Take + care* is also a verb + noun collocation that could be influenced by cultural background of the speakers using it, supporting Hinkel's (2009) findings of topics involving parental roles and family duties impacting use of modal verbs with certain L1 language and cultural backgrounds.

4.2.2.4. *Use + internet (C1/C2)*

Table 34 Occurrences of use + internet collocation by task across proficiency groups

	Frequency	**Speakers**
B1	**19**	**11**
Conversation	4	3
Discussion	15	8
B2	**11**	**9**
Conversation	4	4
Discussion	7	5
C1/C2	**49**	**28**
Conversation	49	28
Total	**79**	**48**

As mentioned in the previous section, Grades 10 and 11 have set topics for the Conversation task for the examiner to choose from, while the Grade 12 exam has no specific subjects for examiners and candidates to engage with. A clear insight into topic-influenced collocation usage by candidates comes from the *use + internet* verb + noun collocation. "Use of the Internet" is a main topic on List A of the Grade 10 examination,

and this includes the core verb and noun of the collocation. As well as this, as seen in Table 34, all the instances of *use + internet* in the C1/C2 group occur in the Conversation task – this is the examiner introduced topic dialogue. 49/79 (62.03%) of the collocations are found within the advanced proficiency speaker group, with the remaining 30 in the B1 and B2 group in the Discussion tasks – where the candidate has introduced the topic. There is one C1/C2 speaker – Candidate 2_IT_33 – who uses the collocation *use + internet* 10 times (accounting for 12.66% of the instances) and there are only 28 speakers in total who use the collocation 79 times, showing a small dispersion in the corpus. This narrower dispersion can also point to a more highly topic-influenced collocation (Huang, 2023).

It can also be seen that there is a notable difference in the relative frequency of the use of this collocation between the groups. In the advanced group, this is 0.13 per 1,000 words, in the B1 group it drops to 0.03 per 1,000 words and this decreases further at B2 level to 0.01 instances per 1,000 words. Within both the B1 and B2 groups, there are only 4 instances each of the collocation used in the Conversation task, such as that in Example (12) from a B2 candidate:

Conversation task; B2 speaker

(12) Candidate 2_7_CH_27: I usually speak Cantonese to my classmates and use Cantonese all the time the only times I use English is when I'm **using the internet** or just just chatting in the English lesson

Use + internet occurring 49 times solely in the Conversation task at C1/C2 is of particular interest as it may not be a frequently occurring collocation within L1 English, typically. To investigate this, the collocation was search for in the BNC2014 in the 10M word subset of informal spoken English (Love et al., 2017). There are only 28 instances in this corpus whereas a similar meaning collocation of *go + online* occurs almost twice as frequently, with 41 instances. *Go + online* would be thought of as a more 'native-like' collocation than *use + internet* and the evidence from the Spoken BNC2014 supports this. Furthermore, another verb + noun collocation with a similar meaning – *surf + internet* – occurs just once in the BNC2014 compared to 16 times in the TLC-L2 and demonstrates differences in how L1 and L2 English speakers express a similar idea.

As well as being topic-influenced, *use + internet* could also be said to be register-influenced due to the impact of its presence form the examiner as an interlocutor. Looking

at concordance lines in the TLC-L2 supports the claim that the examiner is influencing the candidate use of the verb + noun collocation by their language choice when introducing the topic, which is something to keep in mind when designing language examinations. The following examples are all from the C1/C12 group in the Conversation task.

Example (13) and (14) show the exact phrasing of the collocation repeated by the candidate after the topic is introduced by the examiner:

(13) Examiner: now we'll move on to the conversation phase. I'd like to talk about erm **the use of the internet**. Now can you sum up your views on the best and worst aspects of the internet
Candidate 2_IT_23: internet is erm is a positive thing because the the family could er educate their son or their daughter er thanks to **the use of the internet**

(14) Examiner: talk about let 's talk about **the use of the internet** erm are there any dangers to using the internet?
Candidate 2_SP_35: yeah there are I think it's really erm it's really important to keep in mind that the privacy is getting lost with **the use of the internet** specially with young people

However, there are also some instances of the examiner not using the specific construction:

(15) Examiner: do you think it's got out of hand this instant communication?
Candidate SP_110: okay yes you can **use the internet** but you don't where where do you have the internet

Rarely the collocation occurs when the speaker is talking about an entirely different topic. Example (16) is discussing the roles in the family:

(16) Candidate 5_10_CH_5: well I think it won't be because even if nowadays we have a huge internet that can let us even er take for example if you if you now let us imagine that I 'm in China and my parents is in UK and we can **use the internet** to talk to each other

These are rare instances, in fact, just 13 of the 49 instances (26.53%) do not have the examiner use the collocation *use + internet* to begin the conversation topic which, also

demonstrates the impact of topic on how collocations are used by candidates. This verb + noun collocation also shows the influence examiner speech can have on candidates in the interaction, supporting findings from Lazaraton (1996) and Young and Milanovic (1992).

4.3. Frequent verb types

To look at the data from an alternative perspective, the focus of this section is placed on frequently occurring verbs within the verb + noun collocations. Firstly, an overview is given based on the 201 collocation types from the 9,674 collocation tokens that are shared across the proficiency levels. Then, focus moves to investigating the topic-influenced collocations – those combinations likely occurring due to the subject of the interaction – and register-influenced collocations – those likely occurring due to the nature of the examination context including the examiner as an interlocutor. Finally, this section concludes with a focus on select abstract-noun collocations that may help to highlight L2 English development of these speakers, as previous research has found looking at abstract nouns to be fruitful for this purpose (Juknevičienė, 2008).

4.3.1. Overview

Using the narrowed list of 201 collocation types, there are a total of 72 verb types within these verb + noun collocations. Table 35 shows the ranked frequencies of these verb types based on each proficiency group and combined overall.

Table 35 Ranked frequencies of verbs in the verb + noun collocations per group and combined

B1		**B2**		**C1/C2**		**Combined**	
play	334	think	287	think	176	think	672
learn	295	make	268	take	169	take	652
buy	232	take	253	make	135	make	583
take	230	get	225	get	111	play	582
get	210	play	186	use	92	get	546
think	209	like	145	spend	78	buy	405
like	192	use	130	play	62	learn	394
spend	190	see	121	buy	62	spend	384
read	189	watch	120	know	62	like	382

watch	188	read	117	see	60	use	369

Table 35 also shows similar frequency of use of verbs within the collocations for B2 and C1/C2 groups. The top four verbs are the same and these are three core delexical verbs, *get, make* and *take*, as well as the verb *think.* Within the B1 group, *learn* is a unique verb and is likely within the top ten due to the prevalence of the collocation *learn + language*, as discussed in Section 4.2.2.1. The core delexical verb *make* is 11th on the ranked frequency list for the B1 speakers, which explains the position within the Overall column. In C1/C2, the verb *know* is unique to this group of speakers. This could be evidence of the advanced speakers being more confident in the interaction where they are discussing their chosen topics and answering questions about their opinion, such as in the case study from Li and Schmitt (2009) who found the L2 English learner they were longitudinally studying gained confidence in using lexical phrases over time. This may also be instead of using *think* to create what Kwon et al. (2018) call 'uncited generalisations' on the topics under discussion.

4.3.2. Topic-influenced collocations

4.3.2.1. *Watch + TV/television/movie/film/video*

The first verb to be investigated as a possible topic-influenced verb + noun collocation is *watch* + noun. The nouns within these collocations are *TV, television, movie, film* and *video.* It was decided to group these nouns together as collocates of *watch* because semantically they are very similar. *TV* is a shortened version of *television, film* and *movie* are dialect variations of the same concept of a work of visual art, and *video* also ties into this too. Table 36 shows that these collocations mostly occur in the Discussion task and thus this could be evidence that the use is driven by the candidate choice of topic. Therefore, *watch* + noun could be an example of a topic-influenced collocation.

Table 36 Occurrences of watch + TV/television/movie/film/video collocation by task across proficiency groups

	Raw Frequency	**Speakers**
B1	**188**	**135**
Conversation	77	58
Discussion	111	77

B2	**120**	**83**
Conversation	36	30
Discussion	84	53
C1/C2	**44**	**32**
Conversation	26	20
Discussion	18	12
Total	**352**	**250**

Looking at Table 36, it can also be seen there is a fairly even distribution of the collocations across the groups when considering the variability of their token sizes. Therefore, the groups are using the collocations with comparable frequency, so looking further at the concordance lines is also needed to see how these collocations are being used in context (McEnery & Hardie, 2012).

It was of interest to note if there were any specific adjectives occurring with these nouns to highlight any specific topics. It was found that *movie* and *film* were the most likely to have adjectives modifying the collocations and these can be categorised by 'region': *American* and *English*; 'genre/type': *horror, terror, ghost, violent, strange, animation* and *dinosaur*; and finally, 'franchise': *Harry Potter, Hunger Games, Marvel* and *Disney.*

Considering the concordance lines, there are more complex topics being discussed within the C1/C2 group compared to the B1 group, as would be expected, but they are expressing these differing levels of complexity using the same language. In the advanced proficiency group, there are indeed very complex topics that are introduced by 12 different speakers. One speaker – Candidate IT_19 – uses the collocation *watch films* when talking about the history of moving pictures which can be seen in Example (17):

Conversation task; C1 speaker

(17) Candidate IT_19: personally I I think that er **watch films** on a big screen a bigger screen with a higher level of sound is erm quite better because you can appreciate er er the film er more maybe

Another C1/C2 candidate talks about cinematography in their Discussion task.

(18) Candidate 2_SP_49: what I mean by purpose is erm when you **watch a film** erm in the background it has an idea and it wants to transmit you some something

Examples (17) and (18) are two instances where it makes logical sense to have the collocation *watch + film* included based on the topic that is being spoken about. However, there were also mentions of the collocation in seemingly completely unrelated topics such as one C1/C2 candidate talking about The Holy Spirit in Example (19) in their Discussion task:

(19) Candidate 2_ME_11: I was feeling depressed I was feeling alone and I decide to go to a Christian church that's where I meet and oh have you **watch er a movie** I don't know how you calls the movie but it's has a ball in the stomach like a a big ball oh that's h-how the spirit the Holy Spirit feels like

This is not a one-off occurrence, as Example (20) demonstrates with another topic-unrelated use of the collocation where the C1/C2 candidate is talking about the difference between the resume virtues and the eulogy virtues in their Discussion task:

(20) Candidate 2_IN_23: I was **watching a video** about death a month back and I that was the time that I realised that people should think about dying because it is people are scared of the topic of death

In contrast to the above complex topics used by the advanced candidates, many of the B1 speakers are using the collocations to talk about their hobbies and interests, which is a more typical and expected use of the *watch* + noun collocation at this lower level of proficiency. Example (21) shows this B1 speaker using the collocation 7 times within the Discussion task and mostly using the verb node *watch* within the present continuous tense:

(21) Candidate 2_6_IN_121: I have more hobbies ma'am er but er still I er
my favourite hobby is **watching television**
benefits of **watching tv** I can er er what can I say?
while **watching television** we'll er er sit together
when I'm **watching television** I'll make my son to play
starting **watch a tv** programme now he will be addicted to that
he doesn't like **watching television** much

In four of these examples, the speaker uses *watching + television* as the collocation, following on from the introduction of the topic. This could indicate the collocation is fairly fixed in nature for this candidate when considering it within the context of a phraseological approach (Gyllstad & Wolter, 2016). This is in contrast to a further example from another B1 speaker who uses a variation on the collocation 6 times within the Discussion task. They first introduce their topic as 'enjoying films' before using the collocation in different tenses, with varying articles and an inclusion of an adjective 'horror'. This is showing that, even within the same proficiency, there is variation of how fixed the collocations are for the individual speaker:

(22) Candidate 2_6_SP_46: I'm talking about films but my topic is enjoying films
every day I **watch some films**
the next Saturday I'm going to **watch the films** with my friends
I that it's better er **watching films** at cinema
but if I want to to **watch films** at cinema I will must go to Jerez
I I **watch my films** at TV
when my parents was **watching a horror film**

Overall, although *watch* + noun would seem to be a topic-influenced collocation from the outset due to it falling within the theme of 'talking about a hobby', looking at examples from the concordance lines shows that this collocation can be used in unexpected ways unrelated to the specific topic that has been chosen by either the examiner or the candidate demonstrating a level of creativity in the L2 speakers language use (Carter & McCarthy, 2004). The data also shows a breadth of use across the proficiency levels and across the speakers who use it. Although it can occur due to the topic choice, as evidenced by examples from the B1 speakers, it is not the only reason for it to be used within this corpus. Furthermore, even within the same proficiency group, when looking at concordance lines it is clear to see evidence of individual differences with regards to how fixed collocations can be linking back to results from Omidian et al. (2021).

4.3.2.2. *Become + (career)*

Looking at the common collocates with specific verbs, *become + (career)* was found to frequently occur, which warranted further investigation. The careers grouped for this collocation were lawyer (7), teacher (23) and doctor (31). As seen in Table 37, the

majority of the collocations occur within the Discussion task, which could potentially mean that these are candidate inspired, topic-influenced collocations. This is due to the fact the candidate introduces the topic within the presentation and the fact they are using concrete nouns to collocate with the verb i.e., specific professions. Further consideration is needed to decide if these instances are in fact topic-influenced collocations.

It could be said that the use of these *become* + noun collocations are becoming less examiner- and topic-influenced in the more advanced C1/C2 group as there is a decrease in their occurrence in the Discussion task. Overall, fewer candidates in this proficiency group talk about careers, although this could also be due to the smaller group size in general, therefore resulting in there being fewer opportunities for the collocation to occur (Caines & Buttery, 2017). To investigate further, some of the concordance lines from the five instances in the Conversation task are provided in Examples (23) to (25):

Conversation task; C1 speaker

(23) Examiner: erm er I want to talk about er equal opportunities now
Candidate 2_IT_26: we are there are many women who teach who are teachers and a few men who **become teachers**

The above example shows the examiner introducing the topic of equal opportunities and it is the candidate who decides to begin talking about work-related opportunities. Likewise in Example (24) below, the examiner introduces a topic that arguably encourages the *become* + noun collocation, though it is not directly referencing it in the same way that was seen with *use + internet* in Section 4.2.2.4.:

Conversation task; C1 speaker

(24) Examiner: let 's go on to talk about the school curriculum now
Candidate IT_14: I want to go after the high school I want to work to Rome in the erm private university in a Catholic private university er because I think that is the best university to to **become a lawyer**

The following two examples show the topic of 'personal ambition' introduced by the examiner, arguably encouraging the candidate into the choice of using *become + (career)* collocation. Again, this is not unexpected, but it is not a direct influence on the choice of their language in that the candidate is not repeating the specific collocation introduced by the examiner.

Conversation task; C1 speaker

(25) Examiner: let's start with erm ambition er personal ambitions do you set targets for your ambitions

Candidate IT_17: yeah I think that it's very very important to have an ambition in your life at least one ambition er for example er personally er I like to be to mm **become a teacher** at school or also at university

(26) Examiner: how much do you agree with the statement if you want to succeed you must be ambitious

Candidate IT_17: oh so for this statement it's a bit problematic because er as for my fields because I I er like to **become a teacher** because erm er we are a lot a lot o-of people who want become a teacher so we are erm few er possibilities and er more people so we have a minimum in this er field in the in teaching so it's a problematic

Table 37 Occurrences of become + (career) collocations by task across proficiency groups

	Frequency	**Speakers**
B1	**26**	**17**
Conversation	4	4
Discussion	22	13
B2	**27**	**17**
Conversation	10	8
Discussion	17	9
C1/C2	**7**	**5**
Conversation	5	3
Discussion	2	2
Total	**60**	**39**

Table 37 also shows that 12/22 occurrences (54.55%) in the B1 group were from three speakers out of the 13 overall using *become + teacher* or *become + doctor* in the Discussion task. This shows that just a few speakers can have a significant influence on the frequency counts and therefore, distribution always needs to be considered.

Furthermore, looking at just the Discussion task occurrences as these are accounting for two thirds of the overall instances – this is the based on the candidate choice of topic in the Presentation task. The topics for each of the three speaker groups can be broken down as follows: for the B1 candidates, the topics include hometown, teaching job, education in India, passion for hip-hop, future career, job, personal ambition, the gym, family, job, oneself, and supersonic cars. For the B2 group, the topics related to the collocation include teaching as a career, changing career, education in Mexico, personal dream, youngsters and ambitions – general and personal – as well as what is the ideal person and the importance of time management. Finally, the advanced C1/C2 speakers discussed topics such as the scientific process of dreaming (27) and ancient Greek theatrical representations (28):

(27) Candidate IT_20: today I'd like to talk about dreams er dreams are that erm occur in our minds while we sleep
Candidate IT_20: our our deep sleep state is the state where we 're not dreaming at all well I hope I will study this when I **become a doctor** more accurately

(28) Candidate 2_IT_30: my presentation topic is about er the ancient Greek theatrical representations that er as you probably already know er were very important events in Greek society because they were at the same time a political social and religious moment
Candidate 2_IT_30: this kind of reflection lead me to understand what I want to do in my future because erm for instance I would like to **became a doctor** to help people in need

Looking at the topics, there are clear differences between the B1 and B2 levels and the advanced C1/C2 speaker choices of what to present which is somewhat influenced by what general language development when considering the CEFR descriptors (Council of Europe, 2001). The topics within the B levels are very egocentric and thus increases opportunity for the *become + (career)* collocation to arise (Caines & Buttery, 2017). At B1 level, these are also very much aligned with the collocations of *become + (career)* too. An anomaly of this would seemingly being the topic ‘supersonic cars’; however, further analysis shows that this speaker used the collocation in way that made it still relevant to the topic:

(29) Candidate 2_8_IN_17: saw it on internet I saw a video how to how they drive this supersonic cars

Examiner: right so when do you plan to get one? and how do you plan to get one?

Candidate 2_8_IN_17: sir it I want to **become a doctor** so if I if I work out of India I can get one million if I get suppose so I can buy one supersonic car in a year

Overall, candidates at each of the three levels use *become + (career)* as a collocation due to topic-influence of the Discussion task and there is a distinct difference between both the B1 and B2 candidates' choice of topic when compared to the C1/C2 group's presentation topics. This demonstrates that verb + noun collocations in the Discussion and Conversation tasks within the TLC-L2 are topic-influenced and that there are differences in use of these collocations across proficiency levels related to how personal the collocations which links to Jones et al. (2017).

4.3.3. Register-influenced collocations

The following analysis will consider the collocations that are register-influenced. This means they are potentially occurring due to the nature of the examination context and the tasks involved, as well as to help support the interaction between the two interlocutors: candidate and examiner.

4.3.3.1. *Repeat/understand + question*

The first collocation under review is *repeat/understand + question.* The two verbs *repeat* and *understand* are considered together in this section, combined with the noun *question* since although they are semantically different, within the examination context they function in a similar way in that they are requesting clarification. These clarification requests are of interest to this study as Jones et al. (2017) found differences in focus and expression dependant on speaker proficiency level related to strategic competence. For instance, B1 speakers seek clarification of task instructions while C1 speakers focus on the meaning of words (Jones et al., 2017).

Table 38 Occurrences of repeat/understand + question collocation by task across proficiency groups

	Frequency	**Speakers**
B1	**28**	**26**

Conversation	18	16
Discussion	10	10
B2	**12**	**12**
Conversation	10	10
Discussion	2	2
C1/C2	**4**	**4**
Conversation	2	2
Discussion	2	2
Total	**44**	**42**

Table 38 shows there are more instances occurring of the collocation *repeat/understand + question* within the conversation task in both the B1 and B2 speakers, supporting the findings from Jones et al. (2017) regarding lower proficiency speakers asking for more clarifications than higher proficiency speakers. 41 occurrences are from unique speakers, with one speaker using the collocations three times and another twice. Candidate 2_6_IT_101 uses the collocation in both the Conversation task and the Discussion task:

Conversation task; B1 speaker

(30) Examiner: ah okay but do you work for your money at home?
Candidate 2_6_IT_101: er n= can you **repeat the question**?
Examiner: do you work for your money at home?

Discussion task; B1 speaker

(31) Examiner: okay and how what does this do to your brain?
Candidate 2_6_IT_101: can you **repeat the question**?
Examiner: what does this game do to your brain ?

Conversation task; B1 speaker

(32) Examiner: are you controlling your diet ?
Candidate 2_6_IT_101: can you s=
Examiner: are you controlling your diet at the moment? are you
Candidate: I don't **understand the question**
Examiner: are you controlling your diet at the moment?

Candidate: er no because er erm er I erm I eat er I eat er er some food and er er in this moment I don't want to give a diet

Example (32) shows the examiner simply repeating the question, three times in fact, without attempting to reframe it for the candidate. The candidate does eventually seem to understand, although is clearly unsure, as demonstrated by their use of fillers and repetition. Interestingly, in Example (31), the candidate used the *repeat + question* collocation and the examiner did clarify by expanding on their initial phrasing to include the noun *game* to be more specific. These interactions show the candidate using collocations to help support their interaction in this examination context, which links to Mauranen (2004) and previous findings that note the benefit of formulaic language to help with this. Therefore, this verb + noun collocation can be claimed to be evidence of register-influenced collocations within the TLC-L2.

4.3.3.2. *Choose + topic*

Table 39 Occurrences of choose + topic collocation by task across proficiency groups

	Frequency	**Speakers**
B1	**26**	**24**
Conversation	0	0
Discussion	26	24
B2	**54**	**54**
Conversation	1	1
Discussion	53	53
C1/C2	**5**	**5**
Conversation	1	1
Discussion	4	4
Total	**85**	**83**

The second register-influenced verb + noun collocation of interest in this study is *choose + topic*. Table 39 shows that the vast majority of occurrences of *choose + topic* occur within the Discussion task; altogether, 83 out of 85 instances (97.65%) are found here. One possible reason for this collocation to be occurring frequently in the B1 and B2 groups is because these proficiency levels are undertaking the GESE exam grades 7-9 – these exams begin with the topic discussion where the candidate introduces a topic of

their choice and then the two interlocutors engage in dialogue about this. For the C1/C2 candidates, the Discussion task follows the Presentation task, where the candidate begins with a formal presentation of their topic. Therefore, there is less need to use the collocation *choose + topic* as this is likely to have already been stated in the presentation task monologue. It is not that the speaker is not using it at all, but that it is not necessary to be present in the tasks that are included in this analysis for the C1/C2 speakers due to the structure of the exam. This again highlights that the verb + noun collocation is register-influenced. Aside from the fact the B1 and B2 groups are overall much larger than C1/C2, this could also account for the difference in occurrence between the threshold and intermediate and advanced speakers. In many instances, the collocation production by the candidate is likely to have been primed by the examiner stating it first (Lazaraton, 1996). This occurs in a variety of instances in the Discussion task with B1 speakers (Examples (33) to (37)) as well as with B2 speakers (Examples (38) to (43)). However, there is an anomaly of usage in a Conversation task with a C1/C2 speaker in Example (44), demonstrating why looking at concordance lines to investigate frequency counts is so valuable.

Discussion task; B1 speaker

(33) Examiner: my question of course is why did you **choose** to talk to me about **this topic**?
Candidate 5_6_AR_20: okay I **chose this topic** because I think that it's erm it's interesting and it's original as I think

Discussion task; B1 speaker

(34) Examiner: certainly yeah absolutely erm okay why why did you **choose this topic?**
Candidate 2_6_AR_48: erm I **choose this topic** because I think it's amazing

Discussion task; B1 speaker

(35) Examiner: okay so why have you **chosen this topic**?
Candidate B1 5_6_CH_21: I think this topic topic is very interesting er we are healthy people so we're very lucky er and there's a little girl next to my house and she had a cancer mm so I **choose this topic**

Discussion task; B1 speaker

(36) Examiner: wow child labour okay that's a very serious topic okay well why did you **choose this topic**?

Candidate B1 2_6_IN_12: ma'am I **choose this topic** because nowadays erm many topics are there but child abuse most one of the serious problem

Discussion task; B1 speaker

(37) Examiner: okay so yeah why have you **chosen this topic**?

Candidate B1 2_6_IT_103: I **chosen this topic** because I think friendship is very important

Discussion task; B2 speaker

(38) Examiner: okay so what have you **chosen** to talk about today for your **topic**?

Candidate B2 7_ME_8: yes today I 'm going to be talking about recycling okay I er I **chose this topic** er because I 'm interested in it

Discussion task; B2 speaker

(39) Examiner: why did you **choose this topic**?

Candidate B2 5_7_AR_10: well I **choose this topic** because for me it has a lot of nowadays

Discussion task; B2 speaker

(40) Examiner: so we 're going to talk about that okay erm so why did you **choose this topic**

Candidate 2_8_ME_17: well er today I 'm going to talk about cell phones and

Examiner: but hang on wait a minute the question is why did you **choose the topic**??

Candidate 2_8_ME_17: I **choose this topic** because…

Discussion task; B2 speaker

(41) Examiner: oh okay why did you **choose this topic**?

Candidate B2 2_8_SP_36: er well the reason why I've **chosen this topic** is because my wife is Lithuanian

Discussion task; B2 speaker

(42) Examiner: okay erm why did you **choose this topic** to talk to me?

Candidate B2 5_7_AR_7: er and I **choose that this topic** because my grandparents are people er so important for me and I think that I want to talk with I want to talk about him them

Discussion task; B2 speaker

(43) Examiner: so have you **chosen a topic**?

Candidate B2 2_8_IT_33: yes yes I've **chosen topic** and er I brought pictures with me this is my topic a topic I chose for the exam

Conversation task; C1/C2 speaker

(44) Examiner: **choose the topic**

Candidate 5_10_CH_2: yeah list B from list B??

Examiner: er no I **choose the topic**

Candidate 5_10_CH_2: oh I thought I I could **choose**

Examiner: no no

Candidate 5_10_CH_2: **the topic**

Examiner: you **choose** the you **choose** the

Candidate 5_10_CH_2: I just **choose** the list whether list A or list B and you decide the **topic**

Examiner: yes

Candidate 5_10_CH_2: oh

When considering the distribution of this collocation, it should be noted that all but two speakers only use *choose + topic* once. This is further indication that it is a collocation influenced by the register of the interaction; this is both due to the examiner introducing and priming the collocation, sometimes even influencing the tense used by the candidate as seen in Example (43), and because it is a necessary part of the exam for the candidate to explain why they have selected a particular topic. This links to Lazaraton (1996) discussion on examiners priming topics for scaffolding candidate speech.

4.3.4. Abstract-noun collocations

The final part of the verb analysis will focus on abstract-noun collocations; these are defined as collocations that have a noun collocate that is semantically abstract in combination with the verb node. The purpose for considering these is that there is some

evidence that more advanced speakers of English will use more abstract concepts within their speech (Juknevičienė, 2008).

4.3.4.1. *Spend/waste + time*

The first abstract-noun collocation under review is a combination of *spend* and *waste* verbs combined with the noun *time.* While *spend* and *waste* arguably have the same literal meaning (time passing), they have distinct connotations; namely, the former connotes productivity while the latter connotes an opposite meaning of wastefulness. Nevertheless, since these are both abstract concepts, they are combined for the purposes of this analysis.

Table 40 Occurrences of spend/waste + time collocation by task across proficiency groups

	Frequency (rf)	**Speakers (%)**
B1	**63 (0.85)**	**52 (5.57)**
Conversation	28 (0.38)	23 (2.47)
Discussion	35 (0.47)	29 (3.11)
B2	**107 (1.31)**	**82 (10.19)**
Conversation	50 (0.61)	39 (4.84)
Discussion	57 (0.70)	43 (5.34)
C1/C2	**57 (1.51)**	**45 (14.29)**
Conversation	39 (1.03)	33 (10.48)
Discussion	18 (0.48)	12 (3.81)
Total	**227**	**179**

In Table 40, it can be seen that the collocations occur more frequently in the Discussion task for B1 and B2 groups, but the converse is true for the C1/C2 speakers. This shows that the collocation occurs most in the task where the candidate introduces the topic for the B1 and B2 groups (in the Discussion task) and this changes for the more advanced C1/C2 group where the examiner introduces the topic (in the Conversation task). This change could be accounted for with the topic choices at this level for the Conversation task. As mentioned in Section 4.2.2.4, the topic of using the internet is a core subject for the C1/C2 speakers to be introduced to in this exam. It is understandable that this topic choice would give rise for more opportunities to talk about *spending time* and *wasting*

time as it is an activity. The engagement with this particular topic can be seen below in Examples (45) and (46):

Conversation task; C1/C2 speaker

(45) Candidate CH_2: this is not very good I think this is a kind of way to waste **waste their time** and it's not good for their eyes

(46) Candidate IT_65: I see that my sister spends er **spent a lot of time** on PC

As the raw frequencies are difficult to interpret precisely for this collocation, relative frequencies have also been added to Table 40 (per 10,000 words). Considering the relative frequencies, this demonstrates an increase in usage of *spend/waste + time* linearly, as the proficiency levels increase. Interestingly, this linear increase also occurs with the percentage of speakers using the collocation within each proficiency group. For the threshold B1 speakers, 5.57% of the candidates use the collocation, whereas this almost doubles for the B2 speakers at 10.19% with a further increase to 14.29% for the advanced C1/C2 group. This could be an indicator that the increase in English language proficiency is being shown through an increase in the use of collocations, supporting findings from research such as Forsberg and Bartning (2010), but particularly for abstract-noun collocations like *spend + time* and *waste + time* which further links with findings from Du et al. (2022). However, it could also be demonstrating there are more opportunities for candidates to engage in this type of language due to the nature of the examination, echoing findings from Neshkovsha (2019); this may be due to the relative frequency increasing in the Conversation task, in which topics are set by the examiner from a pre-defined list.

4.3.4.2. *Change + mind*

Table 41 Occurrences of change + mind collocation by task across proficiency groups

	Frequency	**Speakers**
B1	**9**	**8**
Conversation	2	1
Discussion	7	7
B2	**16**	**15**
Conversation	8	8

Discussion	8	7
C1/C2	**9**	**7**
Conversation	5	4
Discussion	4	3
Total	**34**	**30**

Table 41 shows the frequency of use of *change + mind* in each proficiency group. It can be seen that the B2 group of speakers uses the collocation most frequently and this is balanced between the two tasks, with 8 instances in each. It could be inferred that the increase of usage from B1 to B2 and beyond (when considering the size of the subcorpora) demonstrates increased proficiency in using abstract concepts in the examination or increased opportunities for using these concepts. Overall, only 4 speakers repeat the collocation, and in these instances the repetition occurs only twice. This shows that the 34 instances that occur in the candidate speech are broadly distributed within the corpus.

To investigate the usage further, Table 42 shows the collocations comparing B1, B2 and C1/C2 groups; by looking at the threshold (B1) and comparing to the advanced speakers (C1/C2), there might be evidence of development in some way.

Table 42 Total instances of change + mind collocation, including highlighted interceding words, listed per group

B1	**Total**	**B2**	**Total**	**C1/C2**	**Total**
Change **my** mind	4	Change **my** mind	5	Change **your** mind	5
Change **their** mind	4	Change **their** mind	3	Change **their** mind	3
Change **his** mind	1	Change **your** mind	2	Change **the** mind	1
		Change **everyone's** mind	1		
		Change **of** mind	1		

		Change **in people's** mind	1		
		Change **people's** mind	1		
		Change **his** mind	1		
		Change **our** mind	1		

For B1 speakers, all the instances of *change* + *mind* include an interceding word that is either a first person or third person pronoun. Among the advanced speakers, candidates also used the second person pronoun *you* within the collocation. This shows that they are engaging with one of the core purposes of the examination, which is to express interactional competence (Plough et al., 2018). They demonstrate this by using the second person pronoun because this is needed for a dialogue i.e., in the form of asking questions. This kind of dialogic exchange is a requirement and expectation at the higher levels of the exam and candidate speech; therefore, these speakers can be said to be engaging with the level of communicative competence level necessary for the examination – an important factor in language testing (Harding et al., 2023). This could also be the speaker developing pragmatic awareness in their use of pronouns as L2 proficiency increases (Belz & Kinginger, 2003).

Discussion task; B1 speaker

(47) Candidate 2_6_ME_51: I told him that I want to read something and he len= he lend me the book er when I saw the book I I thought that it wa= it was a but when I started reading I **change my mind**

Conversation task; B2 speaker

(48) Candidate 2_8_16: the real problem of people and er er they er they are n-next to young people they erm **changed erm their mind** on er some important value

Conversation task; C2 speaker

(49) Candidate 2_SP_20: they go there to to study something that they heard about they don't have an idea clear idea and and maybe they they **change their mind**

The B2 speakers demonstrate much more diversity in the internal modification of the *change + mind* collocation. The most common iteration is *change my mind*, which is similar to the B1 group in that it is showing the speaker is talking about themselves and their opinions and experiences. However, different from the B1 group is the inclusion of the interactive *change your mind* version of the collocation. Furthermore, the collocations also include interceding nouns *everyone* and *people,* which could be an indication that they are sharing their thoughts to others within the world i.e., thinking beyond personal experience and being able to express that in the language they have used. A further interesting collocation is *change of mind*, which is a slightly different construction, as seen in Example (50), that shows some deviation from the standard *change + mind* collocation seen in the lower proficiency levels.

Conversation task; B2 speaker

(50) Candidate 8_IT_28: as a matter of fact there are no er precise laws or which improve this **change of mind** in people

B2 speakers have more internal diversity in the collocation than C1/C2 speakers. This could be due to the larger cohort of speakers, meaning more opportunities for this diversity to occur; however, it is also interesting to note that there is a lack of this diversity comparing to B1 speakers who have a larger group of speakers than B2. Therefore, it may be due to opportunity in the difference between B2 and C1/C2 but could be due to language developed based on the difference between B1 and B2.

The majority of the instances of *change your mind* in the C1/C2 group occur within the Conversation task, which shows that the candidates are working to maintain the interaction with their examiner during this dialogue and is shown by the question in Example (53) to encourage the examiner to expand further on what they have said. In Example (55), the candidate interrupts the examiner with this statement to clarify what they have said and demonstrate they have understood the examiner's main message. There is one instance of *change your mind* occurring in the Discussion task shown in Example (51) and again this is a clarification question used to engage the examiner, further showing the use of this collocation at the highest proficiency level is working to support interactional competence (Galaczi & Taylor, 2018) in the exchange.

Discussion task; C1 speaker:

(51) Candidate 2_ME_23: think that's a big factor knowing you're conscious of what you're doing
Examiner: I see but if you made the decision while you were okay and then later had some mental disorder
Candidate 2_ME_23: mm and you wanna **change your mind**

Conversation task: C1/C2 speakers:

(52) Candidate 2_SP_11: well in Spain it's the same I mean if you are not going to university you don't have to continue until you are eighteen but I mean you should just in case you **change your mind**

(53) Candidate 5_11_CH_2: oh why? so erm is it s= why are you not quite sure now? er have you **change your mind**?

(54) Candidate ME_11: he always fighting for me so when when the Holy Spirit came into your lives even even if anything around you don't doesn't **change but your mind** change and you start to see everything different

(55) Examiner: because I started school very young because of travelling erm so that was I was lucky in that sense but I agree with you I didn't come back and go to university because I had an er an idea about
Candidate 2_SP_20: yeah you **changed your mind**

Overall, *change + mind* is a collocation that is used in different ways by the individual speaker proficiency groups. The threshold B1 candidates use it to express their own opinions, and this later develops in the higher proficiency groups to engage in discussion with the examiner and demonstrate interactional competence (Galaczi & Taylor, 2018), thus exhibiting features of a register-influenced collocation. This abstract-noun verb + noun collocation could therefore be argued as one of several collocations that help to define a register (Hyland, 2008), due to the reasons outlined above.

4.4. Collocational patterns in high frequency delexical verbs: get, make and take

To investigate the patterns in these three high frequency delexical verb + noun collocations, the analysis will be split into two parts. Firstly, only the collocations that are present in all proficiency levels will be considered – the 'shared' collocations – building on the analysis in the previous section. For the second part of the analysis, it was decided to return to the original list of 43,644 collocations to consider what collocations learners were using that were unique to the level. The reason for going back to the original set was that this would be able to show the fullest picture. The cleaning and the processing to find the shared collocations was necessary in order to ensure comparability across the proficiency levels for the previous analyses. However, when looking at unique combinations, there may be errors or other innovations within the data that would be removed with manual processing; therefore, this approach allows for all the data to be looked at and anomalies can then be qualitatively analysed. Specifically, this broader analysis will consider the semantic fields of the high-frequency verb + noun collocations, inspired by previous research from Du et al. (2022).

4.4.1. *Get*

Table 43 Raw frequencies of all get + noun collocations that occur in all groups

	B1	**B2**	**C1/C2**	**Total**
get + job	52	65	32	149
get + money	75	39	30	144
get + thing	15	27	8	50
get + chance	25	16	5	46
get + information	10	16	17	43
get + education	2	23	6	31
get + degree	7	12	4	23
get + disease	12	7	3	22
get + mark	8	10	4	22
get + water	4	10	2	16

Table 43 shows the 10 most frequent *get + noun* collocations across the groups combined. *Job* is the most frequent collocate of these with 149 instances, only narrowly more

frequent than the next collocate, *money*. The collocation *get + job* occurs much more frequently in the Conversation task (114) compared to the Discussion task (35). After these two collocations, the third most frequent is *get + thing*; however, this only accounts for 50 of the instances overall – a significant decrease compared to the previous examples and demonstrating Zipf's Law where "frequency of any word type in [a] corpus [is] inversely proportional to its rank" (Ha et al., 2009, p. 101). This means that the second most frequent word (or collocation in this case) will typically occur half as often as the most frequent. Between the second and third ranked frequencies, total occurrences will halve again, which means "the amount of evidence that we can get from corpora about words diminishes rapidly" (Brezina, 2018, p. 44). Only eight instances of *get + thing* occur in the C1/C2 group (16%); this could potentially show that this advanced speaker group is becoming more precise with their noun choices, rather than using the general placeholder of *thing*. *Get + education* is a notable collocation within the B2 group as this accounts for 23/31 (74.19%) of the overall instances. This could, therefore, be a topic-influenced collocation that is occurring more frequently at this proficiency level due to the subjects that are introduced by the examination. The general themes of these 10 frequent collocations can be grouped by work and school (*job, money, education, degree, mark*) abstract nouns (*thing, chance, information, disease*) with *water* as the final collocate falling outside these two categories.

Table 44 Relative frequencies of top 10 get + noun collocations per 10,000 words

	B1	B2	C1/C2
get + job	0.70	0.80	0.85
get + money	1.01	0.48	0.80
get + thing	0.20	0.33	0.21
get + chance	0.34	0.20	0.13
get + information	0.13	0.20	0.45
get + education	0.03	0.28	0.16
get + degree	0.09	0.15	0.11
get + disease	0.16	0.09	0.08
get + mark	0.11	0.12	0.11
get + water	0.05	0.12	0.05

Table 44 shows the relative frequencies of these top 10 most frequent *get + noun* collocations. Colour shading further highlights the differences between the proficiency levels, with the darkest highlight showing the collocation appearing most relatively frequently within the particular group. This table shows that 3/10 (30%) of the collocations occur most frequently in the B1 group, 5/10 (50%) in the B2 group and 2/10 (20%) in the C1/C2 group; this is evidence that there is not necessarily a linear increase in usage based on proficiency level in the *get + noun* collocations when considering the overall top occurrences. Instead, there is a surge at the B2 level, with half the most frequent collocations occurring in this group. This is of interest to note, as it supports other research that claims a nonlinear development of collocations based on frequency of occurrences (Forsberg & Bartning, 2010; Nizonkiza, 2012, 2017). This also supports previous claims from Vedder and Benigno (2016) regarding a u-shaped curve where learners become more proficient and overuse collocations while attempting to master more complex grammatical constructions. Finally, looking only at frequency here, it cannot be said how the speakers are using the collocations accurately and appropriately; frequency of use alone is not enough of an indicator of collocation development supporting claims from Paquot and Granger (2012). Qualitative analysis can shed further light on usage and so the analysis will look in more detail at one collocation per group that is deemed to be of interest due to the fact it occurs significantly more than the other two groups.

4.4.1.1. *Get + money* (B1)

To begin, the first collocation of interest for *get* is *get + money* for the B1 speakers, as 75 out of the 144 overall occurrences (52.08%) are found within this proficiency group. Furthermore, 63/75 occurrences (84%) are found within the Conversation task, thus the topic that has elicited the collocation has been led by the examiner within the interaction. It is worthwhile to compare this to the more 'native-like' collocation of *earn + money*. As well as its frequency, this verb + noun collocation is also of interest because the verb *get*, though grammatically correct, is not the most salient to occur with *money* when considering a native speaker perspective. Referring to the Macmillan Collocation Dictionary (2023a), as it can be seen in Example (56), shows that the first verb entry when searching for the noun *money* is another delexical verb *make*, as well as *earn* – the latter creating a more restricted collocation of *earn + money* due to the limited noun choices available to collocate with the verb.

(56) make/earn money: The business has made more money this year

The dictionary also includes other verbs within the entry: *spend, cost, borrow* and *save*. However, the first example in use is actually *get,* as seen in Example (57):

(57) No, I can't come – I haven't got any money.

Suzuki (2015) also found evidence of an overuse of *get* + noun collocations (such as *money* and *friend*) in L2 speakers and attributed this to their reliance on the open-choice principle rather than the speakers having knowledge of the more appropriate verb + noun collocations such as *earn* + *money.* Considering the data within the TLC-L2, one factor that may account for this grammatical but less common collocation of *get* + *money* could be the age of the speakers using it. *Make* and *earn* are semantically related to *work, job, career* etc. 896/2,053 (43.64%) speakers in the TLC-L2 are within the age band 8-15 years old and 499/896 (55.69%) of these have undertaken the B1 examination (Gablasova et al., 2019, p.154). This is a large proportion of the corpus overall (43.79%) and of the B1 group data (53.77%). Therefore, the candidates may be using *get* over *make* or *earn* because they are too young to be employed and subsequently not using the more collocable verb that is also more specific. This could be an example of the speakers using a high frequency verb that on the surface looks less native-like, and adhering to the open-choice principle (Erman & Warren, 2000), rather than a more restricted verb like *earn* but has in fact been chosen by the speaker for other reasons. To consider this further, concordance analysis was conducted. Example (58) involves a 15-year-old candidate that is receiving money from their parents. The examiner also potentially influences the candidate's verb + noun collocation choice by using *get* + *money* in the preceding turn.

(58) Examiner: and where do you **get money** now?
Candidate 6_SP_51: I **get money** with my parents

However, there is also evidence of the opportunity to use *earn* with *money* that a speaker does not use. In Example (59), the 14-year-old candidate is talking about a future career:

(59) Candidate 2_6_IN_21: I'm going to be an engineer that I'll **get money**

While in Example (60), the 20-year-old candidate is asking the examiner about their feelings towards their work:

(60) Candidate 2_6_ME_4: and do you do you enjoy to at this job or you do it just for **get money**? (20 years old)

Overall, *get + money* may be demonstrating B1 speakers' overreliance on a delexical verb, adhering to the open-choice principle due to their current stage of collocational development rather than using the more idiomatic *earn + money*, which supports previous evidence from Juknevičienė (2008), Luzón-Marco (2011) and Zinkgräf (2008). However, this also shows the importance of considering individual differences such as speaker experience, as well as other variables like age.

4.4.1.2. *Get + education* (B2)

Moving on to the B2 speakers; within this group the most notably frequent verb + noun collocation is *get + education.* Although it is not the most relatively frequent collocation, there is a difference between the occurrences here (0.28 per 10,000 tokens) compared to B1 speakers (0.03) and C1/C2 speakers (0.16), which is why it is under further consideration. 21 out of the 23 instances (91.3%) occur in the Conversation task and looking at the concordance lines further, the topic was introduced by the examiner with the question "[would you say that] education is more or less important today than previously?" Much like *get + money* as explored in Section 4.4.1.1, *education* as a noun could be said to have a more idiomatic verb combination as it is restricted to specific delexical verbs. One such verb is *have.* To explore whether the B2 speakers are engaging in the open choice or idiom principle (Erman & Warren, 2000), the analysis will consider the adjectival modifiers that occur within the verb + noun collocation.

Firstly, there are 33 instances of *have + education* within the B2 speaker data and 21 of these occurrences (63.64%) are modified with adjectives including *free, good* and *bad.* 13 of the 23 occurrences (56.52%) of *get + education* also have modifiers and there are notable differences as to what these modifiers are. These include *good* but also *standard, proper* and *enough. Good* and *bad* are both referring to quality of education while *standard, proper* and *enough* involve judgements on quality, with an added aspect of suitability. This value/suitability judgement is only found with the verb *get* and not *have* when collocating with *education.* Consequently, although the two verbs could be said to be interchangeable, the modifiers indicate they are being used by speakers for different reasons. This shows some nuance of using delexical verbs with the noun *education* within the TLC-L2 by these intermediate speakers, as they are using appropriate collocations rather than considering them as single words and thus engaging in the idiom principle rather than open choice (Erman & Warren, 2000).

4.4.1.3. *Get + information (C1/C2)*

Finally, the C1/C2 speakers' most notably frequent collocation based on Table 44 is *get + information*, which occurs 0.45 per 10,000 tokens compared to 0.13 for B1 speakers and 0.20 for B2 speakers. There is a split between the usage in the Discussion (seven instances) and Conversation (10 instances) tasks; in comparison, all but two occurrences in both the B1 and B2 data were found in the Discussion task so there is a difference in frequency overall, but also differing use within specific exam tasks. The prevalence in the Conversation task for the C1/C2 speakers is likely due to a proficiency level specific topic introduced here, 'use of the internet', which is further discussed in Section 4.2.2.4. Once again, this use of *get* with the noun could be considered to be less idiomatic and this may be influenced by the speakers' overuse of delexical verbs, demonstrating reliance on the open-choice principle and producing language as singular words rather than collocations. To investigate further, the Macmillan Collocation Dictionary (2023b) was again consulted to find other frequently occurring verbs with the noun *information.* From this, it was found that *obtain* and *collect* can be used synonymously with *get* when combined with *information.*

To look into this further, comparing frequency and usage of the differing verbs with *information* was done. As the noun *information* is low-frequency, the BNC2014 was consulted as a reference corpus to explore how native British English speakers use the collocations of interest. *Get + information* occurs 341 times in the BNC2014 (0.034 per 10,000 tokens), which is less relatively frequent when compared to the 43 times in TLC-L2 (0.10 per 10,000 tokens). Within the TLC-L2, there is no occurrences of *obtain + information* by any candidate. This is rare within British English in general, as the BNC2014 only contains 137 occurrences (0.01 per 10,000 tokens). The same is the case for *collect + information*, with no instances found within the 43,644 verb + noun collocations overall. Again, the BNC2014 shows this is also a rare collocation, with 130 instances (0.01 per 10,000 tokens). Because of the lack of production from the TLC-L2 candidates, exploring concordance lines of synonymous collocations is not possible. Overall, it can be said this collocation is topic-influenced and depends on the proficiency level, due to the prevalence in the C1/C2 group data and the fact it is mostly found in the Conversation task for this group. This latter assertion links to previous research from Alexopoulou et al. (2017), which shows that topics used to elicit language from L2 speakers then also shapes the language used.

4.4.1.4. Unique *get + noun* collocations

For the final part of the analysis, the full 43,644 verb + noun dataset is used to look for unique collocations. These are defined as only occurring within one proficiency level. Every *get* + noun collocation was extracted from each proficiency level dataset and Microsoft Excel was used to find these unique collocations. For the B1 group, there were 643 *get* + noun collocations overall and 120 of these were found to only be used by B1 speakers. This means that 18.66% of all the *get* + noun collocations B1 speakers use are unique to the group. For the B2 speakers, 112 collocations were unique to the group from the 665 used overall. This means the proportion of unique collocations is slightly fewer, at 16.84%, than the B1 group. Finally, the advanced speakers used 59 unique *get* + noun collocations from the overall frequency of 345, meaning C1/C2 specific collocations account for 17.10% of the total. This is slightly more than the B2 group but a smaller proportion than the B1 group. Therefore, there is evidence of nonlinear usage of collocations when looking at unique combinations across proficiency levels supporting findings from Duan and Shi (2021). To investigate this further, the semantic categories of the noun node within each unique collocation are considered. This was done using the UCREL Semantic Analysis System (USAS) tagger (Rayson et al., 2004). This follows on from a similar approach taken from Du et al. (2022) who analysed the semantic noun elements within *make* and *take* + noun collocations.

Table 45 Top five most frequent semantic categories of unique get + noun collocations across proficiency levels

B1	**Percentage (raw frequency)**	**B2**	**Percentage (raw frequency)**	**C1/C2**	**Percentage (raw frequency)**
S - social actions states and processes	15% (18)	A - general and abstract terms	11.61% (13)	I - money and commerce	16.95% (10)
B - the body and the individual	11.67% (14)	O - substances materials	10.71% (12)	S - social actions states and processes	13.56% (8)

		objects and equipment			
I - money and commerce	8.33% (10)	X - psychological actions states and processes	9.82% (11)	X - psychological actions states and processes	11.86% (7)
O - substances materials objects and equipment	7.5% (9)	B - the body and the individual	8.04% (9)	A - general and abstract terms	8.47% (5)
M - movement location travel and transport	7.5% (9)	S - social actions states and processes	7.14% (8)	Q - linguistic actions states and processes	8.47% (5)

Table 45 shows the top five semantic categories for each proficiency level, with the proportion of *get* + noun collocations in each as well as the raw frequency of these. It can be seen in Figure 3 that category S (social actions, states and processes) is highest in B1 accounting for 15% of the unique collocations. This dips to less than half for B2 speakers with 7.14% of the collocations. Finally, there is a strong increase to almost B1 levels for the advanced speaker group at 13.56%. This is in contrast to Du et al. (2022), who found the highest occurrence of this semantic category in the B2 speakers. However, there are some similarities to Du et al.'s findings as category A (general and abstract terms) is the most frequent in the B2 speaker data. This category does not occur at all in top five for B1 speakers. This could be demonstrating collocation development in a nonlinear way based on the semantic types of the collocations, as well as based on frequency – as seen previously in the chapter. Another category of note, due to its more abstract nature, is X (psychological actions, states and processes). Example nouns from this category include *fame* and *warning* (B1), *attitude* and *awareness* (B2) and *recognition* and *purpose* (C1/C2). Category X does not appear in the top five for B1 speakers but does for B2, accounting for 9.82% of the unique *get* + noun collocations. This increases to 11.86% for

the advanced speakers of C1/C2. This linear development is different to Du et al. (2022) who found an increase from B1 but a decrease in usage between B2 and C1/C2.

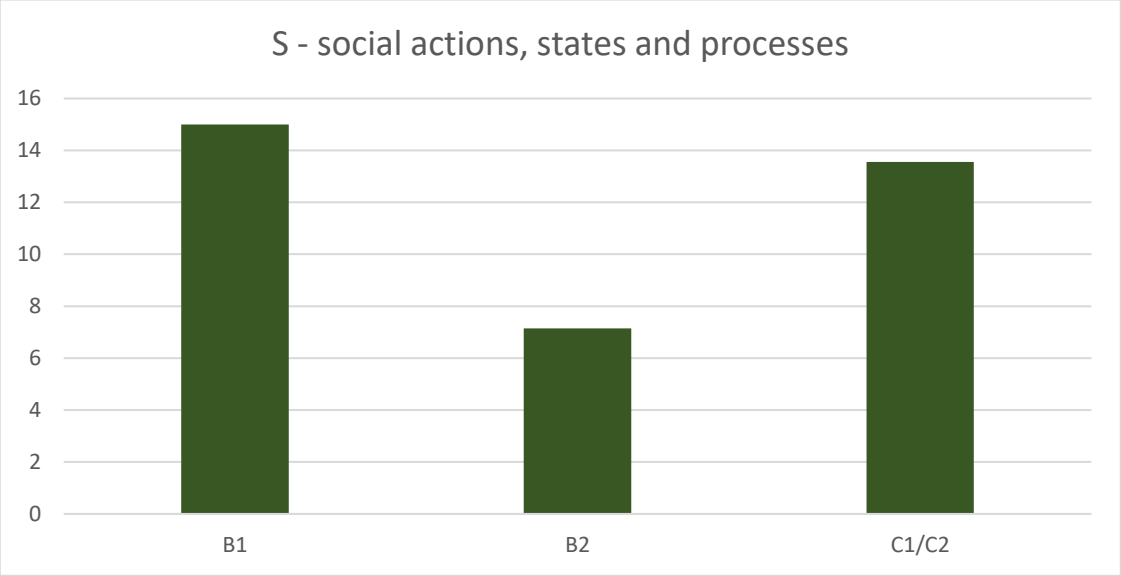

Figure 3 Percentage of S category nouns in high frequency delexical get + noun collocations across proficiency levels

Finally, a category that typically includes more concrete nouns is B (the body and the individual). Example nouns from this category include *shower* and *wipe* (B1), *muscle* and *rash* (B2) and *organ* and *transplant* (C1/C2). The results seen in Figure 4 show category B nouns within the *get* + noun collocations are more frequent in the B1 group (11.67%) than the B2 group (8.04%); this is a similar result to Du et al. (2022). However, the category is less frequent in C1/C2 (6.78%) than either of the B levels, which is different to Du et al. (2022).

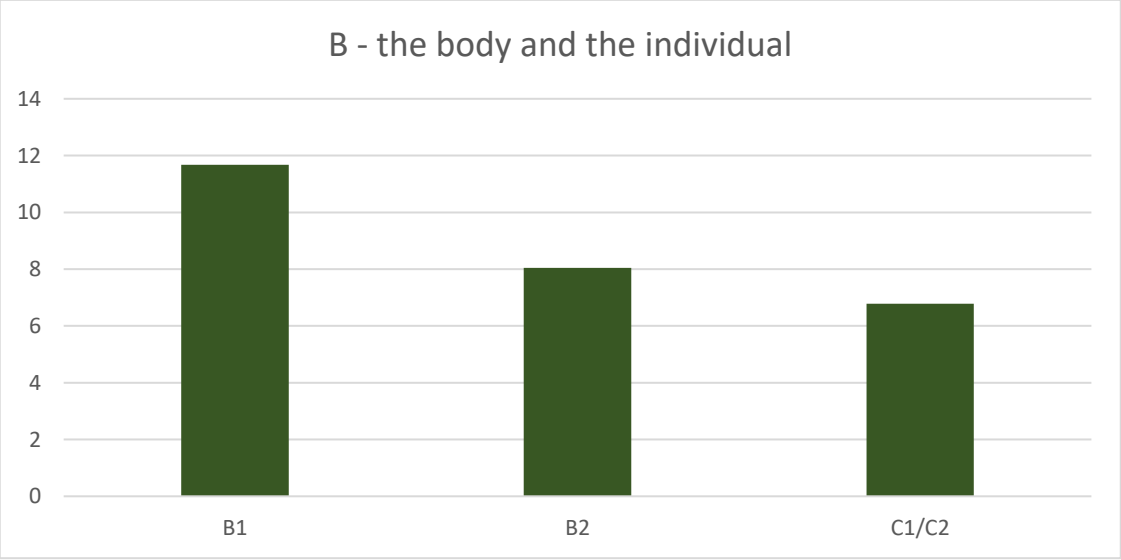

Figure 4 Percentage of B category nouns in high frequency delexical get + noun collocations across proficiency levels

These results indicate there are differences in semantic categories of most frequent unique *get* + noun collocations across the proficiency levels and that this relationship is complex.

4.4.2. *Make*

Make is the second verb node under further investigation. It can be seen in Table 46 that there are 14 different noun collocates that occur across all the proficiency groups, the most frequent being *people*. This is likely the case because *people* is being used in place of a proper noun or pronoun for example in Example (61):

Discussion task; B1 speaker

(61) Candidate 2_6_IN_43: how do you **make people** change?

In these cases, the combination is not a verb + noun collocation based on the phraseological definition and therefore *make* + *people* will not be considered in further detail. The second most frequent collocate is *feel*; this is an abstract noun and occurs most frequently in the B2 group with 0.58 occurrences per 10,000 words (see Table 47). Although this is an abstract-noun collocation, the relative frequency of this decreases for the C1/C2 advanced speakers when compared with both the B1 and B2 levels, suggesting a complex relationship between use of these collocations and language development. This will be explored further in Section 4.4.2.4 when considering the unique *make* + noun collocations.

Table 46 Frequencies of all make + noun collocations that occur in all groups

	B1	**B2**	**C1/C2**	**Total**
make + people	20	57	23	100
make + feel	34	47	12	93
make + friend	43	41	6	90
make + thing	18	32	21	71
make + money	24	17	18	59
make + difference	7	10	13	30
make + decision	3	17	7	27
make + mistake	6	15	6	27

make + sense	2	5	11	18
make + food	5	9	3	17
make + effort	7	2	7	16
make + career	3	5	3	11
make + student	2	5	2	9
make + family	4	3	1	8
make + fun	2	3	2	7

Table 47 Relative frequencies of make + noun collocations per 10,000 words

	B1	**B2**	**C1/C2**
make + people	0.27	0.70	0.61
make + feel	0.46	0.58	0.32
make + friend	0.58	0.50	0.16
make + thing	0.24	0.39	0.56
make + money	0.32	0.21	0.48
make + difference	0.09	0.12	0.34
make + decision	0.04	0.21	0.19
make + mistake	0.08	0.18	0.16
make + sense	0.03	0.06	0.29
make + food	0.07	0.11	0.08
make + effort	0.09	0.02	0.19
make + career	0.04	0.06	0.08
make + student	0.03	0.06	0.05
make + family	0.05	0.04	0.03
make + fun	0.03	0.04	0.05

Looking at Table 47 – with highlighting to show relative frequency in more detail – it can be seen that the most relatively frequent noun collocates of *make* appear in the C1/C2 advanced proficiency group with seven out of the 15 shared noun collocates. This is closely followed by six as the most frequent in the B2 group and only two in the B1 group are the most frequent. These latter two collocations are *make + friend* and *make + family*,

both of which have a more concrete noun that are potentially more tangible for the candidate to understand and thus produce.

4.4.2.1. *Make + friend* (B1)

Make + friend occurs most commonly in the B1 group and there is a significant difference in the frequency of usage between the B1 speakers (0.58) and the C1/C2 speakers (0.16). However, the difference between the B1 and B2 groups is much smaller with the latter having 0.08 fewer occurrences per 10,000 tokens. 13 of the instances include the modifier *new*, as exemplified below.

(62) Candidate 2_6_IT_98: I think it's good to send children there because they can have a lot of fun and they erm can erm er **made new friends**

This decreases to nine uses of *new* in the *make + friend* collocation for B2 speakers, which is still relatively high. This demonstrates a similarity in the use of this collocation across the two lower proficiency levels, with a sharp decrease in the frequency of usage at the advanced level and no speakers in C1/C2 use the adjective *new* when using the collocation. This difference could be due to the topics under discussion and the ages of the participants; for example, younger speakers who account for a substantial proportion of the B1 candidate group may be more frequently in situations where making new friends is common compared to the more advanced speakers, who are generally also older. This is evidenced by just one of the C1/C2 speakers being 13 years old, but the rest are within the range of 18-29 years old. The prevalence of the collocation in this group, therefore, could be due to individual differences in the use of collocations (Halim & Kuiper, 2018).

4.4.2.2. *Make + feel* (B2)

The second collocation of interest is *make + feel*, as it is used most frequently by the B2 speakers (0.58) compared to B1 (0.46) and C1/C2 (0.32). *Feel* is also a noun of interest as it can be more concrete in its semantic field i.e., a physical act but also can be used in a more abstract way to mean being in a particular state or experiencing an emotion. The latter uses of the noun within a more abstract semantic field are more challenging to conceptualise, and therefore using collocations with these kinds of nouns is thought to develop later in language acquisition (Du et al., 2022). Subsequently, it is noteworthy there is evidence here of nonlinear frequency of usage between the three groups, where it increases at B2 level before decreasing for the advanced proficiency group, interestingly in the opposite way to the u-shaped curve observed by other researchers such as Vedder

and Benigno (2016). This is also supporting claims from Larsen-Freeman (2006) regarding the individual fluctuations that occur during language use and eventual development. Considering the concordance lines for the most frequent group usage (B2) and least frequent (C1/C2), there is a difference in the pronoun usage that occur within the *make + feel* collocations. For the B2 speakers, 31/47 (65.96%) use *me* while there are six instances of *you*; for C1/C2 candidates, 5/12 (41.67%) use *me* while only one of the occurrences includes *you* in the collocation. The B2 candidates are engaging with more individually focused discussion using this collocation.

4.4.2.3. *Make + sense (C1/C2)*

The final *make* + noun collocation under consideration is *make + sense.* This occurs least in the B1 group (0.03), with a slight increase for the B2 candidates (0.06). In contrast, the C1/C2 speakers use this collocation significantly more frequently at 0.29 instances per 10,000 tokens. Looking at this collocation further, although *make* is a delexical verb, when combined with *sense* the combination creates a fixed expression to mean easy to understand. Looking at how the collocation is used in context with the C1/C2 candidates, 8/11 (72.73%) of the instances are being used in an overall negative construction, such as in Example (63):

(63) Speaker 2_S_P_37: so it's doesn't **make sense** to to teach or to learn things that are are ou-outdated

There are only two instances of *make + sense* in the B1 group and both are also negative. For the B2 group there is only one instance of the five overall that is negative, with the negation appearing within the collocation:

(64) Speaker 8_IN_4: going to America and eating **makes no sense**

As there are so few instances of *make + sense* in the B1 and B2 speaker data, it is challenging to explore how these speakers may be using the collocation. However, the prevalence of *make + sense* in the advanced speaker group compared to the B1 and B2 candidates could be a demonstration of the C1/C2 speakers using more fixed expressions and thus engaging in the idiom principle with their choice of collocation; they are seeing this combination of verb + noun as a fixed expression for a specific meaning. *Make + sense* is also mostly used in the negative for the C1/C2 speakers. As there is little evidence in the lower proficiency groups, a comparison between the TLC-L2 advanced speakers and the TLC-L1 native speakers was undertaken. In the latter corpus, six instances out of

26 are used in this negative construction (23.08%) which is far fewer than the 72.73% the C1/C2 use. The collocation is also used for a different purpose, as seen in Example (65), to check the examiner's understanding of what the speaker has just said:

(65) Speaker 085: so later down the line like in a couple of years' ti= like a couple months' time whatever they don't have to put the process in to restock it if that **makes sense**

This is a register-influenced collocation and shows a difference between how the L2 speakers use the collocation and the L1 speakers use it to fulfil the purposes of the examination setting i.e., to maintain the interaction between them as the candidate and the examiner (Mauranen, 2004); it is used nine times in this way by the L1 speakers.

4.4.2.4. Unique *make + noun* collocations

Considering the collocations that are unique to each group, for the B1 group there were 555 *make* + noun collocations overall and 136 of these were found to only be used by these B1 speakers; 24.5% are unique to these candidates. For the B2 speakers, 200 collocations were unique to the group from the 817 used overall. This means the proportion of unique collocations is almost exactly the same at 24.48% as the B1 group. Finally, the advanced speakers used 68 unique *make* + noun collocations from the overall frequency of 335, meaning C1/C2 specific collocations account for 20.3% of the total. This is slightly less than both the B1 and B2 groups, meaning there is less diversity of these verb + noun collocations in the most advanced speakers. The B1 group having the largest proportion of unique collocations is paralleled between the *get* and *make* + noun data. These results also contrast with the *get* + noun unique collocations, since with *make* as there is an overall decrease by the C1/C2 level. Frequency of unique collocations does not increase as language proficiency increases for either *get* or *make* + noun collocations supporting findings from Thewissen (2015) regarding a lack of link between production frequency and proficiency level in language learners. Further investigation can be done as to what collocations are used by considering the semantic fields of the nouns within the collocations, once again following on from Du et al. (2022).

Table 48 Top five most frequent semantic categories of unique make + noun collocations across proficiency levels

B1	**Percentage (raw frequency)**	**B2**	**Percentage (raw frequency)**	**C1/C2**	**Percentage (raw frequency)**
O - substances materials objects and equipment	15.44% (21)	O - substances materials objects and equipment	12.5% (25)	A - general and abstract terms	19.12% (13)
F - food and farming	10.29% (14)	B - the body and the individual	10.5% (21)	X - psychological actions states and processes	13.24% (9)
A - general and abstract terms	10.29% (14)	X - psychological actions states and processes	10% (20)	O - substances materials objects and equipment	10.29% (7)
X - psychological actions states and processes	7.35% (10)	M - movement location travel and transport	10% (20)	S - social actions states and processes	10.29% (7)
S - social actions states and processes	7.35% (10)	S - social actions states and processes	7.5% (15)	Q - linguistic actions states and processes	8.82% (6)

Table 48 shows the five most frequent categories assigned to the nouns through the USAS tagging. Looking at the percentages in Figure 5 to see which semantic categories are used

the most, there is a steady increase in X – psychological actions, states and processes from B1 level (7.35%) to B2 level (10%) to C1/C2 level (13.24%). This category typically includes more abstract and complex nouns, and this increase across the proficiency levels shows the candidates are using a larger proportion of these abstract nouns in comparison to the other semantic categories. Looking at Figure 6, there is also a decrease in O – substances, materials, objects and equipment, which typically include concrete nouns such as *sandcastle* and *spike* (B1), *ice* and *compost* (B2) and *powder* and *banner* (C1/C2). Although X and O category nouns were not highlighted specifically in Du et al. (2022), overall, they did find similar evidence in that more advanced learners used collocations tending to belong to semantic fields with abstract social and psychological topics, while the less proficient speakers used collocations with more concrete nouns. Therefore, this pattern of an increase in X category verb + noun collocations and a decrease in O category verb + noun collocations supports previous findings from Du et al. (2022).

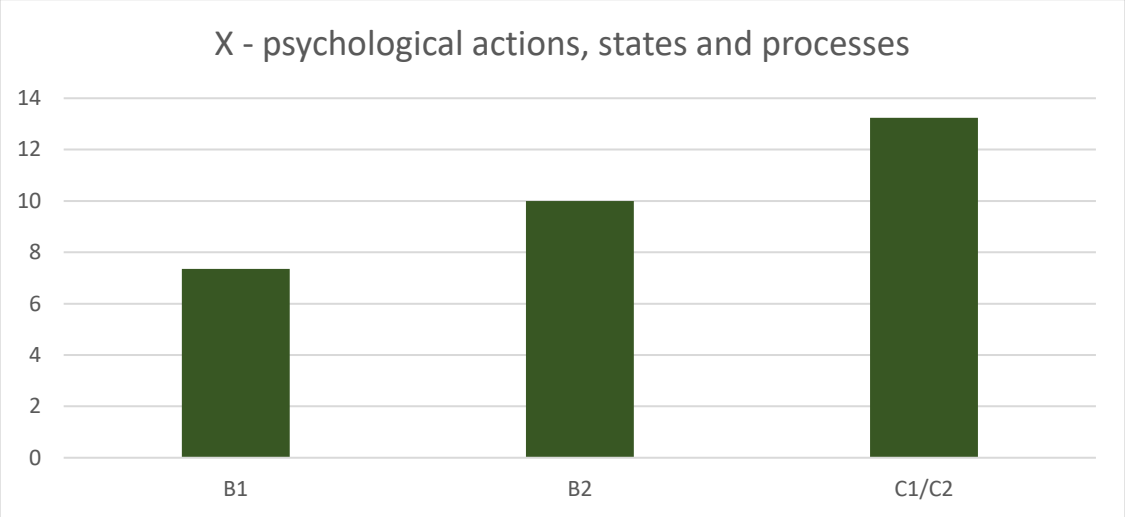

Figure 5 Percentage of X category nouns in high frequency delexical make + noun collocations across proficiency levels

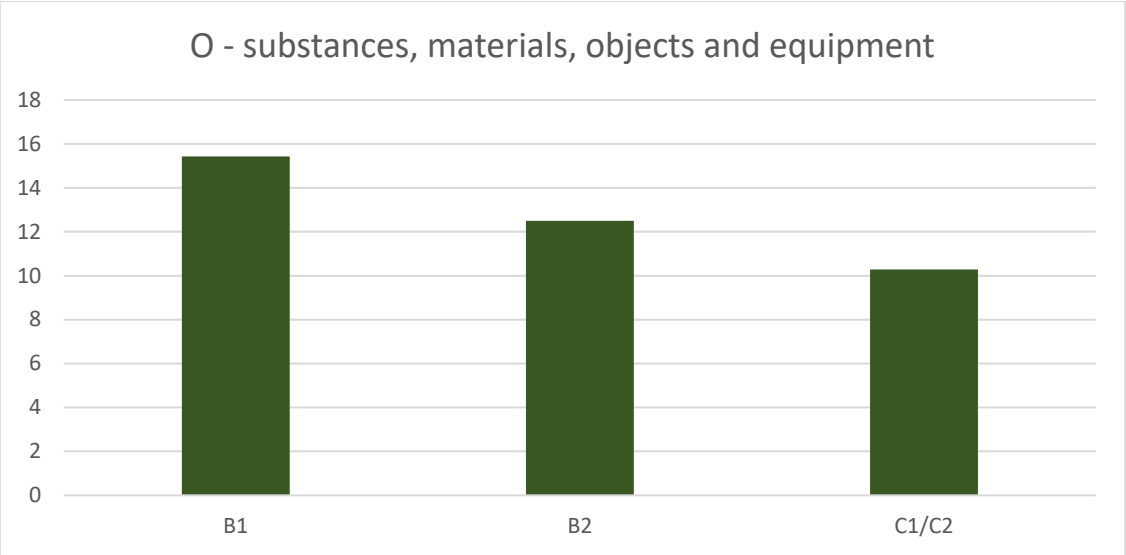

Figure 6 Percentage of O category nouns in high frequency delexical make + noun collocations across proficiency levels

4.4.3. *Take*

Take is the final verb node under further investigation. It can be seen in Table 49 there are nine different noun collocates that occur across all the proficiency groups, the most frequent being *care* and the least frequent *shower.*

Table 49 Frequencies of all take + noun collocations that occur in all groups

	B1	**B2**	**C1/C2**	**Total**
take + care	74	77	104	255
take + photo	46	55	1	102
take + part	42	30	8	80
take + place	22	26	14	62
take + time	15	31	12	58
take + money	20	14	5	39
take + advantage	2	5	14	21
take + drug	1	10	10	21
take + shower	8	5	1	14

Table 50 Relative frequencies of all take + noun collocations per 10,000 tokens

	B1	**B2**	**C1/C2**

take + care	1.00	0.94	2.76
take + photo	0.62	0.67	0.03
take + part	0.57	0.37	0.21
take + place	0.30	0.32	0.37
take + time	0.20	0.38	0.32
take + money	0.27	0.17	0.13
take + advantage	0.03	0.06	0.37
take + drug	0.01	0.12	0.27
take + shower	0.11	0.06	0.03

Table 50 shows the relative frequencies of the most frequent *take* + noun collocations, with highlighting to demonstrate comparison between groups. The most relatively frequent collocation is *take* + *care* for C1/C2 speakers with 2.76 per 10,000 tokens compared to 0.94 occurrences at B2 level and 1.00 occurrences at B1 level. This collocation is also by far the most frequent *take* + noun collocation overall. Finally, there is a fairly even split between which collocations are most frequent in each group – 3/9 in the B1 group, 2/9 in the B2 group with the highest proportion of 4/9 in the C1/C2 group.

4.4.3.1. *Take* + *part* (B1)

This collocation is notable for occurring in the B1 group compared to the other two groups as it occurs 0.57 times per 10,000 tokens, while the B2 speaker group frequency is 0.37 and the C1/C2 less frequent still at 0.21. All but one of the 43 instances for the B1 group use *take* + *part* as a cohesive collocation i.e., there are no interceding modifiers, therefore it could be that this is evidence of an early acquired fixed collocation when considering the phraseological approach to formulaic language (Omidian et al., 2021).

The one instance is likely an error where the speaker was intending on using the fixed collocation:

(66) Speaker 2_6_RU_32: I 'm going to **take a part** in a horse riding competition

Only three of the instances occur within the Conversation task therefore, the remainder are introduced in the Discussion task which is candidate-led. This is in contrast to the C1/C2 speaker usage, which is split equally between the two tasks. Considering this as a

fixed collocation, the study looked into the usage of the semantically similar verb + noun collocation of *be + involved* which occurred in the C1/C2 groups 12 times, while for the B1 group, the number of occurrences was 9. This could indicate that there is a move from the simpler *take + part* to the more difficult construction *be + involved* as language develops. According to the English Vocabulary Profile (Kurtes & Saville, 2008), *involve* with the meaning of *to take + part* is introduced at B2 level on the CEFR. Therefore, it is understandable why there is a preference for B1 speakers to use *take + part* and the more advanced speakers to use *be + involved* as they develop their language proficiency: it is a fixed collocation that is acquired early and then used less as proficiency develops and a more complex collocation is introduced. This ties in with findings from Saito and Liu (2022), who claim that content words used by L2 speakers became more diverse and complex during language development.

4.4.3.2. *Take + time* (B2)

One noteworthy collocation including *take* is *take + time* for the B2 proficiency group. At 0.38 occurrences per 10,000 tokens, this is slightly more frequently used than in the C1/C2 group (0.32) and almost twice as frequent as the B1 group (0.20). Looking at the concordance lines for the B2 speakers, it can be seen that there is frequent modification of *time.* In fact, only three of the 31 instances occurring in the B2 group are the bigram *take time*, the rest include some kind of interceding particle(s) with the most frequent being *a long* to create *take a long time* e.g.:

(67) Speaker 2_8_AR_12: er I think that it will **take a long time**

Take + time as a verb + noun collocation allows for some variation within the construction. The C1/C2 group use it in a similar way, but there is also one instance of the candidate speaking to the examiner outside of the confines of a specific task i.e., it is a register-influenced collocation rather than something that could be said to be more topic-influenced

(68) Speaker IT_61: yeah **take your time**

This inclusion of the second person pronoun shows the speaker engaging in the interactional nature that the GESE focuses on trying to elicit from their candidates, which links to previous research from Jones et al. (2017) into pronoun use in L2 speech.

4.4.3.3. *Take + advantage* (C1/C2)

Looking at the relative frequencies, *take + advantage* is by far more commonly used within the advanced speaker group in the TLC-L2 with 0.37 occurrences per 10,000 tokens compared to B1 (0.03) and B2 (0.06). These instances mostly occur in the Conversation task (9/14). However, on closer inspection of the concordance lines, one candidate – Speaker 2_SP_34 – uses the collocation four times, when it is in fact their self-repair that is being accounted for:

(69) Speaker 2_SP_34: sometimes people take ad= **take advantage** advantage of the situation like oh if I don't work I get er the the the social security money

According to the English Vocabulary Profile (Kurtes & Saville, 2008), this construction is introduced at B1 level; however, it is found very infrequently at B1 within the TLC-L2 and likewise for B2 speakers. This may not necessarily be evidence of the speakers not being able to use the collocation, but that there were minimal opportunities to use it (Caines & Buttery, 2017). Looking at the BNC2014 to see how frequently this collocation occurs within a large L1 corpus, the relative frequency is 0.19 per 10,000 tokens, which is fewer occurrences than found in the C1/C2 candidate groups' language.

One of the instances within the advanced group using *take + advantage* involves some potential priming from the examiner:

(70) Examiner: but do you think they are aware of they are **taking advantage** of those rights?
Speaker 2_ME_20: yes they are taking **taking many advantage** of the rights they are having

Macmillan Collocation Dictionary (2023c) also notes that the *take + advantage + of sb/smt* construction is commonly occurring enough to warrant its own entry – this could indicate that it is a later acquired collocation that L2 speakers produced but may comprehend it sooner, which supports research from Lee (2021) who notes that productive knowledge of collocations seems to be more of a challenge for learners than developing their receptive knowledge of the combinations. This also supports further evidence from Kamarudin et al. (2020).

4.4.3.4. Unique *take + noun* collocations

Firstly, frequency will be considered regarding the unique *take* + noun collocations. Overall, for the B1 group, there are 605 of this type of verb + noun collocation with 130 only used by these speakers; 21.49% are unique to B1 candidates. For the B2 speakers, 165 collocations were unique to the group from the 595 used overall. This means the proportion of unique collocations is 27.73%, which is an increase from the B1 group. Finally, the advanced speakers used 77 unique *take* + noun collocations from the overall frequency of 338, meaning C1/C2 specific collocations account for 22.78% of the total. This is fewer than the B2 group but slightly more than the B1 group. In comparison with the other two high frequency delexical verbs, *take* + noun collocations at B1 level are the least diverse, while for *get* and *make* they are the most. None of the three high frequency delexical collocations have the most unique instances at C1/C2 level, meaning frequency of unique collocations does not increase as language proficiency increases for *get*, *make* or *take* + noun collocations. Finally, extended analysis into what semantic fields the nouns within these collocations can be done to see how they are being used by the speakers.

Table 51 Top five most frequent semantic categories of unique take + noun collocations across proficiency levels

B1	**Percentage (raw frequency)**	**B2**	**Percentage (raw frequency)**	**C1/C2**	**Percentage (raw frequency)**
S - social actions, states and processes	12.31% (16)	A - general and abstract	12.73% (21)	S - social actions, states and processes	15.58% (12)
B - the body and the individual	11.54% (15)	X - psychological actions, states and processes	10.91% (18)	I - money and commerce	15.58% (12)

O - substances, materials objects and equipment	10% (13)	B - the body and the individual	9.09% (15)	X - psychological actions, states and processes	12.99% (10)
A - general and abstract	8.46% (11)	S - social actions, states and processes	8.48% (14)	A - general and abstract	9.09% (7)
I - money and commerce	6.92% (9)	Q - linguistic actions, states and processes	7.27% (12)	Q - linguistic actions, states and processes	7.79% (6)

Table 51 shows the five most frequent categories assigned through the USAS tagging to the nouns used with *take*. Looking at the percentages to see which semantic categories are used within these collocations, there are three patterns of particular interest. Firstly, for S category – social actions, states and process, there is a u-shaped curve in the usage with 12.31% at B1, dipping to 8.48% at B2 before increasing again to 15.58% in the C1/C2 group and this can be seen in Figure 5. This is similar to what is exhibited in *get* + noun collocations (see Section 4.4.1.4, Figure 3). Both findings support the results from Vedder and Benigno (2016) regarding u-shaped development, Siyanova-Chanturia and Spina (2020) explaining language development can get worse before getting better and Larsen-Freeman (2006) claiming that individual fluctuations occur during language use and subsequent development. Examples of some of the nouns in the B1 group are *divorce, team* and *inspiration,* in the B2 group are *contestant, blessing* and *lifestyle* and in the C1/C2 group are *respect, power* and *influence.* Secondly, the B category – the body and the individual only occurs in the top five for B1 and B2. Nouns within this semantic field are typically more concrete. For example, in B1 these include *sleep, pain* and *refreshment* while B2 include *cancer, cough* and *dress.* The decreased occurrences of this semantic category within the advanced C1/C2 candidates indicates they are using fewer concrete nouns with this delexical verb; this can be seen in Figure 8 below and is similar to findings

with *get* + noun in Section 4.4.1.4 (see Figure 4). Finally, much like the unique *make* + nouns, X category – psychological actions, states and processes is of interest. This semantic field does not occur in the top five most frequent categories for B1 but does for B2 (10.91%) and again slightly more frequently for C1/C2 (12.99%); this could be evidence that this category of noun is being used creatively (uniquely) only at the B2 (threshold) level and beyond for this verb. Overall, the evidence from this analysis also supports Du et al.'s (2022) findings of more proficient learners using more abstract nouns in their high frequency delexical verb + noun collocations.

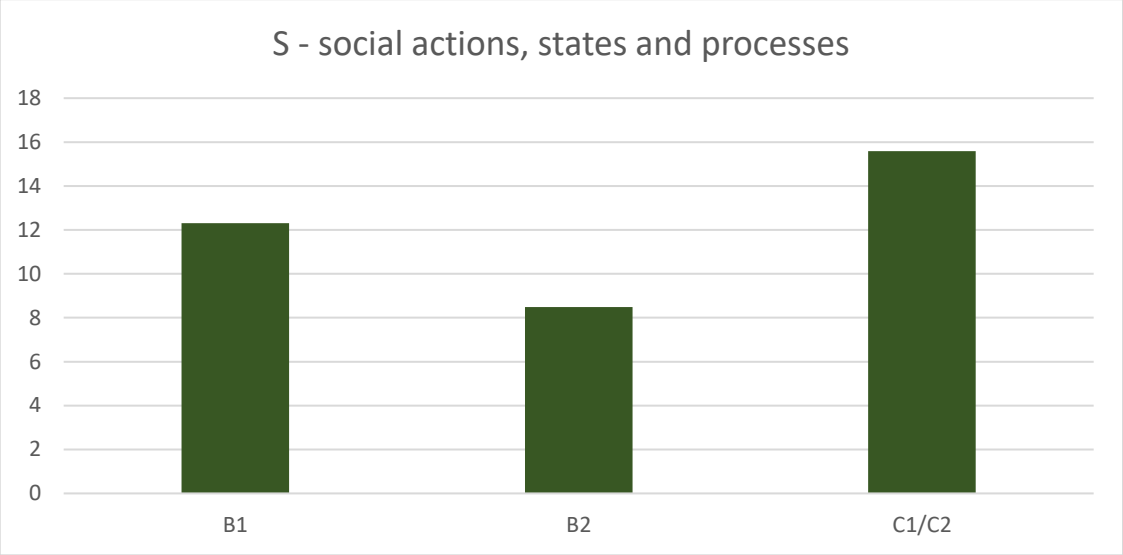

Figure 7 Percentage of S category nouns in high frequency delexical take + noun collocations across proficiency levels

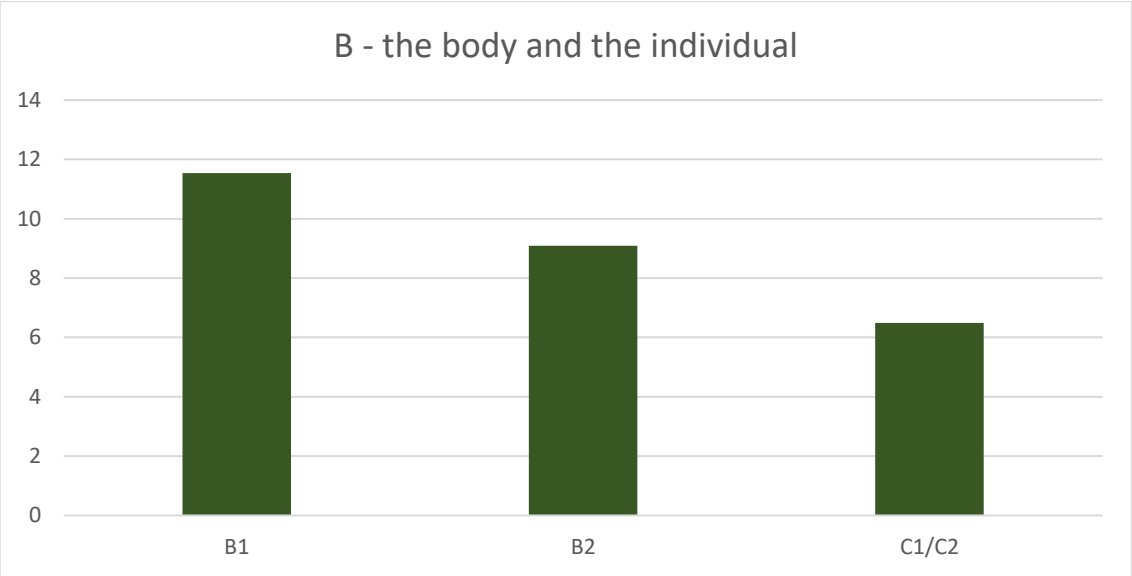

Figure 8 Percentage of B category nouns in high frequency delexical take + noun collocations across proficiency levels

4.5. Summary

This analysis chapter has quantitatively and qualitatively explored answers to the research questions for the TLC-L2. The chapter considered all verb + noun collocations extracted based on a CQL query before focusing on the shared collocations that are present in each of the three proficiency levels within the TLC-L2: B1 (threshold), B2 (intermediate) and C1/C2 (advanced). This found that certain verb + noun collocations, despite occurring throughout the levels, were more frequent in specific levels than others. The implication from this is that these are topic-influenced collocations that are occurring due to the specific subjects posed by the exam that the candidates are thus at a level of language proficiency to engage with successfully. This supports previous findings regarding topic-influenced language occurrence in language learners. As well as considering the shared collocations, the research also explored frequent verb types within these verb + noun collocations. This analysis found there to be significant evidence of, again, topic-influenced collocations but also register-influenced collocations. Furthermore, this aspect of the analysis also began to suggest abstract-noun collocations were of interest to explore further, as examples of register-influenced collocations. The final part of the analysis focused on high frequency delexical verb + noun collocations, specifically highlighting *get*, *make* and *take*. As well as exploring one notably frequent collocation per proficiency level, the analysis also expanded to consider collocations that were unique to each level, based on the verb under review. Within this analysis, the exploration of the abstract-verb

collocations deepened by tagging nouns of the unique collocations with the USAS semantic tagger. This meant the abstractness of the nouns could be explored in a more systematic way and followed previous research from Du et al. (2022). The next chapter will explore the TLC-L1 corpus with a similar approach to the analysis; first, considering frequent verb + noun collocations, collocations unique to the corpus, topic- and register-influenced collocations and finally collocational patterns in high frequency delexical verb + noun collocations.

Chapter 5: TLC-L1 Results and Discussion

This chapter presents the results from the TLC-L1 analysis and introduces these findings to situate them within current literature, thereby answering the study's research questions. First, Section 5.1 explores the most frequent verb + noun collocations in the corpus before occurrences of unique verb + noun combinations are discussed in Section 5.2. Next, Section 5.3 examines frequent verb types within verb + noun collocations, with collocation patterns in high frequency delexical verb + noun collocations investigated through three verb case studies in Section 5.4. Finally, Section 5.5 brings the chapter to a close.

5.1. Most frequent verb + noun collocations

5.1.1. Overview

The TLC-L1 subcorpus of the Discussion and Conversation tasks contains 2,150 verb + noun collocations that had a dispersion of 50 or above in the BNC2014. Of these, there are 1,181 different types. These are the L1 collocations that will be investigated further in this research and later explored in context and compared to the verb + noun collocations found in TLC-L2. Each speaker in the TLC-L1 uses 10.59 of these collocations on average (SD=4.78). The relative frequency in the whole L1 corpus is 7,974.98 per million words.

Considering the most common collocation types, Table 52 presents the top 30 types within the TLC-L1 and includes how often they occur, and how many of the 203 candidate speakers in the corpus use them.

Table 52 Thirty most frequent verb + noun collocation types in the TLC-L1

Collocation	**Frequency**	**Speakers**
make + decision	31	21
make + sense	22	16
eat + meat	21	11
get + people	18	15
get + money	16	14
make + difference	15	13
see + people	14	12
say + thing	14	13

give + people	14	9
take + time	13	13
make + money	13	12
make + choice	12	10
tell + people	11	9
see + thing	11	9
mean + people	11	11
live + life	11	10
make + people	11	8
encourage + people	10	9
get + friend	10	10
commit + crime	10	9
spend + time	9	7
see + point	9	9
say + people	9	8
make + change	9	8
get + time	9	7
ask + question	9	7
make + thing	9	8
take + responsibility	8	7
go + way	8	5
eat + food	8	6

The table shows that *make + decision* is the most frequent verb + noun collocation in the dataset, with 31 instances. 21 speakers used this collocation which is 10.34% of the speakers within the corpus, showing that it is a relatively common combination when considering both frequency and dispersion in the TLC-L1. Furthermore, eight of the collocations include the generally common English nouns *people* and *thing* i.e., *see + people, give + people, tell + people, mean + people, make + people, encourage + people, say + people, say + thing, see + thing* and *make + thing*. Due to the high frequencies of the component nouns and verbs, many of these collocations are predictable and could be expected to appear in any corpus of spoken English. However, as has been explored in Chapter 4 with the TLC-L2 analysis, other collocations show a significant influence from

the context of the examination to the language used by the candidates. These influences can be categorised into i) topic-influenced collocations – collocations that are used because of the theme of the interaction that has been set by either the candidate or the examiner – and ii) register-influenced collocations – these are used because of the situational parameters of the interaction; that is, the exam invites opinions and having candidates explain and defend their positions on topics, as well as maintaining conversation and attempting to demonstrate communicative competence. The analysis in sections 5.1.2. and 5.1.3. will explore topic-influenced and register-influenced collocations, respectively, to provide an in-depth qualitative analysis of these combinations and to exemplify them with the L1 language from this corpus.

5.1.2. Topic-influenced collocations

The following section details three collocations that frequently occur in the TLC-L1 and have been identified to be potentially topic-influenced due to the specific nature of the noun collocates within the combinations. These collocations will be further qualitatively analysed through concordance lines.

5.1.2.1. *Eat + meat*

The first verb + noun collocation under analysis is *eat + meat.* This was chosen as a notable inclusion in the top 30 collocations, initially due to its unexpected presence amongst more commonly used and high frequency delexical verb collocations such as *make + decision.* To check this intuition, *eat + meat* was searched for in the BNC2014 and only 305 instances were found (0.03 per 10,000 tokens) in 0.21% of the texts. With 21 instances, *eat + meat* it is the third most frequent verb + noun collocation in the TLC-L1. However, only 11 speakers (5.4%) throughout the corpus use it and within these 11 speakers, one candidate – Speaker 170 – uses it seven times. This is evidence of how distribution amongst speakers is important to look at as well as frequency, as although it is frequently occurring, 1/3 of the instances can be attributed to one speaker.

This verb + noun collocation demonstrates the influence topic can have on language use. Here, Candidate 170 chose to present the topic “should we eat meat?” in the Presentation task and this led to the interaction in the Discussion task shown in Examples (71), (72) and (73) below, which underlines the introduction of the topic, with the collocations presented in bold.

(71) Candidate 170: okay erm yeah so today I'm just gonna talk about er meat and whether people should eat it or not cos I 've erm **eaten meat** my whole life apart from a f= well until a few months ago I gave up eating meat so I just thought it was something interesting to talk about (Presentation task)

(72) Candidate 170: if you **eat meat** every day youre getting vitamin B twelve every day (Discussion task)

(73) Candidate 170: yeah would be interesting to see but now I've been sort of two three months without **eating meat** I kind of don't want to go back (Discussion task)

Expanding the analysis to the rest of the candidate's exam, Candidate 170 uses the verb *eat* as a lemma 28 times in the sub corpus, with it occurring 10 times in the discussion task. These occurrences represent a large proportion (22.05% in subcorpus and 24.39% in the Discussion task) of the 127 uses of *eat* in the corpus overall. This further demonstrates the influence one topic, and in this case one speaker, can have on the language found within a relatively small corpus. The choice of Speaker 170 to talk about eating meat is likely to have limited the variety of collocations they were able to select from, given the time constraints of the task (5 minutes; Trinity College London, 2021). This in turn may have led to their speech sounding repetitive to the examiner and could account for the lower Discussion task score the candidate received (a C grade) when compared to their high score on the c-test (93%). Their language ability is evident based on the vocabulary test, but this did not necessarily translate to the Discussion task grade.

In comparison, Example (74) demonstrates a candidate choosing a similar topic, veganism, but one that allows for broader linguistic choices due to its discursive scope. This may in turn allow for more opportunity to use more varied language including more diverse collocations leading to perceived mastery of language. Candidate 016 also uses the verb *eat* frequently – indeed, 11 times in the Discussion task, which is higher than Candidate 170 – yet only uses the collocation *eat + meat* once. From the concordance analysis, it can be seen that Candidate 016 is discussing a similar topic, veganism, but there is nuance to the presentation and language use within this, based on the candidate's perspective and framing of the topic:

(74) Candidate 016: I've chosen to talk about veganism erm because I follow the vegan lifestyle and I've been a vegan for two years now erm it's a topic that I'm really passionate about (Presentation task)

Although it is within a similar topic sphere as 'choosing to eat/not eat meat', veganism as a topic encompasses much more than detailing dietary choices as it is a lifestyle which means this a broader topic to engage with. This could be why Candidate 170 graded lower than expected based on what their c-test would indicate; Candidate 016 in comparison had a c-test score of 85%.

Considering verb + noun collocation diversity of these two speakers, Candidate 016 used 8 types once from the 1,181 chosen for further analysis (5.93 per 1k); these are: *commit + crime, kill + animal, take + step, get + vote, hold + people, go + back, get + funding* and *eat + meat.* This is in comparison to Candidate 170 who uses 13 types in 21 instances (14.07 per 1k). When compared to the mean for all speakers – 10.56 – Candidate 016 is using fewer of these core collocations while Candidate 170 is using more. This result may be due to the overall number of verb + noun collocations differing and/or the overall text length. However, both speakers have very similar relative frequencies for verb + noun combinations and tokens (Candidate 016 – 96.44 per 1k; Candidate 170 – 98.46). The difference between the two speakers and their frequency of use of the core collocations could be individual speaker variation (Larsen-Freeman, 2006) or it could be that Candidate 016 chose a topic that allowed for a more diverse range of verb + noun collocation choices. They had similar amounts of verb + noun collocation density but Candidate 170 used far more of the most frequent collocations than Candidate 016 as well as above the mean for all the speakers. From this, it could be that the topic is not only influencing the specific verb + noun collocations used but also the diversity of collocations that are possible. Garner (2020) found that more proficient L2 writers used more diverse collocations. By looking at these L1 speakers, we can see there is also variation in diversity of verb + noun collocations which could be based on individual variation (Lowie & Verspoor, 2015), though the qualitative analysis of the topic suggests this could also have an impact too. Encouraging candidates to expand on topics and/or including broader topics within language testing therefore could be beneficial for supporting varied language choice opportunities for the speakers leading to further success in their examination (Buttery & Caines, 2012).

This absence of the collocation *eat + meat* on the same topic of veganism is also demonstrated by Candidate 066 who again talks about veganism but does not use the collocation *eat + meat.*

(75) Candidate 066: okay so I'm gonna talk about veganism (Presentation task)

Candidate 066 mentions *eat* 4 times, but not as the collocation *eat + meat*. Again, this supports the claim that veganism is a broader topic, compared with choosing to (not) eat meat, and arguably, is does not limit the speaker to a set of collocations. Indeed, we can consider veganism a lifestyle, with many factors to discuss and the different subtopics allow more lexical choices. Furthermore, all instances of *eat* are in the Discussion task, demonstrating that the examiner is asking about dietary choices based on the topic the candidate introduced. Ultimately, the speaker's choice of topic can lead to more opportunities for collocational diversity (Buttery & Caines, 2012; Weigle & Friginal, 2015).

Further analysis showed that five instances of *eat + meat* occurred in the Conversation task, in which the examiner introduces the topic. Specifically, the topics were healthy eating and animal rights. The remainder of the collocations were observed in the Discussion task i.e., the topic was candidate-influenced. This observation demonstrates that the production of collocations can be informed by both candidate-led and examiner-led topic introduction. Furthermore, since these topics principally concern different issues, we can say that there is topic influence on the verb + noun collocations used, but this is not straightforward. Rather, there is nuance to how the topic manifests in the language. This is to be expected and shows how formulaic language use is individualised (Omidian et al., 2021) i.e., that collocations can be used and made relevant across a range of topics. Nevertheless, some topics may lend themselves to more complex language use than others, which could impact the overall grading of their exam and so teachers preparing their students to undertake examinations where there is freedom to choose topics should keep this in mind (Paquot, 2020) Furthermore, item writers need to choose topics appropriate for the level of exam (Huang et al., 2018).

Another finding from studying this collocation is the potential effect of topic priming from the examiner (Lazaraton, 1996). Although it does not specifically mention *meat*, the noun *roadkill* is a near-synonym that is used by Examiner 016 and Candidate 016 within another conversational exchange. It is natural to repeat language in a conversational turn

and in Example (76), the candidate's use of the same verb + noun collocation immediately after the examiner's use of the combination indicates that the examiner as an interlocutor has an influence on the kind of language a candidate uses::

(76) Examiner 016: what 's your viewpoint on the the growing number of people for example that **eat roadkill**?"
Candidate 016: erm I think if people **eat roadkill** it's a completely different matter

Overall, two potential conclusions can come from looking at the collocation *eat + meat* in relation to topic-influence. The collocation is likely to have been influenced by topic based on (1) how two speakers use this in comparison to their use of other verb + noun collocations and (2) how frequent the collocation is within the candidate-led and examiner-led tasks. Finally, there is also some evidence of examiner-priming having an effect on language choices of a candidate, which is a typical feature of examination language as the examiner supports the candidate in the interaction (Lazaraton, 1996). The dialogic patterning between examiner and candidate contributes to the overall register of the exchange and so while there is clear evidence surrounding *eat + meat* as a topic-influenced collocation, there is also instances that could be categorised as register-influenced.

5.1.2.2. *Commit + crime*

The second in this analysis of topic-influenced collocations is *commit + crime.* Much like *eat + meat,* this collocation was selected for further investigation based on it being unusually frequent when compared to more commonly expected combinations. To test this intuition, the BNC2014 was again consulted. 245 instances were found (0.02 per 10k) across 202 texts for *commit + crime,* much less frequent than a more intuitively common combination such as *make + decision* (3,408; 0.34 per 10k; 2,479 texts). This is the first indication that there may be something about the topics covered in the corpus to result in this being a top 30 verb + noun collocation. Within the TLC-L1, there are 10 instances of the *commit + crime* and it is used by nine different speakers (4.43% of speakers), which shows that although it is not used frequently by every speaker, the distribution of occurrences is broad and only one speaker uses it twice. When looking at topics, it can be seen that eight of the instances are based on a discussion of the death penalty as a punishment and deterrent for crime. The other two instances are related to journalism and celebrity phone hacking.

(77) Candidate 143: I don't I don't it won't stop them **committing crimes** cos that's your choice (Conversation task)

As is the case in Example (77), the majority of these collocations (seven) are found within the Conversation task, where the examiner introduces the topic. A further two instances are from the same speaker who chose to talk about capital punishment in their Presentation task, thus leading to the Discussion task including this collocation. This is a further example of how topic can influence the collocations used within these interactions, in both examiner-led and candidate-led tasks. This observation echoes Alexopoulou et al. (2017), who found that "the topics used to elicit the L2 samples shape the language that is represented in the corpus" (p. 181). Furthermore, this this is a case in which issues identified for L2 language study can also be applied to analysis of L1 corpora.

5.1.2.3. *Eat + food*

Eat + food is the third topic-influenced collocation under consideration. In comparison to *eat + meat* and *commit + crime*, it is more commonly found in British English based on frequency counts, with the BNC2014 showing 706 instances (0.07 per 10k over 519 texts). Considering the collocation from a phraseological approach, the elements are somewhat commutable (Nesselhauf, 2005), and this flexibility may account for its inclusion in the top 30 verb + noun collocations of the TLC-L1. More can be revealed with the *eat + food* collocation and its link to topic by considering its internal modifications. In the nine instances present in the corpus, all but one are modified in some way. Examples of this include:

(78) Candidate 169: ultimately if even after being taught the right way to eat and they end up going home to a family where they don't **eat particularly healthy food** that's going to influence them more than what they 've been taught at school (Conversation task)

(79) Candidate 182: so you can like **eat junk food** but like only doing like a day of smoking can like really damage your health (Conversation task)

All the instances of *eat + food* are found in the examiner-led Conversation task, meaning this particular collocation is likely to be occurring due to the pre-set topics from Trinity College London (2023b). Indeed, every instance of *eat + food* is related to the topic of 'healthy eating'. This is another demonstration of how topic can influence the

collocations found within a corpus echoing findings from Paquot (2014); without this topic, the candidates potentially would not have had the opportunity to use *eat + food*.

5.1.3. Register-influenced collocations

The second part of this analysis into context will look at register-influenced collocations. As previously discussed in the literature review, this is important to consider as certain formulaic language sequences are used in certain situations, e.g., stance and discourse organising lexical bundles in speech (Biber et al., 2004). Subsequently, such formulaic language sequences can help to define registers (Hyland, 2008). Looking at register-influenced collocations can give an indication as to whether a speaker is using language appropriate to the context and have an awareness of genre (Gablasova et al., 2017) and this is especially crucial with the TLC-L1, as it is unique in having L1 British English speakers undertaking a language examination – a register these speakers typically would not be familiar with. The collocations discussed in this section are the verb + noun instances that are likely present due to the context of the examination and the language functions the candidates (and examiners) engage with during the process of this interaction, such as asserting and implying (Trinity College London, 2021). To maintain focus, five collocations that are particularly salient for register will be investigated in more detail. These are *make + decision*, *make + sense*, *make + difference*, *make + choice*, and *ask + question*. It is interesting to note that four of these collocations include the verb *make*, a high frequency delexical verb that has been previously investigated by several researchers including Gilquin (2007), Kim (2002) and Sawaguchi and Mizumoto (2022) and that will be considered in more detail in this thesis in Section 5.4.2.

5.1.3.1. *Make + decision*

In the sub corpus, there are 31 instances of *make + decision* from 21 different speakers, which means 10.34 % speakers overall are using this collocation. It is also the most frequent verb + noun core collocation of the TLC-L1 among the 1,818 selected for further analysis here. 24 of these occurrences are within the Conversation task in the exam, where the topic has been introduced by the examiner. These topics have been designed to be discursive and therefore, give rise to differing opinion. Examples include the Brexit referendum, euthanasia, and the death penalty. It follows that such topics are leading candidates to 'make decisions' about their opinions on the topic; however, *make + decision* is not strictly a topic-influenced collocation. Instead, it has been categorised as register-influenced, as its presence demonstrates engagement with the rhetorical purpose

of the exam. The GESE aims to test "authentic communicative skills for real life" through a "genuine discussion format" (Trinity College London, 2023c) and does this through using discursive prompts that can encourage collocations such as *make + decision.* This also demonstrates that the test has been created with a 'bias for best' approach in mind (Fox, 2004), which ensures opportunities for candidates to do well during testing. The collocations of *make + decision* have various internal modifications that include *best, informed, kind of, own, right, conscious* and *any,* showing elements of grading, decision making and subsequently, creating more complex collocations. Combinations that include infrequent, abstract and/or complex words have been found to be strong factors for perceived comprehensibility in L2 speech (Saito, 2020). It is of interest that complexity of the collocations may be arising in the L1 speakers, through the use of internal modifications to evaluate the noun in the combination.

In Example (80) we can see the two interlocutors engaged in the Discussion task and talking about the topic of badger culling. The examiner probes the candidate with a question to test their opinion on the subject they brought to the exam in the Presentation task. This is evidence of the examiner adhering to the register of the exam by challenging the speaker's perspective to have them engage in the language functions of asserting and affirming (Trinity College London Exam Specifications, p. 64). The candidate uses the collocation *make + decision* in reference to the question, which acts to maintain the discussion and supports the purpose of giving a reason for their opinion. This therefore demonstrates the communicative skills that are being tested at the GESE Grade 12 (p. 50) and shows the candidate using language appropriate to this register.

(80) Examiner 014: but surely there's nothing to be ashamed of to be persuaded by evidence
Candidate 014: erm I think that's true but I think I I just think that once they **make a decision** they try and hold onto that for as long as possible

Another example comes from Candidate 085 who has been posed a point for debate by the examiner in the Conversation task. As with Candidate 014, this speaker uses the verb + noun collocation to help support their answer and maintain the interaction with the examiner, which is part of the requirement of the task and thus the register. They also use the internal modifier *kind of* to hedge their answer to show they are being cautious. Here, the candidate is engaging in "softening and downplaying propositions" (Trinity College

London, 2021, p. 51), which is a core language function tests at Grade 12, with *I guess* also working to support this hedging.

(81) Examiner 085: I think she had a point when she said er people feel these days rather than think erm I wonder what the implications of that are
Candidate 085: erm well I guess it's that people **make kind of decisions** based based more off like empathy and f= and and more off like feeling bad

These examples show the candidates using the collocation *make + decision* to help maintain the spoken interaction in response to a question posed by the examiner. There is also evidence of the collocation being used for interaction maintenance with an example of collocation echoing. One instance, as seen in Example (82), has the examiner using the collocation just before the candidate repeats it back and they also restate the topic. Again, this repetition is helping to build the interactional conversational language that is expected in the exam register. Tannen (2007) also speaks to the value of repetition in conversation, stating it is "the central linguistic meaning-making strategy, a limitless resource for individual creativity and interpersonal involvement" (p. 101). This is also evidence of interlocutor influence occurring within an L1-L1 interaction, thus extending findings from L1-L2 interaction studies in a different context (Rosas-Maldonado, 2017).

(82) Examiner 036: part of me thinks well maybe it's because is it **men that are making these decisions**?
Candidate 036: yeah well I think and that's probably the case in the sense that is comes down to that attitude I was er it gets into a revolving circle then doesn't it? is it women are not able to progress because it's the **men at the top making these decisions**?

5.1.3.2. *Make + sense*

Another highly frequent collocation found in the TLC-L1 that includes *make* as the verb is *make + sense.* This collocation has 22 instances from 17 different speakers (8.37%). Compared to *make + decision,* there is much less diversity in the internal modification of the collocation; it is more 'fixed' in this sense, as additions to the collocation only include *any* and *a lot of* when considered from a phraseological perspective. Seven of the collocation instances work as a check for understanding, from the candidate to the examiner. This means the collocation functions differently from simply giving information, as demonstrated with Example (83).

(83) Candidate 028: the probably the biggest message is return and repent and believe from your sins and looking after the world is part of god's design for humanity and so the two are interlinked but they're not exactly one-to-one erm s= **does that make sense**?

Overall, there is an even split between frequency in Conversation (11 instances) and Discussion (11 instances) tasks. Interestingly, the seven instances where the verb + noun collocation works as a clarification request were only in the Discussion task. As this task is candidate led, it shows that the L1 candidates are using these checks for understanding, to engage in the rhetorical purpose of the exam by "actively seek(ing) ways in which to engage the examiner in a meaningful exchange of ideas and opinions" and "take full responsibility for the maintenance of the discussion" (Trinity College London, 2021 p. 50). Therefore, this verb + noun collocation acts as a register-influenced collocation when used for checking for understanding, as it fulfils these communicative skills set out by Trinity College London (2021) for the exam.

5.1.3.3. *Make + difference*

A third collocation within the register-influenced category is *make + difference*, with 15 instances from 13 different speakers (6.40%). Much like *make + sense*, there is some limited internal modification – *big, all the, up the* – suggesting this collocation is more fixed in its structure when considering it from a phraseological approach (Gyllstad & Wolter, 2016). The most common topic for producing this collocation is environmentalism and climate change, with eight of the instances relating to this in some way. Overall, 12/15 instances (80%) are found in the Conversation task, demonstrating the complex and globally relevant nature of the topics that are posed by examiners in this part of the exam. The reason this was deemed to be a register-influenced collocation rather than topic-influenced is that it is not a specific topic that is eliciting *make + difference*; instead, it is the general purpose of the exam task and thus, the register. The GESE Conversation task introduces "an opportunity for a realistic exchange of information, ideas and opinions" (Trinity College London, 2021, p.7) and is designed for more abstract subject areas in the higher grades, such as environmentalism. Using the collocation *make + difference* supports the candidates as they discuss the complex topics. Example (84) shows this in context, where the examiner has introduced the environment as the subject area and has begun the task with the following question to engage the candidate:

(84) Examiner 094: it does make me think isn't it a little too little and too late really ?

Candidate 094: hopefully we can **make a difference** erm because political and f= erm political things are very important but if we don't sort out our environmental situation there's going to be no political er debates anyway

The collocation from the speaker also works to answer this intentionally confrontational question from the examiner – as they are instructed to challenge the opinions of the candidates – thus fulfilling the requirements of the exam to demonstrate interactional competence (Galaczi & Taylor, 2018).

5.1.3.4. *Make + choice*

The final *make* + noun collocation that can be said to be register-influenced is *make + choice.* This collocation has 12 instances from 10 different speakers (4.93%), with internal modifications including *conscious, own, really good* and *few*. *Make + choice* mostly occurs within the Conversation task, which is expected since the purpose of this task is for the examiner to pose subjects for the candidate to comment on with their personal judgements. Furthermore, this collocation is used in relation to topics that have been previously mentioned i.e.: politics, child marriage, euthanasia, healthcare and healthy eating, COVID and veganism. Whether in the Discussion task or the Conversation task, these topics are discursive in nature and the collocation *make + choice* is appropriate to use for these language testing tasks as the examiner is encouraging the candidate to discuss their perspectives on either their chosen topic (Discussion) or the exam-set subject area (Conversation), as seen in Examples (85) and (86).

(85) Candidate 066: I never try to force my views on anybody but I think it's time people started **making conscious choices** (Discussion task; talking about veganism)

(86) Candidate 151: I think a lotta people are **making some really good choices** during this lockdown though (Conversation task; COVID and lockdown habits)

Again, these collocations tend to be register-influenced rather than topic-influenced, because the range of topics they are referring to have been brought about by the context of the examination. In other words, the communicative function of giving opinions on a

topic inherent in the GESE (Boyd & Taylor, 2016) is likely driving the language choices, rather than just the content of the topic itself.

Make + choice is similar in meaning to *make + decision* and the fact there is very little overlap in speakers who use both may indicate it is being used for the same purpose. Only Candidate 150 and Candidate 187 use both *make + decision* and *make + choice* from the 31 speakers combined, and both use it to express their opinions on the subject of euthanasia in the Conversation task. The examiner poses the question as follows, with the prompt remaining the same for both candidates:

> "some people say we have a right to **choose** our way of death we have a right to **choose** where and when we die and say it's a government issue and must be controlled by the government, what's your view?"

(87) Candidate 150: I would never like to be the person in that position who has to **make the choice** erm…
…erm if if people have problems erm with their understanding they might not be **making a erm a straight decision** (Conversation task; euthanasia)

(88) Candidate 187: so I believe that it's down to yourself and your family to **make that choice**…
…I feel the government doesn't personally know you so I feel like they can **make that decision** for everyone (Conversation task; euthanasia)

This shows there could be some interlocutor influence from the examiner on what noun the speaker is choosing to collocate with the verb *make*, as the prompt includes the verb *choose* twice. The two candidates then nominalise *choose* but also use the collocation *make + decision*, which is more frequent overall in the corpus. This is evidence that candidates are making a linguistic choice to adapt their language to suit the context of the exam as they are maintaining the interaction through this repetition. It is also interesting to note that research from Tavakoli and Uchihara (2020) found higher oral proficiency L2 speakers were less likely to use verbatim repeats of borrowed formulaic language, instead changing the language slightly while maintaining the same meaning. The act of nominalisation from Candidates 150 and 187 for *choose* could be evidence of this creativity in L1 English (Carter & McCarthy, 2004), along with the later change in noun to *decision* to avoid repetition in their speech.

5.1.3.5. *Ask + question*

The collocation *ask + question* occurs nine times in the corpus and is used by seven different speakers (3.45%). This has been categorised as a register-influenced collocation, since although there are some instances where the topic has directly led to it being used, these largely come from one speaker talking on one topic. Candidate 163's Presentation task topic was on the importance of asking effective questions, thereby demonstrating the impact the Presentation task topic can then have on the types of collocations used in the Discussion task. Example (89) below shows this collocation occurring twice within the same sentence within the Discussion task:

(89) Candidate 163: **asking a straightforward question** is much more valuable but I think in most environments **asking a speculative question** it's better

Ask + question also mostly occurs in the Discussion task (seven instances) in comparison to the Conversation task (two instances) due to Candidate 163's topic of choice.

The instances where this collocation moves beyond just giving information can be first seen in Example (90). Here, it is used as metacommentary, as Candidate 137 uses it to comment on the exam.

(90) Candidate 137: I was supposed to **ask you more questions** and I I saw I didn't do that (Conversation task)

The collocation is found to also work as a clarification request, as seen in Example (91) below.

(91) Candidate 085: I guess erm oh do I have to **ask you questions** in this (Conversation task)

This is the first instance of one of the chosen register-influenced collocations working in more than one way i.e., in the function of giving information. It could be argued this is an example of one of the core verb + noun collocations working as both topic-influenced and register-influenced, depending on the speaker and the linguistic function it has been chosen for.

5.1.4. Summary

Overall, when looking at the most frequent verb + noun collocations in the TLC-L1, these results show the influence of both the topics discussed within the exam and the

examination context on the language being used. This is evidenced in two ways: firstly, the topic-influenced collocations occur relatively infrequently in the BNC2014, yet clearly align with some of the subject areas Trinity College London expressly state they will test on in the Conversation task; and secondly, there is co-ordination between candidates' Presentation topics and the resulting linguistic choices in the Discussion task. Meanwhile, the register-influence collocations involve linguistic choices to show candidates' decision making and their opinions on different topics. The GESE exam is designed for this as it "simulates real- life communicative events in which the candidate and the examiner exchange information, share ideas and opinions, and debate topical issues" (Boyd & Taylor, 2016, p. 42). These results can also be used as evidence that the TLC-L1 is an appropriate comparison to the TLC-L2 in that the examination context is the same, much like the relationship between the LINDSEI (Gilquin et al., 2010) and the LOCNEC (De Cock, 2004). The L1 speakers were not having a general conversation; instead, the intent of their interaction was specific to the examination context, which has resulted in use of topic-influenced and register-influenced verb + noun collocations.

5.2. Unique combinations

To take a different view to describe the corpus, investigate use of collocations, and to further demonstrate its appropriateness as a comparison corpus, this section of the analysis considers the verb + noun collocations that are unique to the TLC-L1, meaning they do not appear in the BNC2014 – a large representative corpus of British English. After the coding, which is described in Chapter 3, it was found that 316 combinations were deemed unique for further analysis, with an additional 180 combinations occurring just once in the BNC2014. The 316 unique combinations were then checked manually to see which fulfil the criteria of phraseological verb + noun collocation, rather than erroneous results from the search query due to sentence boundary crossing or mistagging. For example, *choose + crochet* would be included but *crochet + group* would not be a valid verb + noun collocation as it is an adjective + noun collocation. Manual checking reduced the list to 271 combinations. Then, for feasibility, 27 (10%) of the collocations were randomly chosen for further analysis and these can be seen in Table 53. Some collocations were likely not in the BNC2014 due to events that occurred after the time frame of data collection of the BNC2014 i.e., the COVID pandemic. The influence of the pandemic is shown in the collocations *enjoy + lockdown*, *escape + vaccination* and *vaccinate + people*. In other cases, accounting for the non-occurrence of collocates in the

BNC2014 is less straightforward, as with *portray + idea* and *skew + image*. Concordance lines show that these are used accurately in the TLC-L1, as in Examples (92) and (93).

(92) Candidate 026: I have found that is a lot of that because there's more awareness more people are getting diagnosed and it seems that they're like **portraying the idea** that this is a new thing (Discussion task)

(93) Candidate 100: well of course but at the same time if it's like **skewing the public image** in different areas (Conversation task)

To account for the absence of grammatically correct and seemingly regular collocations in a large reference corpus such as the BNC2014, it is useful to recall Zipf's law, which states that "frequency of any word type in [a] corpus [is] inversely proportional to its rank" (Ha et al., 2009, p. 101). This is partly why there is a lack of occurrences for some collocations in the BNC2014 compared to the TLC-L1, though the context also plays a part (such as major events influencing language).

To further consider the acceptability of the sample of 27 collocations, their occurrence was also checked in the enTenTen20: an English corpus of 43 billion words built from the web (Jakubíček et al., 2013). Of these, four were found to not appear in this corpus (these can be seen in the table in bold). These are noted as 'unique combinations' rather than collocations as they were not frequently occurring and concordance analysis was undertaken to see if there is anything semantically interesting about these combinations, or if they were language errors made by the speakers (which would offer one explanation as to why they were not found in the reference corpora).

Table 53 Unique verb + noun combinations in the TLC-L1

assist + language	escape + vaccination	portray + idea
assume + intent	facilitate + assistance	penetrate + soil
attend + Olympics	hack + stuff	read + trash
criticise + race	harass + reporter	**re-enter + organisation**
delete + reference	**hit** + **duckling**	run + bathwater
devise + punishment	input + employee	see + drunkenness
differentiate + news	eat + kitten	skew + image
enjoy + lockdown	**foresee** + **path**	**tie** + **mummy**
enforce + neutrality	foresee + pathway	vaccinate + people

5.2.1. *Foresee + path*

There was one instance of the combination *foresee + path* in the TLC-L1 corpus, provided by Candidate 022. As can be seen in Example (94) below, the speaker uses the combination and then later in the same conversational turn, changes the noun to create *foresee + pathway*, which occurs in the enTenTen20 13 times and can therefore be deemed a more usual combination. This is a demonstration of the speaker correcting themselves as they continue in attempting to explain their opinion and engaging in the GESE's language function of asserting (Trinity College London, 2021).

(94) Candidate 022: I'm referring to it as a time bomb is like because it's like it's like a ticking process you know like a process where okay you have to y-you could **foresee your path** you know if a if it was a female they can foresee their pathway their path their life path (Discussion task)

5.2.2. *Re-enter + organisation*

The second unique combination also only occurs once. *Re-enter + organisation* is grammatically sound and used appropriately, as can be seen in Example (95). It is also a viable verb + noun collocation for the purpose of this analysis. The reason for it not occurring elsewhere may be because a more typical collocation would include the noun *workforce* or *employment* – the BNC2014 does have an occurrence of the verb with the latter noun.

(95) Candidate 017: I th= do think if women take time out to have children and they **re-enter the organisation** they 're inevitably going to be a different point on the scale potentially to to men and they've not had that career progression opportunity (Conversation task)

5.2.3. *Tie + mummy*

The third combination that can be said to be unique is *tie + mummy*, which again appears only once in the corpus. As can be seen in Example (96) below, *tie* is part of a phrasal verb *tie up* with *mummy* interceding the parts.

(96) Candidate 130: if for instance you had erm a child saying erm you know er daddy daddy's in the next room he's er he's **tied mummy** up and set fire to the house (Discussion task)

This demonstrates how speakers can use language that is grammatically correct, but very rare. Here, the collocation is from an L1 speaker; however, this links with caution from Deshors et al. (2016) that identifying innovations in L2 English is challenging, as frequency cannot be the only factor in determining what is creative and what is a mistake or error.

5.2.4. *Hit + duckling*

The combination *hit + duckling* is one that could be expected to be rare; however in contrast with the previous three unique combinations that only occur in the TLC-L1, *hit + duckling* does appear in a reference corpus for general English. It does not occur in the enTenTen 20 but does appear in the enTenTen18, from the same family of corpora (albeit, only twice). One of these instances is in fact from the Handbook of Applied Linguistics, as an example to demonstrate semantic roles and grammatical relations as seen in Example (97):

(97) "the farmer is hitting the ducklings" (Liddicoat & Curnow, 2004, p.43)

Within the TLC-L1, the speaker uses it as part of a description of the challenges of timing in photography.

(98) Candidate 062: it does mean that if you want to have the pin-sharp image of the gull **hitting the fledgling duckling** just at that moment (Discussion task)

Ultimately, unique verb + noun combinations in the TLC-L1 demonstrate that novel combinations can be acceptable, even if they do not appear in a large reference corpus. As stated, taking into account Zipf's law and the observation that the frequency of words and phrases reduces proportionally in accordance with rank in a corpus, even a corpus of tens of billions of words may not capture possible combinations. Furthermore, our data can contain evidence of linguistic creativity, which is especially apparent in spoken language within specific contexts (Carter & McCarthy, 2004) and deliberate among native speakers (Howarth, 1998). Researchers must also be wary of the tendency to credit native speakers with innovative and creative language use, or simply a slip of the tongue, while the same utterance within L2 speech might be read as deviating from the norm and marked as an error in language assessment or learner language research (Callies & Götz, 2015). It is also important to remember that use of collocations does not impede the linguistic creativity of the speakers, echoing Pawley and Syder's evaluation that "it

should not be thought that a reliance on ready-made expressions necessarily detracts from the creativity of spoken discourse" (1983, p. 208).

5.3. Frequent verb types

5.3.1. Overview

Table 54 shows the frequency of different verb types among the top 30 most frequent collocations (as seen in Table 52). Overall, there are 15 different verbs used within the top 30 collocations, showing some diversity. While some are to be expected, given their high frequency in English in general, we have seen in previous sections that the presence of others, such as *eat* and *commit*, can be partly accounted for by the topics discussed in this specific corpus.

Table 54 Frequencies of verbs in the most common collocations in the TLC-L1

Verb	**Frequency**
make	8
get	4
see	3
say	2
eat	2
take	2
mean	1
commit	1
encourage	1
ask	1
spend	1
give	1
tell	1
go	1
live	1

The top verb, *make*, appears in 8/30 (26.67%) of the most common verb + noun collocations. This helps us to discern the most common combinations involving *make* and shows that there is some variation. Collocates of *make* include *decision*, *sense*, *difference*, *money*, *choice*, *people* and *change*. We can consider the commutability of each

collocation to interpret the frequency of the combination involving *make* and any alternatives (Nesselhauf, 2005). For instance, speakers could choose *take* + *decision* or *make* + *decision*. However, there is only one instance of *take* + *decision* in the corpus, in contrast with the more frequent *make* + *decision*. This shows that although both options can be seen to be grammatically correct, there is a clear preference for one verb collocate over another due to the frequency and distribution of the results (Almela-Sánchez, 2019). Based on the data, the regular combinations *make* + *sense/difference/change* appear somewhat fixed in that the nouns do not tend to occur with an alternative verb (while carrying the same meaning), indicating a low degree of commutability (Nesselhauf, 2005).

Make + *money* is a collocation where the noun can be substituted for another and maintain the same meaning, demonstrating a high degree of commutability. Although money can be conceptual, it can also be more concrete and this allows for the noun to be substituted with *cash*, *living*, *wage* etc. to retain a similar meaning. The prevalence of *make* + *money* is also of interest when considering the presence of *get* + *money* in the TLC-L2 data, which is further explored in Section 4.4.1.1. The other nouns that collocate with *make* are more abstract. Nevertheless, frequency values indicate that there is a preference for *make* + *money* over the other noun options. Ultimately, even with the possibility of individual speaker creativity and choice, which is especially apparent within speaking vs. writing due to the language processing involved (Staples, 2015), some collocations are more fixed and therefore, expected.

5.3.2. More/less formulaic collocations

To look at the combinations from a different perspective, the analysis also considers which verbs are used within all the collocations. Taking into consideration all the verb + noun combinations that were initially extracted before the coding against the BNC2014 took place for the more formulaic/less formulaic label, this full list comprised 4,306 combinations and the most common verbs used are shown in Table 55.

Table 55 Most commonly occurring verbs within all the verb + noun combinations broken done between all and the more formulaic collocations.

All combinations (less formulaic)	**All collocations (more formulaic)**

Verb	Frequency		Verb	Frequency
get	439		get	299
make	243		make	216
take	169		take	136
see	140		see	110
give	134		give	97
say	107		say	70
use	88		use	58
put	87		put	40
need	59		find	39
mean	58		eat	37

Overall, the top eight most common verbs are in the same rank order before and after the coding took place. For example, *make* is the second most common verb within formulaic verb + noun collocations in the TLC-L1 and in the overall general list of combinations. The only differences in the most common verbs used between the full set and the more formulaic set are *need* (59) and *mean* (58) vs *find* (39) and *eat* (37). This shows a high degree of commonality between the high frequency verbs overall when looking at all the less formulaic combinations and the verbs within the verb + noun more formulaic collocations. This is not surprising as these are highly frequent verbs used in spoken English. However, this helps to validate the use of the TLC-L1 as a reference corpus of spoken L1 English in this examination context and supports the use of it within this research in comparison to the TLC-L2. When compared to informal conversational spoken L1 English, such as can be found in the BNC2014, *eat* is a verb that is a little more unusual to see. However, its prevalence is likely influenced by the couple of candidates that chose to talk about veganism and vegetarianism within their Presentation task, thus leading to the topic being highlighted in the Discussion task too. This is of interest to note, as it shows the influence that topic can have on the language that candidates produce and could be of interest to language testing boards for item writing purposes. As described in 5.1.2.1., this verb also occurs within *eat + food*, a collocation

that occurs only within the Conversation task and demonstrates the influence of the examiner's choice of topic – in this case healthy eating – on the language used.

Table 55 also shows that the top 3 most frequently occurring verbs within the verb + noun collocations are *get*, *make* and *take*. These are high frequency delexical verbs and so it is not surprising they are also highly frequent in the TLC-L1. These are verbs that have "little meaning" (Sinclair, 1990, p. 147) on their own, which results in the semantic load being placed on the noun rather than the verb within verb + noun collocations. The high frequency of these verbs also means they are of interest to investigate in light of the previous analysis involving topic- and register-influenced collocations. Chi et al. (1994) assert that these frequently occurring verbs also tend to have neutral connotations in use, which leads to minimal bias based on register or topic. However, as seen in the previous section, *make* does frequently occur in the register-influenced collocations, so this analysis will look into this claim further. To investigate this specific category of verb + noun collocation further, *get*, *make* and *take* as nodes will be focused on in more detail; these have been chosen based on previous foci for delexical verb collocation research with L2 speakers (e.g., Gilquin, 2007: *make*; Du et al., 2020: *make* and *take*; Ma and Kim, 2013: *get*, *make* and *take*) and to align with the previous analysis undertaken in Chapter 4 with the TLC-L2.

5.4. Collocational patterns in high frequency delexical verbs: *get*, *make* and *take*

5.4.1. *Get*

Table 56 shows all the nouns that collocate with *get* within the TLC-L1 and their frequency of occurrence. The top noun is *people*, with 18 occurrences of *get* + *people* in the dataset. Other notably frequent nouns include *money*, *friend*, *time*, and *dog*. Overall, there are 157 different types of *get* + noun collocations, with 299 occurrences of these in total. Therefore, 52.51% of the collocations are used just once, demonstrating the diversity of this structure across the speakers in the corpus.

Table 56 All nouns collocating with get in TLC-L1

***Get* +**	**299**
people	18
money	16
friend	10
time	9

dog	6
opportunity, point, car, information, job	5
student, thing, support, degree, result, food, idea	4
grade, attention, skill, chance, child, system, impression, amount, contract	3
right, vote, hold, letter, question, mark, service, work, energy, value, way, offer, history, experience, sense, pay situation, character, family, hand, benefit, present, company, problem, promotion, cat, woman, kid meat	2
room, handle, call, holiday, range, home, shot, house, gun, hundred, project, balance, equipment, image, bus, change, sleep, income, teacher, argument, area, issue, back, bike, bread, choice, record, kind, brother, lesson, eye, boat, share, licence, size, like, star, load, taste, loan, cancer, connection, training, word, view, boot, win, access, brain, model, ear, country, race, month, recognition, news, relationship, note, reward, number, road, couple, second, one, sentence, opinion, shape, cross, shop, option, feedback, parent, feeling, park, space, part, cake, daughter, funding, box, tax, permission, text, person, ticket, phone, town, photograph, truth, place, vehicle, plan, voice, discount, head, power, card, doctor, pressure, year, message, million	1

The most frequent *get* + noun collocations are not particularly collocable; this means they are frequent but there is a high degree of commutability in that they are concrete terms rather than abstract concepts, mostly (Nesselhauf, 2008). *Time*, *opportunity* and *point* are the most fixed of the noun collocates in that they have a lower degree of commutability. The nouns that co-occur with *get* are more general than exclusively co-occurring, supporting the view that *get* is a semantically unmotivated verb (Allerton, 1984). The following shows the counts for the *get* + noun per speaker distribution of some of the most frequently occurring collocates: *people* (15 speakers), *money* (14 speakers), *friend* (10 speakers), *time* (seven speakers), and *dog* (three speakers). Other collocates of interest are *opportunity* (five speakers) and *point* (five speakers), as these are also abstract nouns.

To investigate the most frequent *get* + noun collocation, examples are shown below of concordance lines of *get* + *people*. It should be noted that these have been expanded to include the specific verb forms and interceding word(s) that were removed in the initial grouping strategy. This helps to gain a better sense of how the collocations are used within the interaction.

(99) Candidate 002: erm th-there were certainly you will **get people** that'll be too adventurous (Discussion task)

(100) Candidate 009: if we had that in our currently vastly unequal society you'd **get rich people getting poor young people** and taking their blood and living forever (Conversation task)

(101) Candidate 078: erm yeah I'd s= I'd say that it has erm because you can easily share something and **get other people** to share it (Conversation task)

There is a variety of internal modification possible with *get + people*, as seen above in the examples. Other adjectives that can be added to *people* include *rich*, *poor*, *young*, *clever*, *ordinary* and *other*. Furthermore, both *get* and *people* can be substituted for other words of similar meaning i.e., they have a higher degree of commutability (Nesselhauf, 2008). This suggests that they collocation is applied not according to the idiom principle but more so using open choice (Erman & Warren, 2000). In contrast, *get + point* appears more fixed, due to the noun being more abstract. Of the 16 instances of this collocation, there is much less internal modification than for *get + people.* Other than determiner changes, there are only two occurrences of other differences, as seen in the examples below.

(102) Candidate 117: it's got key texts it's **got key erm points** that you would go yeah that's a liberal (Discussion task)

(103) Candidate 183: and with she's **got eighty eight point** five million followers and like there's gonna be just as many people who dislike her (Discussion task)

As can be seen in Example (103), this is an erroneous hit in the corpus as the speaker is discussing the number of Taylor Swift's social media followers and therefore does not fulfil the criteria for a verb + noun collocation in this analysis. Consulting concordance lines in corpus analysis is important as queries can be highly accurate but some anomalies can still be included.

To gain a sense of how the *get* + noun collocations are situated within the TLC-L1, Table 57 shows the top *get* + noun collocations that are evident in the BNC2014 as a comparison. 'Lot' will be excluded from further analysis, as it frequently occurs as 'get a lot of [noun]' and therefore it is irrelevant to this analysis. Subsequently, these are the top 10 nouns that collocate with *get* in the BNC2014.

Table 57 Frequency of nouns as a collocate of get in the BNC2014

Noun	**Raw frequency**
job	1630
thing	1602
time	1402
people	1393
car	1358
money	1205
way	1114
back	929
chance	867
~~*lot*~~	~~*830*~~
work	744

The top noun collocates extracted from the BNC2014 show some similarities to those found in the TLC-L1 with a frequency of three or more occurrences. *People*, *money*, *time*, *car*, *job*, *thing* and *chance* occur in both corpora frequently. These are mostly concrete nouns, highly frequent in general English corpora. The more context specific corpus of the TLC-L1, however, also includes less common concrete nouns such as *friend*, *dog*, *food*, *student*, *degree*, *grade* and *child* which are likely to have been influenced by the topics discussed within the interactions. Furthermore, there are more register-influenced collocates present due to the nature of the examination, such as *opportunity*, *point*, *information*, *idea*, *impression* and *attention*, given that one of the requirements of the GESE is for candidates to demonstrate variety in their language (Trinity College London, 2021).

5.4.2. *Make*

Table 58 shows the nouns that collocate with *make* within the TLC-L1. The top noun is *decision*, with 31 occurrences of the collocation *make + decision* in the dataset. Other notably frequent nouns are *sense, difference, money, choice, people* and *change*. There are 72 different types of *make* collocations, which means 33.33% of these collocations are unique; this is a considerably lower diversity than the *get* + noun collocations

(52.51%), showing that *make* + noun collocations may be more fixed in nature for the L1 speakers in this corpus.

Table 58 All nouns collocating with make in TLC-L1

***Make* +**	**216**
decision	31
sense	22
difference	15
money	13
choice	12
people	11
change	9
thing	8
think, friend	5
point, mistake	4
world, argument, feel	3
life, film, look, effort, progress, judgement, assumption, comment, contact, model, meal,	2
skin, relationship, comeback, commitment, school, impact, stop, joke, video, case, rule, contribution, service, living, statement, deal, subject, market, transition, amends, careers, mind, room, debut, salad, attempt, excuse, billion, situation, news, sport, distinction, step, person, story, phone, family, economy, threat, problem, use, product, way, programme, food, Christmas	1

Many of the most frequent noun collocates of *make* are also highly fixed in the collocation in that they are not fully commutable with other nouns to maintain the same meaning (Nesselhauf, 2008). Further, only a 1/3 of the collocations are unique, suggesting there is significant repetition of the same type of *make* + noun collocations within the corpus. Considering *make* + noun per speaker distribution, *decision* occurs in the exams of 21 (over 10%) speakers, *sense* among 16 speakers; *difference* in 13 speakers; *money* in 12 speakers; *choice* in 10 speakers; *people* in eight speakers; and *change* in eight speakers.

To investigate the most frequent *make* + noun collocation, Examples (104) to (106) show the concordance lines of *make* + *decision*. It should be noted that these have been

expanded to include the specific verb forms and interceding word that were removed in the initial grouping strategy. This helps to gain a better sense of how the collocations are used within the interaction. The examples show the candidates including modifiers such as *conscious* and *informed* to alter the noun and engage with the register demands of the examination, which involves language functions such as asserting and affirming (Trinity College London, 2021).

(104) Candidate 202: I suppose you have to work out what suits you your budget your lifestyle but think about it I suppose **make a conscious decision** rather than just getting the offers in the supermarket that look cheap and cheerful (Conversation task)

(105) Candidate 115: and I think that erm I hear what you say but I think if the student experiences that world of work they can then **make an informed decision** (Discussion task)

(106) Candidate 034: erm so at sixteen I would probably say that I was not in a position where I could have **made an informed decision** voting however given the chance I would have educated myself (Conversation task)

To further contextualise how the TLC-L1 speakers use *make* + noun collocations, Table 59 shows the 10 most frequent noun collocates occurring with *make* in the BNC2014.

Table 59 Frequency of nouns as a collocate of make in the BNC2014

Noun	**Raw frequency**
sense	4639
decision	3351
difference	2821
way	2534
money	1652
use	1576
progress	1552
thing	1357
feel	1290
people	1238

The top frequent collocates in the BNC2014 have much overlap with the collocates found in the TLC-L1 – *sense*, *decision*, *difference*, *money*, *thing*, *feel* and *people* all occur within both corpora. as the high frequencies of *sense*, *decision* and *difference* attest to their common use in general English. Nevertheless, as discussed in Section 5.1.3, concordance analysis of these collocations in the context of the TLC-L1 exam pointed to the influence of register on a number of *make* + noun collocations. Therefore, the speakers are using highly frequent verb + noun collocations in the TLC-L1 due to both the nature of the interaction between the examiner and candidate, and because of their high frequency in general L1 English (based on the BNC2014). These verb + noun collocations are also highly fixed in that the noun cannot be substituted for another noun to maintain the same or similar meaning, thus meaning a lower degree of commutability (Nesselhauf, 2008). In contrast, *choice*, *change*, *point*, *mistake*, *argument* and *world* are all likely to be occurring due to topic choices within the TLC-L1. In the Conversation task in particular, the exam uses globally relevant topics such as the impact of climate change to engage the candidate in the type of language needed to show interactional competence (Galaczi & Taylor, 2018). This is further support for the presence of topic- and register-influenced collocations within the TLC-L1 data.

5.4.3. *Take*

Table 60 shows the nouns that collocate with *take* within the TLC-L1. The top noun is *time*, with 13 occurrences of the collocate in the dataset. Other notably frequent nouns are *responsibility*, *thing*, *action*, and *part*.

Table 60 All nouns collocating with take in TLC-L1

***Take* +**	**136**
time	13
responsibility	8
thing	7
action, part	6
photo, place	5
child, job, people	4
break, life, step, care, picture	3
effort, while, risk, advantage, money, position, day, person, look	2

supplement, profit, work, ground, side, home, walk, hundred, precaution, information, blood, amount, exam, car, film, line, week, average, girl, charge, priority, name, effect, option, role, baby, stance, country, stuff, couple, example, phone, view, back, way, photograph, foot, decision, year, dog, drug	1

Overall, there are 65 different types of *take* collocations and 136 instances combined, making the *take* collocations more diverse in their noun usage than *make* but less diverse than *get* + noun collocations. When considering *take* + noun per speaker distribution, the highest distribution is *time* (13 speakers), followed by *responsibility* (seven speakers), *thing* (six speakers), *action* (five speakers), and *part* (six speakers). With fewer overall occurrences compared to *make* and *get* + noun collocations, *take* also has fewer types. However, there is more uniqueness in the usage of *take* than *make*. Four of the top five nouns are conceptual in their meaning when combined with *take*, namely: *time*, *responsibility*, *action* and *part.*

Considering the top 10 noun collocates in the BNC2014. as seen in Table 61, *place*, *time*, *part*, *step*, *care*, *advantage* and *look* also occur in the TLC-L1. Interestingly, *account*, *week* and *breath* are not present in the examination language corpus. This is likely due to the candidates not having the opportunity to use these collocations, as a result of the topic and register influence of the examination (Buttery & Caines, 2012). This is important to note when analysing the TLC-L2, as lack of the presence of collocations may not mean speakers are unable to use them, but that the opportunity was not present.

Table 61 Frequency of nouns as a collocate of take in the BNC2014

Noun	**Raw frequency**
place	7816
account	3596
time	3450
part	3131
step	2609
week	2326
care	2037
advantage	1918

breath	1666
look	1486

5.5. Summary

This analysis chapter explores answers to the thesis research questions related to the verb + noun collocations in the TLC-L1, specifically in the Discussion and Conversation tasks of the GESE. The analysis took an exploratory route, starting off by looking at all the possible combinations, deciding how to categorise this as more or less formulaic and subsequently generating a final list of 'more formulaic' collocations and less formulaic combinations. From this, the research found evidence of language creativity in the corpus from the speakers, which can be linked to findings from L2 research from Tavakoli and Uchihara (2020) who found high oral proficiency speakers used multi-word expressions more creatively than lower proficiency speakers. The results also show that the language within the TLC-L1 is heavily influenced by topics and the register of the examination, as well as the candidate and the examiner as interlocutors. From this, certain collocations are used frequently due to influence from the topics discussed in the interactions, both introduced by the candidate and the examiner; also highly influential was the register used by the candidates, and encouraged by the examiner, due to the context of the corpus being an English language exam. The final stage of the analysis has narrowed in on three high frequency delexical verb + noun collocations. It has found that there is further evidence of topic and register influence on the use of collocations within the TLC-L1 and, through a ranked-frequency comparison with the BNC2014, these highly frequent verb + noun collocations are in general typical and expected in a corpus of L1 English. This means it can be said the TLC-L1 is an appropriate corpus to use alongside investigations into the TLC-L2. The following chapter will explore some similarities from the data that has been presented in Chapters 4 and 5.

Chapter 6: General Discussion

The aim of the study was to examine verb + noun collocations in L1 and L2 spoken English using a corpus-based approach. Specifically, the investigation looked at frequency and distribution of these collocations in English, focusing on the influence of topic and register and highlighted three high frequency delexical verbs to delve deeper into the analysis. Overall, the research brings an original contribution to the field of investigating verb + noun collocations as it is the first to use both L1 and L2 spoken English corpora of speakers engaging in the Trinity College London GESE.

This chapter works to bring together Chapter 4 and Chapter 5 to look at the descriptive accounts of the TLC-L2 and TLC-L1 in one space. It first presents a summary of the main results of this thesis (Section 6.1), directly answering the proposed research questions concisely. Each of these research questions will then be explored in more detail in Section 6.2 to Section 6.5 to situate the findings in the context of current research surrounding the main research questions.

6.1. Summary of results

The thesis results begin with the general finding that verb + noun collocations were used by every speaker in both the TLC-L2 and TLC-L1 corpora. This was expected, given that it is a common construction in English, but this also supports the fact that it is an appropriate language feature to study because of its extensive presence. Although it was known to be common in English, corpora in general are only language samples of possible texts (McEnery & Brookes, 2021), so it was necessary to confirm the presence of this collocation type before proceeding with the analysis. Moving deeper into the analysis, the thesis answers the proposed research questions as follows:

RQ1: To what extent are there differences in the (1) frequency and (2) distribution of use of verb + noun collocations amongst L2 English speakers at B1, B2 and C1/C2 level?

The findings show some nonlinear development of collocations supporting previous research by others such as Brezina and Fox (2021), Forsberg and Bartning (2010), Nizonkiza (2017) and Paquot et al. (2021). The implication is that phraseological development in language learners is complex with many factors potentially accounting for this complexity, including individual differences (Omidian et al., 2021).

RQ2: To what extent is there evidence of topic influence on speaker choice of verb + noun collocations in the TLC-L1 and the TLC-L2 corpora?

There is evidence in both corpora of topic influence, which supports previous findings from Paquot (20140 and Suzuki (2015), amongst others. This shows a similarity between L1 and L2 speech and one implication is that care should be taken when teaching English to ensure ample opportunities for learners to speak about different topics. Another implication is in designing language examinations to strive for varied topics that elicit the language needed to exemplify the speaker's proficiency level, as topic also impacts L1 speakers during the same language exam.

RQ3: To what extent is there evidence of register influence on speaker choice of verb + noun collocations in the TLC-L1 and the TLC-L2 corpora?

Evidence is of register influence is stronger in the TLC-L1 than in the TLC-L2 which suggests the native speakers are more adept at using language suitable to the specific context. This observation supports findings from others such as Ädel and Erman (2012), Hyland (2012) and Chen and Baker (2016). The implication is that teaching needs to consider the pragmatics of interaction when developing materials to support language learning ahead of an interactive examination like the GESE.

RQ4: How do TLC-L1 and TLC-L2 speakers use high frequency delexical verb + noun collocations in spoken examination language?

There are patterns of interest found in both corpora that supports the findings from RQ1, RQ2 and RQ3. The implication is that high frequency delexical verb + noun collocations are especially valuable to focus on in research because their high frequency means they are found in smaller specialised corpora like the TLC-L1 and they also minimise topic bias (supporting suggestions from Zinkgräf, 2008).

The chapter will now explore some further themes that emerged through the analysis, expanding on the summary of results to answer the four core research questions.

6.2. RQ1: Nonlinear development of L2 English speaker use of verb + noun collocations

The first research question set out to investigate how the three proficiency groups in the TLC-L2 use verb + noun collocations. By looking at the B1, B2 and C1/C2 groups, the analysis took a pseudolongitudinal approach, following previous research from Granger and Bestgen (2014). Previous research has provided complex results about the nature of collocation development in relation to proficiency levels (e.g., Paquot & Granger, 2012), and this research question aimed to add to this conversation using the large spoken language corpus of the TLC-L2.

Overall, the results demonstrate evidence of nonlinear development of collocation use in the TLC-L2 corpus. When considering all instances of verb + noun collocations produced by the L2 speakers, there was a decrease in the relative frequency of verb + noun

collocations per 1,000 words across the three proficiency levels of B1, B2 and C1/C2. However, this was in contrast to a steady increase in these collocations' mean, SD and range. This meant that candidates were using fewer verb + noun collocations overall in the highest proficiency group when compared to the lower proficiency candidates. Conversely, as proficiency increased, some individuals were using more of these collocations with a greater range of use on average. This evidence supports other accounts of the nonlinear development of collocations and language proficiency more widely (e.g., Brezina & Fox, 2021). However, it also contradicts other findings. For example, Laufer and Waldman (2011) and Pauqot and Granger (2012) both found that higher proficiency learners used more collocations than lower proficiency learners, though these studies were investigating L2 writing. The results from this first step in the analysis could be highlighting a difference between L2 speech and writing, further justifying the need for more spoken language research. It also adds to the overall picture of the challenge in finding a clear picture of how language learners develop their use of collocations. It could be that individual differences have an impact on verb + noun collocation due to the increase of mean, SD and range and the decrease in frequency across the proficiency levels (Omidian et al., 2021).

There is further evidence of nonlinear development of collocation use when looking at the shared collocations across the speaker groups; overall, there were 9,674 collocations and 3,700 types shared by all of the three proficiency groups. Looking only at these gave a similar picture regarding relative frequency and means, again demonstrating a suggestion of nonlinear development in collocations. The relative frequency was highest in the lowest proficiency group; this then dipped at B2 before increasing slightly at C1/C2. There is a reduction in the use of collocations in the intermediate B2 group when considering both (1) every instance of a collocation and (2) only collocations present across the proficiency groups. This dip in the B2 group usage could exist because this group is engaging in more complex language topics and grasping these resulted in decreased use of this particular language feature. A u-shaped curve in verb + noun collocation usage follows previous research from Vedder and Benigno (2016) and further supported by Siyanova-Chanturia and Spina (2020), noting that development of phraseological competence may get worse before it improves at higher proficiency levels. It could also be related to another variable beyond this research's scope, such as L1 background or age. With the latter, this could be a matter of literacy development rather

than language development or other age-related factors that can impact language learning in general (Hu, 2016).

The reduced set of shared collocations had similar findings as when considering all the verb + noun collocations regarding the mean usage; the mean was highest in the C1/C2 group and lowest in the B2 group. Again, this demonstrates the nonlinear usage of verb + noun collocations across the proficiency levels and can be linked to the findings from Paquot et al. (2022), who found that phraseological sophistication decreased from B1 to B2 level when studying verb + direct objects. They found that the while proficiency did not necessarily result in the use of a higher number of collocations, there is instead qualitative development of more diverse combinations. It could be that the B1 speakers are attempting to "play it safe" (Paquot et al., 2022, p. 132) – much like the use of lexical teddy bears by less proficient speakers, as noted by Hasselgren (1994) – and using collocations they are confident with, rather than venturing to use more creative and diverse combinations that would be expected as proficiency increased. This also links to Tavakoli and Uchihara (2020), in that the B1 speakers in the TLC-L2 could also be relying more on repeating the same verbatim borrowed collocations, while the higher proficiency speakers in the B2 group are using more creativity in their production of collocations.

In the TLC-L2, the most frequent verb + noun collocations are related to the topic of hobbies or general interests e.g., *read + book* and *learn + language*, with some exceptions to this, such as *take + care* and *make + feel*. Overall, the topic-influenced aspect of the language choices in the corpus is shown to be significant by the ranking of the verb + noun collocations combined across the groups. These ranked collocations also include topics typical for language tests and progress along the core developmental stages in language learning that are reflected within the CEFR (Council of Europe, 2001).

When looking at the relative frequencies of the verb + noun collocations, it can be seen that the top 10 most frequent are present in this view because of their frequency in the B1 group (*read + book*, *learn + language*, *take + care*, *spend + money*, *play + game*, *spend + time*, *earn + money*, *give + money*, *play + football*, *save + money*; see Table 29). The B1 speakers use a smaller set of collocations more frequently, and the diversity of collocations increases as proficiency increases. This finding is supported by the C1/C2 speaker data accounting for more abstract collocations in the top 10 across groups,

e.g., *take + care*. Furthermore, there is no increase in the frequency of these top 10 verb + noun collocations as proficiency increases, showing that the different proficiency groups use different verb + noun collocations.

The above results focus on the frequency of the verb + noun collocations; however, this is just one dimension of collocation (Brezina, 2018). The second is distribution, and to look at this, the use of the verb + noun collocations within individual speakers was also considered within the research. From this, further support was found for the finding in this thesis that the B1 group use a smaller set of collocations more frequently, as there is a higher percentage of the B1 speakers using the verb + noun collocations in the top 10 most frequent set when compared to the other two higher proficiency speaker groups.

A further finding from this research is that one speaker group can have a significant impact on aggregated data; therefore, there is a need to consider differences at differing proficiency levels when looking at L2 speaker data, echoing studies such as Chen and Baker (2016). When considering the relative frequency of specific collocations in the top 10 and focusing on per speaker group rather than aggregating the data, this shows the influence of *learn + language* in the B1 group and *take + care* in the C1/C2 group. This led to further investigation and significant findings regarding two types of verb + noun collocation within both the TLC-L1 and the TLC-L2 corpora, and these have been named as topic-influenced and register-influenced verb + noun collocations. The rest of the discussion chapter will explore topic-influenced and register-influenced verb + noun collocations in the context of previous research.

Looking at the noun collocates within the verb + noun collocations in the TLC-L2 led to interesting interpretations of the findings related to language development, item writing in language tests and demonstration of interactional competence. Firstly, analysis into the use of *spend/waste + time* can be used to explore the first two findings, as this collocation occurred more in the Discussion task for B1 and B2 speakers while occurring more in the Conversation task for C1/C2. This could be accounted for by topic choices at the advanced level chosen by the examiner and put forth by the GESE language testing item writing. Furthermore, there is a linear increase in usage across the proficiency groups when considering both relative frequency and percentage of speakers using the collocations. This means the findings could be due to many different factors. It could be indicating an increase of verb + abstract-noun collocations as language develops in the

candidates (similar to the findings in written English that there is an increase in formulaic language usage as proficiency increases from Laufer & Waldman, 2011). Alternatively, it could be evidence that there are simply more opportunities for the advanced speakers to use this collocation due to the task topics decided by the examiner (Caines & Buttery, 2018). A third possibility for explaining the use of this verb + noun collocation could be related to the test, in that it was designed to allow more opportunities; there is an assumption that advanced speakers would be able to more confidentially use these verb + abstract-noun collocations, which could also be related to co-construction of the proficiency level from the test designer, as discussed by McNamara (2001). It should be noted that only candidates who passed their GESE are part of the TLC-L2 corpus, therefore only the data from successful candidates are included. It could be said that the examination is working strictly as intended if the item writers have taken into account the language development of the candidates and have designed tasks with topics to allow these candidates the opportunities to flourish within the context. This approach would fulfil the so-called 'bias for best' approach to language testing that Trinity College London and others favour (Fox, 2004).

Regarding evidence of interactional competence within the TLC-L2 corpus, c*hange* + *mind* was found to be a collocation of interest. This is because it showed similar usage patterns to the above collocation types i.e., there was an increased use in the higher proficiency group; however, further analysis also shows C1/C2 speakers using second person pronouns to engage their interlocutor, thus demonstrating their interactional competence. This supports previous statements from Plough et al. (2018) about the need to investigate further the interaction of different language competencies, such as interactional competence at differing speaker proficiency levels. Furthermore, this distinction between how the proficiency groups use pronouns within their verb + noun collocations is highlighted by the intermediate candidates' use of pronouns to express their own opinions. It can be said that this later develops in the higher proficiency groups to engage in discussion with the examiner and demonstrate interactional competence. Other research has also found evidence for the development of pragmatic awareness in pronoun use as L2 proficiency increases (Belz & Kinginger, 2003). Although this research focused on L2 German in the classroom, this preliminary evidence from the TLC-L2 works to add to the picture of how pronouns are used pragmatically accurately in an L2.

Furthermore, Jones et al. (2017) also investigated pronoun usage in language learners. They found that the use of *we* changed depending on proficiency levels and with respect to the function and meaning of the pronoun. At the lower level of B1, the speakers tended to use *we* when discussing topics that were somehow connected to themselves. Then, in the progression from B1 to B2, there was more evidence of usage functioning to share experiences with the interlocutor in the discourse. Finally, at the most advanced level in the study, Jones et al. (2017) found that more complex ideas were explored using the pronoun *we,* including within hypothetical situations and going beyond the individual. Although a different pronoun was seen in the TLC-L2 data, there are similarities between Jones et al. (2017) and this research in the shift away from focusing on the individual with the use of *you* as an interceding particle in the collocation *change* + *mind,* to it functioning more as a way to engage with the examiner.

6.3. RQ2: Topic-influenced verb + noun collocations in L1 and L2 spoken English

Both the TLC-L2 and the TLC-L1 corpora showed evidence of their speakers using topic-influenced verb + noun collocations. These collocations can be defined as those that occur due to the influence of one or more of the following: (1) the examination topic pre-set within the GESE; (2) topic chosen by the examiner in the instances where they can decide on appropriate subjects; and (3) topic chosen by the candidate in either the presentation task (for the higher proficiency and L1 speakers) that then fed into the Discussion task or the topics chosen by the candidate for the discussion task for the lower level proficiency speakers. From this finding, three influential factors can impact what collocations are used within a language testing context in this research, supported by previous research. Firstly, the exam itself can influence linguistic variation due to the specific register needed to engage with the pre-set topics within the tasks (Gablasova et al., 2017) as well as the impact from test designers on what language features are elicited from the candidates, based on the topics chosen to engage with (McNamara, 2001). Secondly, the examiner as the interlocutor in the dialogic tasks can influence candidate language from topic-priming used to support candidates' language production in exams (Lazaraton, 1996), as well as other interlocutor effects that occur within speaking tests (McNamara, 1997). This also links with Young and Milanovic (1992), who noted that examiners could have a controlling role in the speaking interaction while candidates take a more reactive approach to the discussion. Finally, candidates themselves are a source of topic-influenced collocations due to individual speaker variation and their choices of what to

talk about (Lowie & Verspoor, 2015). For example, Omidian et al. (2021) also found phraseological development to be a complex process when considering verb + noun collocations, and this could not be attributed only to time or language proficiency but individual differences. This individual variation in language use can be further explored considering CDST(De Bot et al., 2007). Furthermore, this finding is also related to speaker opportunity of use (Buttery & Caines, 2012) in that it may be down to the individual and the individual's unique situation in the moment of language production.

Evidence of these topic-influenced verb + noun collocations was found in every group in the TLC-L2 – B1, B2 and C1/C2 – as well as within the TLC-L1. This is an indicator that topic can have a significant impact on the collocations used within a language examination. Some specific topic-influenced collocations were found to cluster within certain proficiency groups e.g., *learn + language* (B1), *read + book* (B2), *take + care* and *use + internet* (C1/C2), which is likely due to the examination topic pre-set within the GESE. This could indicate that the test writers have chosen topics that are suitable for the level of proficiency of the candidates taking the specific grade they are writing for (Huang et al., 2018). It is important to do this, as McNamara (2001) suggests that candidate proficiency is a co-construction between the test designer, interlocutor and rater – particularly in interactive speaking tests – and he later states that language testing in general "constructs the notion of language proficiency" as a social practice (McNamara, 2001, p. 339).

The topic-influenced collocations also showed the development of complexity within the TLC-L2 corpus. For example, looking further into the *watch* + noun collocations, it was found that adjectives were included within these collocations. However, it was the C1/C2 group that used these in ways beyond the expected topic of 'hobby,' i.e. "I watched tv". Showing more complex topics is a feature in the advanced proficiency candidate group and supports previous research from Saito and Liu (2022).

Other evidence regarding the topic type changing in the advanced speaker group of C1/C2 comes from the *become + career* collocations. This occurs most frequently in the candidate-led Discussion task, partly in the B1 and B2 groups with topics categorised as egocentric, i.e., candidates are talking about their own careers and other topics personal to their own experiences. There is then a decrease in the use of the *become + career* collocations in the more advanced C1/C2 speaker group, while the topics also

become more abstract and less personally related to the speaker. This could be occurring because producing concrete words is easier for lower proficiency level speakers. This finding supports previous research from Crossley et al. (2011), where L2 learners' lexical development over time became more abstract. This also feeds into findings from Saito (2020) regarding L1 judgements of L2 comprehensibility, in that low frequency combinations that included abstract and complex words were strongly correlated with perceived language appropriateness; the evidence from the TLC-L2 data shows the more advanced speakers are also engaging in the use of more complex and abstract words.

This could be due to the relationship between concreteness and semantic access, as Kroll and Tokowics (2001) explored, where concrete words are more likely to share meanings across languages. Therefore, concrete collocations are favoured by L2 speakers in the lower proficiency groups, as the conceptual processing of these collocations is easier than for abstract words. As the proficiency level increases, the ability to express beyond the individual perspective does too. Furthermore, concordance analysis in Chapter 4 shows that candidates are not only repeating back collocations initially uttered by the examiner but spontaneously creating them.

The examiners for Trinity College London expressed – in informal discussions during the data collection of the TLC-L1 – that L2 candidates in the TLC-L2 tended to overuse simplistic language such as 'thing' and 'nice' and overall, produced more general terms rather than specific ones. Alongside this, the examiners also mentioned that the L1 participants were mostly not making mistakes in their language, which can also be seen in the TLC-L1 concordance lines, analysed in chapter 5. L1 candidates were, at times, possibly not choosing the most appropriate word, but there were little or no mistakes or errors in their use of vocabulary or grammatical structures. In contrast, the L2 candidates at this level could be successful at communication throughout the exam but still have fossilised errors within their language; as Laufer and Waldman (2011) point out, "many cases [of] collocation errors may appear to become fossilised" (p. 654). This supports the need to not strive for 'native-like' production in exams but to have successful communication as the goal (Elder et al., 2017).

Considering the L1 speakers further and the topic-influenced collocations within the TLC-L1, there is further evidence that the exam and the tasks influence topic due to the majority of the most frequently occurring of this type of collocation occurring within the

examiner-led task – the Conversation task. The three most commonly occurring collocations of this type are *commit + crime*, where ten instances are used by nine different speakers; this supports the notion that it is an examination-specific topic-influenced collocation, due to the distribution of the usage. The second is *eat + meat*, which shows evidence of the candidate being primed by the examiner to use this collocation based on the topic of discussion (a feature of helpful examination technique initially proposed by Lazaraton, 1996). However, this is also an example of where one speaker can have a significant influence on frequency when looking at data in an aggregated way; idiolectal features of speakers mean it is important to consider dispersion for all but the most frequent vocabulary (Schmitt, 2010), especially when it is within a relatively small but specialised corpus. Finally, all instances of *eat + food* occurred in the Conversation task, showing that the examiner introduced the topic. This is evidence of interlocutor influence, as previously discussed by Rosas-Maldonado (2017).

As the findings above are similar for the TLC-L1 and the C1/C2 speakers within the TLC-L2, language proficiency is not necessarily influencing the presence or absence of these types of topic-specific collocations. Instead, it could be considered that the task is the driving factor in the language choices of the candidates for both the L1 and L2 speakers and any language proficiency influence is because the GESE tasks have been aligned to the candidates' expected level of language proficiency and the CEFR (this process of alignment is explored further in Papageorgiou, 2007). Therefore, it is not about their capabilities, necessarily, but how the task has been created to ensure the language proficiency of the speakers is being tested thus ensuring assessment validity (Cushing, 2021). These results also suggest that it is not necessary to strive to include the 'ideal' topics for discussion within an examination but instead ensure there is enough opportunity for the speakers to engage in a variety of topics (Paquot, 2020). In other words, if candidates choose a topic for their Presentation or Discussion task and this topic is also in the Conversation task, the examiner should pick a different prompt to ensure the speech opportunities are varied for the candidate (Buttery & Caines, 2012; Weigle & Friginal, 2015).6.4. RQ3: Register-influenced verb + noun collocations in L1 and L2 spoken English

As defined by Biber and Conrad (2009, p. 6), "a register is a variety associated with a particular situation of use (including particular communicative purposes)". In both the TLC-L1 and TLC-L2 corpora, there is evidence of register-influenced collocations; these

are collocations that are likely occurring due to the examination situation of use the candidates are speaking within and the communicative purpose of language testing. In the TLC-L2, these register-influenced verb + noun collocations are *repeat + question*, *understand + question* and *choose + topic*. In contrast, the TLC-L1 candidates use a wider variety of this type of collocation, with seven types commonly occurring: *make + decision, make + sense, make + difference, take + time, make + choice, ask + question* and *take + responsibility*. The fact that the L1 speakers use a wider variety of these verb + noun collocations could indicate that using collocations appropriate to the register necessary for the interaction is a feature of an expert speaker. This is supported by Gablasova et al. (2017b), who suggest that "even advanced L2 users struggle to adjust their linguistic choices according to the context or genre of the discourse" (p. 613) and base this statement on several previous studies including Hinkel (2005) and Gilquin & Paquot (2008).

Looking into more detail of the use of these in the TLC-L2, the collocations r*epeat/understand + question* were found to be more frequent in B1 and B2, with a wide dispersion of occurrence amongst speakers. This wide distribution indicates that these collocations are related to register (the examination context) rather than topic because the topic changes for each speaker while the context stays the same. The fact that these collocations are found more frequently in the B levels supports research from Jones et al. (2017), who found clarification requests for B1 speakers were typically on task instructions and B2 speakers needed to clarify vague questions from the examiner. In contrast, the advanced C1 speakers asked about vocabulary. Furthermore, concordance analysis shows the purpose of these collocations is to support interaction between two interlocutors, which Fulcher (2010) argues can aid the candidate's speech in speaking tests. In comparison, the TLC-L1 register-influenced collocations included *ask + question*, which at times also worked as a clarification requisition regarding the task instructions (see Examples 90 and 91). However, there were also instances where the speaker was using the verb + noun collocation to talk about their topic This shows there is influence on the language used from both the topic and the register in the TLC-L1 corpus and concordance analysis is invaluable for interpreting these instances.

Another verb + noun collocation that was found to be influenced by register in the TLC-L2 is c*hoose + topic*. As well as frequently occurring in the Discussion task, which is candidate-led rather than examiner-led, the collocation introduces set topics planned by

the exam item writers; it was also shown to be a core part of the beginning of the interaction as candidates were required to explain their decided subject for this task. Furthermore, concordance analysis found that examiners frequently introduce the collocation in the preceding utterance, a requirement of the examination context. Therefore, its use has been dictated by the register. Then, the candidate repeats this back to the examiner. This links to research from Gómez González (2018) into lexical cohesion in interaction and how linguistic features can help with cohesive ties. They found that a highly frequent strategy for managing the interaction was repetition. This research was conducted by using the International Corpus of English-GB, thus showing that repetition is a feature in native spoken language and necessary for managing turn-taking and topic maintenance. The TLC-L2 speakers engage in this strategy when they are repeating collocations, which is an effective way to maintain the interaction. Fung (2018) further supports this with research on self-repetition in speaking, while Mauranen (2004) notes the benefits of 'prefabs' to help speakers maintain interactive discourse. Interestingly, this collocation does not occur in the most frequent collocations for the TLC-L1, demonstrating that the L2 speakers may perhaps be overusing *choose + topic* when compared to the L1 speakers. However, the results and discussion in Chapter 4 have found the benefits of this collocation in aiding the cohesion of the interaction; therefore, L2 speakers 'overusing' collocations when compared to L1 speakers may not be working as a detriment to the overall interaction. Instead, it could be a positive feature of maintaining interaction within L2 spoken English.

With the TLC-L1, there is further evidence that verb + noun collocations can be categorised as register-influenced. Again, supporting previous research from Gablasova et al. (2017), it can be said that more L1 speakers use these collocations than L2 speakers because they have more awareness of the register, and they are more skilled in adapting to the register. Thus, the realisations of this strategy can be used to create teaching materials for L2 speakers to learn to adapt their language to the necessary register of the task and the exam. Some suggestions for possible pedagogical interventions using the TLC-L1 corpus data are explored in Section 7.3.

The findings also suggest the TLC-L2 speakers are using more topic-influenced verb + noun collocations, and this could be due to the language learners focusing more on the content of their speech, rather than using the appropriate socio-pragmatic function (see also Saito & Liu, 2021). Socio-pragmatic function is an essential consideration for

interactional competence (Plough et al., 2018). Once again, this evidence could be used to demonstrate appropriate interactive language choices for the exam within teaching activities for L2 speakers.

In contrast to the L2 speakers engaging in frequent topic-influenced collocations, the L1 speakers use more register-influenced collocations, for example, with m*ake + decision.* 24/33 instances were in the Conversation task, which is examiner-led, and so these topics are designed to be able to give differing opinions and potentially be polarising, with the examiner focusing on leading the candidate to consider differing points and come to a conclusion. Therefore, part of the function of the Conversation task is for the candidates to 'make a decision' about something, which engages with the examination's rhetorical purpose, (Trinity College London, 2021). With the L2 speakers, there are 27 instances – fewer than occurring in the comparatively smaller TLC-L1 – and the majority (21) of the occurrences are found in the Discussion task, further demonstrating a difference in usage between the L1 and L2 speakers. This difference in usage has been seen elsewhere, with fewer collocations in general between L1 and L2 speakers noted by Laufer and Waldman (2011), and this is also evidence of register-based linguistic variation between L1 and L2 speakers using collocations in the spoken exams (Gablasova et al., 2017).

Further evidence for the finding that L2 speakers use fewer register-influenced collocations comes from looking at the concordance lines to m*ake + sense.* Here, the L1 speakers use the collocation evenly between the two tasks, and the purpose of it is typically to ensure that the interlocutor is following the expression of ideas. This is not the case for the L2 speakers, who are only using the verb + noun collocation for content purposes rather than to support the interaction in a functional way. Again, supporting the interaction is a core component of interactional competence deemed as important in overall language proficiency (Plough et al., 2018).

6.5. RQ4: High frequency delexical verb + noun collocations in L1 and L2 spoken English

Part of the analysis focused on high frequency delexical verb + noun collocations, as these "verbs with little meaning" (Sinclair, 1990, p. 147) tend to have neutral connotations in use leading to minimal bias from other contextual aspects like topic or register (Chi et al., 1994). It was decided to focus on a small number of specific collocations to be better able

to compare the two corpora, as these delexical collocations are evident in both. Due to previous literature focusing on specific delexical verb + noun combinations (e.g. Du et al., 2022, looking at *make* and *take* + noun collocations; Gilquin, 2007 researching *make* collocations; Brezina, 2018, using *make*, *take* and *do* + noun collocations to study language learning; Sawaguchi & Mizumoto, 2022 investigating *make* + noun collocations in writing) and taking into account the most frequent verbs in the TLC-L2, it was decided that focus would be placed on three delexical verbs in particular as case studies – *get, make* and *take.* These three verbs are in the top four most frequent verbs nodes within the verb + noun collocations across the proficiency groups. The remaining verb for the B1 and B2 speakers is *think,* while for the C1/C2 speakers, the fourth verb is *know*. This is interesting to note, as it could be evidence of a shift in stance by the candidates from the more hesitant *think* to the more concrete *know* of advanced speakers (Crossley et al., 2010). This also supports previous suggestions from Salsbury and Bardovi-Harlig (2000) that speakers at a lower proficiency use phrases like *I think* to show stance and that these are easier to acquire due to their singular meaning.

The first high frequency delexical verb in the analysis for each corpus was *get.* In the TLC-L2, there was further support for the evidence of nonlinear development of collocation usage in language learners. This is because there was no linear increase in using the topmost frequent *get* + noun collocations. Instead, there was a surge at the B2 level, with the most top ranked collocations being used by these intermediate candidates. This supports recent findings from Brezina and Fox (2021) and also Paquot et al. (2022) with regard to the nonlinear development of phraseological competence and a particular focus on the nonlinearity of B2 level (intermediate) learners of English. Within the TLC-L2, the collocations can generally be grouped thematically, including work and school with abstract nouns. There is also possible evidence of C1/C2 speakers becoming more precise in their noun usage due to decreased instances of *thing.*

There was further evidence of topic-influenced (*friend, dog, food, student, degree, grade* and *child)* and register-influenced collocations (*opportunity, point, information, idea, impression* and *attention*) within the TLC-L1, with certain collocates of *get* also being similar to those found in the BNC-2014 (*people, money, time, car, job, thing* and *chance*). This demonstrates the suitability of the corpus for further use as a comparison corpus for other L2 spoken language data and as a standalone corpus for further L1 spoken English study (Gilquin, 2022).

Adding to this evidence on high frequency delexical verb + noun collocations, *feel* is the second most frequent collocate co-occurring with the node *make*, and this occurs most frequently in the B2 group in the TLC-L2. This is an abstract-noun collocation, but relative frequency decreases for the C1/C2 group. Once again, this is evidence of the non-linear development of collocations in that not all abstract nouns linearly increase in usage as language proficiency increases thus supporting the notion that formulaic language development is likely to be a complex dynamic system (Duan & Shi, 2021) with individual variation crucial within this (Lowie & Verspoor, 2015). This contradicts suggestions of usage of *spend/waste + time* in Section 4.3.4.1, which demonstrate a more complex relationship. In the language learner corpus, *feel* occurs frequently; however, this noun only co-occurs three times with *make* in the L1 corpus. Looking at the breakdown of usage by proficiency group, the most speakers using *make + feel* are the B2 candidates, and there is a decrease in frequency in the C1/C2 speakers. This could indicate the advanced speakers' usage becoming more similar to the L1 speakers as, instead, the most frequent noun collocate for *make* used in the TLC-L1 corpus is *decision.* It could also be further evidence of topic-influence because it occurs in one specific group supporting findings from Suzuki (2015) that topic can impact linguistic choices. Overall, the collocates used by the native speakers are also more fixed in nature when compared to those used with *get* with notably frequent nouns, including *sense*, *money*, *difference*, *choice* and *change* – again demonstrating that different verbs influence the fixedness of the collocation overall.

Finally, the research also considered the development of high frequency delexical verb + noun collocations concerning the semantic categories the nouns are aligned with, following previous research from Du et al. (2022). This was done by looking at unique collocations to each proficiency level and then categorising them with the USAS semantic tagger (Rayson et al., 2004). This part of the analysis combines RQ1 and RQ4 within the investigation. Looking at all three high frequency delexical verb + noun collocations together, a pattern emerges. *Get*, *make* and *take* + noun collocations show evidence of L2 speakers engaging in using more abstract noun categories in their collocations. This evidence comes from the presence of the B category – the body and the individual – being the most frequent for each of the verbs within the B1 category. For *get* and *take*, the category does not occur in the top five most frequent for advanced speakers. Adding to this is the increase in the X category of psychological actions, states and processes that

occurs for all three verbs. For both *get* and *take*, this does not occur at all in the top five most frequent categories and enters in B2 before increasing at C1/C2, while for *make*, there is a steady increase across all levels. Finally, the analysis shows some verb specific patterns that adhere to this concept of language learners using more abstract concepts as they develop their proficiency. Firstly, the A category of general and abstract terms first occurs in the B2 level for *get*. In contrast, for *make*, the O category of substances, materials, objects and equipment decreases over the proficiency levels. As well as demonstrating this pattern within high frequency delexical verb + noun collocations as was set out to be uncovered with RQ4, the analysis also adds further support to answering RQ1 regarding the frequency of verb + noun collocations across proficiency levels and evidence of nonlinear development of collocations as explored further in Section 6.2. This comes from the u-shaped curve present in both *get* and *take* + noun collocations when looking at the S category nouns of social actions, states and processes. Here, the candidates use these nouns more frequently at B1 and C1/C2 levels, with a dip in usage at the B2 level. This echoes findings from Vedder and Benigno (2016) regarding language proficiency development and, if considering frequency as evidence of development, adding support to Siyanova-Chanturia and Spina (2020) that phraseological production may become worse before it increases at higher proficiency levels. Overall, the findings support evidence from Du et al. (2022) that language learners use more abstract nouns in their collocations as they become more advanced. The findings in this thesis further support this regarding the initial abstract-noun collocation results in Section 4.3.4.

Regarding collocational diversity, in the TLC-L1, the collocates for *take* are more diverse than *make* but less diverse than *get*. Both topic and register influence from the examination context led to differences in collocates for the node *take* in the TLC-L1 when compared to the BNC2014. This demonstrates that language not present might not be due to a lack of proficiency or understanding but a lack of opportunity to engage with that language feature, which connects to speaker opportunity of use research from Buttery and Caines (2012). This should be taken into account when analysing learner language as a lack of some element of language may not necessarily be an inability to produce that element; Kreyer (2021) also supports this idea, noting that it is crucial to consider both task descriptions and opportunity of use when conducting analyses of individual collocations (p. 116).

Chapter 7: Conclusion

This thesis adds novel contributions to corpus linguistics and language learning research in three ways: theoretical, methodological, and pedagogical. This chapter concludes these contributions by first briefly restating the theoretical contributions of the study through a review of the major findings in Section 7.1 (see Chapter 6 for a more thorough discussion). Section 7.2 concludes methodological contributions and Section 7.3 explores some pedagogical implications of the study, before limitations are considered in Section 7.4, leading to opportunities for further research. Finally, the thesis is concluded in Section with 7.5 with closing remarks.

7.1. Theoretical contributions – brief review of main findings

The thesis adds novel findings to the field of language learner research by answering four core research questions. Firstly, there are differences in the (1) frequency and (2) distribution of the use of verb + noun collocations amongst L2 English speakers at B1, B2 and C1/C2 levels and L1 English speakers. Secondly, there is evidence of topic influence on speaker choice of verb + noun collocations in the TLC-L1 and the TLC-L2 corpora. Thirdly, there is evidence of register influence on speaker choice of verb + noun collocations in the TLC-L1 and the TLC-L2 corpora. Finally, there are patterns in how TLC-L1 and TLC-L2 speakers use high frequency delexical verb + noun collocations in spoken examination language, which include differing frequency counts and the influence of topic and register on these collocations. The theoretical implications support previous findings about the challenges with attributing collocational usage based on proficiency level. Using new data, this further supports the argument that more research is needed. The findings demonstrate the complex descriptive picture of collocation development in L2 speech and the individual variation that can be found in both learners and native speakers too.

7.2. Methodological contributions

7.2.1. A new corpus

This thesis is based on a new corpus of L1 spoken British English, the TLC-L1, which is a counterpart to the TLC-L2. Not only is this new data, which is beneficial to furthering the field, but it is also noteworthy that this is a spoken corpus; spoken corpora are considerably more challenging to build due to the time and costs involved in the transcribing process of the corpus creation (Reppen, 2010), resulting in fewer available datasets and therefore adding additional value to the creation of this corpus. Furthermore,

corpora of this kind of spoken interlanguage are especially rare (Gablasova et al., 2017). At 833,878 tokens from 203 speakers, the size of this corpus also holds value when considering its specialised content. The size is comparable to other specialised corpora (e.g., LINDSEI, Gilquin et al., 2010); however, the TLC-L1 is the first corpus of this size – that the author is aware of – that uses native speakers of English engaging in a language testing situation. This means that the corpus can be used for multiple research purposes in a variety of fields, such as SLA and language testing, as well as adding further data to research investigating L1 spoken British English. It is also the only corpus to have collected spoken data from L1 English speakers within the well-defined environment of the GESE language examination from Trinity College London. Some of the many further research possibilities are discussed in Section 7.4 of this thesis.

This thesis was the first investigation of the TLC-L1 and, although not a core focus for the research questions, the results were able to establish its potential as a comparable corpus to the TLC-L2 by way of presence of collocations. Within the TLC-L1 corpus, there is evidence of the collocations used by the speakers to be predictable and expected to be seen in any corpus of spoken English, based on the most frequent verb + noun collocations present being those that generally include common nouns (see Table 54 in Section 5.3.1). The corpus contains 2,150 verb + noun collocations that were categorised as 'more formulaic', with 1,181 types overall. *Make + decision* is the most frequent verb + noun collocation in the dataset, with 33 instances and 21 speakers using it; this means roughly 10% of all speakers in the TLC-L1 used the collocation. This presence of common verbs was the same for the TLC-L2 (see Table 35 in Section 4.3.1). Much like the TLC-L2 corpus, there is evidence of notable influence from two different factors – topic and register – both of which are caused by the context of the language examination setting.

Regarding comparability with other native speaker corpora, there were few instances of unique collocations (see Section 5.2) i.e., that were not present in the larger native speaker corpora of the BNC2014 and the EnTenTen20. Some differences were to be expected, given the fundamentally creative nature of language; nevertheless, the language found, with regards to collocations, was not vastly different to these reference corpora. Therefore, it can be said that the TLC-L1 corpus can be used as a new dataset for investigating L1 spoken English independently. A further advantageous feature of the corpus is that it includes both monologic and dialogic speech – the latter of which has

been noted to be particularly rare (Liesenfeld & Dingemanse, 2022) and so the representation of a more socially interactive type of language is therefore a valuable contribution to the field.

Another benefit of the TLC-L1 corpus is the accessibility of the rich metadata recorded during creation, which is essential to include for interpreting corpus findings (Granger & Lefer, 2020). This metadata was collected based on what information was available for the TLC-L2 to ensure some comparability of variables regarding demographics such as age and gender. However, as a stand-alone corpus, this too is a considerable advantage to those wanting to research L1 spoken British English from a sociolinguistic perspective, as other variables such as occupation and location have been recorded and aligned with the BNC2014 (Brezina et al., 2021; Love et al., 2017) therefore opening further possibilities of research into linguistic register using the TLC-L1 as a comparison to more informal spoken language. Furthermore, the data collected also has the potential to aid in corpus-based pragmatic research due to the metadata available (e.g., Aijmer, 2014).

The motivations of the L1 and L2 speakers involved in each corpus are likely to be very different, which needs to be highlighted when considering the comparability. The L2 candidates are likely to be more motivated due to the stakes involved with the examination – these are 'real-life' exams with subsequent consequences, and they will have likely practised frequently in an educational context with a teacher to ensure the best possible chance for success in the exam. Meanwhile, the L1 participants were taking part in linguistic research for a small sum of money to thank them for their time. The stakes are, of course, much lower i.e., there is some motivation to 'do well' due to a certain level of an intrinsic want to do their best, but there is no consequence after completion. However, the researcher and the examiners noted a level of nervousness and low self-esteem when entering the exam from many of the L1 participants, most likely because they were doing something outside their comfort zone. This means that there is likely some level of nerves present in both the L1 and L2 speakers undertaking the examinations, albeit for different reasons; the L1 participants due to unfamiliarity with the situation and the L2 candidates due to the stakes involved. As such, there is reason to think that the language performance of both L1 and L2 speakers may be (negatively) affected in the context of the examination, which could in turn maximise their comparability.

7.2.2. A new L1 norm?

Along with the value the TLC-L1 brings to linguistic research as outlined above, this new corpus also has the potential to offer insights into native speaker language use within a specific setting: a language examination. This offers a chance to present an alternative norm to English language learners, with potential for significant impact on teaching and testing research and development.

The 'norm' within language teaching and testing has been a much-debated subject, with Gilquin (2022) offering one of the most recent perspectives to the discussion. Learner language has often been seen as the 'other' to the 'norm' of the 'native speaker' within learner corpus research, setting up a parallel that the non-native speaker must achieve proficiency of nativelikeness. However, some argue that being a proficient L2 speaker would be a more appropriate goal to aim for, and this needs to be reflected in language teaching (Cook, 2007). This also aligns with questions raised by Beaulieu (2018) regarding what is considered the appropriate target for L2 speakers when descriptive, textbook and pedagogical norms all differ. Furthermore, Pennycook (2017) highlights that in the current world context, the majority of English speakers are not native speakers. Instead, English as a Lingua Franca is a major speaking context for communication; thus, striving for 'native-like' proficiency is problematic and unnecessary.

There is a need for a norm for language teaching and testing, as learners and educators need something to aim towards. However, there needs to be clarity regarding how language is used when talking about the norm and, encouragingly, describing differences has become more common in more recent years (e.g., Sawaguchi & Mizumoto, 2022) rather than highlighting 'problems' that learners have with language acquisition (Chi et al., 1994). This thesis accepts that a norm can be used to help describe the features of a learner corpus in looking at similarities and differences, but to do this accurately and in a way that is fair to those learners in what they should be striving to achieve with their language learning, this 'norm corpus' needs to be comparable. Hence, the TLC-L1 was created for this purpose.

The TLC-L1 is a native speaker corpus and its design involves expert speakers (the candidates) engaging in the examination tasks, a speaking context they are unfamiliar with. There is no prototypical, expected response, and as seen in the analysis, the engagement with the task varies considerably according to each individual. This means that just as there is no one typical L1 speaker, there should not be the expectation of a

typical language learner. This further supports research into work on the native speaker as a social construct (Seargeant, 2013) and the difficulty of defining language proficiency (Leung, 2022). In addition, the examiners found that the L1 candidates were challenged with some tasks in the exam. Although they had received training before the examination and were fully competent in English, they still had some issues with fulfilling the criteria they were expected to meet in the exam. This shows the importance of speaking within the appropriate register of the situation, rather than simply the mastery of language; in other words, there needs to be a mastery of the pragmatic decisions in employing this language within a specific setting.

Further investigations could position the TLC-L1 corpus as another norm that can be used to gain perspective on L1 individual variation and use it for L2 spoken English analysis in the TLC-L2 and other corpora. Furthermore, this corpus is unique in design because the speakers are not expert test takers in the same way written corpora can include expert writers. They are native speakers, but these speakers have not practised the tasks involved. The tasks may involve competencies that the speakers use sparingly, i.e., engaging in interaction on current affairs topics (Conversation task) or defending their beliefs about a chosen subject (Discussion task). The TLC-L1 had speakers in an unfamiliar setting using their native language. At the same time, the TLC-L2 candidates were likely to have received more comprehensive training for the situation, even though their mastery of the language, such as using correct grammar and vocabulary, was less proficient than the native speakers. As it is more effective and reasonable to compare novice learners to L1 speakers that are novices in a particular genre or type of interaction (Gilquin, 2022), the TLC-L1 presents a new, fairer comparison for L2 language examination language.

7.3. Pedagogical implications

This section will offer some general pedagogical implications that have arisen from the analysis. These implications are considered in relation to main trends within current work to link research to practice, as discussed by Szudarski (2023), before an example of a corpus-informed activity that can be used and adapted for L2 English learners in the classroom is introduced (see Appendix).

The results from this investigation suggest that attention within the language classroom could be directed to the context of collocation production, namely register and topic. The analysis has shown these two contextual factors to have an influence on the types of collocation used. There is no single approach for teaching collocations that is the most

appropriate, as this depends on the context of the classroom and the students; as such, taking a multi-faceted approach to teaching collocations is beneficial (Liu, 2021). Using corpus methods is one such approach and Szudarski (2023) outlines four main trends within the application of corpus linguistics to language teaching that can be considered in the context of this research.

Firstly, Szudarski (2023) notes the importance of the indirect impact of corpora and how this can aid the teaching of L2 collocations (p. 59), for example, through the use of syllabus design and material creation. Szudarski summarises the benefit of this, supported by Curry et al. (2022), in terms of corpus-based research reflecting how language is changing, which can be used to inform L2 teaching materials. For the findings in this thesis, both the TLC-L2 and the TLC-L1 can be utilised to help identify typical collocations within the GESE language exam from Trinity College London. As well as register-influenced collocations, this research has also found that many of the verb + noun collocations used are influenced by topic. There is evidence of more advanced speakers using more abstract concepts, and thus collocations, particularly in the spoken tasks that are candidate-led such as the Discussion test. Therefore, it would be beneficial for those studying to take the GESE to understand the impact topic can potentially have on their use of language. Interpreting this into the language classroom, it would be helpful for teachers to engage in idea generation activities to encourage candidates to consider choosing appropriate topics to engage with in the Presentation task (if at C1/C2 level) or the Discussion task, if working with learners at a lower proficiency. Furthermore, it would be beneficial to havw teachers select different topics within the classroom discussions, to ensure there are ample opportunities to use a variety of different collocations, alongside educators supporting students in spontaneous spoken tasks in the classroom. This is to ensure they are demonstrating the language skills needed for the GESE within these tasks. Moreover, additional exploration of semantic categories within the corpora could help to identify topics that are particularly fruitful in allowing opportunities for abstract-noun collocations. Extending beyond the classroom, the results could also be used for Trinity College London item writing to ensure that a wide range of topics are put forth to the candidates, thereby creating opportunities for complex and varied collocation usage, and for other language examination boards to consider in their own item writing process.

Secondly, corpus-based lists of collocations are highlighted as beneficial to language teaching (Szudarski, 2023, p. 63), with a suggestion that “they should be constructed for

particular purposes and particular groups of L2 learners" (p. 65). This construction could begin with the initial register-influenced collocations described in this thesis to create a list of general interactive conversational collocations based on the GESE. Further research could expand this with more investigation into these collocations, developing beyond the verb + noun collocations highlighted here to other types of formulaic language such as adverb + verb collocations. This would be of particular benefit as many collocation lists are mostly based on written language (such as the Academic English Collocation List from Lei & Liu, 2018 and the PHRASE List from Martinez & Schmitt, 2012) rather than interactive spoken language, thus aligning the corpus-based list to the specific purpose of the speaking examination.

The third major theme from Szudarski (2023, p. 63) relates to corpus- and technology-enhanced L2 teaching resources. The author gives examples of these with IdiomsTube (Lin, 2022) and ColloCaid (Frankenberg-Garcia et al., 2019), the latter of which makes real-time collocation suggestions during writing. Taking inspiration from the latter and with advancing technology, the Trinity Lancaster Corpora could be used to develop a similar tool for spoken interactive language, with focus on highlighting collocations amongst other formulaic language features while students are producing their task answers. Real time feedback could be given to encourage collocations that are beneficial for maintaining and controlling the conversation e.g., checking for understanding (such as *make + sense*) or clarification requests (such as *repeat + question*).

Finally, Szudarski (2023) highlights the importance of Data Driven Learning (DDL) in relation to teaching collocations, a topic which is explored more thoroughly in this thesis in Section 2.4.4.1. This "direct use of corpus data in language teaching and teacher training" (p. 67) inspired the sample worksheet in the Appendix of this thesis, which involves teaching for register. The TLC-L1 analysis showed L1 speakers using more collocations that were related to maintaining the interaction of the examination with their examiner interlocutor (see Section 6.4). Interactional competence is a major focus of the GESE and a core goal for language learners in general. Therefore, highlighting these collocations that are used to maintain the cohesion of the interaction could benefit English language students as they work to achieve higher levels of proficiency in their language use, as not only the content of their spoken conversation is of importance but also on the pragmatics involved in the interaction.

The results from this research found a variety of verb + noun collocations that occurr due to the register of language examination, including *choose + topic*, *make + decision* and *make + sense*, which have likely arisen due to the nature of the GESE, and the language involved in the Discussion task and Conversation task. The topic chosen for the Discussion task (with is either presented alone for the B1 group or based on the Presentation task in the B2 and C1/C2 groups) is supposed to be discursive in nature so this naturally creates language opportunities for the examiners to be asking questions about the candidates reasonings while exploring the topic. The research found that some collocations present in both corpora engage with this rhetorical purpose of the exam and help to maintain the conversation; the examiner is asking about the candidates' judgements while the candidate asks clarification questions (*ask + question* in Section 5.1.3.5 and *repeat/understand + question* in Section 4.3.3.1), checks for understanding (*make + sense* in Section 4.4.2.4) and signposts their responses (*choose + topic* in Section 4.3.3.2).

The above guided the creation of the example worksheet in the Appendix; this is targeted for C1 learners wanting to develop their proficiency to C2 level and uses one of the register-influenced collocations from the thesis results, *make + sense*. The activities use #LancsBox X (Brezina & Platt, 2023), taking inspiration from Liu's (2021) study using #LancsBox V.2. The worksheet is also set up so as not to posit the data from the TLC-L1 as the 'correct' way to use the collocation, but instead to have learners look at how both datasets of speakers use the collocation and notice patterns in this way. By comparing to the more informal native speaker corpus of the BNC2014, the intention is that the activity will raise awareness of how to use the collocation within the language testing context.

Overall, the findings from this study can contribute to L2 English pedagogy in a variety of ways to ensure collocations selected for implicit or explicit teaching are relevant when considering why they are being taught and who they are being taught to.

7.4. Limitations and further research opportunities

As discussed in Section 7.2.1, this thesis contributes to corpus linguistics methodology by introducing a new dataset – the TLC-L1. This means only a small part of the new corpus could be investigated within the scope of this research. Further research could extend to other types of frequently occurring phraseological collocations used by the L1 speakers in this language examination setting. Such collocations have already begun to

be researched in an L2 speaker setting, such as adjective + noun (e.g., Brezina & Fox, 2021, which then compared to the BNC2014 for the L1 comparison) or adverb + adjective (e.g., Lee & Shin, 2021, which investigated recognition and recall of these and other collocations for L2 speakers). This would contribute to a richer picture of both L1 usage and L2 language development. The corpus can also be used to compare the usage of L1 speakers to previously investigated aspects of L2 language within the TLC-L2 corpus, such as certainty adverbs (Pérez-Paredes & Díez-Bedmar, 2019), lexical backchannels (Castello & Gesuato, 2019), and filled pauses (Götz, 2019).

Further to expanding on research into different types of linguistic feature, the TLC-L1 has the potential to be used in other areas of research. As detailed in Section 3.2.3, a broad range of rich metadata was recorded during the data collection process in this project, focusing on a variety of social categories of the participants. This could be used, for example, to consider linguistic differences based on variables such as educational background (see Section 3.2.3.1.3) or language learning experience (see Section 3.2.3.1.6). Investigating such variables are beyond the scope of this project, but the introduction of the TLC-L1 in this thesis elicits many exciting opportunities for further research in other fields such as sociolinguistics.

Some limitations should be acknowledged in this research, with some related to the data collection process. Firstly, during the debriefing discussions after data collocation, one examiner mentioned feeling more lenient at the beginning of the data collection due to the 'chattier' nature of the L1-L1 interactions. They noted that the shared language background meant there was some degree of pragmatic understanding that may not always occur with the L2 candidates, even at the highest-grade exam. The examiner noted they were not necessarily trying to push the candidate's language because it was already there, and so they did not feel like they needed to. Approaching the examination differently because of the language background of the candidate could have influenced the language elicited by the examiner in the TLC-L1. However, only one examiner mentioned this when asked, and all examiners were briefed to try to avoid this, where possible, during the data collection.

A limitation that also works as useful feedback for teachers preparing their students for Trinity College London's GESE is that all examiners noted that L1 participants utilised some of the topics in the Conversation and Interactive tasks better than others in the TLC-

L1. In fact, these topics elicited more functions and language appropriate to the level. This indicates that topic is critical to consider when developing tasks for the GESE from the view of the examiner and could be an interesting avenue for further research using the L1 corpus.

Regarding the data analysis done within this thesis, a limitation is that the research direction did not give rise to enough space to consider the impact of language and cultural background on language production within the TLC-L2, as this has been acknowledged to be a factor in collocation production. Although this was beyond the scope of the current work, preliminary studies have begun to use the TLC-L2 for investigating the role of the L1 in language learning (for example Götz, 2019 looking at filled pauses as a language feature; Castello and Gesuato, 2017 looking at lexical backchannels). More focus on the role of linguistic background in collocation use within the TLC-L2 would be of value to consider alongside the new TLC-L1 as well.

Furthermore, again in relation to the scope of this thesis, only one type of phraseological collocation was investigated in this research – verb + noun collocations. It would be beneficial to extend what has been done here to other types of collocation to get a sense of the picture across phraseological collocations in general. Some well-researched collocation types that would benefit from this new data include adjective + noun collocations; this would act to extend the work done previously in Brezina and Fox (2021) to include the TLC-L1 as a comparison.

Finally, further research could also take a different approach than this primarily descriptive account of the corpora. Description is vital for new corpora to explore the data within them, but further investigations could take these descriptive accounts and also look at linguistic aspects of learner language, such as erroneous collocations like Kreyer (2021), Nesselhauf (2005), Siyanova and Schmitt (2008) and Laufer and Waldman (2011), to investigate atypical collocations. The analysis here briefly touched on unique verb + noun collocations within the TLC-L1 that did not appear in other larger reference corpora and also looked at the unique occurrences of high frequency delexical verb + noun collocations in the TLC-L2. However, this could be further extended to consider the creativity of collocation production in both L1 and L2 speech.

Furthermore, the scope of the study here did not allow for exploration of association measures, which would be a valuable addition to future investigations taking the blended

phraseological and frequency-based approach. It would be of interest to see how frequency (measured by t-scores) and exclusivity (measured by MI scores) of verb + noun collocations in both the Trinity Lancaster Corpora compare to previous research and whether this perspective could further unpick the complexity of collocation development L2 spoken language.

7.5. Closing remarks

This thesis introduces the new TLC-L1, a corpus of L1 spoken English examination language and details a descriptive study aimed to contribute to learner corpus research by exploring the nature of verb + noun collocations in L1 and L2 spoken English using corpus methods. The investigation found evidence of nonlinear development of L2 verb + noun collocations with an influence of topic and register on the types of verb + noun collocations used by both L1 and L2 English speakers, particularly in collocations using high frequency delexical verbs. Looking ahead, it is hoped the thesis is used as a next step methodologically with the introduction of a new corpus to investigate various linguistic phenomena and theoretically in further uncovering the complex nature of formulaic language in spoken L1 and L2 English.

Appendix 1 – Consent Form (over 18 years old)

CONSENT FORM

Project Title: Development of a Native Speaker Corpus of Spoken British English

Name of Researchers: Dana Gablasova, Vaclav Brezina, Tony McEnery, Lorrae Fox

Email: l.m.fox@lancaster.ac.uk

Please tick each box

1. I confirm that I have read and understand the information sheet for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily	☐
2. I understand that my participation is voluntary and that I am free to withdraw at any time during my participation in this study and within 2 weeks after I took part in the study, without giving any reason. If I withdraw within 2 weeks of taking part in the study my data will be removed.	☐
3. I understand that any information given by me may be used in future reports, academic articles, publications or presentations by the researcher/s, but my personal information will not be included and I will not be identifiable.	☐
4. I understand that the corpus (database) created from the language samples, with fully anonymised data, will be available to other researchers for re-use.	☐
5. I understand that my name will not appear in any reports, articles or presentation without my consent.	☐
6. I understand that any interviews will be audio-recorded and transcribed and that data will be protected on encrypted devices and kept secure.	☐
7. I understand that data will be kept according to University guidelines for a minimum of 10 years after the end of the study.	☐
8. I agree to take part in the above study.	☐

_____ _____ _____

Name of Participant Date Signature

I confirm that the participant was given an opportunity to ask questions about the study, and all the questions asked by the participant have been answered correctly and to the best of my ability. I confirm that the individual has not been coerced into giving consent, and the consent has been given freely and voluntarily.

Signature of Researcher /person taking the consent_____

Date_____ Day/month/year

Appendix 2 – Consent Form (Parental or Guardian)

PARENTAL OR GUARDIAN CONSENT FORM

Project Title: Development of a Native Speaker Corpus of Spoken British English

Name of Researchers: Dana Gablasova, Vaclav Brezina, Tony McEnery, Lorrae Fox

Email: d.gablasova@lancaster.ac.uk

Please tick each box

9. I confirm that I have read and understand the information sheet for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily	☐
10. I understand that my child or dependent's participation is voluntary and that I am free to withdraw my consent at any time during their participation in this study and within 2 weeks after they took part in the study, without giving any reason. If I withdraw consent for their participation within 2 weeks of taking part in the study, their data will be removed.	☐
11. I understand that any information given by my child or dependent may be used in future reports, academic articles, publications or presentations by the researcher/s, but their personal information will not be included and they will not be identifiable.	☐
12. I understand that the corpus (database) created from the language samples, with fully anonymised data, will be available to other researchers for re-use.	☐
13. I understand that my child or dependent's name will not appear in any reports, articles or presentation without my or their consent.	☐
14. I understand that any interviews will be audio-recorded and transcribed and that data will be protected on encrypted devices and kept secure.	☐
15. I understand that data will be kept according to University guidelines for a minimum of 10 years after the end of the study.	☐
16. I agree for my child or dependent to take part in the above study.	☐

_____ _____ _____

Name of Legal Guardian Date Signature

I confirm that the participant's legal guardian was given an opportunity to ask questions about the study, and all the questions asked by the legal guardian have been answered correctly and to the best of my ability. I confirm that the individual has not been coerced into giving consent, and the consent has been given freely and voluntarily.

_____ _____

Signature of Researcher /person taking the consent Date

One copy of this form will be given to the legal guardian and the original kept in the files of the researcher at Lancaster University

Appendix 3 – Participant Information Sheet (over 18 years old)

Participant Information Sheet

Participant information sheet
I am a researcher at Lancaster University and I would like to invite you to take part in a research study about the current use of British English. Please take time to read the following information carefully before you decide whether or not you wish to take part.

What is the study about?
This study aims to collect samples of spoken language from native speakers of British English. These samples will then be used to create a large corpus (electronic database) of British English that can be used to study patterns in current English use. Findings from this corpus can be compared with those of the British National Corpus which records and represents British English from early 1990s. The findings can be also compared to the results from a corpus representing English from learners of English to better understand how learners of English differ from native speakers.

Why have I been invited?
We are interested in recording speech from native speakers of English. I would be very grateful if you would agree to take part in this study.

What will I be asked to do if I take part?
If you decided to take part, this would involve the following:

- *Completing a brief questionnaire about your background and language use (10 min.)*
- *Reading a brief informational handout about the interview (10 min.)*
- *Taking part in an interview (25 min).*
- *Preparing a five-min. presentation on a topic of your interest to be used in the interview.*
- *Completing a vocabulary test (15 min).*

What are the possible benefits from taking part?
Participating in the study will allow you to become part of a project that will help our understanding of how English is used and how learners of English differ from native speakers.

Do I have to take part?
No. It's completely up to you to decide whether or not you take part. Your participation is voluntary. If you are a student and you decide not to take part in this study, this will not affect your studies and the way you are assessed on your course

What if I change my mind?
If you change your mind, you are free to withdraw at any time during your participation in this study. If you want to withdraw, please let me know, and I will extract any data you contributed to the study and destroy them. However, it is difficult and often impossible to

take out data from one specific participant when this has already been anonymised or pooled together with other people's data. Therefore, you can only withdraw up to 2 weeks after taking part in the study.

What are the possible disadvantages and risks of taking part?
It is unlikely that there will be any major disadvantages to taking part. Taking part will mean investing about an hour of your time for the study (e.g. to take part in the interview, fill in the questionnaire and take part in the information session).

Will my data be identifiable?
During the data collection and transcription, only the members of the research team will have access to the data. The only other person who will have access to the recording of your interview is a professional transcriber who will listen to the recordings and produce a written record of what you have said. The transcriber has signed a confidentiality agreement.

We will keep all personal information about you (e.g. your name and other information about you that can identify you) confidential, that is we will not share it with others. Any personal information will be removed from the written record of your contribution.

After the data are transcribed and the corpus is created, the corpus will be made available to other researchers as well. No personal data or contributions will be identifiable in the corpus which will include data from many other participants.

How will we use the information you have shared with us and what will happen to the results of the research study?

We will use the information you have shared with us in the following ways:

- We will use it for research purposes only. This will include, for example, academic publications or recommendations for teachers of English. We will also present the results of the study at academic and practitioner conferences. We may share the results with other relevant institutions (e.g. Trinity College London).
- The corpus will be made available for future use by other researchers.
- If anything you tell us in the interview suggests that you or somebody else might be at risk of harm, we will be obliged to share this information with the Head of the Department of Linguistics and English Language. If possible we will inform you of this breach of confidentiality.

How my data will be stored
Your data will be stored in encrypted files (that is no-one other than the researcher team will be able to access them) and on password-protected computers. We will store hard copies of any data securely in locked cabinets in a university office. We will keep data that can identify you separately from non-personal information (e.g. your questionnaire). In accordance with University guidelines, we will keep the data securely for a minimum of ten years.

What if I have a question or concern?
If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact myself, Lorrae Fox (l.m.fox@lancaster.ac.uk)

If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact:

Professor Uta Papen,
Head of Department, Linguistics and English Language Professor,
Faculty of Arts and Social Sciences, Linguistics and English Language

Tel:+44 1524 593245 E-Mail: u.papen@lancaster.ac.uk

This study has been reviewed and approved by the Faculty of Arts and Social Sciences and Lancaster Management School's Research Ethics Committee.

For further information about how Lancaster University processes personal data for research purposes and your data rights please visit our webpage: www.lancaster.ac.uk/research/data-protection

Thank you for considering your participation in this project.

Appendix 4 – Participant Information Sheet (Parental or Guardian)

Parental Information Sheet:

Permission for Participation of a Child or Dependent in a Research Study

Participant information sheet
I am a researcher at Lancaster University and I would like to invite you to consider allowing your child or dependent to take part in a research study about the current use of British English. Please take time to read the following information carefully before you decide whether or not you wish your child or dependent to take part.

What is the study about?
This study aims to collect samples of spoken language from native speakers of British English. These samples will then be used to create a large corpus (electronic database) of British English that can be used to study patterns in current English use. Findings from this corpus can be compared with those of the British National Corpus which records and represents British English from early 1990s. The findings can be also compared to the results from a corpus representing English from learners of English to better understand how learners of English differ from native speakers.

Why has my child or dependent been invited?
We are interested in recording speech from native speakers of English, of all ages. The corpus so far doesn't include speakers under 18 years old and so I would be very grateful if you would agree for your child or dependent to take part in this study to help expand the scope of this research.

What will my child or dependent be asked to do if they take part?
If you decided to agree for your child or dependent to take part, their participation would involve the following:

- *Completing a brief questionnaire about their background and language use (10 min.)*
- *Taking part in an information session about the interview (15 min).*
- *Taking part in an interview (30 min).*
- *Preparing a five-min. presentation on a topic of their interest to be used in the interview.*
- *Completing a vocabulary test (15 min).*

What are the possible benefits from taking part?
Participating in the study will allow your child or dependent to become part of a project that will help our understanding of how English is used and how learners of English differ from native speakers.

Do I have to allow my child or dependent take part?
No. It's completely up to you to decide whether or not your child or dependent will take part. Their participation is voluntary. If you don't want your child or dependent to take

part in this study, this will not negatively affect any part of their school life. Equally, they will not gain any scholarly advantage from the study, only the benefits outlined above.

What if I change my mind?
If you change your mind, you are free to withdraw your consent at any time up to two weeks after this study. If you want to withdraw your child or dependent's data, please let me know, and I will extract any data contributed to the study and destroy them. However, it is difficult and often impossible to take out data from one specific participant when this has already been anonymised or pooled together with other people's data. Therefore, you can only withdraw up to 2 weeks after taking part in the study.

What are the possible disadvantages and risks of taking part?
It is unlikely that there will be any major disadvantages to taking part. Taking part will mean investing about an hour for the study (e.g. to take part in the interview, fill in the questionnaire and take part in the information session).

Will my child or dependent's data be identifiable?
During the data collection and transcription, only the members of the research team will have access to the data. The only other person who will have access to the recording of the interview is a professional transcriber who will listen to the recordings and produce a written record of what your child or dependent has said. The transcriber has signed a confidentiality agreement.

We will keep all personal information about your child or dependent (e.g. their name and other information about them that can identify them) confidential, that is we will not share it with others. Any personal information will be removed from the written record of your child or dependent's contribution.

After the data are transcribed and the corpus is created, the corpus will be made available to other researchers as well. No personal data or contributions will be identifiable in the corpus which will include data from many other participants.

How will we use the information your child or dependent has shared with us and what will happen to the results of the research study?

We will use the information you have shared with us in the following ways:

- We will use it for research purposes only.This will include, for example, academic publications or recommendations for teachers of English. We will also present the results of the study at academic and practitioner conferences. We may share the results with other relevant institutions (e.g. Trinity College London).
- The corpus will be made available for future use by other researchers.
- If anything, your child or dependent tells us in the interview suggests that they or somebody else might be at risk of harm, we will be obliged to share this information with the Head of the Department of Linguistics and English Language. If possible, we will inform you of this breach of confidentiality.

How the data will be stored
The data will be stored in encrypted files (that is no-one other than the researcher team will be able to access them) and on password-protected computers. We will store hard copies of any data securely in locked cabinets in a university office. We will keep data that can identify your child or dependent separately from non-personal information (e.g. their questionnaire). In accordance with University guidelines, we will keep the data securely for a minimum of ten years.

What if I have a question or concern?
If you have any queries or if you are unhappy with anything that happens concerning your child or dependent 's participation in the study, please contact myself, Lorrae Fox, l.m.fox@lancaster.ac.uk

If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact:

Professor Uta Papen,
Head of Department, Linguistics and English Language Professor,
Faculty of Arts and Social Sciences, Linguistics and English Language

Tel:+44 1524 593245 E-Mail: u.papen@lancaster.ac.uk

This study has been reviewed and approved by the Faculty of Arts and Social Sciences and Lancaster Management School's Research Ethics Committee.

For further information about how Lancaster University processes personal data for research purposes and your data rights please visit our webpage:
www.lancaster.ac.uk/research/data-protection

Thank you for considering allowing your child or dependent to participate in this project.

Appendix 5 – Participant Information Sheet (under-18 years old)

Under-18 Participant Information Sheet

Participant information sheet
I am a researcher at Lancaster University and I would like to invite you to take part in a research study about the current use of British English. Please take time to read the following information carefully before you decide whether or not you wish to take part.

What is the study about?
This study aims to collect samples of spoken language from native speakers of British English. These samples will then be used to create a large corpus (electronic database) of British English that can be used to study patterns in current English use. Findings from this corpus can be compared with those of the British National Corpus which records and represents British English from early 1990s. The findings can be also compared to the results from a corpus representing English from learners of English to better understand how learners of English differ from native speakers.

Why have I been invited?
We are interested in recording speech from native speakers of English. The corpus so far doesn't include speakers under 18 years old and so I would be very grateful if you would agree to take part in this study to help expand the scope of this research.

What will I be asked to do if I take part?
If you decided to take part, this would involve the following:

- *Completing a brief questionnaire about your background and language use (10 min.)*
- *Taking part in an information session about the interview (15 min).*
- *Taking part in an interview (30 min).*
- *Preparing a five-min. presentation on a topic of your interest to be used in the interview.*
- *Completing a vocabulary test (15 min).*

What are the possible benefits from taking part?
Participating in the study will allow you to become part of a project that will help our understanding of how English is used and how learners of English differ from native speakers.

Do I have to take part?
No. It's completely up to you to decide whether or not you take part. Your participation is voluntary. If you don't want to take part in this study, this will not negatively affect any part of your school life. Equally, you will not gain any scholarly advantage from the study, only the benefits outlined above.

What if I change my mind?
If you change your mind, you are free to withdraw at any time during your participation in this study. If you want to withdraw, please let me know, and I will extract any data you

contributed to the study and destroy them. However, it is difficult and often impossible to take out data from one specific participant when this has already been anonymised or pooled together with other people's data. Therefore, you can only withdraw up to 2 weeks after taking part in the study.

What are the possible disadvantages and risks of taking part?
It is unlikely that there will be any major disadvantages to taking part. Taking part will mean investing about an hour of your time for the study (e.g. to take part in the interview, fill in the questionnaire and take part in the information session).

Will my data be identifiable?
During the data collection and transcription, only the members of the research team will have access to the data. The only other person who will have access to the recording of your interview is a professional transcriber who will listen to the recordings and produce a written record of what you have said. The transcriber has signed a confidentiality agreement.

We will keep all personal information about you (e.g. your name and other information about you that can identify you) confidential, that is we will not share it with others. Any personal information will be removed from the written record of your contribution.

After the data are transcribed and the corpus is created, the corpus will be made available to other researchers as well. No personal data or contributions will be identifiable in the corpus which will include data from many other participants.

How will we use the information you have shared with us and what will happen to the results of the research study?

We will use the information you have shared with us in the following ways:

- We will use it for research purposes only. This will include, for example, academic publications or recommendations for teachers of English. We will also present the results of the study at academic and practitioner conferences. We may share the results with other relevant institutions (e.g. Trinity College London).
- The corpus will be made available for future use by other researchers.
- If anything you tell us in the interview suggests that you or somebody else might be at risk of harm, we will be obliged to share this information with the Head of the Department of Linguistics and English Language. If possible, we will inform you of this breach of confidentiality.

How my data will be stored
Your data will be stored in encrypted files (that is no-one other than the researcher team will be able to access them) and on password-protected computers. We will store hard copies of any data securely in locked cabinets in a university office. We will keep data that can identify you separately from non-personal information (e.g. your questionnaire). In accordance with University guidelines, we will keep the data securely for a minimum of ten years.

What if I have a question or concern?
If you have any queries or if you are unhappy with anything that happens concerning your participation in the study, please contact myself, Lorrae Fox (l.m.fox@lancaster.ac.uk)

If you have any concerns or complaints that you wish to discuss with a person who is not directly involved in the research, you can also contact:

Professor Uta Papen,
Head of Department, Linguistics and English Language Professor,
Faculty of Arts and Social Sciences, Linguistics and English Language

Tel:+44 1524 593245 E-Mail: u.papen@lancaster.ac.uk

This study has been reviewed and approved by the Faculty of Arts and Social Sciences and Lancaster Management School's Research Ethics Committee.

For further information about how Lancaster University processes personal data for research purposes and your data rights please visit our webpage: www.lancaster.ac.uk/research/data-protection

Thank you for considering your participation in this project.

Appendix 6 – Training Sheet: What to Expect on the Day

Interview: what to expect?

Before the interview: very short preparation

There is a bit of preparation before the interview. Here is a list of three simple things that need to be done before the interview.

1) Prepare a 5 min presentation on a topic of your own choice (e.g. climate change, the role of women in Shakespeare, arranged marriage, Brexit, free childcare, advantages of train travel, etc.)
2) Make a few notes (bullet points) about the presentation outlining the main points you want to mention for the interviewer.
3) Send the notes to l.m.fox@lancaster.ac.uk the evening before the interview at the latest. We

On the day: up to 1 hour

The interview has five tasks, each lasting approx. 5 minutes. The tasks are *Topic presentation, Discussion, Interactive task, Listening task* and *Conversation.* You do not have to remember the sequence; the interviewer will keep an eye on the structure and announce each speaking task.

1. Topic presentation (5 min): This task allows the participant to demonstrate language use when talking without interruption on a personalised topic.

Participant:

- Have prepared a 5 min presentation on a topic of their choice; you can bring your own notes and refer to them (but the presentation must NOT be written out as sentences; the notes should be brief).
- The presentation should have a clear structure and you should say at the beginning what you are going to talk about.
- You should finish by asking the interviewer whether he/she has any questions or comments.
- The "presentation" will be semi-formal, seated, and there is no expectation for any visual aids. You will simply be talking uninterrupted and presenting a topic for 5 minutes.

Interviewer: Will take notes about the content for questions in the 'Discussion'.

2. Discussion (5 min): The purpose is to have an authentic discussion on the ideas and opinions given in the presentation.

Participant:

- Should discuss opinions, ideas.
- Should also be proactive and ask the interviewer questions and offer comments on his/her opinions; you can also challenge the interviewer's ideas and opinions.
- Be ready to justify and elaborate on the ideas and opinions from the presentation.

Interviewer: Will ask different questions; at some point, the interviewer may disagree with the participant and may ask challenging questions or offer challenging comments.

---

3. Interactive task (5 min): The purpose of the task is to demonstrate the participant's ability to take control and maintain interaction.

Interviewer: Will present a 'prompt' describing a situation. The following is an example of prompts.

Grade 12 Interactive prompts
1. The concept of a world without borders may seem like an impossible dream, but I feel it's one that's worth pursuing.
2. One effect of increased globalisation is that minority languages are gradually becoming extinct. Many people regret this but I'm not sure it's such a bad thing.
3. Some people claim that sensationalist journalism simply reflects a society and doesn't shape it. I wonder if this is really the case.
4. Some schools encourage competitiveness in their students, while others generally discourage it. It's clearly a controversial issue.

Participant: It is important that, once the interviewer has set up the situation, you take responsibility for the interaction by asking questions and commenting on the interviewer's responses. You should ask questions to find out more about what the interviewer thinks, what his/her position is, offer comments, suggestions and opinions. Keep the conversation moving forward. This is the task in which you should be most proactive in the conversation. You can imagine this as developing a conversation with a stranger on the train ☺.

---

4. Listening task (5 min): Listen to a short text and answer a question about it, thus showing understanding of spoken English.

Interviewer: Will read three short texts; each of the texts will finish with a question and the participant answers it briefly; the responses are expected to be very short.

---

5. Conversation (5 min): The purpose is to give the participant an opportunity to take part in a genuine exchange of information, ideas and opinions.

Interviewer: Will initiate conversation on two topics. The interviewer will always first introduce the topic and then start a conversation on it. Some possible broad topics are: social issues, stress management, the rights of the individual, the media, etc.

Participant: Engages in a discussion, contributing ideas and opinions, asking about and commenting on interviewers' ideas and opinions. You can also develop the topic further and in new directions.

---

6. After the interview

The interview will be followed by a brief questionnaire and vocabulary test.

For more information about the interviewers and to see examples of the interviews see the Trinity College London website: http://www.trinitycollege.co.uk/site/?id=3109

You can also watch a video of the whole interview on this website. Please note that the examples show both, more and less successful communication. The interview that was closest to the target communication is that of Jakub (Grade 12).

For more information about the project or if you have any questions, please feel free to get in touch with Lorrae Fox (l.m.fox@lancaster.ac.uk) To confirm you have read this information, please follow this link.

Appendix 7 – C-test

*Read each of the **four** passages below and fill in the missing letters, so that the words created fit grammatically and logically into each text as a whole.*

The second half of every other word has been left out. If the word is three letters long (e.g., "the"), then two letters are missing (leaving "t_ _"), for example.

You should spend no more than five minutes per passage.

①

Nothing beats the heat like a refreshing dip in a swimming pool. But wh_ _ it co_ _ _ to wa_ _ _ , both ki_ _ and adu_ _ _ need t_ be car_ _ _ _.

Susan King's daug_ _ _ _ _ — Alison, 12, a_ _ Christy, 9 — a_ _ in th_ _ _ grandparents' po_ _ every d_ _. King's gi_ _ _ have ma_ _ pool ru_ _ _, including n_ _ being all_ _ _ _ in t_ _ pool ar_ _ without a_ adult, n_ jumping i_ the sha_ _ _ _ end, n_ running around the pool and no holding each other under water.

"Kids drown quickly and quietly," cautions Jen Costello of the National Safe Kids Campaign. Even less than an inch of water can be enough.

②

The global dominance in word processing software held by Microsoft is under threat from a new coalition. The Silicon Valley-ba_ _ _ Google and Sun Microsystems ha_ _ announced a formi_ _ _ _ _ alliance. Th_ _ plan t_ make wo_ _ processing a_ _ spreadsheet prog_ _ _ _ available o_ the Inte_ _ _ _, in a dir_ _ _ challenge t_ Microsoft. Indu_ _ _ _ observers s_ _ increased compe_ _ _ _ _ _ in t_ _ global soft_ _ _ _ market wi_ _ be go_ _ for cons_ _ _ _ _.

The comp_ _ _ _ _ could n_ _ say wh_ _ Google wo_ _ _ begin carr_ _ _ _ Sun's

technology,
including OpenOffice, which was launched in 2000.

TURN OVER

③

There are many possible causes of insomnia. Sometimes th_ _ _ is o_ _ main ca_ _ _ , but
of_ _ _ several fac_ _ _ _ interacting toge_ _ _ _ will ca_ _ _ a sl_ _ _ disturbance. T_ _ causes o_ insomnia inc_ _ _ _: psychological, phys_ _ _ _ or temp_ _ _ _ _ factors. A la_ _ of
a go_ _ night's sl_ _ _ can le_ _ to var_ _ _ _ problems a_ _ a vic_ _ _ _ circle co_ _ _ develop. Profes_ _ _ _ _ _ counselling fr_ _ a doc_ _ _, therapist o_ sleep specialist can help individuals cope with these conditions.

④

A popular form of recreation in Britain is attendance at dog racing. The fi_ _ _ impression o_
the ar_ _ _ is attra_ _ _ _ _. However, t_ _ races thems_ _ _ _ _ are uninte_ _ _ _ _ _ _ — a f_ _ dogs cha_ _ _ _ a tin ha_ _ — but thi_ _ _-two mil_ _ _ _ people att_ _ _ them annu_ _ _ _. Out o_ two ho_ _ _, barely fi_ _ to t_ _ minutes a_ _ usually dev_ _ _ _ to t_ _ actual rac_ _ _. There wo_ _ _ be n_ interest i_ it were not for the betting. Many of the audience pay little attention to the racing, but have their eyes fixed on a board which gives the number of the winners.

From http://www.sz.uni-stuttgart.de/englisch/einstufungstest.html

Appendix 8 – Background Questionnaire

Background Questionnaire

1. Name [We ask for your name only in order to match your responses to the interview data. All data will be anonymised which means your name will not appear anywhere and will be kept confidential.]

_____

2. Gender

_____

3. Age

_____

4. Education: What is the highest level of education you have <u>completed</u>? [Please tick ☑ one.]

Primary education	☐
Secondary education	☐
Tertiary education (university, college etc.) – Bachelor's degree	☐
Tertiary education (university, college etc.) – Master's degree	☐
Tertiary education (university, college etc.) – Doctoral degree	☐

5. Education: If you are still studying, what degree, programme and year are you in?

_____

6. Employment: If you are in employment, what is your occupation? How long have you been working in your current role?

_____

7. Country of origin [This is the country where you were born.]

_____

8. Country where you have spent most of your life

_____

9. What is your first language? [This is the language you spoke at home and therefore you learnt it first. If you grew up speaking two or more languages, please enter all of them.]

_____

10. Which part of the UK are you from?

---

11. Would you say that the English you speak is related to a particular region in the UK?

---

12. In what ways do you use English in an academic setting?

---

13. Have you learned any other languages? Which?

---

14. Do you use any languages other than English? Which?

---

15. Have you ever taken an oral language exam similar to this interview (in English or another language)? If yes, can you give some details.

---

16. Have you ever taken a commercial English language test (e.g., IELTS, TOEFL, PTE Academic)? If yes, can you give some details (date, scores).

Appendix 9 – Sample Teaching Materials

Learning from assessment corpora:

Using real-life language to learn English

New vocabulary!

A corpus is a big database of naturally occurring language (**corpora** is the plural). This activity uses two corpora of scripts from speakers in the Graded Examinations in Spoken English (GESE). The TLC-L2 includes language from English learners while the TLC-L1 candidates are native English speakers. Looking at both, we can learn more about how to successfully communicate in English.

Register is the type of language that is associated with a particular situation or context like a language exam. Part of learning a language is also learning how to use different registers depending on the situation.

Collocations are combinations of words that often occur together in language and are especially helpful in speech for producing natural sounding language.

These activities will help you develop your use of register-appropriate collocations within an examination.

Activity

1. Read the two extracts from the TLC-L1 and TLC-L2 corpora and underline the collocations that use the verb *make.*
2. What do you notice about the way the speakers use these collocations?
3. Using #LancsBox X, search for *make sense* in the BNC2014.
4. What is different or the same about how *make sense* is used in the BNC2014 compared to the other two corpora?
5. Create your own sentence using *make sense* that could be used in a language exam like the GESE.

References

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*(2), 81–92. https://doi.org/10.1016/j.esp.2011.08.004

Aijmer, K. (2011). *Well I'm not sure I think…* The use of *well* by non-native speakers. *International Journal of Corpus Linguistics, 16*(2), 231-254.

Aijmer, K. (2014). Pragmatic markers. In K. Aijmer & C. Rühlemann (Eds.), *Corpus Pragmatics: A Handbook* (pp. 195–218). Cambridge University Press.

Alderson, J. C. (1996). Do corpora have a role in language assessment. In J. A. Thomas & M. H. Short (Eds.), *Using Corpora for Language Research* (pp. 248–259). Longman.

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task Effects on Linguistic Complexity and Accuracy: A Large-Scale Learner Corpus Analysis Employing Natural Language Processing Techniques. *Language Learning*, *67*(S1), 180–208. https://doi.org/10.1111/lang.12232

Allerton, D. J. (1984). Three (or four) levels of word co-occurrence restriction. *Lingua, 63*(1), 17–40.

Altenberg, B. & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics, 22*(2), 173–195.

Almela-Sánchez, M. (2019). Collocation and Selectional Preferences: A Frame-Based Approach. *Journal of English Studies, 17,* 3-41. http://doi.org/10.18172/jes.3905

Antle, J. B. (2013). Teaching collocations. *Jalt 2012 Conference Proceedings*, 346–354.

Ash, S. (2004). Social class. In K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The Handbook of Language Variation and Change*, (pp. 402– 422). Blackwell. https://doi.org/10.1002/9780470756591.ch16

Aston, G. , & Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press.

Babanoğlu, M. P. (2014). A Corpus-based Study on the Use of MAKE by Turkish EFL Learners. *International Journal of Education and Literacy Studies*, *2*(2), 43–47. https://doi.org/10.7575/aiac.ijels.v.2n.2p.43

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford University Press.

Bahns, J. (1993). Lexical collocations: A contrastive view. *ELT Journal*, *47*(1), 56–63. https://doi.org/10.1093/elt/47.1.56

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System*, *21*(1), 101–114. https://doi.org/10.1016/0346-251X(93)90010-E

Beaulieu, S. (2018). What is the target for L2 learners when descriptive, textbook, and subjective norms widely differ? *The Canadian Modern Language Review*, *74*(4), 548–574. https://doi.org/10.3138/cmlr.2017-0008

Belz, J. A., & Kinginger, C. (2003). Discourse Options and the Development of Pragmatic Competence by Classroom Learners of German: The Case of Address Forms. *Language Learning*, *53*(4), 591–647. https://doi.org/10.1046/j.1467-9922.2003.00238.x

Bestgen, Y. & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing, 26,* 28-41. https://doi.org/10.1016/j.jslw.2014.09.004

Bestgen, Y. & Granger, S. (2018). Tracking L2 writers' phraseological development using collgrams: Evidence from a longitudinal EFL corpus. In S. Hoffmann, A. Sand, S. Arndt-Lappe, & L. M. Dillmann (Eds.), *Corpora and Lexis* (pp. 277-301). Brill. https://brill.com/abstract/title/36203

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, *8*, 9–37.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* Longman.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405.

Biber, D. & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511814358

Biskup, D. (1992). L1 Influence on Learners' Renderings of English Collocations: A Polish/German Empirical Study. *Vocabulary and Applied Linguistics*, 85–93. https://doi.org/10.1007/978-1-349-12396-4_8

Boers, F., Eyckmans, J., Kappel, J., Stengers, H. & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: putting a Lexical Approach to the test. *Language Teaching Research, 10*(3), 245–261. https://doi.org/10.1191/1362168806lr195oa

Boers, F. & Webb, S. (2018). Teaching and learning collocation in adult second and foreign language learning. *Language Teaching 51*(1), 77-89. https://doi:10.1017/S0261444817000301

Boulton, A., & Cobb, T. (2017). Corpus Use in Language Learning: A Meta-Analysis. *Language Learning*, *67*(2), 348–393. https://doi.org/10.1111/lang.12224

Boyd, E., & Taylor, C. (2016). Presenting Validity Evidence: The Case of the GESE. In J. Banerjee & D. Tsagari (Eds.), *Contemporary Second Language Assessment :*

Contemporary Applied Linguistics, Volume 4 (pp. 37-60). Bloomsbury Academic. https://doi.org/10.5040/9781474295055.ch-002

Brazil, D. (1995). *A Grammar of Speech.* Oxford University Press.

Brezina, V. (2018). *Statistics in Corpus Linguistics. A practical guide.* Cambridge University Press.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139–173. https://doi.org/10.1075/ijcl.20.2.01bre

Brezina, V. & Fox, L. (2021). Adjective + Noun Collocations in L2 and L1 Speech: Evidence from the Trinity Lancaster Corpus and the Spoken BNC2014. In: S. Granger (Ed.), *Perspectives on the L2 Phrasicon.* (pp. 152-177). Multilingual Matters. https://doi.org/10.21832/9781788924863-008

Brezina, V., Hawtin, A. & McEnery, T. (2021). The Written British National Corpus 2014 – design and comparability. *Text & Talk, 41*(5-6), 595-615. https://doi.org/10.1515/text-2020-0052

Brezina, V. & Platt, W. (2023). *#LancsBox X* [software], Lancaster University, http://lancsbox.lancs.ac.uk

Burgos, E. G. (2018). Occurrences of formulaic sequences in personal descriptions. *Onomazein, 40*, 103–118. https://doi.org/10.7764/onomazein.40.06

Buttery, P., & Caines, A. (2012). Normalising frequency counts to account for "opportunity of use" in learner corpora. In Y. Tono (Ed.), *Developmental and Crosslinguistic Perspectives in Learner Corpus Research* (pp. 187–204). John Benjamins. https://doi.org/10.1075/tufs.4.16but

Buysse, L. (2017). The pragmatic marker *you know* in learner Englishes. *Journal of Pragmatics, 121,* 40-57. http://dx.doi.org/10.1016/j.pragma.2017.09.010

Caines, A., & Buttery, P. (2018). The Effect of Task and Topic on Opportunity of Use in Learner Corpora. In V. Brezina & L. Flowerdew (Eds.), *Learner Corpus Research : New Perspectives and Applications* (pp. 5–27). Bloomsbury Academic. https://doi.org/10.5040/9781474272919.0007

Callies, M., & Götz, S. (2015). *Learner Corpora in Language Testing and Assessment*. John Benjamins. http://benjamins.com/catalog/books/scl

Cao, D. & Badger, R. (2021). Cross-linguistic influence on the use of L2 collocations: the case of Vietnamese learners. *Applied Linguistics Review, 14*(3), 421-446. https://doi.org/10.1515/applirev-2020-0035

Carlsen, C. (2012). Proficiency level – A fuzzy variable in computer learner corpora. *Applied Linguistics, 33*(2), 161–183. https://doi.org/10.1093/applin/amr047

Carter, R., & McCarthy, M. (2004). Talking, Creating: Interactional Language, Creativity, and Context. *Applied Linguistics*, *25*(1), 62–88. https://doi.org/10.1093/applin/25.1.62m

Castello, E., & Gesuato, S. (2019). Holding up one's end of the conversation in spoken English Lexical backchannels in L2 examination discourse. *International Journal of Learner Corpus Research*, *5*(2), 231–252. https://doi.org/10.1075/ijlcr.17020.cas

Chen, Y.-H., & Baker, P. (2010). Lexical Bundles in L1 and L2 Academic Writing. *Language Learning and Technology*, *14*(2), 30–49. http://llt.msu.edu/vol14num2/chenbaker.pdf

Chen, Y. H., & Baker, P. (2016). Investigating Criterial Discourse Features across Second Language Development: Lexical Bundles in Rated Learner Essays, CEFR B1, B2 and C1. *Applied Linguistics*, *37*(6), 849–880. https://doi.org/10.1093/applin/amu065

Chi, A. M. L., Wong, K. & Wong, M. (1994). Collocational problems amongst ESL learners: a corpus-based study. In L. Flowerdew & A. Tong (Eds): *Entering Text* (pp. 157-165). Hong Kong University of Science and Technology. https://hdl.handle.net/1783.1/1088

Choi, W. (2019). A Corpus-Based Study on "Delexical Verb + Noun" Collocations Made by Korean Learners of English. *The Journal of Asia TEFL, 16*(1), 279-293. http://dx.doi.org/10.18823/asiatefl.2019.16.1.18.279

Cobb, T. (2003). Analyzing late interlanguage with learner corpora: Qu´ebec replications of three European studies. *Canadian Modern Language Review/La Revue canadienne des langues vivantes, 59(*3), 393–423.

Cobb, T. (2018). From corpus to CALL: The use of technology in teaching and learning formulaic language. In A. Siyanova-Chanturia & A. Pellicer-Sanchez (Eds.), *Understanding Formulaic Language: A Second Language Acquisition Perspective* (pp. 192-210). Routledge.

Conklin, K. & Schmitt, N. (2008). Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers? *Applied Linguistics*, *29*(1), 72–89. https://doi.org/10.1093/applin/amm022

Cook, V. J. (2007). The goals of ELT: Reproducing native-speakers or promoting multicompetence among second language users? In J. Cummins and C. Davison (Eds.), *International Handbook of English Language Teaching* (pp. 237-248). Springer.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*, 397-423. https://doi:10.1016/j.esp.2003.12.001

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Cowie, A. P. (1994). Phraseology. In R.E. Asher (Ed.) *The Encyclopedia of Language and Linguistics* (pp. 3168–71). Pergamon Press.

Cowie, A. P. (1998). *Phraseology: Theory, Analysis and Applications.* Oxford University Press.

Cross, J. & Papp, S. (2008). Creativity in the use of verb + noun combinations by Chinese learners of English. In G. Gilquin, S. Papp and M. Belén Díez-Bedmar (Eds.), *Linking Up Contrastive and Learner Corpus Research* (pp. 57-81). Rodopi. https://doi.org/10.1163/9789401206204_004

Crossley, S. A., Salsbury, T. & McNamara, D. S. (2010). The development of word co-referentiality in second language speakers: A case for Latent Semantic Analysis. *Vigo International Journal of Applied Linguistics, 7, 55*–74.

Crossley, S. A., Salsbury, T., McNamara, D. S. & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing 28*(4), 561–580. https://doi.org/10.1177/0265532210378031

Culpeper, J. V., & Gillings, M. (2018). Politeness Variation in England: A North-South Divide? In V. Brezina, R. Love, & K. Aijmer (Eds.*), Corpus Approaches to Contemporary British Speech: Sociolinguistic Studies of the Spoken BNC2014* (pp. 33-59). (Routledge Advances in Corpus Linguistics). Routledge.

Curry, N., Love, R., & Goodman, O. (2022). Adverbs on the move: Investigating publisher application of corpus research on recent language change to ELT coursebook development. *Corpora*, *17*(1), 1-38. https://doi.org/10.3366/cor.2022.0233.

Cushing, S. T. (2021). Corpus linguistics and language testing. In G. Fulcher & L. Harding (Eds.), *The Routledge Handbook of Language Testing* (pp. 545–560). Taylor and Francis. https://doi.org/10.4324/9781003220756-42

Daskalovska, N. (2015). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning*, *28*(2), 130–144. https://doi.org/10.1080/09588221.2013.803982

De Bot, K., Lowie, W., & Verspoor, M. (2007). A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition*, *10*(1), 7–21. https://doi.org/10.1017/S1366728906002732

De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on Computer* (pp. 64-76). Routledge.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL), New Series*, 2, 225-246.

Dechert, H. (1983). How a story is done in a second language. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 175-196). Longman.

Deshors, S. C., Götz, S., & Laporte, S. (2016). Linguistic innovations in EFL and ESL. *International Journal of Learner Corpus Research*, *2*(2), 131–150. https://doi.org/10.1075/ijlcr.2.2.01des

Díaz-Negrillo, A. (2012). Learner corpora: the case of the NOSE corpus. *Systemics, Cybernetics and Informatics, 10*(1), 42–48.

Du, X., Afzaal, M., & Al Fadda, H. (2022). Collocation Use in EFL Learners' Writing Across Multiple Language Proficiencies: A Corpus-Driven Study. *Frontiers in Psychology*, *13*, 1–10. https://doi.org/10.3389/fpsyg.2022.752134

Duan, S., & Shi, Z. (2021). A longitudinal study of formulaic sequence use in second language writing: Complex dynamic systems perspective. *Language Teaching Research*, 1-34. https://doi.org/10.1177/13621688211002942

Durrant, P. (2008). *High frequency collocations and second language learning*. [Unpublished doctoral thesis]. University of Nottingham, UK.

Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations - a meta-analysis. *International Journal of Corpus Linguistics*, *19*(4), 443–477. https://doi.org/10.1075/ijcl.19.4.01dur

Durrant, P. & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching,* 47(2), 157–177. https://doi.org/10.1515/iral.2009.007

Durrant, P., & Mathews-Aydinli, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, *30*(1), 58-72. https://doi.org/10.1016/j.esp.2010.05.002

Egbert, J. (2017). Corpus linguistics and language testing: Navigating uncharted waters. *Language Testing*, *34*(4), 555–564. https://doi.org/10.1177/0265532217713045

Eguchi, M., & Kyle, K. (2020). Continuing to Explore the Multidimensional Nature of Lexical Sophistication: The Case of Oral Proficiency Interviews. *Modern Language Journal*, *104*(2), 381–400. https://doi.org/10.1111/modl.12637

Ellis, N. C. (1996). Sequencing in SLA: Phonological Memory, Chunking and Points of Order. *Studies in Second Language Acquisition, 18,* 91-126.

Ellis, N. C., & Sinclair, S. (1996). Working Memory in the Acquisition of Vocabulary and Syntax: Putting Language in Good Order. *Quarterly Journal of Experimental Psychology, 49*,(1), 234-250. http://dx.doi.org/10.1080/713755604

Ellis, N., Simpson-Vlach, R. & Maynard, C. (2008). Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL. *TESOL Quarterly*, *42*(3), 375–396.

Elder, C., McNamara, T., Kim, H., Pill, J., & Sato, T. (2017). Interrogating the construct of communicative competence in language assessment contexts: What the non-language specialist can tell us. *Language and Communication*, *57*, 14–21. https://doi.org/10.1016/j.langcom.2016.12.005

Erman, B. & Warren, B. (2000). The idiom principle and the open-choice principle. *Text & Talk, 20*(1), 29–62.

Erman, B., Denke, A., Fant, L., & Forsberg Lundell, F. (2015). Nativelike expression in the speech of long-residency L2 users: A study of multiword structures in L2 English, French and Spanish. *International Journal of Applied Linguistics*, *25*(2), 160–182. https://doi.org/10.1111/ijal.12061

Evert, S. (2005). *The statistics of word co-occurrences: Word pairs and collocations.* [Unpublished doctoral dissertation]. University of Stuttgart, Germany.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling and M. Kytö (Eds.), *Corpus Linguistics. An International Handbook* (pp. 1212-1248). Mouton de Gruyte.

Farghal, M., & Obiedat, H. (1995). Collocations: a neglected variable in EFL. *International Review of Applied Linguistics in Language Teaching*, 33(4), 315–332.

Firth, J. R. (1957). *Papers in Linguistics 1934-1951*. Oxford University Press.

Forsberg, F., & Bartning, I. (2010). Can linguistic features discriminate between the communicative CEFR-levels? A pilot study of written L2 French. In M. Martin & I. Vedder (Eds.), *Communicative proficiency and linguistic development. Intersections between SLA and language testing research* (pp. 133–157). EUROSLA Monograph Series.

Forsberg, F., & Fant, L. (2012). Idiomatically Speaking: Effects of Task Variation on Formulaic Language in Highly Proficient Users of L2 French and Spanish 1. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 47–70). Bloomsbury.

Fox, J. (2004). Biasing for the Best in Language Testing and Learning: An Interview With Merrill Swain. *Language Assessment Quarterly*, *1*(4), 235–251. https://doi.org/10.1207/s15434311laq0104_3

Frankenberg-Garcia, A., Lew, R., Rees, G. et al. (2019). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, *31*(10), 23–39. https://doi.org/10.1017/S0958344018000150.

Fritz, E., Dormer, R., Sumi, S., & Kudo, T. (2022). The acquisition of formulaic sequences in EFL email writing. *English for Specific Purposes*, *65*, 15–29. https://doi.org/10.1016/j.esp.2021.08.003

Gablasova, D. (2020). Corpora for Second Language Assessments. In P. Winke & T. Brunfaut (Eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing* (pp. 45–53). Taylor and Francis. https://corpus.mml.cam.ac.uk/efcamd

Gablasova, D., & Brezina, V. (2015). Does speaker role affect the choice of epistemic adverbials in L2 speech? Evidence from the Trinity Lancaster Corpus. In J. Romero-Trillo (Ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015* (pp. 117-136). Springer. https://doi.org/10.1007/978-3-319-17948-3_6

Gablasova, D., Brezina, V., & McEnery, T. (2017a). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing and Interpreting the Evidence. *Language Learning, 67*(1), 155-179. https://doi.org/10.1111/lang.12225

Gablasova, D., Brezina, V., McEnery, T., & Boyd, E. (2017b). Epistemic Stance in Spoken L2 English: The Effect of Task and Speaker Style. *Applied Linguistics, 38*(5), 613–637. https://doi.org/10.1093/applin/amv055

Gablasova, D., Brezina, V., & McEnery, T. (2019). The Trinity Lancaster Corpus Development, description and application. *International Journal of Learner Corpus Research, 5*(2), 126-158. https://doi.org/10.1075/ijlcr.19001.gab

Galaczi, E., & Taylor, L. (2018). Interactional Competence: Conceptualisations, Operationalisations, and Outstanding Questions. *Language Assessment Quarterly, 15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

Garner, J. (2020). The cross-sectional development of verb-noun collocations as constructions in L2 writing. *International Review of Applied Linguistics in Language Teaching*, *60*(3), 909-935. https://doi.org/10.1515/iral-2019-0169

Gass, S. M., Selinker, L., & Plonsky, L. (2013). *Second language acquisition: An introductory course* (4th ed.). Taylor & Francis Group.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCAMDAT). *Proceedings of the 31st Second Language Research Forum.* Cascadilla Proceedings Project.

Gilquin, G. (2007). To err is not all: What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift Fur Anglistik Und Amerikanistik*, *55*(3), 273–291. https://doi.org/10.1515/zaa.2007.55.3.273

Gilquin, G. (2022). One norm to rule them all? Corpus-derived norms in learner corpus research and foreign language teaching. *Language Teaching, 55*(1), 87-99. https://doi:10.1017/S0261444821000094

Gilquin, G., De Cock, S. & Granger, S. (2010). *Louvain International Database of Spoken Learner Interlanguage: Handbook and CD-ROM*. Presses universitaires de Louvain.

Gilquin, G., & Paquot. M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction, 1*. 41–61

Gitsaki, C. (1999). *Second Language Lexical Acquisition: A Study of the Development of Collocational Knowledge.* International Scholars Publications.

González Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have? *International Journal of Applied Linguistics*, *166*(1), 94–126. https://doi.org/10.1075/itl.166.1.03fer

Götz, S. (2019). Filled pauses across proficiency levels, L1s and learning context variables: A multivariate exploration of the Trinity Lancaster Corpus Sample. *International Journal of Learner Corpus Research*, *5*(2), 159–180. https://doi.org/10.1075/ijlcr.17018.got

Goulart, L., Gray, B., Staples, S., Black, A., Shelton, A., Biber, D., Egbert, J., & Wizner, S. (2019). Linguistic Perspectives on Register. *Annual Review of Linguistics*, *6*(7), 7.1-7.21. https://doi.org/10.1146/annurev-linguistics-011718-012644

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: collocations and formulae. In A. Cowie (Ed.), *Phraseology: theory, analysis and applications* (pp. 145–160). Oxford University Press.

Granger, S. (2018). Formulaic sequences in learner corpora: Collocations and lexical bundles. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding Formulaic Language: A Second Language Acquisition Perspective* (pp. 228-247). Routledge.

Granger, S. (2020). Commentary: Have Learner Corpus Research and Second Language Acquisition Finally Met? In B. Le Bruyn & M. Paquot (Eds.), *Learner Corpus Research Meets Second Language Acquisition* (pp. 243–257). Cambridge University Press. https://doi.org/10.1017/9781108674577.012

Granger, S., Dagneaux, E., & Meunier, F. (2002). *International Corpus of Learner English.* UCL Presses Universitaires de Louvain

Granger, S., & Thewissen, J. (2005). The contribution of error-tagged learner corpora to the assessment of language proficiency. Evidence from the International Corpus of Learner English. *27th Language Testing Research Colloquium.* http://hdl.handle.net/2078.1/75894

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In: S. Granger, F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 27–49). John Benjamins.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching, 52*(3) 229-252. https://doi.ord/10.1515/iral-2014-0011

Granger, S., & Lefer, M-A. (2020). Introduction: A two-pronged approach to corpus-based crosslinguistic studies. In S. Granger & M-A. Lefer, *The Complementary Contribution of Comparable and Parallel Corpora to Crosslinguistic Studies – Special issue of Languages in Contrast 20:2* (pp. 167-183). John Benjamins. https://doi.org/10.1075/lic.20.2

Gries, S. Th. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 3–25). John Benjamins. http://doi:10.1075/z.139

Gries, S. Th. (2013). 50-something years of work on collocations. *International Journal of Corpus Linguistics*, *18*(1), 137–166. https://doi.org/10.1075/ijcl.18.1.09gri

Gyllstad, H. (2007). *Testing English Collocations. Developing Receptive Tests for Use with Advanced Swedish Learners*. [Doctoral Thesis (monograph), English Studies] Lunds universitet. http://lup.lub.lu.se/record/599011/file/2172422.pdf

Gyllstad, H., & Wolter, B. (2016). Collocational Processing in Light of the Phraseological Continuum Model: Does Semantic Transparency Matter? *Language Learning*, *66*(2), 296–323. https://doi.org/10.1111/lang.12143

Gyllstad, H., & Schmitt, N. (2018). Testing Formulaic Language. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding Formulaic Language: A Second Language Acquisition Perspective* (pp. 174–191). Routledge.

Ha, L. Q., Hanna, P., Ming, J., & Smith, F. J. (2009). Extending Zipf's law to n-grams for large corpora. *Artificial Intelligence Review*, *32*(1–4), 101–113. https://doi.org/10.1007/s10462-009-9135-4

Halim, H. A., & Kuiper, K. (2018). Individual differences in the acquisition of restricted collocations. *International Journal of Education, Psychology and Counseling*, *3*(16), 36–49. https://doi.org/10.4324/9780203075630-13

Halliday, M. A. K. (1966). Lexis as a linguistic level. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, & R. H. Robins (Eds.), *In memory of J. R. Firth* (pp. 148–162). Longman.

Harding, L., Macqueen, S., & Pill, J. (2023). Assessing communicative competence. In M. Kanwit & M. Solon (Eds.), *Communicative Competence in a Second Language: Theory, Method and Applications* (pp. 187–207). Routledge. https://doi.org/10.4324/9781003160779-14

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics, 4*(2), 237–258.

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing, 9*(1), 122–159. https://doi.org/10.1016/j.asw.2004.06.001

Hill, J. (2000). Revising priorities: from grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp. 47–69). Language Teaching Publications.

Hinkel, E. (2005). Hedging, inflating, and persuading in L2 academic writing. *Applied Language Learning (15)*1, 29–53.

Hinkel, E. (2009). The effects of essay topics on modal verb uses in L1 and L2 academic writing. *Journal of Pragmatics, 41*(4), 667–683. https://doi.org/10.1016/j.pragma.2008.09.029

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics, 19*(1), 24–44.

Hsu, J. (2007). Lexical Collocations and their Relation to the Online Writing of Taiwanese College English Majors and Non-English Majors. *Electronic Journal of Foreign Language Teaching*, *4*(2), 192–209.

Hu, R. (2016). The Age Factor in Second Language Learning. *Theory and Practice in Language Studies*, *6*(11), 2164-2168. https://doi.org/10.17507/tpls.0611.13

Huang, P. (2023). A frequency coverage, ad dispersion analysis of the academic collocation list in university student writing. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2023-0129

Huang, H. T. D., Hung, S. T. A., & Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing*, *35*(1), 27–49. https://doi.org/10.1177/0265532216677106

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4-21. https://doi.org/10.1016/j.esp.2007.06.001

Hyland, K. (2012). Genre and Discourse Analysis in Language for Specific Purposes. In C. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Wiley-Blackwell. https://doi.org/10.1002/9781405198431.wbeal0452

Jackendoff, R. (1997). *The architecture of the language faculty.* MIT Press.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen Corpus Family. *7th International Corpus Linguistics Conference*, 125–127.

Jarvis, S. (2000). Methodological rigor in the study of transfer: identifying L1 influence in the interlanguage lexicon. *Language Learning 50*(2), 245–309. https://doi.org/10.1111/0023-8333.00118

Johns, T. (1991). From print out to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Johns & P. King (Eds.*), ELR Journal 4: Classroom concordancing* (pp. 27–46). University of Birmingham.

Jones, C., Byrne, S., & Halenko, N. (2017). Conclusion. In C. Jones, S. Byrne, & N. Halenko (Eds.), *Successful spoken English: Findings from learner corpora* (pp. 159–172).

Juknevičienė, R. (2008). Collocations with high-frequency verbs in learner English: Lithuanian learners vs native speakers. *KALBOTYRA, 59*(3), 119-127.

Kamarudin, R., Abdullah, S. & Abdul Aziz, R. (2020). Examining ESL Learners' Knowledge of Collocations. *International Journal of Applied Linguistics & English Literature, 9*(1), 1-6.

Keshavarz, M. H. & Salimi, H. (2007). Collocational competence and cloze test performance: a study of Iranian EFL learners. *International Journal of Applied Linguistics, (17)*1, 81-92.

Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing*, *34*(1), 23–48. https://doi.org/10.1177/0265532215595666

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, M., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography, 1*(1), 7–36. https://doi:10.1007/s40607-014-0009-9

Kim, M.-H. (2002). Uses of Make in Korean EFL Learner Writing: A Corpus-Based Study. *English Teaching, 57*(4), 297-314.

Kreyer, R. (2021). Collocations in learner English: A true-longitudinal perspective. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond Concordance Lines: Corpora in language education* (pp. 97–120). https://doi.org/10.1075/scl.102.05kre

Kurtes, S. & Saville, N. (2008). The English Profile Programme – An overview. *Research Notes, 33*: 2–4. Cambridge: Cambridge ESOL.

Kwon, M. H., Staples, S., & Partridge, R. S. (2018). Source work in the first-year L2 writing classroom: Undergraduate L2 writers' use of reporting verbs. *Journal of English for Academic Purposes*, *34*, 86–96. https://doi.org/10.1016/j.jeap.2018.04.001

Kyle, K. & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly (49)*4, 757-786. https://doi.org/10.1002/tesq.194

Labov, W. (1966). *The Social Stratification of English in New York City*. Center for Applied Linguistics.

Lakoff, R. (1975). *Language and Woman's Place*. Harper and Row.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics, 27*(4), 590-619.

Larsen-Freeman, D. (2012). Complex, dynamic systems: A new transdisciplinary theme for applied linguistics? *Language Teaching, 45*, 202–214.

Laufer, B. & Waldman, T. (2011). Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. *Language Learning, 61*(2), 647-672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing*, *13*(2), 151–172. https://doi.org/10.1177/026553229601300202

Lazaraton, A. (2012). Discourse Analysis in Language Assessment. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics,* (pp. 1-5). Blackwell Publishing Ltd. https://doi:10.1002/9781405198431

Leblanc, C., & Fujieda, M. (2012). Investigating Effects of Topic Control on Lexical Variation in Japanese University Students' In-class Timed-writing. *Kwansei Gakuin University Humanities Review*, *17*, 241–253.

Lee, S. (2019). L1 transfer, proficiency, and the recognition of L2 verb-noun collocations: A perspective from three languages. *International Review of Applied Linguistics in Language Teaching*, *59*(2), 1–28. https://doi.org/10.1515/iral-2018-0220

Lee, S. (2021). L1 transfer, proficiency, and the recognition of L2 verb-noun collocations: A perspective from three languages. *International Review of Applied Linguistics in Language Teaching*, *59*(2), 181–208. https://doi.org/10.1515/iral-2018-0220

Lee, S., & Na, Y.-H. (2015). The Use of the Verb Make in the Corpora of American NS Students and Korean EFL Learners. *English21*, *28*(1), 401–426. https://doi.org/10.35771/engdoi.2015.28.1.019

Lee, S., & Shin, S. Y. (2021). Towards Improved Assessment of L2 Collocation Knowledge. *Language Assessment Quarterly*, *18*(4), 419–445. https://doi.org/10.1080/15434303.2021.1908295

Leech, G. (2000). Grammars of Spoken English: New Outcomes of Corpus-Oriented Research, *Language Learning*, *50*(4), 675-724.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English* (1st ed.). Routledge.

Lei, L. & Liu, D. (2018). The academic English collocation list. *International Journal of Corpus Linguistics*, *23*(2), 216–243. https://doi.org/10.1075/ijcl.16135.lei

Leung, C. (2022). Language proficiency: from description to prescription and back? *Educational Linguistics*, *1*(1), 56–81. https://doi.org/10.1515/eduling-2021-0006

Lewis, M. (2008). *The idiom principle in L2 English* [Unpublished doctoral dissertation]. Stockholm University, Sweden.

Li, J. & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing*, *18*(2), 85–102. https://doi.org/10.1016/j.jslw.2009.02.001

Liddicoat, A. J., & Curnow, T. J. (2004). Language Descriptions. In A. Davies & C. Elder (Eds.), *The Handbook of Applied Linguistics* (pp. 25-53). Blackwell Publishing Ltd. https://doi.org/10.1002/9780470757000.ch1

Liesenfeld, A., & Dingemanse, M. (2022). Building and curating conversational corpora for diversity-aware language science and technology. *Proceedings of the 13th Conference on Language Resources and Evaluation*, 1178–1192.

Lin, C. H., & Lin, Y. L. (2019). Grammatical and Lexical Patterning of Make in Asian Learner Writing: A Corpus-Based Study of ICNALE. *3L: The Southeast Asian Journal of English Language Studies*, *25*(3), 1–15. https://doi.org/10.17576/3L-2019-2503-0

Lin, P. (2018). Formulaic language and speech prosody. In A. Siyanova-Chanturia & A. Pellicer-Sanchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 78-94). Routledge.

Lin, P. (2022). Developing an intelligent tool for computer-assisted formulaic language learning from YouTube videos. *ReCALL, 34*(2), 185–200. https://doi.org/10.1017/S0958344021000252

Liu, T. (2021). Data-driven learning: Using #LancsBox in academic collocation learning. In P. Pérez-Paredes & G. Mark (Eds.), *Beyond Concordance Lines: Corpora in language education* (pp. 177–206). https://doi.org/10.1075/scl.102.08liu

Lorenz, G. (1999). *Adjective Intensification–Learners Versus Native Speakers: A Corpus Study of Argumentative Writing*. Rodopi.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22*(3), 319–344.

Lowie, W., & Verspoor, M. (2015). Variability and Variation in Second Language Acquisition Orders: A Dynamic Reevaluation. *Language Learning*, *65*(1), 63–88. https://doi.org/10.1111/lang.12093

Luzón-Marco, M. J. (2011). Exploring atypical verb+noun combinations in learner technical writing. *International Journal of English Studies*, *11*(2), 77–95. https://doi.org/10.6018/ijes/2011/2/149651

Ma, J. H., & Kim, Y. (2013). Korean High School English Learners' Knowledge of Collocations: Focusing on Delexical Verbs: make, get, and take. *Language Research*, *49*(1), 45–71.

Macmillan Collocation Dictionary. (2023a, June 13). *Money*. https://www.macmillandictionary.com/dictionary/british/money

Macmillan Collocation Dictionary. (2023b, June 13). *Information.* https://www.macmillandictionary.com/dictionary/british/information

Macmillan Collocation Dictionary. (2023c, June 13). *Advantage*. https://www.macmillandictionary.com/dictionary/british/advantage

Magogwe, J. M., & Oliver, R. (2007). The relationship between language learning stratifies, proficiency, age and self-efficacy beliefs: A study of language learners in Botswana. *System, 35*(3), 338-352. https://doi.org/10.1016/j.system.2007.01.003

Marín Cervantes, I. (2019). *Second language speakers' use of multi-word verbs in spoken communication: evidence from the Trinity Lancaster corpus.* [Doctoral Thesis]. Lancaster University. https://doi.org/10.17635/lancaster/thesis/668

Martinez, R., & Schmitt, N. (2012). Phrasal expressions list. *Applied Linguistics*, *33*(3), 299–320. https://doi.org/10.1093/applin/ams010

Mauranen, A. (2004). Spoken corpus for an ordinary learner. In J. M. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 89–105). John Benjamins.

McEnery, A., & Brookes, G. (2021). Corpus Linguistics Across the Generations: Remembering Geoffrey Leech. *Text and Talk, 41*(5-6), 589-593. https://doi.org/10.1515/text-2021-0135

McEnery, T., & Hardie, A. (2012). *Corpus linguistics method, theory and practice*. Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics: An Introduction* (2nd ed.). Edinburgh University Press.

McNamara, T. (1997). "Interaction" in second language performance assessment: Whose performance? *Applied Linguistics*, *18*(4), 446–466.

McNamara, T. (2001). Language assessment as social practice: Challenges for research. *Language Testing*, *18*(4), 333–349. https://doi.org/10.1177/026553220101800402

Moon, R. (1998). *Fixed expressions and idioms in English: a corpus-based approach*. Clarendon Press; Oxford University Press.

Myles, F., & Cordier, C. (2017). Formulaic sequence(fs) cannot be an umbrella term in SLA: Focusing on psycholinguistic FSs and their identification. *Studies in Second Language Acquisition, 39*(1), 3–28. https://doi.org/10.1017/S027226311600036X

Namvar, F. (2012). The relationship between language proficiency and use of collocation by Iranian EFL students. *3L: The Southeast Asian Journal of English Language Studies, 18(*3), 41-52.

Neary-Sundquist, C. (2013). Task Type Effects on Pragmatic Marker Use by Learners at Varying Proficiency Levels. *L2 Journal, 5,* 1-21. https://doi.org/10.5070/L25212104

Neshkovska, S. (2019). What do advanced ESL/EFL students' need to know to overcome "collocational" hurdles? *Thesis – Kolegii AAB*, *7*(2), 53–74.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. John Benjamins. https://doi:10.1075/scl.14

Nguyen, T. M. H., & Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research*, *21*(3), 298–320. https://doi.org/10.1177/1362168816639619

Nizonkiza, D. (2012). Quantifying controlled productive knowledge of collocations across proficiency and word frequency levels. *Studies in Second Language Learning and Teaching, 2(*1), 67–92. https://doi.org/10.14746/ssllt.2012.2.1.4

Nizonkiza, D. (2017). Predictive power of controled productive knowledge of collocations over L2 proficiency. In J. Evers-Vermeul & E. Tribushinina (Eds.), *Usage-Based Approaches*

to Language Acquisition and Language Teaching (pp. 263–286). De Gruyter Mouton. https://doi.org/10.1515/9781501505492-012

Office for National Statistics. (2010). *ONS Occupation Coding Tool.* https://onsdigital.github.io/dp-classification-tools/standard-occupational-classification/ONS_SOC_occupation_coding_tool.html

Ohlrogge, A. (2009). Formulaic expressions in intermediate EFL writing assessment. In R. Corrigan, E. A. Moravcsik, H. Ouali and K. Wheatley (Eds.), *Formulaic Language, Volume 2: Acquisition, Loss, Psychological Reality, and Functional Explanations* (pp. 387–404). John Benjamins.

Omidian, T., Siyanova-Chanturia, A., & Spina, S. (2021). Development of Formulaic Knowledge in Learner Writing: A Longitudinal Perspective. In: S. Granger (Ed.), *Perspectives on the L2 Phrasicon.* (pp. 178-205). Multilingual Matters. https://doi.org/10.21832/9781788924863-009

Papageorgiou, S. (2007). *Relating the Trinity College London GESE and ISE examinations to the Common European Framework of Reference Final Project Report, February 2007*. Trinity College London. https://www.trinitycollege.com/resource/?id=2261

Paquot, M. (2014). Cross-linguistic influence and formulaic language. *EUROSLA Yearbook*, *14*, 240–261. https://doi.org/10.1075/eurosla.14.10paq

Paquot, M. (2018). Phraseological competence: A missing component in university entrance language tests? Insights from a study of EFL learners' use of statistical collocations. *Language Assessment Quarterly*, *15*(1), 1–15.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145. https://doi.org/10.1177/0267658317694221

Paquot, M., & Granger, S. (2012). Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics*, *32*, 130–149. https://doi.org/10.1017/S0267190512000098

Paquot, M., Naets, H., & Gries, S. Th. (2021). Using syntactic co-occurrences to trace phraseological complexity development in learner writing: verb + object structures in LONGDALE. In B. Le Bruyn & M. Paquot (Eds.), *Learner Corpus Research Meets Second Language Acquisition* (pp. 122–147). Cambridge University Press. https://doi:10.1017/9781108674577.007

Paquot, M., Gablasova, D., Brezina, V., & Naets, H. (2022). Phraseological complexity in EFL learners' spoken production across proficiency levels. In A. Leńko-Szymańska & S. Götz (Eds.), *Complexity, Accuracy and Fluency in Learner Corpus Research* (pp. 115–136). John Benjamins. https://doi.org/10.1075/scl.104.05paq

Park, K. (2014). Corpora and language assessment: The state of the art. *Language Assessment Quarterly*, *11*(1), 27–44. https://doi.org/10.1080/15434303.2013.872647

Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and

nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–226). Longman.

Pennycook, A. (2018). Towards a critical pedagogy for teaching English as a worldly language. In *The Cultural Politics of English as an International Language* (pp. 295–327). Routledge. https://doi.org/10.4324/9781315843605-9

Pérez-Paredes, P. (2022). How Learners Use Corpora. In R. R. Jablonkai & E. Csomay (Eds.), *The Routledge Handbook of Corpora and English Language Teaching and Learning* (pp. 390–405). Taylor & Francis. https://doi.org/10.4324/9781003002901-31

Pérez-Paredes, P., & Díez-Bedmar, M. B. (2019). Certainty adverbs in spoken learner language: The role of tasks and proficiency. *International Journal of Learner Corpus Research*, *5*(2), 253–279. https://doi.org/10.1075/ijlcr.17019.per

Pérez-Paredes, P., & Mark, G. (2021). What can corpora tell us about language learning? In A. O'Keeffe & M. J. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (2nd ed., pp. 313–327). Routledge. https://doi.org/10.4324/9780367076399-22

Plough, I., Banerjee, J., & Iwashita, N. (2018). Interactional competence: Genie out of the bottle. *Language Testing, 35*(3), 427–445. https://doi.org/10.1177/0265532218772325

Raatz, U., & Klein-Braley, C. (1982). The C-test—A modification of the cloze procedure. In T.Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.), *University of Essex Occasional Papers: Vol. 4. Practice and problems in language testing* (pp. 113–138). Colchester, UK: Department of Language and Linguistics, University of Essex.

Raupach, M. (1984). Formulae in second language speech production. In H. W. Dechert, D. Mole, & M. Raupach (Eds.), *Second language productions* (pp. 114– 137). Gunter Narr Verlag.

Rayson, P., Archer, D., Piao, S. & McEnery, A. M. (2004). *The UCREL Semantic Analysis System*. Lancaster University.

Read, J. (2022). Test Review: The International English Language Testing System (IELTS). *Language Testing, 39*(4), 679-694. https://doi.org/10.1177/02655322221086211

Reppen, R. (2010). *Using corpora in the language classroom.* Cambridge University Press.

Revier, R. L. (2009). Evaluating a new test of whole English collocations. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp.125–138). Palgrave Macmillan.

Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. *Language Testing, 34(*4), 477–492. https://doi.org/10.1177/0265532217711431

Römer, U., & Garner, J. (2019). The development of verb constructions in spoken learner English: Tracing effects of usage and proficiency. *International Journal of Learner Corpus Research, 5*(2), 207-230. https://doi.org/10.1075/ijlcr.17015.rom

Rosas-Maldonado, M. (2017). Use of communication strategies in an interactional context: The interlocutor influence. *Poznan Studies in Contemporary Linguistics*, 53(4) 563–592. https://doi.org/10.1515/psicl-2017-0021

Saito, K. & Liu, Y. (2022). Roles of collocation in L2 oral proficiency revisited: Different tasks, L1 vs. L2 raters, and cross-sectional vs. longitudinal analyses. *Second Language Research, 38*(3), 531-554. https://doi.org/10.1177/0267658320988055

Saito. K. (2020). Multi- or Single-Word Units? The Role of Collocation Use in Comprehensible and Contextually Appropriate Second Language Speech. *Language Learning*, *70*, 548-588. https://doi.org/10.1111/lang.12387

Salsbury, T., & Bardovi-Harlig, K. (2000). Oppositional talk and the acquisition of modality in L2 English. In B. Swierzbin, F. Morris, M. E. Anderson, C. A. Klee, & E. Tarone (Eds.), *Social and cognitive factors in second language acquisition: Selected proceedings of the 1999 second language research forum* (pp. 57-76). Cascadilla Press.

Sawaguchi, R. & Mizumoto, A. (2022). Exploring the use of make noun collocations by Japanese EFL learners through a bilingual essay corpus. *Corpora, 17*(Supplement), 61–77. https://doi.org/10.3366/cor.2022.0247

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, processing and use* (pp. 127-151). John Benjamins.

Schneider, K. P., & Barron, A. (Eds.) (2008). *Variational pragmatics: A focus on regional varieties in pluricentric languages*. John Benjamins.

Seargeant, P. (2013). Ideologies of nativism and linguistic globalization. In S. A. Houghton and D. J. Rivers (Eds.), *Native-Speakerism in Japan: Intergroup Dynamics in Foreign Language Education* (pp. 231–242). Multilingual Matters.

Shih, H.-H. R. (2000). Collocation deficiency in a learner corpus of English: From an overuse perspective. *Proceedings of the 14th Pacific Asia Conference on Language, Information and Computation*, 281–288.

Shin, D. & Nation, P. (2008). Beyond single words: The most frequent collocations in spoken English. *ELT Journal 62*(4), 339-348.

Sinclair, J. M. (1990). *Collins Cobuild English grammar.* Harper Collins.

Sinclair, J. M. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review, 64*(3), 429–458. https://doi.org/10.3138/cmlr.64.3.429

Siyanova-Chanturia, A. & Spina, S. (2020). Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study. *Language Learning 70*(2), 420–463. https://doi.org/10.1111/lang.12383

Siyanova-Chanturia, A. (2015). Collocation in beginner learner writing: A longitudinal study. *System, 53*, 148–160. https://doi.org/10.1016/j.system.2015.07.003

Sonbul, S. (2015). Fatal mistake, awful mistake, or extreme mistake? Frequency effects on off-line/on-line collocational processing. *Bilingualism: Language and Cognition, 18*(3),419–437.

Spöttl, C. & McCarthy, M. (2004). Comparing knowledge of formulaic sequences across L1, L2, L3, and L4. In N. Schmitt (Ed.), *Formulaic Sequences: Acquisition, processing and use* (pp. 190-225). John Benjamins.

Staples, S. (2015). Spoken discourse. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics*, (pp. 271–291). Cambridge University Press.

Staples, S., Egbert, J., Biber, D., & Conrad, S. (2015). Register Variation: A Corpus Approach. In D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), *The Handbook of Discourse Analysis* (2nd ed., pp. 505–525). John Wiley & Sons. https://doi.org/10.1002/9780470753460.ch10

Suzuki, Y. (2015). The Uses of Get in Japanese Learner and Native Speaker Writing: A Corpus-based Analysis. *Komaba Journal of English Education*, *6*, 3–18.

Szudarski, P. (2023). *Collocations, Corpora and Language Learning.* Cambridge University Press.

Takač, V. P. and Lukač, M. (2013). How word choice matters: An analysis of adjective-noun collocations in a corpus of learner essays. *Jezikoslovlje 14*(2–3), 385–402.

Tan, K. H., & Azmi, N. A. (2021). Collocational Competence as a measure of ESL/EFL Competency: A Scoping Review. *3L the Southeast Asian Journal of English Language Studies, 27*(1), 115–128. https://doi.org/10.17576/3l-2021-2701-09

Tannen, D. (1990). *You Just Don't Understand: Women and Men in Conversation.* Virago.

Tannen, D. (2007). *Talking Voices: Repetition, Dialogue and Imagery in Conversational Discourse* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511618987

Tavakoli, P. & Uchihara, T. (2020). To What Extent Are Multiword Sequences Associated with Oral Fluency? *Language Learning (70)*2, 506-547. https://doi.org/10.1111/lang.12384

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, *28*(1), 127–152. https://doi.org/10.1017/S0305000900004608

Thewissen, J. (2008). The phraseological errors of French-, German- and Spanish-speaking EFL learners: evidence from an error-tagged learner corpus. In *Proceedings of the 8th Teaching and Language Corpora Conference* (pp.1-15) http://hdl.handle.net/2078.1/75986

Thewissen, J. (2015). *Accuracy Across Proficiency Levels: A Learner Corpus Approach*. Presses universitaires de Louvain.

Thompson, P., & Nesi, H. (2001). The British Academic Spoken English (BASE) Corpus Project. *Language Teaching Research, 5*(3), 263–264.

Trinity College London (2021). *Exam Information: Graded Examinations in Spoken English (GESE).* https://www.trinitycollege.com/resource/?id=5755

Trinity College London. (2023a). *Register your organisation.* https://www.trinitycollege.com/local-trinity/UK/english-language/centre-registration

Trinity College London. (2023b). *GESE levels and resources.* https://www.trinitycollege.com/qualifications/english-language/GESE/GESE-levels-and-resources

Trinity College London. (2023c). *GESE – Graded Examinations in Spoken English.* http://www.trinitycollege.com/site/?id=368

Uchihara, T., Eguchi, M., Clenton, J., Kyle, K. & Saito, K. (2021). To What Extent is Collocation Knowledge Associated with Oral Proficiency? A Corpus-Based Approach to Word Association. *Language and Speech, (65)*2, 311-336. https://doi.org/10.1177/00238309211013865

Vedder, I., & Benigno, V. (2016). Lexical richness and collocational competence in second-language writing. *International Review of Applied Linguistics in Language Teaching*, *54*(1), 23–42. https://doi.org/10.1515/iral-2016-0015

Wang, Y. (2016). *The Idiom Principle and L1 Influence: A Contrastive Learner-corpus Study of Delexical Verb + Noun Collocations*. John Benjamins.

Wang, Y., & Shaw, P. (2008). Transfer and universality: Collocation use in advanced Chinese and Swedish learner English. *ICAME Journal*, *32*, 201–232.

Wardhaugh, & Fuller, J. M. (2015). *An introduction to sociolinguistics* (7th ed.). John Wiley & Sons.

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning, 63*(1), 91–120. https://doi:10.1111/j.1467-9922.2012.00729.x

Wei, M. (2011). A comparative study of the oral proficiency of Chinese learners of English across task functions: A discourse marker perspective. *Foreign Language Annals, 44*(4), 674–691. https://doi.org/10.1111/j.1944-9720.2011.01156.x

Weigle, S. C., & Friginal, E. (2015). Linguistic dimensions of impromptu test essays compared with successful student disciplinary writing: Effects of language background, topic, and L2 proficiency. *Journal of English for Academic Purposes*, *18*, 25–39. https://doi.org/10.1016/j.jeap.2015.03.006

Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of l1 intralexical knowledge. *Applied Linguistics*, *32*(4), 430–449. https://doi.org/10.1093/applin/amr011

Wood, D. (2015). *Fundamentals of formulaic language: An introduction.* Bloomsbury.

Wray, A. (1999). Formulaic language in learners and native speakers. *Language Teaching, 32* (4), 213-231.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.

Xi, X. (2017). What does corpus linguistics have to offer to language assessment? *Language Testing*, *34*(4), 565–577. https://doi.org/10.1177/0265532217720956

Yamanishi, H., Mizumoto, A. & Someya, Y. (2013). Kansai daigaku bilingual essay corpus project: Sono gaiyou to kyoiku kenkyu nikansuru tenbou. [Kansai university bilingual essay corpus project and prospects for research and pedagogical applications]. *Journal of Foreign Language Studies, 9,* 117–139.

Yamashita, J. & Jiang, N. (2010). L1 influence on the acquisition of L2 collocations: Japanese ESL users and EFL learners acquiring English collocations. *TESOL Quarterly*, *44*(4), 647–668. https://doi.org/10.5054/tq.2010.235998

Yamashita, J. (2018). Possibility of semantic involvement in the L1-L2 congruency effect in the processing of L2 collocations. *Journal of Second Language Studies 1*(1), 60-78. https://doi.org/10.1075/jsls.17024.yam

Yan, H. (2010). Study on the Causes and Countermeasures of the Lexical Collocation Mistakes in College English. *English Language Teaching*, *3*(1), 162–165. https://doi.org/10.5539/elt.v3n1p162

Yoon, H. J. (2021). Interactions in EFL argumentative writing: effects of topic, L1 background, and L2 proficiency on interactional metadiscourse. *Reading and Writing*, *34*(3), 705–725. https://doi.org/10.1007/s11145-020-10085-7

Young, R., & Milanovic, M. (1992). Discourse Variation In Oral Proficiency Interviews. *Studies in Second Language Acquisition*, *14*(4), 403–424. https://doi.org/10.1017/S0272263100011207

Zareva, A., Schwanenflugel, P. & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition, 27*, 567–95. https://doi:10.1017/S0272263105050254

Zhang, W. Z., & Chen, S. C. (2006). EFL learners' acquisition of English adjective-noun collocations—A quantitative study. *Foreign Language Teaching and Research*, *38*(4), 251–258.

Zinkgräf, M. (2008). V+N Miscollocations in the Written Production of University Level Students. *ELIA: Estudios de Lingüística Inglesa Aplicada*, *8*, 91–116.