FACULTEIT LETTEREN
EN WIJSBEGEERTE

Exploring Phraseological Sophistication

A study utilizing the New Academic Collocation List

Author: **Jiacheng Shen**
Promoter: **Dr. Prof. Orphée De Clercq**
Academic Year: **2023-2024**
Major and specialization: **Master of Arts in Advanced Studies in Linguistics: Natural Language Processing: Theory and Practice**

UNIVERSITEIT
GENT

Table of Contents

1
Introduction

The acquisition, fluency, and competency of a language have been demonstrated to be significantly impacted by the use of various types of word combinations, including idioms, restricted collocations, proverbs, fixed expressions, and similes (e.g., Coxhead, 2008). In the domain of learner corpus research, the importance of phraseological competence for achieving fluent and idiomatic language use has been widely acknowledged (Paquot, 2018) and numerous researchers conducted various research to investigate how learners use such word combinations. For instance, a study by Durrant and Schmitt (2009) revealed that non-native writers of English heavily relied on high-frequency collocations, but they exhibited a tendency to underuse less frequent, strongly associated collocations. Similarly, another study by Siyanova-Chanturia and Schmitt (2008) focused on adjective-noun collocations in the English writings of Russian learners and suggested that despite their ability to produce a significant number of frequent and strongly associated English word combinations, even advanced learners' underlying intuitions and fluency with collocations did not match those of native speakers. These findings were further supported by Bestgen and Granger's (2018) longitudinal investigation of the phraseological development of English as Foreign Language (EFL) learners, which demonstrated an increase in the proportion of collocations with high pointwise mutual information scores over time. Collectively, these studies highlight that learners of different proficiency levels may exhibit varying degrees of mastery in phraseology, and the use of phraseology can serve as a useful indicator for examining L2 performance.

Recent research has focused on examining the complexity of L2 phraseological performance in terms of phraseological complexity, which Paquot (2019:126) referred to as “the range and degree of sophistication of phraseological units used in language production”. In Paquot (2019), phraseological units were operationalized as binary dependency relationships between a head and its dependent. The operationalization of phraseological complexity included two dimensions: phraseological diversity and

phraseological sophistication, which respectively capture the breadth and depth of phraseology. Three types of dependency relations (i.e., adjectival modifiers, adverbial modifiers, and direct objects) were used to operationalize phraseological diversity as the root type-token ratios of dependencies. Phraseological sophistication was assessed using two methods. The first method employed the mean pointwise mutual information score, while the second method involved the calculation of type- and token-based ratios. These ratios were calculated by comparing the quantity of sophisticated dependencies produced by leaners to the total number of dependencies produced by learners. The study used a corpus of linguistics essays (the Varieties of English for Specific Purposes dAtabase; VESPA; Paquot et al., 2022) written by French EFL learners. The result of employing the second methodology to measure phraseological sophistication unveiled that although the ratios of phraseological sophistication increased with learners' overall proficiency levels according to the Common European Framework of Reference (CEFR; Council of Europe, 2001), no statistically significant difference was observed.

Research on phraseological sophistication within second language (L2) writing, operationalized through the utilization of both type- and token-based ratios derived from the division of the count of sophisticated dependencies by the total count of dependencies, as exemplified in the investigations conducted by Paquot (2018, 2019), has employed Ackermann and Chen's (2013) Academic Collocation List (ACL) as sophisticated phraseological units to operationalize the dimension of phraseological sophistication. Findings have revealed that more proficient L2 academic writers tend to employ higher portions of sophisticated phraseological units, but these differences have not been statistically significant. This could be attributed to the very nature of the ACL. The ACL was designed to assist EFL learners in improving their collocational competence and to aid EAP teachers in lesson planning. However, it was not designed for research purposes. As a result, the list has limited coverage and is relatively small, with academic collocations only accounting for 1.4% of the source corpus used in the study of Ackermann and Chen (2013). Therefore, its use as a measure of phraseological sophistication is limited.

With the objective of creating a novel academic collocation list for examining phraseological sophistication of L2 writings, Shen (2023) created the New Academic Collocation List (NACL). The NACL was created from the ground up for research purposes, i.e., examining phraseological sophistication in L2 writings, instead of serving as resources of English pedagogy. It contains 3,756 collocations and it has a coverage of 15.21% within its source corpus. With a larger scope and a larger coverage, the NACL was deemed by Shen an ideal source for examining phraseological sophistication of L2 writings.

Hence, the primary objective of this research is to replicate the phraseological sophistication study in Paquot (2019) while using the New Academic Collocation List. The collocations in the NACL will be regarded as sophisticated collocations and the portion of

such sophisticated collocations will be calculated and tested whether a statistically significant increase will be observed from adjacent lower proficiency levels to higher proficiency levels.

2
Phraseology: Definition and Classification

2.1 Defining Criteria

According to Knappe (2004), phraseology emerged as an academic method for studying language in the 20th century. Initially, the study of phraseology began in the former Soviet Union and other Eastern European countries (Cowie, 1998). In recent years, there has been a significant surge in interest regarding the field of phraseology, which includes both the study of the phenomenon itself and its intersection with other disciplines such as learner corpus research and second language acquisition. Recently, phraseology has come to an intersection of various linguistic disciplines, including lexicology (De Cook, 2000), psycholinguistics and language acquisition (Wray, 2002), second language acquisition (Nation, 2001), and English for academic purposes (Cowie, 1997).

Cowie (1994) provided a general definition of phraseology as the study of the structure, meaning, and use of word combinations, leaving much room for ambiguity as mere explanation or specification has been made for the concept “word combinations”. Researchers have proposed various terms to refer to the object of study, including word-combinations (Cowie, 1994), multi-word units (De Cock, 2003), fixed expressions (Alexander, 1984), and formulaic sequences (Wray, 2002). Despite the many proposed definitions, there is still no clear agreement on the definition of phraseology, with some definitions being opaque or conflating different terms (Gries, 2008).

Gries (2008:2) proposed a contemporary definition for phraseologism or phraseological units, wherein they are described as “the co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as

one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance".

This section is designed to review scholarly literature pertaining to phraseology. The subsequent sections present a comprehensive examination of diverse aspects of phraseology. Section 2.1 adopts a specific perspective by focusing on several defining criteria for phraseology, with the aim of delving into its fundamental nature. In Section 2.2, various categorizations of phraseology proposed by prior researchers are explored.

2.1.1 Polylexicality

Polylexicality in phraseology generally refers to the fact that multi-word units are made up of more than one lexical item (i.e., a word or a group of words with a lexical meaning). Polylexicality is a distinguishing feature of all types of multi-word expressions, including idioms, collocations, and phrasal verbs.

For example, the idiom *spill the beans* is a polylexical unit, as it is made up of the verb *spill* and the noun *beans*. Similarly, the collocation *read a book* is a polylexical expression, as it is made up of the verb *read* and the noun *book*.

Polylexicality is a key characteristic of phraseology. The fact that multi-word units are made up of more than one lexical item constitutes a basic characteristic of phraseology.

2.1.2 Compositionality

Barkema (1996:138) provided a definition of compositionality as "the degree to which the interpretation of a multi-word unit can be determined by the combination of the basic or derived meanings of the lexical items contained within the unit and the syntactic relationships within the constituent that encompasses these lexical items". Prior to introducing this concept, Barkema first differentiated between three types of senses for a lexical item, e.g., basic sense, extended sense, and derived sense.

The concept of lexical sense can be categorized into three types, as identified by Barkema (1996:138). The first type is the basic sense, which refers to the primary and most salient meaning of a lexical item that comes to mind when it is encountered as a standalone unit, rather than within a larger sentence. This sense forms the foundation for all other senses that may be associated with the lexical item. The second type is the extended sense, which arises as a result of violating a restriction rule in a specific context, leading to an expansion of the fundamental meaning. The third type is the derived sense, which denotes a secondary meaning of a lexical item that has become established in the speaker's mind. Typically, this sense is the second or third interpretation that comes to

mind when the speaker encounters the lexical item as a single word. The subsequent paragraph presents a collection of illustrative instances concerning the three distinct types of senses.

The basic sense of the lexical term *house* typically refers to a structure intended for human habitation. However, the extended sense of *house* is exemplified in the sentence "But most Czech believed until now that Slovakia would still be in favor of a federal ***house***" (The Economist, 1/8/1992). Here, the term *house* transcends its literal meaning and refers to a *country*. Furthermore, the derivational sense of *house* is illustrated in the sentence "Sir Marcus divided the ***house*** and did not divide without organization. Hughes's motion was roundly defeated" (New Statesman & Society, 2/8/1991). According to the Oxford English Dictionary, in this context, *house* signifies *a group of people who meet to discuss and make the laws of a country*.

The concept of compositionality was approached via such three senses of a lexical item. In other words, the meaning of a multi-word unit is compositional if it can be predicted based on the meanings of its constituent words. For example, the meaning of the multi-word unit *red apple* can be derived from the meanings of the individual words *red* and *apple*. The meaning of the multi-word unit is compositional because it reflects the combination of the meanings of the two words. However, not all multi-word units are compositional. Some multi-word units have a meaning that cannot be predicted from the meanings of their individual words. For example, the multi-word unit *kick the bucket* means to die, but this meaning cannot be derived from the meanings of the words *kick* and *bucket* alone.

Barkema (1996) developed a classification scheme that distinguishes between different levels of compositionality in language, which can be categorized into four levels: fully compositional, pseudo-compositional, fully non-compositional, and partly non-compositional. These levels are elaborated upon as follows:

- Fully compositional constructions are those in which "the meaning of the construction is the combinatorial result of the basic senses and the syntactic relations of the lexical items within the construction" (Barkema, 1996: 138). For instance, the multi-word unit *a black dog* is fully compositional since it can be understood by combining the basic senses of *black* and *dog* with the syntactic relationship between them.
- Pseudo-compositional constructions involve lexical items in which "the basic senses of the lexical items in the constructions play a role in, but form only parts of, their meanings" (Barkema, 1996: 139). For example, the multi-word unit *bed and breakfast* refers to a specific type of accommodation that provides a systematic service, including but not limited to providing a bed and breakfast.

- Fully non-compositional constructions are those in which "the constructions do not have lexical items with basic senses that form (part of) their meanings" (Barkema, 1996: 139). An example of this type of construction is "put the cat among the pigeons".
- Partly non-compositional constructions contain at least one lexical item with a basic sense, but not all the basic and derived senses of the lexical items contribute to the overall meaning of the construction (Barkema, 1996: 140). For example, "a blind valley" is partly non-compositional since the basic sense of "blind" contributes to the overall meaning of the construction, while the basic sense of "valley" does not. This differs from pseudo-compositional constructions, where some of the basic or derived senses of the lexical items contribute to the overall meaning.

Compositionality plays a crucial role in the field of phraseology, as it enables researchers to differentiate between various types of phraseological units. Katz (1996) distinguished idioms from other phraseological units, as their meaning "is not a compositional function of the meanings of the idiom's elementary grammatical parts" (p. 275). However, the degree of compositionality may vary across different types of phraseological units. For example, Barkema's (1996) theory suggests that the compositionality of two adjectival modifiers, "black dog" and "red herring," is entirely different, with the former being fully compositional, whereas the latter is fully non-compositional. While "black dog" denotes a dog whose color is black, "red herring" refers to a misleading statement, question, or argument intended to divert a conversation from its original topic.

Svensson (2008) identified four defining dichotomies associated with non-compositionality, namely motivation/non-motivation, transparency/opacity, analyzability/unanalyzability, and literal/figurative meaning. The term “motivation” refers to the possibility of comprehending the meaning of each word in an expression once the meaning of the entire expression is known (Svensson, 2008:83). To describe an expression in which understanding the meaning of each word contributes to understanding the complete expression, Svensson suggested the term "motivatable." Conversely, an expression is unmotivated if it is impossible to comprehend the meaning of the expression by making sense of the meaning of each word included in it. For example, the expression *white wedding* is motivatable since the color white is commonly associated with purity and innocence.

Transparency denotes the attribute of an expression that enables a language user to understand it without any difficulties or previous knowledge, except for the comprehension of each separate word that constitutes the expression. In contrast, opacity describes the inability to reconstruct the meaning of an expression from the meanings of

its composing elements. Similes, such as *as white as snow*, are typical examples of transparent expressions.

Analyzability is the characteristic of an expression where each individual element contributes to the meaning of the expression as a whole. The notion of "decomposition" is also used to describe analyzability. For instance, the expression *pop the question* is analyzable as each individual element in the expression contributes to the meaning of the expression as a whole. In this case, *pop* can be taken to mean to ask, and *question* can be taken to mean wedding proposal.

Another significant dichotomy of non-compositionality is literal and figurative meaning. Svensson argued that it is rather problematic to provide a definition to literal and figurative meaning. Nevertheless, she believed that there is a relation between literal and compositional meaning as well as between figurative and non-compositional meaning. Gross (1996:11) argued that there are many sequences that a foreigner cannot interpret literally in any language, even if he or she knows the common meanings of all words that make them up. She stated that the “ordinary” meaning of the sentence “La moutarde lui monte au nez [‘The mustard goes up his nose’] does not allow one to conclude that the whole sentence means that a person is getting angry. Thus, we say that this sentence does not have a compositional meaning. Svensson then concluded that the feature ‘not having a literal interpretation’ must be equivalent to ‘having a figurative interpretation’.

Although the concept of compositionality has garnered considerable attention within scholarly investigations, the task of categorizing multi-word units according to their varying degrees of compositionality remains complex. In this exposition, two prototypical challenges shall be expounded upon: ambiguity and context-dependence.

Determining the precise degree of compositionality for certain phraseological units can be difficult due to the potential ambiguity in their meanings. Some expressions may have multiple interpretations, making it challenging to assign a specific compositionality level accurately. Take the multi-word unit *break a leg* for example. It is a commonly used idiom in English, typically used to wish someone good luck. Due to the ambiguity, two degrees of compositionality can be recognized: non-compositional and partially compositional. In the first interpretation, *break a leg* is treated as a non-compositional multi-word unit with a high degree of idiomaticity. The literal meaning of breaking one's leg doesn't relate directly to the intended good luck wish. Therefore, this interpretation suggests that the phrase is largely opaque and not easily analyzable based on its individual components. In the second interpretation, *break a leg* is considered partially compositional. It acknowledges that the literal meaning of "break a leg" is unrelated to the idiomatic usage but points out that the phrase still contains familiar and recognizable words. The degree of compositionality here might be seen as intermediate, as the phrase involves a well-

known action ("break") and body part ("leg"), even though the specific idiomatic meaning isn't immediately obvious.

Determining the precise degree of compositionality for certain phraseological units can also be difficult due to the degree of compositionality of context dependent. Take the multi-word unit *kick the bucket* for example. Depending on the context, two degrees of compositionality can be identified. If the phrase is used in a context where someone is actually kicking a physical bucket, the degree of compositionality is high. The expression is being used literally, and the individual components (*kick* and *bucket*) contribute directly to the overall meaning of the action. In most other contexts, the phrase is used idiomatically to refer to someone's death. In this case, the degree of compositionality is low. The literal meaning of the individual components doesn't align with the intended idiomatic meaning, and the phrase becomes less transparent.

Evidently, the extent of compositionality undergoes influences from a multitude of factors, among which two (ambiguity and context dependency) have been elaborated upon earlier. This elucidation underscores the intricate nature of classifying multi-word units according to their level of compositionality, revealing the array of challenges inherent in such categorization.

2.1.3 Restricted Collocability

Restricted collocability is a phraseological phenomenon where specific words tend to occur together within particular contexts or collocational patterns. These patterns demonstrate the predictable and regular ways in which words combine with one another in language use. According to Barkema (1996:45), collocability can be defined as "the extent to which a lexical item from an open class can be replaced in a construction with an alternative from the same class, such as a noun with another noun, a verb with another verb, and so on."

The collocability of multi-word units varies according to the degree to which their component elements can be replaced by synonyms, near-synonyms, or antonyms. Barkema (1996) classified multi-word units into three types based on their collocational openness: collocationally open, closed, and limited. For instance, the multi-word unit *increase dramatically* is an example of a collocationally open unit where substituting the word *dramatically* with synonyms such as *significantly* or *excessively* does not significantly change the meaning. In contrast, the components of some units, like *red tape*, are collocationally closed. This means that substituting the word *tape* with a synonym like *ribbon* alters the entire construction's meaning and may only result in a literal interpretation. Finally, some units are collocationally limited, indicating that while their

components can be alternated to some extent, there are limits to the available options. For instance, *generally speaking* can be replaced with *strictly speaking*, but alternatives like *elaborately speaking* are not typically used.

In summation, this section has undertaken an examination of the trio of pivotal criteria that delineate phraseology, namely polylexicality, compositionality, and restricted collocability. Remarkably, these three attributes collectively embody the essence of the phraseological domain. Importantly, the concept itself presents a multi-dimensional construct, and the attempt to classify phraseology based on a sole criterion such as compositionality is fraught with a diverse range of challenges.

2.2 Categorization of Phraseological Units

This section provides an overview of the categorization schemes delineated by prior researchers within the realm of phraseology. Broadly, two primary methodological approaches to phraseology have emerged: the phraseological approach and the distributional approach.

2.2.1 The phraseological approach

In the former Soviet Union, the study of phraseology was pioneered by scholars such as Vinogradov, who proposed a conceptual framework that characterizes phraseological units as a continuum ranging from highly rigid and impenetrable to flexible and transparent. This approach, commonly referred to as the "phraseological approach," utilizes linguistic criteria to distinguish various types of phraseological units. The phraseological approach of classifying multi-word units involves focusing on their internal structure and composition. Additionally, multi-word units are categorized based on their inherent semantic and syntactic properties, and their classification is guided by the principles of grammar and meaning.

Vinogradov (1953) classified phraseological units into phraseological fusions, phraseological unities, and phraseological combinations based on the relationship between the meaning of the unit and the meaning of its component parts. Similarly, Cowie (1981) categorized phraseological units and created the following continuum based on the degree of opacity and fixedness, with opacity and fixedness increasing along the spectrum:

- Free combinations (e.g., *drink tea*): The restriction on substitution can be specified on semantic grounds and all component elements of the word combination are used in a literal sense.
- Restricted collocations (e.g., *perform a task*): It enables certain with arbitrary limitations; At least one element has a non-literal meaning, and at least one element is used in its literal sense.
- Figurative idioms (e.g., *do a U-turn*): Mere substitution of the elements is pertained; The combination has a figurative meaning, but preserves a current literal interpretation.
- Pure idioms (e.g., *blow the gaff*): No substitution of the elements is pertained; The combination has a figurative meaning and does not preserve a current literal interpretation.

Due to the lack of a consensus on the terminology used in phraseology, researchers have encountered difficulties in defining and categorizing multi-word units. Different typologies have been developed for diverse purposes, including lexicographic, pedagogical, and psycholinguistic. Regarding the categorization from the perspective of the phraseological approach, Cowie (2001) proposed a typology of word combinations, which included composites and formulae. Composites were divided into restricted collocations, figurative idioms, and pure idioms, with decreasing degrees of transparency and fixedness. Restricted collocations were defined by their limited collocability, such as adjective-noun word combinations (e.g., *dramatic increase*). Figurative idioms had figurative meanings and were characterized by non-transparency and a certain degree of fixedness, such as *have a seat*. Pure idioms, such as "the early bird gets the worm", were the least transparent and non-compositional, meaning that the elements in the idioms could not be replaced with other elements. Cowie further categorized formulae into routine formulae, which were expressions tied to recurrent social situations (e.g., *good morning*), and speech formulae, which were expressions used to organize discourse and indicate speakers' attitudes (e.g., *I beg your pardon).* A diagram illustrating Cowie's typology is presented in Figure 2.1.

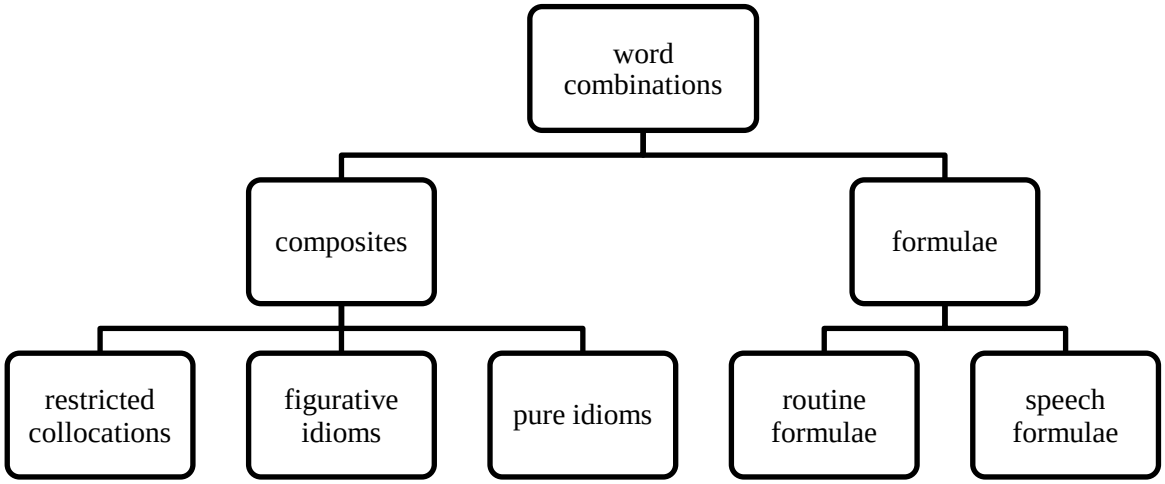

Figure 2.1 Cowie's (1998, 2001) typology of phraseological units

Mel'čuk (1998) introduced an influential taxonomy that distinguished set phrases or phrasemes into two categories: semantic phrasemes and pragmatic phrasemes or pragmatemes. According to Mel'čuk, semantic phrasemes refer to set phrases whose meaning is freely chosen but the expression for that meaning is not. This category comprises full phrasemes or idioms, quasi-phrasemes or quasi-idioms, and semi-phrases or collocations. Full phrasemes contain semantic phrasemes whose signified does not include the signified of their components in a semantically dominant position. For instance, the signified of *spill* and *bean* are not included in the signified of *spill the bean*. Quasi-phrasemes include the signified of their components plus an additional signified. An example of a quasi-phraseme or quasi-idiom is *bed and breakfast*, which includes not only the signified of its components, but also additional signified such as heat, electricity, and the service of hotel waiters, among others. Semi-phrasemes or collocations are semantic phrasemes in which the global signified is derived from the signified of one of their components A and a signified 'X,' with the other component B only expressing 'X' contingent on A. For instance, the collocation heavy rain is constructed out of the signified of rain and a signified X that the adjective heavy only expresses contingent to the noun rain. In contrast, pragmatic phrasemes or pragmatemes are context-dependent and their meaning corresponds to Cowie's formulae. Thus, they are "non-compositional pragmatically" (Mel'čuk, 1998:29). For instance, *good morning* is an example of a pragmatic phraseme.

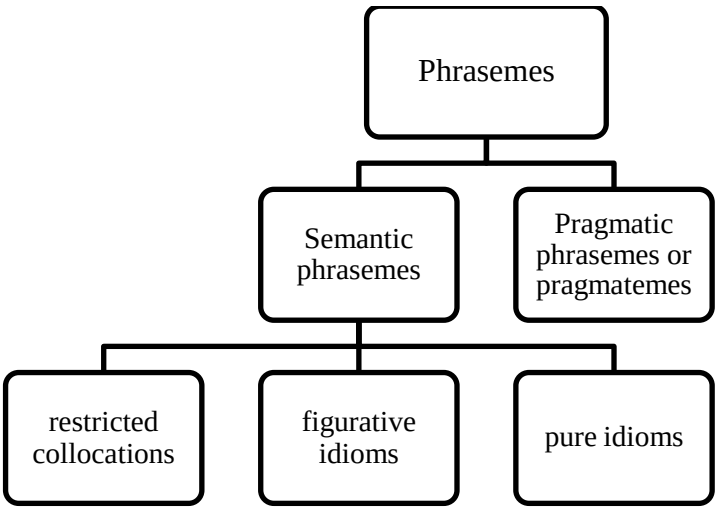

Figure 2.2 Mel'čuk's (1998) typology

2.2.2 The distributional approach

In contrast to the phraseological approach of categorizing multi-word units, a more recent approach, known as the "distributional approach" or "frequency-based approach," has emerged. This approach uses a bottom-up, corpus-driven methodology to examine phraseological units. Rather than pre-categorizing different types of units, this approach examines the lexical co-occurrence of words in phrases. This approach, as demonstrated by Altenberg (1998), expands the possibilities for phraseology research, including units that were previously outside the domain of phraseology in the traditional approach. Altenberg analyzed recurrent word-combinations in the London-Lund Corpus of Spoken English (Svartvik and Quirk, 1980) by extracting all n-word ($n \geq 2$) word-combinations occurring more than once in identical form and focusing specifically on 3-word combinations that occurred at least ten times in the corpus. The distributional approach, which examines the usage patterns of phraseological units, and as illustrated in Figure 3, has gained considerable recognition in the field.

In this categorization, two distinct categories are defined, namely n-gram or cluster analysis, and co-occurrence analysis. The n-gram or cluster analysis focuses on studying continuous or adjacent sequences of two or more words. These extracted sequences are defined as "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" (Biber et al., 1999:990). The researcher determines the length of the sequences, frequency threshold, and other relevant parameters depending on the research objectives. For example, Simpson-Vlach and Ellis (2010) set the length of the sequences to 3, 4, and 5 words in the Michigan Corpus of Academic Spoken English (Simpson et al., 2002), plus selected British National Corpus (Davies, 2008; BNC) files that occurred at least 10 times per million words in the target corpus, to create a list of exclusively academic word sequences.

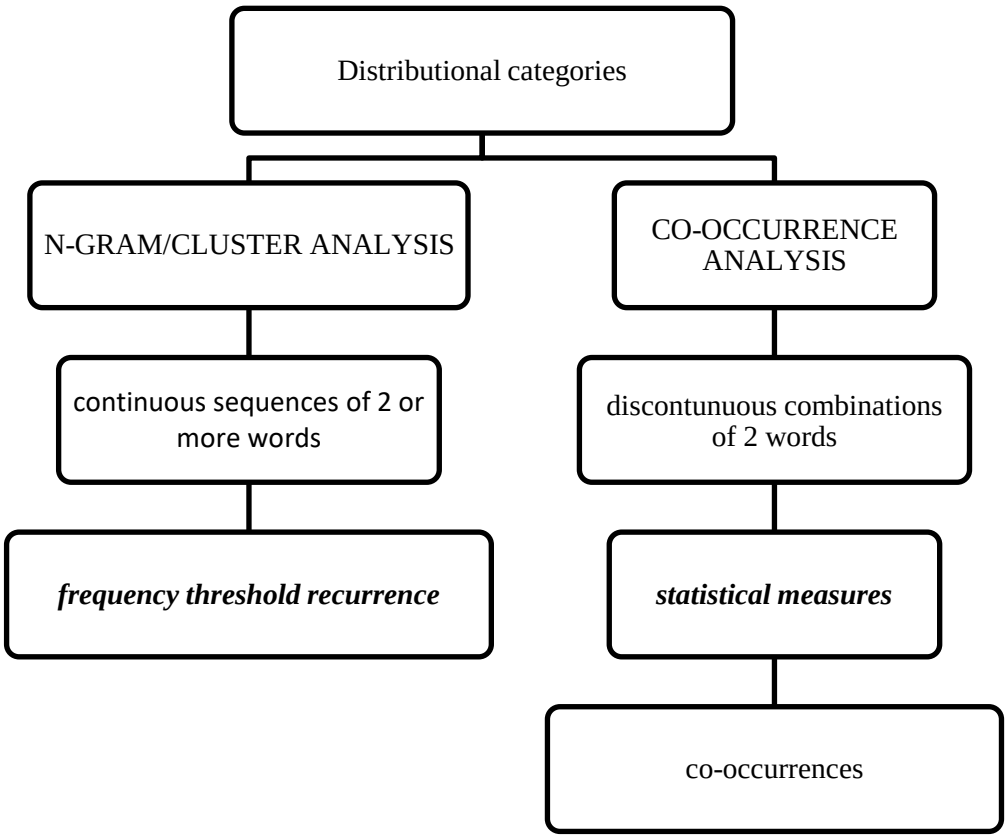

Figure 2.3 Distributional categories

In contrast, co-occurrence analysis focuses on studying non-adjacent combinations of two words using statistical measures such as mutual information (Rogers et al., 2021) or log likelihood ratio (Durrant, 2009), which are referred to as 'collocations' or 'collocates.' For instance, Durrant (2009) created a corpus that was comprised of approximately 25 million tokens from five disciplines, i.e., arts and humanities, life sciences, science and engineering, social-administrative, and social-psychology. The texts of these five disciplines were collected from the corresponding department of the university where the author taught. He then extracted academic collocations while taking into account

variations across disciplines. He followed Jones and Sinclair's (1974) widely used precedent of limiting 'co-occurrence' to occurrences within a four-word span and requiring any potential collocation included in the final list to have a mutual information score of at least 4. Mutual information (MI) is a measure borrowed from information theory that compares the probability of observing word pairs together with the probabilities of observing them independently. It is used to quantify the strength of association between two words in a text corpus. It assesses how much more (or less) likely two words are to occur together than if they were independent of each other. Generally, the higher the MI score, the higher the association between the two words. Specifically, it is calculated using Formula 2.1.

$$\mathrm{MI}(w_1, w_2) = \log_2 \left(\frac{P(w_1, w_2)}{P(w_1)P(w_2)} \right)$$

Formula 2.1. The calculation of mutual information scores of the word pair (W1, W2)

$P(W_1, W_2)$ is calculated by dividing how many times the word pair (W_1, W_2) appears in the text by the number of words in the text. P(W1) and P(W2) are calculated by dividing how many times the words W_1 and W_2 appear in the text by the number of words in the text, respectively.

To sum up, the study of phraseology has evolved over time, from the phraseological approach that classified phraseological units based on their rigidity and opacity, to the distributional approach that uses a corpus-driven methodology to examine lexical co-occurrence in phrases. Different typologies have been proposed, each with its own set of categories and criteria, which highlights the difficulties in defining and categorizing phraseological units.

3
Phraseology in Learner Language

In addition to the study of phraseology produced by native speakers, the use of phraseology by EFL learners have sparked the interest of linguistic researchers. Specifically, it is becoming an essential part of L2 complexity research. Over the years, the utilization of complexity measures as indicators of L2 development holds a significant place in the realm of second language research. In conjunction with accuracy and fluency, complexity remains a pivotal factor in assessing L2 proficiency (Bulté & Housen, 2012). Despite the diverse methods available for operationalizing linguistic complexity, commonly employed measures often concentrate solely on the lexical or syntactic facets of a text (De Clercq & Housen, 2019). This limitation is regrettable, given that numerous phraseological studies have highlighted the crucial role of phraseological units, which stand between the syntactic and lexical facets of a text, encompassing collocations, phrases, and similar constructs, emphasizing their integral role in EFL development (e.g., Granger & Bestgen, 2014). The aim of this section is to critically examine the investigation of phraseological units utilized by EFL learners and assess the extent to which such usage serves as an effective indicator of EFL learners' proficiency levels. This examination is structured into three main sections. Firstly, Section 3.1 provides an overview of the datasets employed for analyzing the utilization of phraseological units by EFL learners. Following this, Section 3.2 delves into the various types of phraseological units that have been the focus of researchers' investigations. Lastly, Section 3.3 offers a comprehensive summary of the methodologies utilized in these studies along with their major findings.

3.1 Learner corpora data used

Over the last several years, more research has been conducted on phraseological units used by EFL learners. This can be contributed to the compilation of various learner corpora, which Paquot and Granger (2012) defined as "electronic collections of texts produced by foreign or L2 learners". Due to the complexity of learners, e.g., age, L1 backgrounds, time spent learning English, etc., and tasks, e.g., mode, register, timing, etc., various types of learner corpora were compiled for different research purposes.

Several learner corpora have played significant roles in the studies of phraseological units by EFL learners, with some standing out prominently. For instance, many studies have utilized the International Corpus of Learner English (ICLE), which was among the first learner corpora made available for research purposes (Granger et al., 2020). The ICLE comprises approximately 5 million tokens (9,000 texts) of argumentative essays authored by learners representing 26 different native language backgrounds. Other notable corpora of EFL learner writing include the Uppsala Student English Corpus (University of Oxford, 2003) and the Varieties of English for Specific Purposes Database (Paquot et al., 2022). In addition to written learner corpora, researchers have also compiled corpora consisting of spoken transcriptions. For example, the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin, De Cock, & Granger, 2010) comprises around 800,000 words of oral data produced by learners from 11 different native language backgrounds. The purpose of the LINDSEI corpus was to provide a spoken counterpart to the ICLE, thereby offering a comprehensive resource for analyzing learner language in both written and spoken forms.

The size of corpora such as ICLE or LINDSEI offers a major advantage, primarily in terms of their representativeness and reliability for phraseological studies. Larger corpora increase the representativeness of data, enhancing the reliability of analyses, especially concerning multi-word phraseological units. Analyzing phraseological units based on small corpora is generally considered less acceptable in terms of reliability as Durrant (2009:159) noted that "since most collocations are relatively rare, in comparison to individual words, a large corpus is required in order to find such items". However, this does not imply that small corpora lack value in phraseological analysis. For instance, Wang and Shaw (2008) conducted research on verb + noun collocations with *have*, *do*, *take*, and *make*, using self-compiled learner corpora generated by Chinese-speaking and Swedish-speaking learners of English. Although these learner corpora were relatively small, containing approximately 40,000 tokens in total, all the learner texts focused on argumentative essays, allowing for control over the influence of the topic on lexical choice. Despite no standards have been established on how large a corpus should be to analyze L2 phraseology, the prevailing

trend in phraseological studies within learner corpus research leans towards utilizing larger corpora.

In addition to categorizing learner corpora based on learner characteristics such as age and gender, it is imperative to consider whether the learner texts in the corpus were gathered at a specific juncture or across a span of time. Corpora comprising learner texts collected at a particular moment are termed as cross-sectional corpora, while those encompassing texts amassed over an interval are denoted as longitudinal corpora. At present, the predominant inclination in learner corpora is towards cross-sectional designs. Nonetheless, there exists a burgeoning interest in exploring phraseology through longitudinal learner corpora, as elaborated further in section 2.3. A noteworthy illustration of a longitudinal corpus is the Longitudinal Database of Learner English (Meunier, 2008; LONGDALE). This corpus endeavors to compile learner productions over a minimum duration of three years, with data collections orchestrated at least annually.

3.2 Types of phraseological units investigated and methodologies

In learner corpus research, a significant area of focus revolves around verb + noun collocations, particularly those involving high frequency delexical verbs like *get* and *make*. These studies (e.g., Altenberg and Granger, 2001; Howarth, 1996; Nesselhauf, 2005) highlight the constrained co-occurrence of elements, where certain combinations allow limited substitution within a specific grammatical structure. For instance, *perform a task* is acceptable, whereas *make a task* is not. However, learner corpus research extends beyond this narrow focus, encompassing the entirety of verb + noun combinations produced by learners. This includes both restricted collocations, as mentioned previously in section 1.2.1, and free combinations such as *read a book* and *cook some food* (see, for example, Howarth, 1996; Nesselhauf, 2005). Learners have been shown to encounter difficulties with various types of verb + noun collocations, including both restricted collocations and free combinations (see, for example, Laufer & Waldman, 2011).

Learner corpus researchers also focus on another category of phraseological units: phrasal verbs. Learners often demonstrate difficulties by either using the correct verb with the wrong particle (e.g., "this is harmful for [be harmful to] our children") or using the correct particle with the wrong verb (e.g., "We tried to come back to [go back to] Los Angeles").

Learner corpus researchers have devoted significant attention to pragmatic speech formulae, which are sequences of words used to organize the speaker's discourse. For instance, Aijmer (2009) undertook a comparison of the usage of *I don't know* in the Swedish segment of the LINDSEI corpus (Gilquin, De Cock, and Granger, 1995) and a comparable native speaker corpus. The investigation reveals divergent functions of this expression in native versus learner speech contexts. Swedish EFL learners predominantly employ *I don't know* as a conversational management cue, whereas native speakers primarily utilize it to indirectly refrain from asking questions.

In addition to pragmatic speech phraseological units, pragmatic phraseological units in learner writing have also garnered attention. Corpus linguistic tools and methodologies have played a vital role in establishing connections between phraseological units and text organization (see, for example, Conrad & Biber, 2004). Pragmatic phraseological units in writing are frequently utilized for structuring text content and fulfilling rhetorical functions such as comparison, summarization, and conclusion (e.g., *it has been suggested that*, *as a result*). This array of textual sequences has been explored in corpora of learner writing to identify challenges faced by learners and to inform the development of teaching materials for English for Academic Purposes (see, for example, Gilquin, Granger, & Paquot, 2007; Paquot, 2008).

Various techniques have been employed to extract phraseological units from learner corpora, all of which share a common characteristic: they involve some degree of automation. A fundamental method for retrieving formulaic sequences entails utilizing the concordance feature of a program, such as WordSmith Tools (Scott, 2024), which displays all instances of a linguistic item within their respective contexts. Each occurrence is presented on a single line, with the search item positioned in the center and the surrounding context on either side. This approach has been employed to extract all instances of continuous word combinations (see, for instance, Aijmer, 2009). By organizing the context both to the left and right of the search items, consistent patterns can be discerned.

In addition to merely identifying phraseological units in learner language, certain language tools, such as MonoConc Pro (Barlow, 2000), offer further insights into the extracted phraseological units, including MI scores. Lorenz (1999) conducted one of the earliest learner corpus studies focusing on statistically significant co-occurrences. He utilized the mutual information statistic to pinpoint strongly associated intensifier-adjective combinations in a corpus of essays authored by German EFL learners. The MI score highlights co-occurrences that, while not necessarily frequent, demonstrate close association, making them likely to be highly noticeable to native speakers.

As some recent research increasingly emphasizes more extensive types of phraseological units, such as collocations following specific grammatical structures (e.g.,

adjective + noun collocations; Paquot (2019), Paquot et al., (2020)), relying solely on corpus linguistic tools (e.g., WordSmith Tools) for analysis becomes inadequate. Instead, various programming languages, such as Python, Java, Perl, etc., offer researchers a more flexible approach to language processing. In particular, Natural Language Processing (NLP) tools like NLTK (Bird, Edward, and Ewan, 2009) and spaCy (Honnibal and Montani, 2017) enable the automatic processing of large learner language corpora, facilitating tasks that would be excessively time-consuming if performed manually, such as high-accuracy part-of-speech tagging, dependency parsing, and named entity recognition.

Recent research in learner corpus analysis has heavily relied on such methodologies. For instance, Paquot (2018) compared learners' proficiency in utilizing phraseological units within a broader framework of complexity, encompassing lexical and syntactic aspects. Paquot examined a corpus of argumentative essays authored by French EFL learners, comprising 336,749 tokens, and investigated three types of phraseological units: adjectival modifiers (adjective + noun, e.g., *clear evidence*), adverbial modifiers (adverb + adjective, verb, or adverb, e.g., *statistically significant*), and direct objects (verb + noun, e.g., *make efforts*). These phraseological units were extracted using the Stanford NLP suite of tools in Java, which enabled automatic and high-accuracy lemmatization, part-of-speech tagging, and syntactic parsing. Given the size of the learner corpus, manual processing would be nearly impossible to execute efficiently; thus, automatic NLP tools offer a more effective solution.

To date, the predominant approach to exploring phraseological units in learner corpora has revolved around comparing the outcomes of learner corpus analysis to those from scrutinizing equivalent native corpora. This methodology seeks to discern patterns and errors in the way learners deploy phraseological units. Termed contrastive interlanguage analysis (CIA; Granger, 1996), this method is widely utilized. For instance, Chen and Baker (2010) scrutinized and compared the usage of contiguous four-word combinations (e.g., "it is obvious that") between Chinese EFL learners and native English speakers. Moreover, the CIA technique might entail contrasting different varieties of learner language, typically distinguished by their native language backgrounds. For instance, Waibel (2008) scrutinized the usage of phrasal verbs in two sets of argumentative essays written by German and Italian EFL learners to gauge how the learners' first languages influence their adoption of both literal and idiomatic verb-particle pairings.

3.3 Major findings

In this section, our attention is directed towards the overarching patterns observed in studies of phraseological units in learner language explicitly outlining their methodologies, thus enabling meaningful summaries of phraseological units in learner corpus research. Building upon the framework proposed by Paquot and Granger (2012), we differentiate between two aspects of phraseology: co-occurrence and recurrence. It is essential to examine these aspects separately because learners demonstrate distinct challenges when confronted with these two types of phraseological units.

3.3.1 Co-occurrence

Co-occurrence stands as one of the two fundamental patterns central to phraseology. As described by Paquot and Granger (2012), co-occurrence denotes the joint selection of (typically) two lexical items, which may or may not be adjacent. In the phraseological framework outlined in section 2.2, co-occurrence items are termed collocations or restricted collocations. These represent lexically restricted combinations permitting limited substitution within a specific grammatical framework. For instance, in English, one can aptly say *perform a task* but not *make a task*. These lexical constraints are arbitrary and vary significantly across languages. Consequently, learners must commit these constructions to memory individually, which presents challenges. In contrast to restricted collocations, free combinations impose no restrictions on substitution and can be defined based on semantic considerations. All constituent elements of the word combination are employed in a literal sense. For example, it is possible to *drink coffee* or any other liquid, but solids like bread cannot be "drunk".

The collective findings from studies based on learner corpora indicate that learners' utilization of co-occurring combinations is marked by a blend of underuse, overuse, and misuse. Lee (2006) investigated the usage of amplifier collocations in English by both native English speakers and Korean EFL learners. The results revealed that in comparison to native English speakers, Korean EFL learners employ a limited range of high-frequency amplifiers in their writing. Similarly, Laufer and Waldman (2011) illustrated that Hebrew-speaking EFL learners produce approximately half the number of verb + noun collocations compared to young adult native speakers. Moreover, additional research (e.g., Chen & Baker, 2010) has suggested that this overall underutilization is accompanied by an excessive reliance on specific high-frequency collocations, such as *have a look*. Nesselhauf

(2005) hypothesized that learners may gravitate towards these expressions due to a perceived confidence in their usage.

The combination of the overuse of high-frequency collocations and the underuse of less common collocations aligns with findings from statistical analyses of co-occurrence in learner corpora. For instance, Durrant and Schmitt (2009) explored the usage of adjacent adjective + noun and noun + noun co-occurrences among non-native and native English speakers. In their study, the associational strength of the two lexical items in a co-occurrence was assessed using mutual information scores, while frequency was evaluated using t-scores. The findings revealed that non-native writers tend to underutilize less frequent but strongly associated co-occurrences, which are likely salient for native speakers. However, they exhibit a propensity to heavily rely on high-frequency co-occurrences such as "good example" and "hard work." Durrant and Schmitt (2009: 175) provided an explanation, stating that "this is an intuitively satisfying result: learners are quick to pick up highly frequent collocations, but less common, strongly associated items take longer to acquire."

In addition to contrasting the usage of phraseological units between native English speakers and non-native speakers, researchers have also examined the usage of phraseological units across different proficiency levels among learners. Granger and Bestgen (2014) adopted a methodology similar to that used by Durrant and Schmitt (2009) to compare the usage of phraseological units not between native and non-native speakers, but between intermediate and advanced learners of English. They analyzed adjacent noun + noun, adjective + noun, and adverb + adjective co-occurrences, employing mutual information scores to gauge the associational strength of the two lexical items and t-scores to measure frequency. Their findings indicated that intermediate learners tend to overuse high-frequency co-occurrences (such as *hard work*) and underuse low-frequent but strongly-associated co-occurrences (such as *immortal souls*). However, it remains challenging to conclude that as learners progress to higher proficiency levels, they will utilize more strongly associated but low-frequency co-occurrences, as the learner texts used in this study were written by different groups of learners. Thus, it seems more reasonable to compare the usage of phraseological units by the same group of learners over a period of time during which they enhance their overall English proficiency levels.

Therefore, as a follow-up study, Bestgen and Granger (2018) tracked the phraseological development of the same learners over an extended period, utilizing the Longitudinal Database of Learner English. They examined 178 texts written by 89 French EFL learners, with each learner contributing an argumentative essay in their first year (T1) at university and another in their third year (T3) on the same topic. Their analysis focused on three types of adjacent two-word (bigram) co-occurrences: noun + noun, adjective + noun, and adverb + adjective. They employed t-scores to measure frequency and mutual information

to measure association. The results indicated a general tendency for T3 texts to contain fewer non-collocational bigrams (i.e., bigrams with an MI score of less than 3 and a t-score of less than 2) and fewer high-scoring t-score co-occurrences, but more high-scoring MI co-occurrences. This outcome resembled those obtained when applying the same method to cross-sectional data (e.g., Granger & Bestgen, 2014). This study thus stands among the first to track the phraseological development of EFL learners, laying the groundwork for discussions on phraseological complexity, as will be explored in Chapter 3.

Researchers exploring phraseological units in learner language have also uncovered that collocations are frequently implicated in errors as well. Nesselhauf (2005) utilized dictionaries, corpora, and input from native speaker informants to evaluate the acceptability of approximately 2,000 verb-noun collocations produced by German EFL learners. Her findings revealed that approximately one third of these collocations could be deemed unacceptable. Furthermore, she identified the most prevalent types of errors in verb-noun collocations, which primarily involved the incorrect selection of verbs (e.g., *make a try*). Other error types included prepositional errors (e.g., *discuss about*) and determiner errors (e.g., *eat a breakfast*).

Several studies have highlighted a greater usage of co-occurrence phraseological units, such as collocations and phrasal verbs, in higher-rated learner writing. For instance, Laufer and Waldman (2011) observed a statistically significant disparity in the number of verb-noun collocations used by advanced and elementary students. However, no significant differences were found between intermediate learners and either of the other two groups, indicating a relatively slow progression in this aspect of language proficiency. Furthermore, the researchers calculated the ratio of errors to well-formed collocations and found no correlation between this ratio and the students' proficiency levels. Approximately one third of all collocations were deemed erroneous, regardless of the learners' level. However, when the number of collocation errors was measured as a proportion of the total number of words, advanced and intermediate learners were found to produce significantly more deviant collocations than basic learners. Thewissen (2008) suggests that this paradox may indicate increased phraseological richness: higher-level learners attempt a broader range of lexical phrases, albeit with some errors. While there might not be substantial differences in the overall number of errors produced by more and less advanced learners, the types of errors do vary. The most advanced learners tend to produce a greater number of near hits compared to their lower intermediate counterparts, who generate a higher number of completely inaccurate expressions.

3.3.2 Recurrence

Recurrence, the second primary category of phraseological units, is defined as "the repetition of continuous strings of words of a specified length (e.g., bigrams, trigrams)" (Paquot & Granger, 2012). Biber et al. (1999) referred to this category of phraseological units as lexical bundles, which may encompass both grammatically complete (e.g., trigrams like *I don't know*) and incomplete strings (e.g., trigrams like *I think that*).

In the realm of learner corpus research, various researchers have delved into lexical bundles of differing lengths and in diverse contexts, rendering direct comparisons of results challenging. For instance, Reppen (2009) scrutinized the 20 most common three-word sequences in learner corpora, while Groom (2009) examined all two- to five-word lexical bundles with a minimum frequency of 10 occurrences per 250,000 words. Additionally, Chen and Baker (2010) investigated all four-word lexical bundles occurring at least 4 times in their learner corpus.

However, despite the complexities in comparing findings across studies, we can still discern some general trends. As previously mentioned in the co-occurrence section, learners of English often exhibit a narrower range of co-occurrences compared to native speakers, yet they demonstrate a broader utilization of recurrences. For instance, Ping (2009) found that Chinese EFL learners employed four times as many four-word lexical bundles as native speakers. Similarly, Bo and Shutang (2005) observed a higher frequency of three- to six-word bundles in the Chinese Learner English Corpus, with a particular emphasis on longer lexical bundles. In contrast to the usage of collocations, the overall frequency of lexical bundles tends to decrease as proficiency in the language advances (e.g., Reppen, 2009).

In addition to variation in terms of the quantity of lexical bundles utilized, learners have also been found to employ lexical bundles with distinct functional patterns compared to native speakers. This observation has emerged from studies comparing the discourse functions of lexical bundles in learner and native corpora. For instance, Chen and Baker (2010) conducted a comparison of four-word lexical bundles in a corpus of Chinese EFL learners' academic writing, a corpus of British student academic writing, and a corpus of expert writing. They employed the structural classification of lexical bundles provided in the Longman Grammar of Spoken and Written English (Biber et al., 1999) to categorize them. Additionally, they adopted the functional categorization of lexical bundles proposed by Biber and Barbieri (2007), classifying them into referential bundles (e.g., *at the same time*), stance bundles (e.g., *it could be argued that*), and discourse organizers (e.g., *on the one hand*). Their findings indicated that Chinese EFL learners made relatively limited use of lexical bundles with passive verb forms followed by a preposition (e.g., *can be used for*)

and quantifying bundles (e.g., *to a certain extent*). Moreover, they employed very few hedging expressions (e.g., *is considered to be*).

Most studies on lexical bundles in learner language have primarily focused on written communication, with relatively few investigating their usage in spoken language. One notable exception is De Cock (2004), who analyzed two- to six-word lexical bundles in the French section of the LINDSEI corpus (Gilquin, De Cock, and Granger, 1995) and compared their frequency and usage with a comparable native speaker corpus, i.e., Louvain Corpus of Native English Conversation (De Cock, 2004). The study revealed that learners' preferred lexical bundles were less interactive compared to those of native speakers and contained few markers of vagueness (e.g., *kind of*, *sort of*). Additionally, learners exhibited a tendency to favor some rather assertive lexical bundles, such as *yeah of course*, which might lead them to sound overly emphatic or even impolite. Overall, French learners were found to lack the habitual patterns of interaction and rapport-building with their conversational partners, as well as the ability to incorporate appropriate levels of imprecision and vagueness into their speech.

4
Studies of Phraseological Complexity

As reviewed in the previous chapter, a substantial body of research within the field of learner corpus research has concentrated on examining the usage of various types of phraseological units produced by EFL learners from diverse backgrounds, including variations in proficiency levels and first languages. These studies have recognized that phraseological units, such as collocations, have emerged as crucial indicators of EFL learners' linguistic proficiency. Despite that, phraseological competence has not received much attention from most language testers, as noted by Paquot (2018:29), who stated that "its development has not received the attention it deserves in the CEFR". To address this issue, Paquot (2019) proposed the concept of phraseological complexity, which connects L2 phraseology research with L2 complexity research. This construct, based on Ortega's (2003:124) definition, measures "the range of phraseological units that surface in language production and the degree of sophistication of such phraseological units", and was operationalized as phraseological diversity and phraseological sophistication. Paquot conducted a study using the Varieties of English for Specific Purposes dAtabase (VESPA; Paquot et al., 2022), a corpus of linguistics term papers written by French L2 EFL learners to explore phraseological complexity and evaluated three types of phraseological units that have been found to be challenging for L2 learners: adjectival modifiers (adjective + noun, e.g., *clear evidence*), adverbial modifiers (adverb + adjective, verb, or adverb, e.g., *statistically significant*), and direct objects (verb + noun, e.g., *make efforts*). These three syntactic relationships are abbreviated as *amod*, *advmod*, and *dobj*, respectively, according to Stanford typed dependency (De Marneffe & Manning, 2013).

Phraseological diversity, the first construct of phraseological complexity, was used to measure the breadth of learners' phraseological competence. Specifically, it represents the number of unique phraseological units to the total number of phraseological units, indicating the variety of different phraseological units used by learners. It was

operationalized as the root type-token ratio of the three types of syntactic relations, which represented the number of unique phraseological units to the total number of phraseological units, indicating the variety of different phraseological units used by learners.

Phraseological sophistication, the second construct of phraseological complexity, was used to measure the depth of learners' phraseological competence and referred to the ability of learners to use advanced and appropriate phraseological units that are specific to the topic and style of writing, rather than general everyday vocabulary. The study employed two methods to measure phraseological sophistication. The first method involved defining sophisticated phraseological units. In Paquot (2019), sophisticated phraseological units were defined as academic collocations and were operationalized as word combinations that appear in Ankermann and Chen's (2013) Academic Collocation List (ACL). The ACL includes 2,468 highly frequent and pedagogically relevant lexical collocations extracted from academic writing and was evaluated using a corpus-driven and expert-judged approach. Phraseological sophistication was measured as ratios of the number of sophisticated *amod*, *advmod*, and *dobj* tokens to the total number of *amod*, *advmod*, and *dobj*, both on a token-based and type-based level. The study also employed pointwise Mutual Information (PMI) for *amod*, *advmod*, and *dobj* dependencies, Three mean MI scores were calculated for each learner text based on all the different word pairs found in the *amod*, *advmod*, and *dobj* dependencies.

The result of phraseological diversity indicated that there was no significant difference in phraseological diversity across proficiency levels. The result of phraseological sophistication indicated that the PMI-involved phraseological sophistication exhibited a significant increase across proficiency levels. The ACL-involved phraseological sophistication exhibited a systematic ratio increase across proficiency levels of all dependencies. This demonstrates that as English EFL learners progress to higher proficiency levels, there is a systematic increase in the proportion of 'sophisticated collocations' (i.e., collocations included in the ACL), both type- and token-based, suggesting that the ACL-involved phraseological sophistication is an effective measure of learners' phraseological competence and can be used to evaluate their ability to use appropriate phraseological units in academic writing. Nevertheless, no statistical significance was found in any intergroup ratio increase.

For the statistical insignificance amid systematic ratio increase resulted from the ACL-involved phraseological sophistication, Paquot believed that the small size of the ACL along with the small coverage of the ACL in the source corpus is one of the causes. The ACL contains 2,496 collocations, and covers 1.4 % of total tokens in the source corpus. Compared to the coverage of other lists used in examining learner writing sophistication

(e.g., The Academic Word List, which covers approximately 10% of academic texts), this coverage rate is relatively small.

In addition to utilizing phraseological complexity to describe L2 performance at different proficiency levels, Paquot (2019) also compared phraseological complexity with traditional measures of syntactic and lexical complexity. The outcomes of the analysis revealed that while no significant difference in lexical or syntactic complexity across proficiency levels was observed, a marked increase in PMI-involved phraseological sophistication was observed as proficiency levels progressed, suggesting that the construct of phraseological complexity can be a useful index of L2 proficiency at the upper levels of proficiency.

Paquot (2019) has opened new avenues for future research on phraseological complexity in L2 settings, thereby creating endless possibilities for investigating L2 phraseology and complexity. Building on this foundation, Paquot et al. (2020) conducted a longitudinal study on phraseological complexity, with a specific focus on direct object constructions, using the Longitudinal Database of Learner English (Meunier, 2008; LONGDALE). The research investigated the impact of proficiency and time spent on learning English on the phraseological sophistication of verb + object relations in the writing of French EFL learners. The phraseological complexity was measured using mutual information at three different data collection points. It is worth noting that no ACL-derived phraseological sophistication method was used. The study employed a mixed-effects model to control for topic, time, and proficiency, and the results indicated that learner proficiency was a better predictor of phraseological complexity than the variable of time spend learning English.

Although the majority of the research on phraseological complexity discussed earlier centered on operationalizing it in relation to the written English of English EFL learners, there have been several attempts to expand the scope of phraseological complexity in a broader sense, starting with examining different languages and with the objective of investigating whether the phraseological complexity measures which were originally developed by Paquot (2018, 2019) for L2 English, would also be predictive of proficiency in other languages. Vandeweerd et al (2021) conducted replication research of Paquot (2019) in a corpus of L2 French argumentative essays, in which phraseological complexity was operationalized as the diversity (root type-token ratio; RTTR) and sophistication (pointwise mutual information; PMI) of three types of grammatical dependencies: adjectival modifiers, adverbial modifiers and direct objects. Utilizing an L2 French corpus, the Leerdercorpus Frans (Vanderbauwhede, 2012), Vandeweerd operationalized the two approaches of phraseological complexity and rephrase: observed a significant increase in the mean PMI of direct objects and the RTTR of adjectival modifiers across proficiency levels. In addition, phraseological complexity's relationship between other traditional

complexity measures, i.e., morphological, lexical, and syntactic complexity, were investigated. Contrary to Paquot (2019), this research found that the most important predictors of learners' L2 French performance also included traditional measures of complexity. It is worth noting that in this phraseological complexity replication study conducted on L2 French, no measure of phraseological sophistication based on a list of sophisticated academic collocations was used.

Similarly, Rubin et al (2021) conducted a replication study of Paquot (2019) in a corpus of L2 Dutch, i.e., the written portion of the Certificaat Nederlands als Vreemde Taal (CNaVT; Certificate of Dutch as a Foreign Language), revealing that the measures of L2 Dutch phraseological complexity contribute substantially to the prediction of Dutch learners' proficiency assessing regression model, suggesting that "complexity measures tapping into phraseological phenomena will help to better model the full range of learner proficiency" (p. 120). It is also worth noting that in this replication study conducted on L2 Dutch, no measure of phraseological sophistication based on a list of sophisticated academic collocations was used.

Despite the significant role of phraseology in L2 complexity research, previous studies were mostly cross-sectional and focused only on the written mode. To address these limitations, Vandeweerd et al (2022) conducted a longitudinal and multitask-based research of phraseological complexity on L2 French. The study involved recruiting a specific cohort of L2 French learners who were instructed to complete a task consisting of a written argumentative essay, a semi-guided oral interview, and a picture-based oral narrative at three distinct time points over a duration of 21 months. Syntactic co-occurrences of adjectival modifiers and direct objects were extracted to analyze phraseological complexity. Results indicated that phraseological complexity performed differently across oral and written tasks, and no significant increase in phraseological diversity was observed over the 21-month period. There was a slight increase in phraseological sophistication, but only for direct objects. These findings emphasized the importance of task characteristics in measuring phraseological complexity.

To date, the only investigation of phraseological sophistication through the lens of an academic collocation list stems from Paquot (2019). However, subsequent explorations into the domain of phraseological sophistication within the contexts of L2 French and L2 Dutch have not used the academic collocation list-based methodology introduced by Paquot (2019). This departure can be attributed to the absence of dedicated lists for this analytical framework. It is noteworthy that the Academic Collocation List (Ackermann & Chen, 2013) employed in Paquot's study was designed not for research purposes, but rather as a pedagogical resource catering to the needs of English for Academic Purposes (EAP) students and educators. Consequently, a research gap emerges — a lack of a purpose-built academic collocation list catering to research needs, particularly in its

application to approximating L2 EFL writings in studies investigating phraseological sophistication.

While several studies on phraseological sophistication have not yet employed the sophisticated phraseological units-based method, this does not imply inefficacy. Indeed, within the domain of L2 complexity research, there has long been a tradition of utilizing sophisticated lists — whether they pertain to sophisticated words or sophisticated phraseological units — to investigate linguistic complexities across various dimensions. For instance, Douglas (2013) examined the vocabulary usage of novice university students using the Academic Word List (Coxhead, 2000), given the low-frequency nature of academic words in everyday English. The study concluded that the use of academic words correlates with higher-quality academic writing. Similarly, Kyle and Grossley (2015) developed the Tool for the Automatic Analysis of Lexical Sophistication (TAALES), wherein a significant component is the incorporation of the Academic Word List (Coxhead, 2000) as sophisticated words. They applied this tool to a corpus of spoken learner language and found that the coverage of the AWL is more pronounced in learner texts of higher L2 speaking proficiency levels. These examples illustrate the effectiveness of employing sophisticated lists in L2 sophistication research, thereby inspiring the utilization of sophisticated phraseological unit lists to examine phraseological sophistication.

In response to concerns regarding the size and coverage limitations of the Academic Collocation List (Ackermann & Chen, 2013), the authors developed the New Academic Collocation List (NACL; Shen, 2023) as part of a dissertation project in 2023. The NACL is designed for research purposes, especially for examining the phraseological competence of writings by L2 EFL learners. Extracted from the selected 9 disciplines of the Louvain Corpus of Research Articles which contains approximately 18 million tokens covering a wide range of disciplines of social sciences and humanities, the new list focuses on three syntactic dependencies (i.e., adjectival modifiers, adverbial modifiers, and direct object) and the constructs of frequency and dispersion due to a lack of measuring phraseological sophistication by these constructs. Finally, the New Academic Collocation List (NACL) contains 3,756 collocations (1,497 amod collocations, 1361 advmod collocations, and 898 dobj collocations). Compared to the Academic Collocation List (ACL; Ackermann & Chen, 2013), the NACL has the following two most prominent features. First, the NACL is a larger collocation list. The NACL contains 3,756 collocations whereas the ACL contains 2,496 collocations. Second, the NACL contains a larger coverage. In the LOCRA where the NACL is extracted from, the coverage of the NACL is three times that of the ACL. By far, as far as the author knows, the NACL is the best resource for the study of phraseological sophistication. Table 4.1 summarizes the compositional differences between the NACL and the ACL.

	NACL	ACL	Examples
amod	1,497 (39.9%)	1,835 (74.3%)	*high level*
advmod	1,361 (37.2%)	294 (11.9%)	*significantly differ*
dobj	898 (23.9%)	340 (13.8%)	*have effect*
Total	3,756 (100%)	2,468 (100%)	

Table 4.1 Compositional Difference between the NACL and the ACL

Consequently, a research gap emerges — a compelling need to replicate the study of Paquot (2019) on phraseological sophistication using the New Academic Collocation List. This study aims to answer the following research question:

- RQ1: To what extent performs the New Academic Collocation List (Shen, 2023) in terms of assessing phraseological sophistication of leaner writing?
- RQ2: To what extent differs the performance of the Academic Collocation List (Ackermann & Chen, 2013) from the New Academic Collocation List (Shen, 2023) in terms of assessing phraseological sophistication of leaner writing?

The hypotheses for the research questions are as follows. Firstly, utilizing the NACL as a benchmark for sophisticated phraseological units, it is anticipated that the proportion of such units will notably increase as proficiency levels rise. Secondly, in comparison to the ACL, the NACL is posited to be a more effective tool for evaluating the phraseological sophistication in L2 writings.

5
Data

To replicate the Paquot (2019) study on phraseological sophistication, the International Corpus of Learner English (ICLE; Granger et al., 2020) was utilized. The ICLE is an extensive, multi-national database containing written and spoken English produced by learners from various native language backgrounds. This corpus is designed to provide researchers with a comprehensive source of data for investigating second language acquisition and learner English.

To assess phraseological sophistication, the proficiency ratings of learner texts are crucial. Paquot et al. (2020) demonstrated that proficiency information is a better predictor of phraseological complexity than the actual point in time when the essay was written. Most studies on phraseological complexity (e.g., Paquot, 2018, 2019; Vandeweerd et al., 2022) use the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001), which outlines six proficiency levels: A1, A2, B1, B2, C1, and C2 (from lowest to highest).

However, the ICLE does not comprehensively include such CEFR ratings. Researchers often request learner texts from the ICLE and recruit professional raters to assess these texts for their research purposes (e.g., Granger and Bestgen, 2014; Bestgen and Granger, 2018). Fortunately, the Crowdsourcing Language Assessment Project (CLAP), currently underway at the Centre for Corpus Linguistics at the Catholic University of Louvain, provided this master dissertation with a subset comprising learner texts that were rated by professional raters according to CEFR.

A total of 246 learner texts, spanning B1, B2, C1, and C2 proficiency levels, were used. Detailed information about these learner texts is presented in Table 5.1.

	# texts	# tokens						
		Total	Mean	SD	Median	IQR	Min	Max
B1	24	13,108	546.2	84.1	542.5	199.0	423	754
B2	96	56,169	585.1	103.4	561.5	154.0	406	812
C1	71	43,589	614.0	112.3	596.0	189.5	405	832
C2	55	35,188	639.8	88.4	626.0	115.5	440	815
Total	246	148,054	601.8	105.0	589.0	163.8	405	832

Table 5.1 Learner texts information

The learner texts consist of argumentative essays written by EFL learners from 30 countries, representing 26 different mother tongues. The majority of these learners are between 19 and 26 years old, with an average age of 22 years (minimum age 17, maximum age 51). The total number of tokens in these learner texts is 148,054, with each text averaging approximately 602 tokens.

	# Texts	Total # words	Means
B2	25	86,472	3,588
C1	62	216,283	3,488
C2	11	33,994	3,090
Total	98	336,749	3,436

Table 5.2 Learner texts used in Paquot (2019)

In the ACL-based phraseological sophistication study of Paquot (2019), the author used a collection of 98 research articles written by French EFL learners by three CEFR levels, i.e., B2, C1, and C2. A detailed description of the learner texts is shown in Table 5.2. In comparison, the learner texts in this replication study are different in that they are argumentative essays, whereas Paquot's study focused on research articles. Additionally, the learner texts in this study are comparatively shorter, with a mean length of 602 words, compared to the mean length of 3,436 words for the research articles in Paquot (2019).

6
Methodology

Similar to the approach in Paquot (2019), the phraseological sophistication in this study is examined through word combinations used in three grammatical relationships: adjectival modifiers (amod: adjective + noun), adverbial modifiers (advmod: adverb + adjective, adverb, or verb), and direct objects (dobj: verb + noun). These relationships are analyzed using Stanford typed dependencies. As illustrated in examples (1), (2), and (3), a Stanford typed dependency represents a binary grammatical relation between a governor and a dependent (see De Marneffe and Manning, 2013).

(1)	amod	adjectival modifier	
		*This is a **big apple***.	amod (apple +NN, big +JJ)
(2)	advmod	adverbial modifier	
		*This is **very big** apple*.	advmod (big +JJ, very +RB)
		*I **run quickly***.	advmod (run +VBZ, quickly +RB)
		*I run **more quickly***.	advmod (quickly +RB, more +RB)
(3)	dobj	direct object	
		*I **eat** an **apple***.	dobj (apple +NN, eat +VV)

The concept of phraseological complexity in this study aligns with the definition used in Paquot (2019). Phraseological complexity measures "the range of phraseological units that surface in language production and the degree of sophistication of such phraseological units" (Paquot, 2019, p. 124). Paquot (2019) defined phraseological complexity through two constructs: phraseological diversity and phraseological sophistication.

Phraseological diversity, the first construct, measures the breadth of learners' phraseological competence. It is operationalized as the Root Type-token Ratio (RTTR) of the three above mentioned types of syntactic relations. Phraseological sophistication, the second construct, measures the depth of learners' phraseological competence. It refers to the ability of learners to use advanced and appropriate phraseological units specific to the

topic and style of writing, rather than general everyday vocabulary. Paquot (2019) used two methods to measure phraseological sophistication. The first method involved calculating the MI scores of the aforementioned phraseological units. The second method identified phraseological units from the Academic Collocation List as sophisticated and examined the proportion of token- and type-based phraseological units in learner texts across different proficiency levels.

In this dissertation, only the method based on sophisticated phraseological units was investigated to measure phraseological sophistication. Since both research questions of this study focus on using sophisticated phraseological units to assess phraseological sophistication, phraseological diversity and the mutual information-based method for phraseological sophistication were not included in this study.

Indexes	Phraseological sophistication	Formula
LS1amod	Phraseological sophistication-1(amod)	Namods (NACL) / Namod
LS1advmod	Phraseological sophistication-1(advmod)	Nadvmods (NACL)/ Nadvmod
LS1dobj	Phraseological sophistication-1(dobj)	Ndobjs (NACL) / Ndobj
LS2amod	Phraseological sophistication-2(amod)	Tamods (NACL) / Tamod
LS2advmod	Phraseological sophistication-2(advmod)	Tadvmods (NACL) / Tadvmod
LS2dobj	Phraseological sophistication-2(dobj)	Tdobjs (NACL) / Tdobj

Table 6.1 Measures of the NACL-based phraseological sophistication

To address the first research question, phraseological units in the New Academic Collocation List (NACL) were defined as sophisticated phraseological units. The New Academic Collocation List was developed by Shen (2023) as part of his graduation project. The NACL comprises 3,756 phraseological units encompassing amod, advmod, and dobj syntactic dependencies. With a coverage of 15.21% in the source corpus, the NACL was regarded as an ideal tool for evaluating phraseological sophistication. Table 6.1 lists the six measures of phraseological sophistication based on the NACL.

As can be observed in Table 6.1, two numbers are used to differentiate the measures: 1 and 2. LS1amod, LS1advmod, and LS1dobj are token-based ratios of the number of sophisticated adjectival modifiers (amod), adverbial modifiers (advmod), and direct object (dobj) tokens (i.e., tokens that appear in the NACL) to the total number of amod, advmod, and dobj tokens, respectively. Similarly, LS2amod, LS2advmod, and LS2dobj are type-based ratios of the number of sophisticated adjectival modifiers (amod), adverbial modifiers (advmod), and direct object (dobj) types (i.e., the ones that appear in the NACL) to the total number of amod, advmod, and dobj types, respectively. To calculate LS1amod for a learner text, all the amod phraseological units are extracted from the text. Each

extracted amod collocation is then checked against the NACL, regardless of how many times it is repeated throughout a text. For instance, if the phraseological unit *play role* appears twice in the learner text and is listed in the NACL, these two instances are counted as two sophisticated phraseological units. LS1amod is then calculated by dividing the total number of sophisticated phraseological units by the total number of amod phraseological units extracted from the text. The calculations for LS1advmod and LS1dobj follow the same pattern.

In addition to token-based measures, three type-based measures are also incorporated in this study in alignment with Paquot (2019): LS2amod, LS2advmod, and LS2dobj. Unlike token-based measures, these type-based measures only consider unique phraseological units. To calculate LS2amod for a learner text, all amod phraseological units are first extracted, and any repeated units are counted as a single unit. Combined with unique phraseological units, the repeated ones are considered only once. Each of these unique phraseological units is then checked against the NACL. LS2amod is calculated by dividing the number of sophisticated phraseological units (non-repeated) by the total number of unique phraseological units. The calculations for LS2advmod and LS2dobj follow the same pattern.

To address the second research question, this study also incorporated the analysis of phraseological sophistication using the Academic Collocation List (ACL; Ackermann & Chen, 2013). The procedures were identical to those used with the NACL, except that the sophisticated phraseological units were defined according to the ACL. Other measures remained the same.

A crucial step in this analysis was the extraction of phraseological units from learner texts. The author utilized the spaCy package from Python (Honnibal and Montani, 2017) with the "en_core_web_trf" pipeline to automatically tokenize, lemmatize, part-of-speech (POS) tag, and syntactically parse each learner text. spaCy is an advanced natural language processing library in Python that uses a pipeline of components to process text. The pipeline begins with tokenization, which splits text into individual tokens such as words and punctuation. Following tokenization, lemmatization then reduces tokens to their base or root forms, enabling normalization of different word forms. Part-of-speech (POS) tagging assigns grammatical categories to each token, helping in syntactic and semantic analysis. Parsing constructs a syntactic structure of the sentence, identifying dependencies and relationships between tokens. This comprehensive pipeline allows for efficient and accurate text processing and analysis. Using the POS and syntactic attribute information provided by the spaCy library, the amod, advmod, and dobj phraseological units were extracted from each learner text.

To ascertain the accuracy and reliability of the automatic POS tagging and syntactic parsing, a pilot study was undertaken. Specifically, a sample of learner text (proficiency

level B2) was subjected to manual part-of-speech tagging and parsing. This approach aimed to evaluate the performance of the automated processes used, ensuring the reliability of the data for subsequent analysis and interpretation. The pilot study aimed to evaluate the reliability of using spaCy for POS tagging and parsing in a corpus linguistics analysis. According to the result of the automatic extraction, this sample consists of 355 words and 266 dependencies. To validate the accuracy of the results, a manual verification was conducted, involving manual POS tagging and parsing of the text. The manual inspection revealed 14 errors in the POS tagging process, indicating an accuracy rate of approximately 96%. Moreover, 12 mistakes were identified in the parsing, resulting in a parsing accuracy of around 95%. These findings demonstrate that spaCy provides a reliable tool for POS tagging and syntactic parsing in corpus linguistics analysis, with a high level of accuracy and overall performance.

After calculating the NACL- and ACL-based phraseological sophistication indexes for each learner across the four proficiency levels (B1, B2, C1, and C2), the distributions were systematically checked for normality using the Shapiro-Wilk test. For frequency counts that were normally distributed, comparisons were made using ANOVAs followed by Tukey contrasts. If the frequency counts were not normally distributed, Kruskal-Wallis rank sum tests were used instead. The significance level for all statistical tests was set at 0.05. All statistical analyses were conducted using R. Table 6.2 summarizes the different steps of the learner texts preprocessing and statistical computing workflow.

	Tools	Learner texts analyzed
1. Lemmatization	spaCy	246 learner texts from the Internation Corpus of Learner English
2. Part-of-speech (POS) tagging		
3. Syntactic parsing		
4. Extraction of dependencies		
5. Sophisticated phraseological unit identification	In-house Python programs	
6. Statistical analysis	R	

Figure 6.2 Learner texts preprocessing and statistical computing workflow.

7
Results

This chapter presents the results of the phraseological sophistication studies based on sophisticated phraseological units. It is divided into three sections: Section 7.1 provides an overall analysis of the phraseological units from the three dependency relationships (i.e., amod, advmod, and dobj) extracted from the learner texts. Section 7.2 reports the results of the phraseological study based on the New Academic Collocation List (NACL), addressing the first research question. Section 7.3 presents the results of the phraseological study based on the Academic Collocation List (ACL; Ackermann & Chen, 2013), addressing the second research question.

7.1 Extracted phraseological units' analysis

The first step in analyzing the total counts of phraseological units from the three aforementioned syntactic relationships (i.e., amod, advmod, and dobj) was to check for normality. Normality was assessed using the Shapiro-Wilk test, with the significance level set at 0.05. The results are presented in Table 7.1.

	Shapiro-Wilk (W)	p-value
token-amod	0.9879	0.0360
token-advmod	0.9785	0.0008
token-dobj	0.9865	0.0199
type-amod	0.9883	0.0422
type-advmod	0.9785	0.0009
type-dobj	0.9882	0.0409

Table 7.1 Shapiro-Wilk tests of the total phraseological unit counts

According to Table 7.1, the token- and type-based distributions of the total phraseological unit counts for the three syntactic relationships (i.e., amod, advmod, and dobj) did not follow normal distributions. Therefore, the median and interquartile range (IQR) were used to report the central tendency and dispersion of these distributions. The results are listed in Table 7.2. Additionally, Figure 7.1 shows the trend of median number of different types of phraseological units across proficiency levels.

	B1		B2		C1		C2	
	Median	IQR	Median	IQR	Median	IQR	Median	IQR
token-amod	28.5	14.5	31	15.5	37	17	37	13
token-advmod	17	8	20	9	22	10	23	10
token-dobj	26	10.5	23	11	23	9	23	10.5
type-amod	25.5	13	27	15	32	15	33	13.5
type-advmod	15.5	7.25	19.5	8.25	22	8	23	9
type-dobj	22.5	7.25	21	8.25	22	9	22	9

Table 7.2 Overview of total counts of phraseological units produced by learners

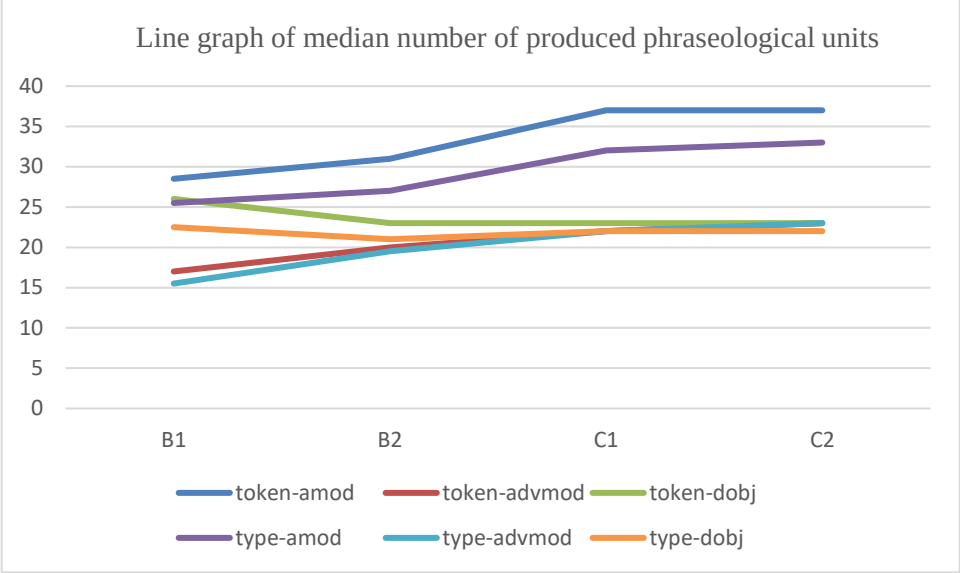

Figure 7.1 Line graph of median number of produced phraseological units

Table 7.2 presents the results of phraseological sophistication for token-based measures (amod, advmod, and dobj). Regarding token-based adjectival modifier phraseological units (row 1), a clear pattern emerges in terms of the median. As the proficiency levels of the learner texts increased from B1 to C2, the median number of phraseological units produced by learners also increased. Specifically, learners at the B1 proficiency level produced a median of 28.5 amod phraseological units per text. This number increased to 31 amod phraseological units per text for the B2 proficiency level, representing an increase of 2.5 units per text. The median then rose to 37 amod phraseological units per text at the C1 level, showing an increase of 6 units per text, and remained at 37 units per text for learners at the C2 level, with no further increase. Overall, the median number of amod phraseological units per text rose from 28.5 at the B1 level to 37 at the C2 level, an overall increase of 8.5 units. In terms of variance, the interquartile range deviation for the four groups of amod phraseological units was 14.5, 15.5, 17, and 13, respectively, remaining approximately around 15.

Similarly, the medians of advmod phraseological units also exhibited an upward trend as proficiency levels increase. Specifically, the median number of token-based advmod phraseological units produced by learners at the B1 proficiency level was 17. This median rose to 20 for learners at the B2 level, indicating an increase of 3 units, then to 22 for learners at the C1 level, reflecting an increase of 2 units, and finally to 23 for learners at the C2 level, showing an increase of 1 unit. Overall, the median number of advmod phraseological units increased from 17 at the B1 level to 23 at the C2 level, an overall increase of 6 units. Compared with token-amod, token-advmod showed a relatively smaller increase in the median number of phraseological units produced from B1 to C2. In terms of variance, the interquartile range for the four groups of advmod phraseological

units was 8, 9, 10, and 10, respectively, remaining approximately around 9. Compared to LS1amod, LS1advmod exhibited a relatively smaller variance.

Unlike the token-based amod and advmod measures, which showed a general trend of increasing phraseological units with higher proficiency levels, the token-based dobj measure yielded different results. The median number of dobj phraseological units produced by learners at the B1 level was 26. Interestingly, this median decreased to 23 for learners at the B2 level, showing a reduction of 3 units per text. This median of 23 remained stable across learners at the B2, C1, and C2 levels. Therefore, learners at the B1 level produced a relatively higher median number of dobj phraseological units, while learners at higher proficiency levels (B2, C1, and C2) maintained a smaller and steady median of 23 units. In terms of variance, the interquartile range for the four groups of dobj phraseological units was 10.5, 11, 9, and 10.5, respectively, averaging around 10.25. Compared to token-based amod and advmod measures, the token-based dobj measure exhibited variance similar to token-based advmod and smaller than token-based amod.

Regarding type-based phraseological extraction, type-based amod, type-based advmod, and type-based dobj exhibited patterns similar to their corresponding token-based measures. For type-based amod, the median number of phraseological unit types produced increased gradually from 25.5 for B1 learners to 33 for C2 learners, showing an increase of 7.5 units per text, which paralleled the increase observed in its token-based counterpart (8.5 units). In terms of type-based advmod, the median number of phraseological unit types also increased gradually from 15.5 at the B1 level to 22 at the C2 level, reflecting an increase of 7.5 units, mirroring the pattern seen in type-based amod. For type-based dobj, the median number of phraseological unit types produced by B1 learners was 22.5. It decreased to 21 for B2 learners and then remains steady at 22 for both C1 and C2 learners.

Here is a brief summary of the above analysis. For token-based amod phraseological units:

- Median: Increased from 28.5 at B1 to 31 at B2, then to 37 at C1, and remained at 37 at C2, showing an overall increase of 8.5 units.
- Interquartile Range (IQR): The IQR for amod units was 14.5, 15.5, 17, and 13 for B1, B2, C1, and C2, respectively, averaging around 15.

For token-based advmod phraseological units:

- Median: Rose from 17 at B1 to 20 at B2, then to 22 at C1, and finally to 23 at C2, showing an overall increase of 6 units.
- IQR: The IQR for advmod units was 8, 9, 10, and 10 for B1, B2, C1, and C2, respectively, averaging around 9, indicating smaller variance compared to amod.

For token-based dobj phraseological units:

- Median: Decreased from 26 at B1 to 23 at B2, and remained stable at 23 for C1 and C2.
- IQR: The IQR for dobj units was 10.5, 11, 9, and 10.5 for B1, B2, C1, and C2, respectively, averaging around 10.25, similar to advmod but smaller than amod.

For type-based measures:

- Type-based amod: The median number of phraseological unit types increased from 25.5 at B1 to 33 at C2, an increase of 7.5 units, similar to the token-based increase of 8.5 units.
- Type-based advmod: The median number of phraseological unit types rose from 15.5 at B1 to 22 at C2, also an increase of 7.5 units, mirroring the pattern of type-based amod.
- Type-based dobj: The median number of phraseological unit types decreased from 22.5 at B1 to 21 at B2, then remained steady at 22 for both C1 and C2.

In summary, the median number of both token-based and type-based amod and advmod phraseological units increased with proficiency levels from B2 to C2. Conversely, for dobj units, there was a decrease from B1 to B2, followed by stability at higher proficiency levels, with no significant increase or decrease observed.

7.2 NACL-based phraseological sophistication study

To assess the normality of the sophistication index, which divides the number of phraseological units appearing in the NACL by the total number of phraseological units in each learner text, the Shapiro-Wilk tests were conducted. Table 7.3 presents the results of the Shapiro-Wilk tests for NACL-based phraseological sophistication measures, with a significance level set at 0.05.

	Shapiro-Wilk (W)	p-value
LS1amod	0.9669	0.0000**
LS1advmod	0.9746	0.0002
LS1dobj	0.9720	0.0000**
LS2amod	0.763	0.0004
LS2advmod	0.9763	0.0004
LS2dobj	0.9761	0.0004

Table 7.3 Shapiro-Wilk tests of NACL-based phraseological sophistication measures

According to Table 7.3, a series of Shapiro-Wilk tests of the six distributions of phraseological sophistication indexes showed a p-value of 0.0000**, 0.0002, 0.0000**, 0.0004, 0.0004, and 0.0004, respectively, indicating that not a single distribution was normal. Therefore, Kruskal-Wallis rank sum tests were conducted to compare the phraseological sophistication indexes across various proficiency levels. The significance level was set at 0.05. The results of these tests are presented in Table 7.4. Similarly, Figure 7.2 shows the trend of median NACL-based phraseological sophistication indexes across proficiency levels.

	B1		B2		C1		C2		Between-group
	Median	IQR	Median	IQR	Median	IQR	Median	IQR	comparisons
LS1amod	0.20	0.13	0.18	0.11	0.17	0.10	0.16	0.11	$KWX^2 = 1.80$; p = 0.61
LS1advmod	0.15	0.19	0.13	0.14	0.14	0.12	0.15	0.11	$KWX^2 = 5.83$; p = 0.12
LS1dobj	0.09	0.11	0.14	0.13	0.13	0.11	0.14	0.15	$KWX^2 = 4.48$; p = 0.21
LS2amod	0.19	0.15	0.19	0.10	0.17	0.12	0.16	0.10	$KWX^2 = 3.14$; p = 0.37
LS2advmod	0.15	0.12	0.12	0.13	0.14	0.11	0.16	0.12	$KWX^2 = 5.91$; p = 0.11
LS2dobj	0.10	0.06	0.14	0.13	0.13	0.11	0.14	0.12	$KWX^2 = 5.47$; p = 0.14

Table 7.4 Measures of phraseological sophistication based on the New Academic Collocation List

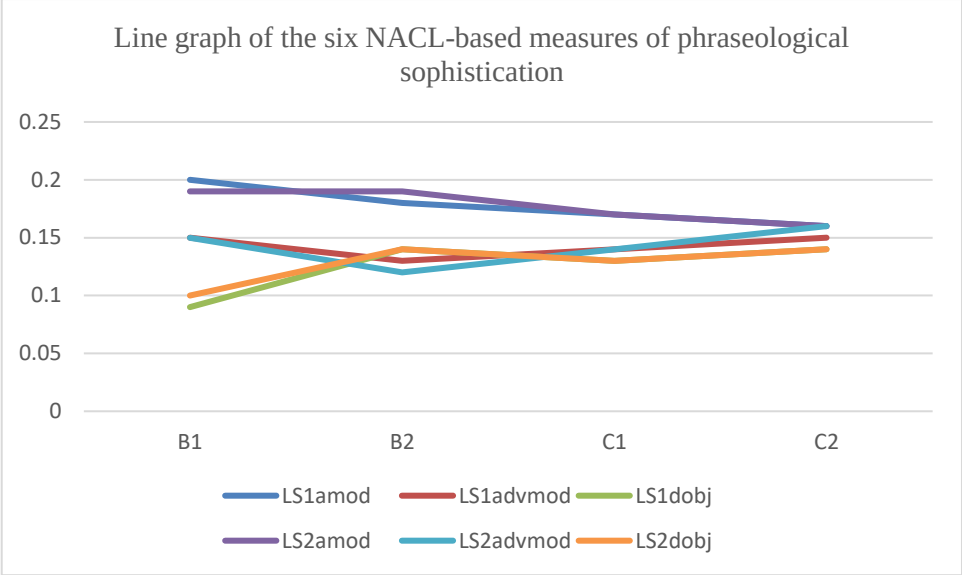

Figure 7.2 Line graph of the six NACL-based measures of phraseological sophistication

Table 7.4 presents the results of phraseological sophistication based on the NACL for both token- and type-based measures. The trends are shown in the line graph in Figure 7.2. For LS1amod, the median phraseological sophistication index started at 0.20 for learners at the B1 level. It decreased slightly to 0.18 for B2 learners, then further to 0.17 for C1 learners, and finally to 0.16 for C2 learners. The Kruskal-Wallis test yielded a chi-squared value of 5.83 with a p-value of 0.61, indicating no statistically significant differences across proficiency levels. For LS1advmod, the median phraseological index started at 0.15 for B1 learners. It decreased slightly to 0.13 for B2 learners, remained steady at 0.13 for C1 learners, and increased slightly to 0.14 for C2 learners. The Kruskal-Wallis test resulted in a chi-squared value of 3.14 with a p-value of 0.12, also showing no statistical significance across proficiency levels. For LS1dobj, the median phraseological sophistication index began at 0.09 for B1 learners, increased to 0.14 for B2 learners, decreased slightly to 0.13 for C1 learners, and then increased again to 0.14 for C2 learners. The Kruskal-Wallis test yielded a chi-squared value of 4.48 with a p-value of 0.21, indicating no statistically significant differences across proficiency levels. In summary, the token-based measures (LS1amod, LS1advmod, LS1dobj) did not show statistically significant increases across proficiency levels based on the Kruskal-Wallis tests.

The type-based measures in Table 7.4 exhibited patterns similar to their token-based counterparts. For LS2amod, the median phraseological sophistication index started at 0.19 for B1 learners, remained steady at 0.19 for B2 learners, decreased to 0.17 for C1 learners, and slightly decreased further to 0.16 for C2 learners. The Kruskal-Wallis test resulted in a p-value of 0.37, indicating no statistically significant differences across proficiency levels. For LS2advmod, the median phraseological sophistication index began at 0.15 for B1 learners, decreased to 0.12 for B2 learners, increased to 0.14 for C1 learners, and further

increased to 0.16 for C2 learners. The Kruskal-Wallis test yielded a p-value of 0.11, suggesting no statistical significance across proficiency levels. For LS2dobj, the median phraseological sophistication index started at 0.1 for B1 learners, increased to 0.14 for B2 learners, decreased slightly to 0.13 for C1 learners, and then increases again to 0.14 for C2 learners. The Kruskal-Wallis test resulted in no statistically significant differences across proficiency levels with a p-value of 0.21. In summary, similar to the token-based measures, the type-based measures (LS2amod, LS2advmod, LS2dobj) did not show statistically significant increases across proficiency levels based on the Kruskal-Wallis tests.

Here is a brief summary of the above analysis. For token-based measures:

- LS1amod: The median phraseological sophistication index decreased from 0.20 at B1 to 0.18 at B2, then to 0.17 at C1, and finally to 0.16 at C2. The Kruskal-Wallis test showed no significant differences ($\chi^2 = 5.83$, $p = 0.61$).
- LS1advmod: The median index started at 0.15 for B1, dropped to 0.13 at B2, remained at 0.13 for C1, and slightly increased to 0.14 at C2. The Kruskal-Wallis test also showed no significant differences ($\chi^2 = 3.14$, $p = 0.12$).
- LS1dobj: The median index increased from 0.09 at B1 to 0.14 at B2, decreased to 0.13 at C1, and increased again to 0.14 at C2. The Kruskal-Wallis test indicated no significant differences ($\chi^2 = 4.48$, $p = 0.21$).

For type-based measures:

- LS2amod: The median index remained at 0.19 from B1 to B2, then decreased to 0.17 at C1 and to 0.16 at C2. The Kruskal-Wallis test showed no significant differences ($p = 0.37$).
- LS2advmod: The median index started at 0.15 for B1, dropped to 0.12 at B2, increased to 0.14 at C1, and further increased to 0.16 at C2. The Kruskal-Wallis test indicated no significant differences ($p = 0.11$).
- LS2dobj: The median index increased from 0.1 at B1 to 0.14 at B2, decreased slightly to 0.13 at C1, and increased again to 0.14 at C2. The Kruskal-Wallis test showed no significant differences ($p = 0.21$).

In summary, none of the token-based measures (LS1amod, LS1advmod, LS1dobj) showed statistically significant increases across proficiency levels. Similar to the token-based measures, the type-based measures (LS2amod, LS2advmod, LS2dobj) did not show statistically significant increases across proficiency levels.

7.3 ACL-based phraseological sophistication study

To compare how the NACL and the ACL perform in terms of assessing phraseological sophistication of L2 writing (research question 2), this study also used the ACL to assess phraseological sophistication of selected learner texts. To evaluate the normality of the ACL-based phraseological sophistication indexes, which divides the number of phraseological units appearing in the ACL by the total number of phraseological units in each learner text, Shapiro-Wilk tests were conducted. Table 7.5 presents the results of these tests, with a significance level set at 0.05.

	Shapiro-Wilk (W)	p-value
LS1amod	0.6917	0.0000**
LS1advmod	0.4123	0.0000**
LS1dobj	0.5350	0.0000**
LS2amod	0.8470	0.0000**
LS2advmod	0.2952	0.0000**
LS2dobj	0.5524	0.0000**

Table 7.5 Shapiro-Wilk tests of ACL-based phraseological sophistication measures

According to Table 7.5, the Shapiro-Wilk tests' results of the six ACL-based measures of phraseological sophistication showed a series of p-values which was smaller than 0.0001, indicating none of the ACL-based phraseological sophistication measures, similar to NACL-based phraseological sophistication measures, exhibited a normal distribution. Consequently, Kruskal-Wallis rank sum tests were conducted to compare the phraseological sophistication indexes across various proficiency levels. The significance level was set at 0.05. The results of these tests are presented in Table 7.6.

	B1		B2		C1		C2		Between-group comparisons
	Median	IQR	Median	IQR	Median	IQR	Median	IQR	
LS1amod	0	0.037	0.028	0.050	0.026	0.047	0.032	0.067	$KWX^2 = 3.2$; $p = 0.37$
LS1advmod	0	0	0	0	0	0	0	0	$KWX^2 = 4.51$; $p = 0.21$
LS1dobj	0	0.031	0	0	0	0	0	0	$KWX^2 = 1.87$; $p = 0.60$
LS2amod	0	0.042	0.028	0.050	0.026	0.046	0.32	0.062	$KWX^2 = 3.12$; $p = 0.37$
LS2advmod	0	0	0	0	0	0	0	0	$KWX^2 = 4.72$; $p = 0.19$
LS2dobj	0	0.035	0	0	0	0	0	0	$KWX^2 = 2.62$; $p = 0.45$

Table 7.6 Measures of phraseological sophistication based on the Academic Collocation List

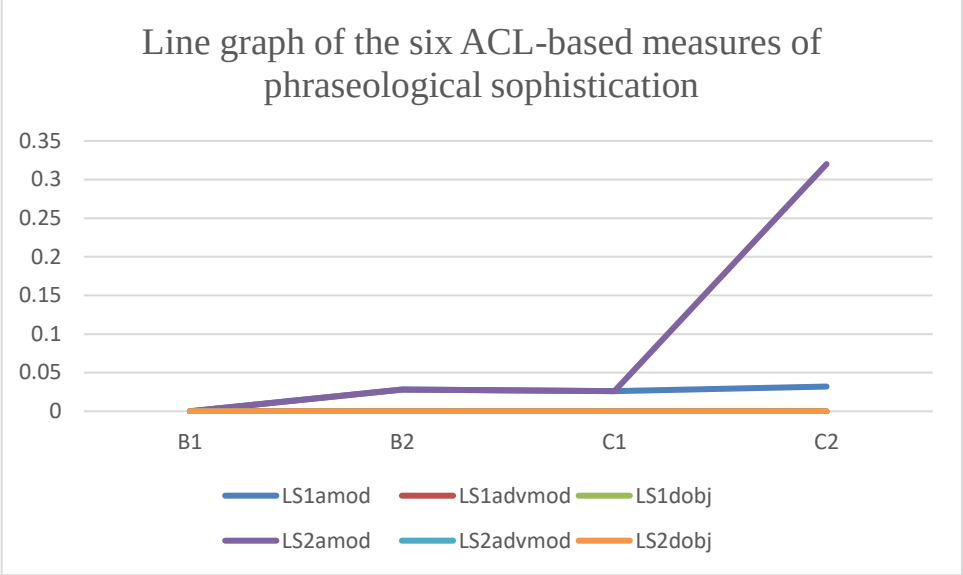

Figure 7.3 Line graph of the six NACL-based measures of phraseological sophistication

Table 7.6 presents the ACL-based results of phraseological sophistication measures. The trends are shown in the line graph in Figure 7.3. Table 7.6 reveals that advmod and dobj phraseological units showed a median portion of 0 in terms of appearance in ACL. Specifically, for LS1amod, the median phraseological sophistication index started at 0 for B1 learners, increases to 0.028 for B2 learners, decreased to 0.026 for C1 learners, and slightly rose to 0.032 for C2 learners. Conversely, LS1advmod and LS1dobj consistently showed a median index of 0 across all proficiency levels. Kruskal-Wallis tests indicated non-statistical significance with p-values of 0.37 for LS1amod, 0.21 for LS1advmod, and 0.60 for LS1dobj, suggesting no significant differences in phraseological sophistication indexes across proficiency levels for these token-based measures based on ACL appearances. In summary, the token-based measures (LS2amod, LS2advmod, LS2dobj) did not show statistically significant increases across proficiency levels based on the Kruskal-Wallis tests.

The ACL-based type-based measures mirror the findings of token-based measures, with advmod and dobj phraseological units showing a median portion of 0 in terms of their appearance in ACL. Specifically, for LS2amod, the median phraseological sophistication index started at 0 for B1 learners, increased to 0.028 for B2 learners, decreased to 0.026 for C1 learners, and slightly rose to 0.032 for C2 learners. Similarly, LS2advmod and LS2dobj consistently exhibited a median index of 0 across all proficiency levels. Kruskal-Wallis tests revealed p-values of 0.37 for LS2amod, 0.21 for LS2advmod, and 0.60 for LS2dobj, indicating no statistically significant differences in phraseological sophistication indexes across proficiency levels for these type-based measures based on ACL appearances. In summary, similar to the token-based measures, the type-based measures

(LS2amod, LS2advmod, LS2dobj) did not show statistically significant increases across proficiency levels based on the Kruskal-Wallis tests.

Here is a brief summary of the above analysis. For token-based measures:

- LS1amod: The median phraseological sophistication index started at 0 for B1 learners, increased to 0.028 for B2 learners, decreased to 0.026 for C1 learners, and rose slightly to 0.032 for C2 learners.
- LS1advmod and LS1dobj: Consistently showed a median index of 0 across all proficiency levels.
- Kruskal-Wallis Tests: Indicated non-statistical significance with p-values of 0.37 for LS1amod, 0.21 for LS1advmod, and 0.60 for LS1dobj, suggesting no significant differences in phraseological sophistication indexes across proficiency levels for these token-based measures based on ACL appearances.

The ACL-based type-based measures mirror the findings of token-based measures. For type-based measures:

- LS2amod: The median phraseological sophistication index started at 0 for B1 learners, increased to 0.028 for B2 learners, decreased to 0.026 for C1 learners, and slightly rose to 0.032 for C2 learners.
- LS2advmod and LS2dobj: Consistently exhibited a median index of 0 across all proficiency levels.
- Kruskal-Wallis Tests: Revealed p-values of 0.37 for LS2amod, 0.21 for LS2advmod, and 0.60 for LS2dobj, indicating no statistically significant differences in phraseological sophistication indexes across proficiency levels for these type-based measures based on ACL appearances.

In summary, the token-based measures (LS1amod, LS1advmod, LS1dobj) did not show statistically significant increases across proficiency levels based on the Kruskal-Wallis tests. Similar to the token-based measures, the type-based measures (LS2amod, LS2advmod, LS2dobj) did not show statistically significant increases across proficiency levels based on the Kruskal-Wallis tests.

8
Discussion

This chapter presents a discussion on the efficacy of the New Academic Collocation List by Shen (2023) and the Academic Collocation List by Ackermann and Chen (2013) in assessing phraseological sophistication. The discussion will meticulously examine the extent to which the NACL can evaluate phraseological sophistication. Additionally, it would compare the NACL with the ACL, providing a comprehensive analysis of how these two lists of phraseological units differ within the context of this study. The chapter is structured into two sections, corresponding to the two research questions outlined in Chapter 4.

RQ1: To what extend performs the New Academic Collocation List in terms of assessing phraseological sophistication of learner writing?

In this study of phraseological sophistication, a corpus of 246 learner texts, rated as B1, B2, C1, and C2 according to the CEFR, was utilized. As reported in the previous chapter, when employing the NACL as the reference for sophisticated phraseological units, the phraseological sophistication indexes for the four levels of learner texts did not follow a normal distribution. To compare differences between groups, Kruskal-Wallis rank sum tests were performed. This study replicated the work of Paquot (2019), which used a corpus of 98 learner texts rated as B2, C1, and C2 according to the CEFR. In Paquot's study, the phraseological sophistication indexes were also not normally distributed, leading the author to use Kruskal-Wallis rank sum tests for between-group comparisons.

The Kruskal-Wallis rank sum tests revealed a series of p-values for both token- and type-based phraseological sophistication measures (i.e., LS1amod, LS1advmod, LS1dobj, LS2amod, LS2advmod, and LS2dobj) as 0.61, 0.12, 0.21, 0.37, 0.11, and 0.14, respectively. None of these p-values was below the previously established significance level of 0.05. Therefore, we can conclude that no significant increase in phraseological sophistication was observed from B1 to C2 for any of the measures. This result aligns with the findings of

Paquot (2019), the study upon which this replication was based, where no statistically significant between-group differences were observed.

However, a series of increases in phraseological sophistication indexes from lower proficiency levels to adjacent higher proficiency levels within the same measure of phraseological sophistication were observed. Table 8.1 summarizes all such adjacent increases observed.

LS1advmod	B2→C1→C2	
	0.13→0.14→0.15	
LS1dobj	B1→B2	C1→C2
	0.09→0.14	0.13→0.14
LS2advmod	B2→C1→C2	
	0.12→0.14→0.16	
LS2dobj	B1→B2	C1→C2
	0.10-0.14	0.13-0.14

Table 8.1 All instances of increase of phraseological sophistication within adjacent levels observed

As illustrated in Table 8.1, increases in phraseological sophistication indexes within adjacent levels were primarily observed in LS1advmod, LS1dobj, LS2advmod, and LS2dobj. For LS1advmod, the median phraseological sophistication index consistently increased from 0.13 at the B2 level to 0.14 at the C1 level and further to 0.15 at the C2 level. Similarly, the median index for LS2advmod increased from 0.12 at the B2 level to 0.14 at the C1 level, and finally to 0.16 at the C2 level. In terms of LS1dobj, an increase in the phraseological sophistication index was observed in two adjacent intervals: from 0.09 at the B1 level to 0.14 at the B2 level, and another interval from 0.13 at the C1 level to 0.14 at the C2 level. However, no increase was observed between B2 and C1 for LS1dobj; instead, the index dropped from 0.14 at B2 to 0.13 at C1. Similarly, two adjacent level intervals showed increases for LS2dobj: from 0.10 at the B1 level to 0.14 at the B2 level, and from 0.13 at the C1 level to 0.14 at the C2 level. Overall, a series of increases in phraseological sophistication indexes were observed, aligning with Paquot (2019), which also noted some increases in phraseological sophistication indexes, though not universally.

In addition to increase, two obvious decrease patterns were observed as illustrated in Table 8.2.

	B1	B2	C1	C2
LS1amod	0.20	0.18	0.17	0.16
LS2amod	0.19	0.19	0.17	0.16

Table 8.2 Two obvious patterns of decrease

As can be observed from Table 8.2, for LS1amod, the median phraseological sophistication index decreases consistently across proficiency levels: starting from 0.20 at the B1 level, dropping to 0.18 at the B2 level, then to 0.17 at the C1 level, and finally to 0.16 at the C2 level. A similar pattern is observed for LS2amod, where the median index begins at 0.19 at the B1 level, remains constant at the B2 level, then decreases to 0.17 at the C1 level, and finally to 0.16 at the C2 level. In summary, both token- and type-based measures of phraseological sophistication for amod phraseological units show a consistent decrease across proficiency levels. This result is particularly interesting because it contradicts our hypothesis that the proportion of sophisticated phraseological units would increase as learners advance to higher proficiency levels. Instead, the results for LS1amod and LS2amod indicate a constant decrease. In contrast, Paquot (2019) did not observe any pattern of decrease. While some instances of indices remaining constant were noted (e.g., the mean indexes of B2 and C1 levels both being 0.03 for LS1amod), there were no instances of adjacent level decreases.

To evaluate the role of the New Academic Collocation List in assessing phraseological sophistication, no statistically significant increases were observed. Consequently, the author concludes that the NACL is not an ideal resource for this research objective. However, the results do provide insights into how the NACL could be improved to eventually become a suitable tool for examining phraseological sophistication. Given that a few cases of increases between adjacent proficiency levels were observed, the author suggests that the NACL is somewhat effective in capturing academic phraseological units in learner texts. This partial success explains the observed increases in phraseological indexes for several adjacent proficiency level intervals. The lack of statistically significant increases across proficiency levels may be attributed to the composition of the NACL, which includes some phraseological units that are not exclusively characteristic of academic English. The creation process of the NACL may have contributed to this inclusion of less specialized units.

During the creation of the New Academic Collocation List, the author utilized the Louvain Corpus of Research Articles, which comprises research articles from prestigious journals across 11 social sciences and humanities disciplines: anthropology, business, economics, education, law, literature, linguistics, medicine, political science, psychology, and sociology. The LOCRA contains approximately 18 million tokens, averaging about 1.6 million tokens per discipline. The author extracted all amod, advmod, and dobj

phraseological units from this corpus. Two statistical constructs were considered for selecting phraseological units in the final NACL: 1) relative frequency and 2) range. To meet the frequency criterion, candidate phraseological units needed to have a relative frequency of at least 1 per million tokens in the LOCRA. Additionally, these units had to satisfy the range criterion by appearing in each of the 11 disciplines, ensuring their wide-ranging applicability. In contrast to other studies that created phraseological units primarily for pedagogical purposes, the NACL's creation process involved specific statistical constructs. Table 8.3 lists various studies along with the statistical constructs they employed.

Studies	The Corpora Used	Methodology	Constructs			Result
			Frequency	Association	Dispersion	
Shin & Nation (2008)	BNC spoken section with a size of approximately 10 million tokens.	Collocation is used to refer to a group of two or more adjacent words that occur frequently together. A collocation is made of two parts – a pivot word which is the focal word in the collocation and its collocate(s), the word or words accompanying the pivot word. The pivot words must occur in the most frequent 1,000 content words according to Leech, Rayson, and Wilson (2001) and were set as a noun, a verb, an adjective, or an adverb. The eligible pivot words were searched by the WordSmith Tool 3.0 and acquired the corresponding collocate(s).	Eligible collocations must occur minimally 3 times per million tokens (PMT) globally in the whole corpus.	No requirement.	No requirement.	The final list includes 4,698 collocations.
Durrant (2009)	An academic corpus which the author built exclusively for this research and is comprised of 25 million tokens of learner texts from various disciplines. Additionally, the non-academic sections of	Academic collocations were defined as word pairs that co-occur at least moderate frequency across a wide range of disciplines, but which are not often found in non-academic language. They were operationalized as word pairs which co-occurred within a four-word span of each other.	Eligible collocations must have a minimal relative frequency (RF) of 1 time PMT globally.	Eligible collocations must have a minimal mutual information (MI) score of 4.	Eligible collocations must occur in each of the five sub-corpora.	The top 1,000 collocations with the largest difference of LL scores were included in the final list.

	BNC was used as reference corpus.					
Simpson-Vlach & Ellis (2010)	Michigan Corpus of Academic Spoken English (1.7 million tokens) plus BNC files of academic speech (1.2 million tokens)	The Academic Formulas List includes formulaic sequences that are subsumed in 3-, 4-, and 5- phrases, identifiable as frequent recurrent patterns in written and spoken corpora that are significantly more common in academic discourse than in non-academic discourse and which occupy a range of academic genres. They were accordingly operationalized as 3-grams, 4-grams, and 5-grams.	Eligible formulaic sequences must have a minimal RF of 10 times PMT globally.	The authors combined the metrics of frequency and MI and created a "formula teaching worth". However, no specific MI score requirement was set.	No requirement.	The Academic Formulas List contains 207 core formulaic sequences with the highest FTW scores.
Martinez & Schmitt (2012)	BNC (100 million tokens)	The phrasal expressions were operationalized as 2-grams, 3-grams, and 4 grams in the BNC. The n-gram function of WordSmith Tools 5.0 was used to extract collocations.	The expressions in the final list must occur with a minimal RF of 0.05 times PMT globally.	No requirement.	No requirement.	The PRASE List includes 505 most frequent non-transparent multi-word expressions.
Liu (2012)	Academic writing sections of the COCA and the BNC (totally 98.24 million tokens)	This study aimed to find the most common MWCs of a broad variety, including LBs, idioms, and phrasal/prepositional verbs. The operationalization for MWCs was done with the corpora search engine on https://www.english-corpora.org/ using the academic sections of BNC and COCA. Liu	Eligible MWUs must have a minimal RF of 20 times PMT globally.	No requirement.	Eligible MWUs must appear in six out of the eight academic divisions in COCA or five out of six academic divisions in BNC.	The final list includes 228 most common MWCs.

		used published sources as a database for query which includes lists of LBs proposed by previous researchers and lexicographers, e.g. Academic Formula List by Simpson-Vlach and Ellis (2010) and Oxford idioms dictionary (2001).				
Chon & Shin (2013)	The British Academic Spoken English Corpus (BASE) and the Academic Corpus (totally 5.1 million tokens)	Collocations were used to refer to a group of two or more adjacent words that occur frequently together. A collocation is made of two parts – a pivot word which is the focal word in the collocation and its collocate(s), the word or words accompanying the pivot word. In the research, the pivot words were set as the top 20 ranking academic words retrieved from each of the BASE and the Academic Corpus, and a pivot word must be either a noun, a verb, an adjective, or an adverb. The Concord function of the WordSmith Tools 3.0 was used to extract collocations.	The collocations included in the final list must occur with a minimal RF of 1.875 times PMT in the spoken corpus and a 1.741 times PMT in the written corpus.	No requirement.	No requirement.	The final list includes 934 written and 460 spoken collocations.
Ackermann & Chen (2013)	The written section of the Pearson International Corpus of	Collocation was defined as a single word (node word) that tends to co-occur in the span of ±3 words from the reference word. MonoConc Pro 2.2 was	The content pivot words must occur at least 5 times PMT and must occur in at least 5	Eligible collocations must have a minimal MI score of 3 and a	Eligible collocations must appear in each field of study. Additionally, the RF	The ACL includes 2,468 collocations.

	Academic English (25.6 million tokens)	used to obtain a list of content words in the corpus as note words. The list of node words was then used to extract collocations. This results in over 130,000 potential collocations, which were then lemmatized, and part-of-speech (POS) tagged. Finally, collocations with the following POS combinations were included in the final list: verb + noun, adjective + noun, adverb + adjective, and adverb + verb.	different texts. The collocations included in the final list must have a minimal RF of 1 PMT globally and a minimal RF of 0.2 in each in each field of study.	minimal t-score of 4.	in each field of study should be 0.2 or higher.	
Hsu (2014)	College Textbook Corpus (CTC) which is comprised of 200 college textbooks and totals 25 million tokens	The multi-word sequences were operationalized as 2-grams, 3-grams, 4 grams, and 5-grams in the CTC. The n-gram function of *Collocate* was used to extract collocations.	The sequences included in the final list must occur with a minimal RF of 5 times PMT globally.	No requirement.	The sequences included in the final list must occur in each of the 40 disciplines and in at least 100 out of the 200 compulsory textbooks.	The final list includes 475 opaque formulaic sequences of 2-5 words.
Rogers et al (2021)	The academic section of the COCA, totally 83 million tokens	MWUs were operationalized as a pivot word and a collocate. The *AntWordPairs* were used to extract MWUs.	The MWUs included in the final list must occur with a minimal RF of 1 time PMT.	The MWUs included in the final list must have a minimal MI score of 3.	No requirement.	The final list includes 5,057 collocations that were judged as useful for academic learners.

Table 8.3 Summary of phraseological unit lists construction

According to Table 8.3, the statistical constructs used in the creation of various phraseological unit lists fall into three main categories: frequency, dispersion, and association.

- **Frequency Requirements**: These requirements ensure that the phraseological units included in the final list have a significant presence in the source corpus. The higher the frequency parameter, the more likely the phraseological units will be observed within the corpus.
- **Dispersion Requirements**: These requirements ensure that the phraseological units are distributed throughout the source corpus rather than being concentrated in specific sections, thus ensuring their general applicability across different texts.
- **Association Requirements**: These requirements ensure that the elements of the phraseological units are statistically associated, indicating that their combination is not random but occurs with a certain degree of regularity.

During the construction of the NACL, the constructs of frequency and dispersion were considered, but not association. To evaluate the association between the elements of the phraseological units in the NACL, mutual information (MI) is a useful metric. Figure 8.1 presents a histogram of the mutual information scores of the NACL phraseological units, providing insight into the degree of association between their elements.

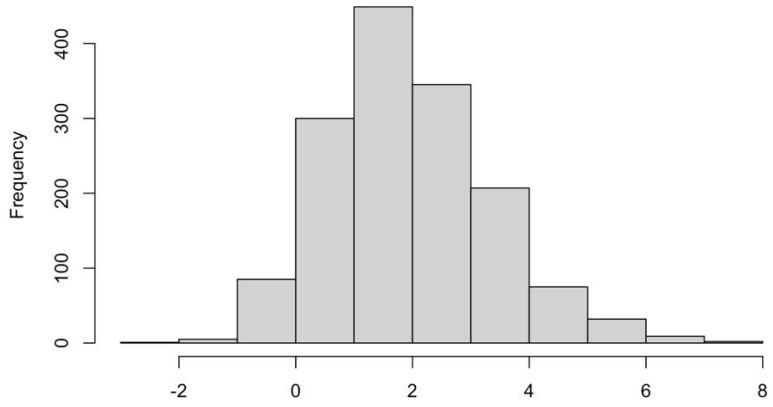

Figure 8.1 Histogram of MI scores of phraseological units in the NACL

A Shapiro-Wilk test confirmed that the distribution was not normal with a W = 0.9830 and a p-value less than 0.001. As a result, the median was used to report central tendencies. The distribution has a median of 1.44. According to Hunston (2002, p. 75), a mutual information score of 3 or above is considered indicative of strong association. Some studies, such as Durrant (2009), have used a mutual information score of over 3 as a benchmark. Therefore, we can conclude that the association strength of the phraseological units in the NACL is generally not strong.

Granger and Bestgen (2014) investigated the extent to which native English speakers and EFL learners differ in their use of phraseological units. They found that EFL learners tend to underuse high-association but low-frequency phraseological units compared to native speakers. Examples of such phraseological units with high mutual information scores include "ozone layer" (MI = 13.8), "steering wheel" (MI = 13.2), "nitrous oxide" (MI = 17.4), "vicious circle" (MI = 12.2), "incurably ill" (MI = 11.8), and "absolutely imperative" (MI = 8.6). An analysis of these units reveals that, despite their high mutual information scores, their frequency and the elements of the phraseological units within the corpus are relatively low. Consequently, the combinations of these elements are not random; rather, they demonstrate a tendency to co-occur in a way that forms phraseological units characterized by strong association.

The lack of strong association among the phraseological units in the NACL suggests that their combinations may be somewhat random, albeit characterized by a high frequency of both elements. This raises the concern that some units in the NACL may lack the uniqueness typically associated with academic discourse, indicating that certain phraseological units do not exclusively belong to the academic domain. Analysis has revealed that some phraseological units exhibit a higher relative frequency in a general reference corpus than in the source academic corpus (LOCRA). Ideally, a list of exclusively academic phraseological units should demonstrate a higher relative frequency in academic contexts compared to general corpora. This would indicate a greater likelihood of encountering these units in academic texts, thereby affirming their academic nature. Table 8.5 presents examples of phraseological units in the NACL that have a lower relative frequency when compared to their relative frequency in the British National Corpus (BNC; Davies, 2004), a comprehensive English corpus encompassing a wide range of domains, including academic contexts.

Phraseological units	Relative frequency in the LOCRA	Relative frequency in the BNC
several time	6.00	11.39
important part	8.07	9.82
few year	10.52	29.72
long time	15.30	38.50
other problem	2.65	3.15
have time	11.16	13.41

Table 8.5 Table of example phraseological units with a higher frequency in the BNC compared to the LOCRA

According to Table 8.5, some phraseological units in the NACL exhibit a lower relative frequency compared to the BNC. This indicates that these phraseological units are not exclusively found in the academic domain and are more likely to appear in general contexts. Furthermore, the components of these units—such as *important*, *few*, *other*,

time, *problem*, and *year*—are not low-frequency words and are commonly encountered across various domains. This finding aligns with Granger and Bestgen (2014), suggesting that these examples also do not demonstrate a high degree of association.

In an ideal scenario, a study of phraseological sophistication should focus on phraseological units that are uniquely exclusive to the academic domain. However, as previously analyzed, the phraseological units utilized in this study (i.e., the NACL) exhibit a lack of distinctiveness in terms of association within the academic context. Consequently, the phraseological units identified in the learner corpus of this study possess less uniqueness to the academic domain, which may lead to phraseological sophistication indexes that are not entirely "accurate," as they include units with varying degrees of academic relevance. This could account for the statistical insignificance observed in the results.

RQ2: To what extent differs the performance of the Academic Collocation List (Ackermann & Chen, 2013) from the New Academic Collocation List (Shen, 2023) in terms of assessing phraseological sophistication of leaner writing?

The sum tests of ACL-based phraseological sophistication measures revealed a series of p-values for both token- and type-based measures (i.e., LS1amod, LS1advmod, LS1dobj, LS2amod, LS2advmod, and LS2dobj) as follows: 0.37, 0.21, 0.60, 0.37, 0.19, and 0.45, respectively. None of these p-values fell below the established significance level of 0.05. Consequently, we can conclude that no significant increase was observed from B1 to C2 for any measure of phraseological sophistication. This finding also aligns with the results of Paquot (2019), the study upon which this replication is based, which similarly found no statistical significance between groups.

However, when comparing the results from the NACL and the ACL in this study, a distinct pattern emerges. A substantial portion of the results for the phraseological sophistication measures based on the ACL exhibited a median of 0. To further analyze the performance of the ACL concerning evaluating phraseological sophistication, an additional step was taken. Table 8.6 shows the results of using the ACL to evaluate phraseological sophistication in Paquot (2019).

	B2		C1		C2	
	Mean	SD	Mean	SD	Mean	SD
LS1amod	0.03	0.02	0.03	0.03	0.04	0.02
LS1advmod	0.003	0.004	0.007	0.01	0.01	0.02
LS1dobj	0.009	0.01	0.009	0.01	0.02	0.02
LS2amod	0.03	0.02	0.03	0.02	0.04	0.02
LS2advmod	0.004	0.0005	0.006	0.0007	0.01	0.01
LS2dobj	0.007	0.0007	0.009	0.0009	0.01	0.01

Table 8.6 Measures of phraseological sophistication based on the ACL in Paquot (2019)

The measures of phraseological sophistication based on the ACL in this study were reported in Table 7.6 of Chapter 7. However, the central tendencies were reported by means of medians, not means as in Paquot (2019). Considering that no median information was reported in Paquot (2019), the author also reported the mean phraseological sophistication indexes in this study in Table 7.7. Note that the standard deviations of the measures were reported in Table 7.6, no standard deviations were reported in Table 7.7.

	B1	B2	C1	C2
	Mean	Mean	Mean	Mean
LS1amod	0.05	0.04	0.03	0.04
LS1advmod	0	0.007	0.008	0.006
LS1dobj	0.01	0.01	0.01	0.01
LS2amod	0.03	0.03	0.03	0.03
LS2advmod	0	0.003	0.006	0.005
LS2dobj	0.02	0.01	0.01	0.01

Table 8.7 Measures of means of phraseological sophistication based on the ACL in this study

A comparison between Table 7.6 and Table 7.7 reveals that the results of B2, C1, and C2 parts share similarities to certain extend and no sharp deviations (e.g., a ten-time difference between the results of a measure based on two datasets) were observed. This shows the consistency of the ACL concerning assessing phraseological sophistication of learner writings. However, when comparing to the results of the NACL-based measures, as shown in Table 7.8, interesting differences emerge.

	B1	B2	C1	C2
	Mean	Mean	Mean	Mean
LS1amod	0.20	0.20	0.19	0.18
LS1advmod	0.16	0.13	0.14	0.16
LS1dobj	0.11	0.15	0.13	0.14
LS2amod	0.20	0.19	0.19	0.16
LS2advmod	0.15	0.13	0.14	0.16
LS2dobj	0.10	0.14	0.13	0.14

Table 7.8 Measures of means of phraseological sophistication based on the NACL in this study

It is obvious from Table 7.8 that all measures of phraseological sophistication based on the NACL is at least 0.10 (LS2dobj). The measures remain between 0.10 and 0.20, indicating that approximately 10% to 20% of various phraseological units produced by learners fall into the NACL. This indicates that the ACL is less effective in capturing sophisticated phraseological units produced by learners. This finding supports Paquot's (2019:136) suggestion that “a larger list be designed as a reference tool for investigating phraseological sophistication and academic language development.” When comparing the NACL and the ACL in terms of assessing the phraseological sophistication of learner writing, it is evident that the NACL is more effective in capturing sophisticated phraseological units, as none of the measures based on the NACL exhibited a median of 0.

In summary, this chapter examined the results from the study of phraseological sophistication based on both the NACL and the ACL. While a selective portion of median phraseological sophistication increases was observed across proficiency levels, no statistical significance was found. The author posits that this may be attributed to the NACL's lack of uniqueness, particularly concerning association. Furthermore, compared to the NACL, the ACL is less effective in capturing sophisticated phraseological units produced by learners.

9
Conclusion

This study replicated Paquot's (2019) investigation into phraseological sophistication. It utilized a corpus of 246 learner texts across B1, B2, C1, and C2 proficiency levels, totaling 148,084 tokens. The New Academic Collocation List (Shen, 2023) and the Academic Collocation List (Ackermann & Chen, 2013) were employed to identify sophisticated phraseological units. Three types of phraseological units—adjectival modifiers (amod), adverbial modifiers (advmod), and direct objects (dobj)—were extracted and analyzed using the Natural Language Processing tool spaCy. The primary focus was on phraseological sophistication indexes, representing the proportion of sophisticated phraseological units produced by learners relative to the total number of phraseological units. Six measures of phraseological sophistication were calculated to facilitate between-group comparisons: LS1amod, LS1advmod, LS1dobj (token-based measures), and LS2amod, LS2advmod, LS2dobj (type-based measures).

The results indicated no statistically significant differences between groups when using the New Academic Collocation List as the measure of sophisticated phraseological units. This finding is consistent with Paquot's (2019) study, which also reported non-statistically significant differences between groups. However, several systematic and consistent increases in phraseological sophistication indexes were observed. Specifically, both LS1advmod and LS2advmod increased consistently from B2 to C1 and from C1 to C2. Additionally, LS1dobj and LS2dobj showed increases from B1 to B2 and from C1 to C2.

The study further examined the reasons for the lack of statistically significant differences. It was noted that during the composition of the NACL, only frequency and range parameters within its source corpus were considered, without taking association parameters into account. Additionally, this study compared the performance of the ACL in assessing the phraseological sophistication of L2 writings. The results indicated that the

NACL is more effective than the ACL in capturing academic phraseological units produced by learners and serves as a better tool for analyzing L2 phraseological sophistication.

For future research, it is recommended that the NACL undergo further revision to include association parameters, thereby creating an academic collocation list that considers frequency, association, and range. This revised NACL would be a more robust tool for analyzing L2 phraseological sophistication and capturing academic phraseological units produced by learners.

10
Bibliography

.

Ackermann, K., & Chen, Y. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes, 12*(4), 235-247.

Aijmer, K. (2009). "So er I just sort of I dunno I think it's just because...": A corpus study of "I don't know" and "dunno" in learner spoken English. In A. Jucker, D. Schreier, & M. Hundt, *Corpora: Pragmatics and discourse* (pp. 151-166). Amsterdam, the Netherlands: Rodopi.

Altenberg , B. (1984). Causal linking in spoken and written English. *Studia Linguistica, 38*, 20-69.

Altenberg, B. (1998). On the phraseology of spoken English: the evidence of recurrent word combinations. In A. Cowie, *Phraseology* (pp. 101-122). Oxford: Oxford University Press.

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of "make" in native and non-native student writing. *Applied Linguistics, 22*, 173-194.

Barkema , H. (1996). Idiomatic and terminology: A multi-dimensional descriptive model. *Studia Linguistica, 50*(2), 125-160.

Barlow, M. (2000). Monoconc Pro 2.0. Athelstan.

Bestgen, Y., & Granger, S. (2018). Tracking L2 writers' phraseological development using Collgrams: Evidence from a longitudinal EFL corpus. In H. Sebastian, A. Sand, A. Sabine, & L. Dillmann, *Corpora and lexis* (pp. 277–301). Brill: Brill Rodopi.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*, 263-286.

Biber, D., Johansson, S., & Leech, G. (1999). *The Longman Grammar of Spoken and Written English.* London: Longman.

Bird, S., Edward , L., & Ewan , K. (2009). *Natural Language Processing with Python.* O'Reilly Media Inc.

Bo, G., & Shutang, Z. (2005). corpus-based contrastive study of recurrent word combinations in English essays of Chinese college students and native speakers. *CELEA Journal, 28*, 37–48.

Bulté, B., & Housen, A. (2012). Defining and operationalizing L2 complexity. In A. Housen, I. Vedder, & F. Kuiken, *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21-46). John Benjamins.

Chen, Y., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30-49.

Conrad, S., & Biber, D. (2004). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica, 20*, 56-71.

Council of Europe. (2001). *Common Framework of References for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Cowie , A. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics, 2*(3), 223–235.

Cowie , A. (1994). Phraseology. In R. Asher, *The Encyclopaedia of Language and Linguistics* (pp. 3168-3171). Oxford: Pergamon.

Cowie, A. (1997). Phraseology in formal academic prose. In J. Aarts, I. De Mönnink, & H. Wekker, *Studies in English language and teaching : in honor of Flor Aarts* (pp. 43-56). Amsterdam and Atlanta: Rodopi.

Cowie, A. (1998). *Phraseology: Theory, analysis and applications.* Oxford: Oxford University Press.

Cowie, A. (2001). Exploring native-speaker knowledge of phraseology: informant testing orcorpus research? In H. Burger, A. Buhofer, & G. Greciano, *Flut von Texten -Vielfalt der Kulturen. Ascona 2001 zur Methodologie und Kulturspezifik der Phraseologie* (pp. 73-81). Essen: Sclmeider Verlag Hogengehren GmbH.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Coxhead, A. (2008). Phraseology and English for academic purposes: Challenges and opportunities. In F. Meunier, & S. Granger, *Phraseology in foreign language learning and teaching* (pp. 149-161). Amsterdam/Philadelphia: John Benjamins.

Davies, M. (2008). British National Corpus. Oxford University Press. Retrieved from https://www.english-corpora.org/bnc/

De Clercq, B., & Housen, A. (2019). The development of morphological complexity: A cross-linguistic study of L2 French and English. *Second Language Research, 35*(1), 71-97.

De Cock, S. (2003). Recurrent sequences of words in native speaker and advanced learner spoken and written English: A corpus-driven approach. *Unpublished PhD thesis*. Louvain-la-Neuve: Universite Catholique de Louvain.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures (BELL)*, 225-246.

De Cook, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. *Corpus Linguistics and Linguistic Theory*, 51-68.

De Marneffe, M.-C., & Manning, C. (2017, Februaray). *Stanford typed dependencies manual.* Retrieved from http://nlp.standford.edu/software/dependenceis_manual.pdf

Douglas, R. (2013). The lexical breadth of undergraduate novice level writing competency. *Canadian Journal of Applied Linguistics, 16*, 152-170.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Academic Purposes, 28*(3), 157-169.

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.

Gilquin, G., De Cock, S., & Granger, S. (2010). The Louvain International Database of Spoken English Interlanguage. *Handbook and CD-ROM*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes, 6*, 319-335.

Granger, S. (1996). From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson, *Languages in Contrast. Text-based cross-linguistic studies* (pp. 37–51). Lund: Lund University Press.

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.

Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. (2020). The International Corpus of Learner English. Version 3. Louvain-la-Neuve, Belgium: Presses universitaires de Louvain.

Gries, S. (2008). Disentangling the phraseological web. In S. Granger, & F. Meunier, *Phraseology: An interdisciplinary perspective* (pp. 27–49). Amsterdam: John Benjamins.

Groom, N. (2009). Effects of second language immersion on second language collocational development. In A. Barfield, & H. Gyllstad, *Researching collocations in another language* (pp. 21-33). Basingstoke, UK: Palgrave Macmillan.

Gross, G. (1996). Les expressions figées en français: noms composés et autres locutions. *L'information grammaticale, 2*, 57.

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Howarth, P. (1996). *Phraseology in English Academic Writing: Some Implications for language learning and dictionary making.* Tubingen: Max Niemeyer Verlag.

Jones , S., & Sinclair, J. (1974). English lexical collocations. *Cahiers deLexicologie, 24*, 15-61.

Kyle, K., & Crossley, S. (2015). Automatically assessing lexical sophistication: Indicies, tools, findings, and application. *TESOL Quarterly, 49*, 757-786.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*, 647-672.

Lee, S. (2006). A corpus-based analysis of Korean EFL learner's use of amplifier collocations. *English Teaching, 61*(1), 3-17.

Lorenz, G. (1999). *Adjective intensification-Learners versus native speakers. A corpus study of argumentative writing.* Amsterdam, the Netherlands: Rodopi.

Mel'čuk, I. (1998). Collocations and lexical functions. In A. Cowie, *Phraseology: Theory, analysis, and applications* (pp. 23-53). Oxford: Oxford Univerisity Press.

Meunier, F. (2008). Introduction to the LONGDALE project. In E. Castello, K. Ackerley, & F. Coccetta, *Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment* (pp. 123-126). Bern: Peter Lang.

Nation, A. (Cambridge). *Learning vocabulary in another language.* 2001: Cambridge University Press.

Nesselhauf, N. (2005). *Collocations in a learner corpus.* Amsterdam, the Netherlands: John Benjamins.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 profiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492-518.

Paquot, M. (2008). Exemplificatiton in learner writing: A cross-linguistic perspective. In F. Meunier, & S. Granger, *Phraseology in foreign language learning and teaching* (pp. 101-119). Amsterdam, the Netherlands: John Benjamins.

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), 29-43.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research, 35*(1), 121-145.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32*, 130-149.

Paquot, M., Larsson, T., Hasselgård, H., Ebeling, S., De Meyere, D., Valentin, L., . . . van Vuuren, S. (2022). The Varieties of English for Specific Purposes dAtabase (VESPA): Towards a multi-L1 and multi-register learner corpus of disciplinary writing. *Research in Corpus Linguistics, 10*(2), 1-15.

Paquot, M., Naets, H., & Gries, S. (2020). Using syntactic co-occurrences to trace phraseological complexity development in learner writing: Verb object structures in LONGDALE. In B. Le Bruyn, & M. Paquot , *Learner Corpus Research Meets Second Language Acquisition* (pp. 122-147). Cambridge: Cambridge University Press.

Ping, P. (2009). A study of the use of four-word lexical bundles in argumentative essays by Chinese English majors-A comparative study based on the WECCL and LOCNESS. *CELEA Journal, 32*, 25-45.

Reppen, R. (2009). Exploring L1 and L2 writing development through collocations: A corpus-based book. In A. Barfield, & H. Gyllstad, *Researching collocations in another language* (pp. 49-59). Basingstoke, UK: Palgrave Macmillan.

Rogers , J., Müller, A., Daulton, F., Dickinson, P., Florescu, C., Reid, G., & Stoeckel, T. (2021). The creation and application of a large-scale corpus-based academic multi-word unit list. *English for Specific Purposes, 62*, 142-157.

Rubin, R., Housen, A., & Paquot, M. (2021). Phraseological complexity as an index of L2 Dutch writing proficiency: A partial replication study. In S. Granger, *Perspectives on the L2 Phrasicon: The View from Learner Corpora* (pp. 101-125). Bristol: Multilingual Matters.

Scott, M. (2024). WordSmith Tools version 9 (64 bit version). Stroud: Lexical Analysis Software.

Shen, J. (2023). Developing Resources for the Study of Phraseological Sophistication. *Unpublished MA Dissertation*. Louvain-la-Neuve, Belgium. Retrieved from https://dial.uclouvain.be/memoire/ucl/en/object/thesis%3A42087

Simpson, R., Briggs, S., Ovens, J., & Swales, J. (2002). *The Michigan Corpus of Academic Spoken English.* Ann Arbor, MI: The Regents of the University of Michigan.

Simpson-Vlach, R., & Ellis, R. (2010). An academic formulas list: new methods in phraseology research. *Applied Linguistics, 31*(4), 487-512.

Siyanova-Chanturia, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review, 64*(3), 429-458.

Svartvik, J., & Quirk, R. (1980). A Corpus of English Conversation. *Lund Studies in English, 56*.

Svensson, M. (2008). A very complex criterion of fixedness: Non-compositionality. In S. Granger , & F. Meunier, *Phraseology: An interdisciplinary perspective* (pp. 81–93). Amsterdam: John Benjamins.

Thewissen , J. (2008). The phraseological errors of French-, German-, and Spanish-speaking EFL learners: Evidence from an error-tagged learner corpus. *Proceedings from the 8th Teaching and Language Corpora Conference* (pp. 300-306). Lisbon: Associação de Estudos e de Investigação Cientifica do ISLA-Lisboa.

University of Oxford. (2003). The Uppsala Student English Corpus (USE). Oxford Text Archive.

Vanderbauwhede, G. (2012). Le déterminant démonstratif en français et en néerlandais: Théorie, description, acquisition [The demonstrative determiner in French and Dutch corpora: Thery, description and acquisition. *Unpublished Doctoral Dissertation*. Katholieke Universiteite Leuven and Unviersite Paris Ouest Nanterre La Defense.

Vandeweerd, N., Housen, A., & Paquot, M. (2021). Applying phraseological complexity measures to L2 French: A partial replication study. *International Journal of Learner Corpus Research, 7*(2), 197-229.

Vandeweerd, N., Housen, A., & Paquot, M. (2022). Comparing the longitudinal development of phraseological complexity across oral and written tasks. *Studies in Second Language Acquisition*, 1-25.

Vinogradov, V. (1953). On some issues of Russian historical lexicology. *Bulletin of the USSR Academy of Sciences, Language and Literature Section, 12*(3), 185-210.

Waibel, B. (2008). *Phrasal verbs: German and Italian learners of English compared.* Saarbrücken, Germany: VDM.

Wang, Y., & Shaw, P. (2008). Transfer and universality: Collocation use in advanced Chinese and Swedish learner English. *ICAME Journal, 32*, 201-232.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.