Phraseological sophistication as a multidimensional construct:

Exploring the relationship between association, register specificity and frequency of word combinations

Magali Paquot and Hubert Naets

Centre for English Corpus Linguistics, UCLouvain and CENTAL, UCLouvain

Since Paquot (2019), several unresolved issues have persisted regarding the operationalization of phraseological sophistication in L2 complexity research. One of the most crucial concerns relates to the extent to which the commonly used measures of phraseological sophistication (MI scores) fully represent the intended construct. In this study, we draw upon insights from L2 phraseological research to reexamine the conceptualization and operationalization of phraseological sophistication. We conduct new analyses on the learner corpus used in Paquot (2019), using alternative operationalisations of phraseological sophistication that represent different dimensions of sophistication (based on the register specificity of word combinations and their frequency). Results show that measures representing the dimensions of association (MI scores) and register specificity (ratios of academic collocations) correlate with each other. Frequency-based measures, however, pattern very differently, which we attribute to some issues in the way we operationalized frequency of co-occurrence.

Keywords: Phraseological complexity, phraseological sophistication, association, frequency, register specificity

1. Introduction

The significance of word combinations – be they framed in terms of phraseological units, formulaic sequences, collocations, constructions or collostructions – is widely recognized in language acquisition, processing, fluency, idiomaticity and change (e.g., Bybee & Beckner, 2012; Ellis, 1996; Goldberg, 2006; Römer, 2009; Schmitt, 2004; Sinclair, 1991; Wray, 2002). However, prior to Paquot's work in 2019, second language (L2) complexity research largely overlooked these theoretical and empirical developments. No L2 complexity study seemed to recognize that language largely relies on strings of co-selected words (Sinclair, 1991). From a frequency-based approach, such combinations constitute the phraseology, or phrasicon, of a language (Granger & Paquot, 2008).

As a result, Paquot (2019) introduced the construct of phraseological complexity with a view to theorize and operationalize linguistic complexity at the level of word combinations. In her initial proposal, Paquot (2019: 124) identified "the range [diversity] of phraseological units that surface in language production and the degree of sophistication of such phraseological units" as two major dimensions of the construct of phraseological complexity. Following this conceptualization, a learner text with a wider diversity of (target-like) word combinations and a higher proportion of sophisticated units is considered to be more complex than one where the same set of (more frequent) word combinations is often repeated. Paquot (2019) focused on relational co-occurrences (e.g. adjective + noun, adverb + adjective/adverb/verb, verb + direct object), an area that had received limited attention in L2 research. Several theoretical and practical reasons guided this choice. First, relational co-occurrences exemplify the interaction between lexis and grammar, showcasing lexical selection within grammatical patterns. Second, while previous research typically used positional co-occurrences, such sequences represent a

variety of structural relations with significantly different frequency distributions (Evert, 2004: 19). Analyzing a more homogeneous set of items allows for clearer and more meaningful quantitative results. Additionally, grounding the analysis in linguistic understanding, rather than relying solely on computational methods, further enhances the clarity and relevance of the findings (Evert, 2004: 19). Third, while some previous L2 phraseological research had explored patterns such as adjective + noun, there had been a lack of systematic, frequency-based approaches to verb + object dependencies. More qualitative studies, however, had often reported that these structures were particularly difficult for L2 English learners (e.g. Laufer & Waldman, 2011; for a survey, see Paquot & Granger, 2012).

Since Paquot (2019), there has been a growing number of (replication) studies that have sought to explore whether increased phraseological complexity can be used as an index of increased foreign language proficiency or foreign language development in a variety of foreign languages (e.g., Hu et al., 2022; Jiang et al., 2023; Kim & Crossley, 2023; McCallum, 2020; Paquot et al., 2021; Rubin et al., 2021; Vandeweerd et al., 2021; Vandeweerd et al., 2022; Vandeweerd et al., 2023). In these studies, while results for phraseological diversity were mixed, phraseological sophistication emerged as a reliable indicator of language proficiency. For example, all studies mentioned above, except Jiang et al. (2023), which is a study that targeted lower proficiency learners, have consistently reported a significant linear increase of phraseological sophistication scores (average MI scores) based on verb + object dependencies across proficiency levels. These results align with Paquot's (2019) initial finding that verb + object dependencies are robust discriminators at higher proficiency levels.

Yet, there are a number of unresolved issues with the way phraseological sophistication is currently operationalized in L2 complexity research. One of the most fundamental issues relates to content validity, i.e. the degree to which the measure typically used fully represents the construct that it is designed to measure. So far, phraseological sophistication has been

conceptualized as the presence of relatively sophisticated or advanced word combinations in learner samples and measured with average pointwise mutual information (MI) scores (i.e. a measure of the strength of association between words; Gries, 2022), on the grounds that, when calculated on a large reference corpus, this measure highlights word combinations that are highly exclusive. Because of their exclusivity, word combinations with a high MI tend to be more topic-specific, have more distinctive meanings, and tend to involve more specialized vocabulary (Gablasova et al., 2017); these combinations are therefore considered 'sophisticated'. However, it could be argued that similar to current perspectives on lexical sophistication (e.g., Kim et al., 2018; Kyle & Crossley, 2015; Kyle et al., 2018), phraseological sophistication should be best conceptualized as a multidimensional construct that taps into both the depth and breadth of phraseological knowledge. While association is undoubtedly an important aspect of phraseological sophistication, other dimensions such as frequency and register-related constraints on use should also be considered (Nation, 2001). Importantly, these dimensions have already received substantial attention in the existing literature on L2 phraseological competence. However, they have typically been examined separately and based on different types of word combinations. It is necessary to consolidate this diverse body of research and build a unified framework.

The main objective of this study is therefore to revisit the construct of phraseological sophistication and how it can be used to describe L2 proficiency. We begin by reviewing phraseological studies that have explored words combinations in learner language, with particular attention to association strength, frequency, and (typically register-specific) constraints on use. Building on the insights gained from this literature review, we then carry out new analyses on the learner corpus used in Paquot (2019), employing different operationalisations of phraseological sophistication that capture different dimensions of sophistication. As initiated in Paquot (2019), we seek to determine whether the same patterns

of variation (for example, across proficiency levels and across modes) are observable when phraseological sophistication is conceptualized multidimensionally and operationalized linguistically in different ways.

2. Statistical collocations, lexical bundles, *n*-grams and dimensions of phraseological sophistication in L2 phraseological studies

Over the past decades, a substantial body of research has embraced a frequency-based approach to phraseology, delving into the exploration of word combinations in learner corpora through semi-automated methods (Paquot & Granger, 2012). These studies have primarily investigated learners' use of (1) statistical collocations in the form of adjacent word pairs that occur more frequently than expected on the basis of chance, (2) lexical bundles (i.e., the most frequent recurring sequences of three or more words in a register; (Biber et al., 1999, Chapter. 13; see also Biber et al., 2004), and (3) *n*-grams, i.e., repeated sequences of *n* words (Paquot & Granger, 2012). Unlike in Paquot's work on phraseological complexity, thus, most of these studies have adopted a positional approach to word combinations, where words are said to co-occur when they appear within a certain distance from each other (Evert, 2004: 18). These studies, however, have produced a wealth of relevant findings that will allow us to formulate hypotheses for our current research.

2.1. The more advanced the learners, the more strongly associated their collocations

The emergence of the first research strand is commonly attributed to Durrant and Schmitt (2009), who grounded their work on a positional definition of collocation as "the relationship a

lexical item has with items that appear with greater than random probability in its (textual) context" (Hoey, 1991: 7). In their study, Durrant and Schmitt (2009) assigned to each adjacent noun/adjective + noun sequence, extracted from a corpus of L2 texts, its t-score and mutual information (MI) score as computed based on a large native reference corpus.[1] The rationale behind the use of these two association measures lies in their supposed complementarity in emphasizing rather different sets of collocations (e.g., Bartsch, 2004). Specifically, rankings based on t-scores tend to highlight highly frequent collocations (to the extent that they resemble rankings based on raw frequency, prompting arguments in favour of using frequency instead, Gries, 2022), whereas MI has often been reported to prioritize "word pairs which may be less common, but whose component words are not often found apart" (Durrant & Schmitt, 2009: 167; but see Gries, 2022 for an empirically based discussion of MI properties). Durrant and Schmitt (2009) then analysed the results by constructing a scale of collocational strength. They showed that non-native writers use collocations with very high t-scores at least as frequently as native speakers. Since non-native writers also tend to repeat certain preferred collocations, when considering collocation tokens rather than types, they exhibit a significant overuse of these high frequency and strongly associated collocations in comparison to native norms. By contrast, L2 writers of English tend to underuse strongly associated items as identified by high MI scores (e.g. *densely populated*, *bated breath*, *preconceived notions*) in comparison to native writers. This difference is even more pronounced when analyzing collocation types. The method has been further refined in a series of studies by Granger and Bestgen (2014) and popularised as the *Collgram* method (e.g. Bestgen & Granger, 2014; Bestgen & Granger, 2017; Bestgen & Granger, 2018; Granger & Bestgen, 2014). Research employing the *Collgram* method to compare learners' collocation usage at different proficiency levels has consistently reported that the same difference observed between learners and native speakers could also be

[1] It is important to note that we do not consider the t-score as a measure of phraseological sophistication. We refer to Gries (2022) for a detailed discussion of why the t-score should not be used as a measure of association.

observed across learners at different proficiency levels. Lower proficiency learners tend to underuse the more strongly associated word pairs identified with high MI scores compared to higher proficiency learners (Granger & Bestgen, 2014; Xia et al., 2022a).

Despite their significance, dependency-based collocations (as used in phraseological complexity research) have received comparatively less attention in L2 studies, possibly due to the greater difficulty in automatically extracting them. Nonetheless, they have already been shown to outperform adjacent word pairs (i.e., bigrams or collgrams) in explaining variance in holistic writing quality scores (Kyle & Eguchi, 2021).

2.2. The more advanced the learners, the more register-specific their recurrent sequences of words

Studies that have explored the use of lexical bundles in learner corpora have produced a plethora of findings related to their structures and functions in learner language (e.g., Ädel & Erman, 2012; Kim & Kessler, 2022; Staples et al., 2013; Xia et al., 2022b). Of particular relevance is the observation that more proficient learners seem to be able to select lexical bundles that align more closely with the intended register of their writing (Chen & Baker, 2016; Yoo & Shin, 2022). Thus, in the context of argumentative essays, proficient learners tend to favour phrasal bundles, often incorporating post-modifiers into longer sentences, thereby approximating academic prose conventions. By contrast, the essays of lower proficiency learners are characterised by linguistic traits typical of spoken and informal registers. These learners tend to employ more clausal bundles, often including first-person pronouns and relying on a restricted repertoire of verbs such as *want* and *be* (Yoo & Shin, 2022).

2.3. Frequency matters too

Recurrent word sequences or *n*-grams have also been investigated within a different framework, serving as indices of lexical sophistication. For example, the *Tool for Automatic Analysis of Lexical Sophistication* (TAALES; Kyle et al., 2018) offers a range of indices of lexical sophistication related to the frequency, association and range of bigrams and trigrams across a variety of corpora/registers. Numerous studies have provided evidence for the efficacy of some of the TAALES *n*-gram frequency and association measures (though not the *n*-gram range measures) in explaining the variance in proficiency scores or writing quality scores (e.g., Crossley et al., 2012; Garner et al., 2019; Kim et al., 2018; Kyle & Crossley, 2015; Zhang & Ouyang, 2023). For example, Kyle et al. (2018) found that specific measures of *n*-gram association strength and frequency were strong predictors in explaining the variance in lexical proficiency scores (three specific *n*-gram indices – COCA News Bigram Association Strength (DP), COCA Magazine Trigram Proportion 80K, and COCA Magazine Trigram Bigram to Unigram Association Strength (MI) – accounted for approximately 28% of the variance). Zhang et al. (2023) reported similar results in a study examining the effect of various linguistic indices on EFL independent and integrated writing quality. However, the *n*-gram indices varied across task types, with frequency indices significantly predicting narrative writing scores and association strength indices significantly predicting argumentative writing scores.

2.4. Cumulative research point to the need for a multidimensional approach to phraseological knowledge

The collective findings of the studies reviewed above indicate a significant correlation between proficiency in L2 writing and various facets of word combinations, including association strength, register-specificity and frequency. However, these conclusions are predominantly derived from research examining specific properties of different types of word combinations separately, typically focusing on the association strength of adjacent word pairs and the frequency and register-specificity of larger sequences of words. This limitation is also evident in current studies of phraseological complexity, which have largely adopted a limited view on phraseological sophistication, often narrowing it down to association alone, and operationalizing it through mutual information scores.

In this study, we therefore use different operationalisations of phraseological sophistication to conduct new analysis of the *Varieties of English for Specific Purposes Database* (VESPA; Paquot et al., 2022b) corpus, which was previously utilized in Paquot (2019). These operationalizations represent what we consider distinct dimensions of sophistication, specifically register-specificity (more particularly, the presence of academic collocations reflecting adherence with the norms and conventions of academic writing), and frequency. To this end, we develop new frequency lists of English general and academic collocations. Our objective is to investigate whether the patterns observed in previous phraseological complexity research, particularly those concerning sophistication as association, are reproduced when the construct is redefined to incorporate register specificity and frequency in a multidimensional framework.

Based on the current body of L2 phraseological research reviewed above, despite most studies employing a positional approach to word combinations instead of the relational approach adopted here, we expect the following:

- Hypothesis 1: Values for metrics of phraseological sophistication that tap into the dimension of register specificity (academic collocations) will increase linearly with increased proficiency level: the more advanced the learner text, the more frequent use of register-specific academic collocations.
- Hypothesis 2: Values for metrics that operationalize the frequency dimension of phraseological sophistication will also increase linearly with increasing proficiency: the more advanced the text, the more low-frequency co-occurrences; and hence,
- Hypothesis 3: There will be positive correlations between measures of phraseological sophistication that represent its three dimensions: association, register specificity and (low) frequency.

Focusing on the same types of relational cooccurrences as used in previous phraseological complexity research also allows us to predict that much like measures of phraseological sophistication by association, measures of phraseological sophistication by register specificity and frequency for different dependency types (adjective + noun, adverb + adjective/adverb/verb, verb + direct object), will not correlate with proficiency in the same way (Hypothesis 4).

3. Data and method

3.1. Learner corpus

The learner texts used in this study are the same as used in Paquot (2019). These texts originate VESPA and consist of 98 research papers submitted by French English as a Foreign Language (EFL) learners as part of the requirements for linguistics courses at the University of Louvain,

Belgium, between 2009 and 2013 (henceforth VESPA-FR-LING). The students are 58 females and 20 males enrolled in Bachelor or Master programs in Modern Languages, which involve the study of two languages (i.e. English along with either Dutch, Spanish, German, French or Italian). They have been studying English at university for an average of 3.5 years (minimum = 1, maximum = 6), and the large majority of them (i.e. 75) are 19–26-years-old (average = 22.7; minimum = 19; maximum = 47).

VESPA-FR-LING texts were evaluated by two professional raters (three in case of disagreement) based on the *Common European Framework of References for Languages* (CEFR; Council of Europe, 2001) descriptors for linguistic competence. These descriptors focus on grammatical accuracy, vocabulary control, vocabulary range, orthographic control, and coherence and cohesion, delineating proficiency across six levels: A1 and A2, B1 and B2, C1 and C2. As a result, each text received a global proficiency score, ranging from B2 (upper intermediate) to C2 (near-native) (for further details on the assessment procedure, see Paquot (2018)). Table 1 shows the distribution of learner texts per CEFR levels. It also provides the total number of words in each learner sub-corpus and descriptive statistics at the text level.

Table 1. VESPA-FR-LING

	Number of texts	Total number of words	Mean	SD	Median	IQR	min	max
B2	25	87,372	3,495	832	3,568	1012	1,812	5,261
C1	62	218,184	3519	1061	3,390	973	2,000	8,755
C2	11	34,438	3131	824	3,059	1181	1,993	4,583
Total	98	339,994	3469	982	3,386	976	1,812	8,755

3.2. Operationalization of phraseological sophistication

Phraseological sophistication is approached in this study via co-occurrences used in three dependency relations, i.e. adjectival modifiers (*amod*: adjective + noun), adverbial modifiers

(*advmod*: adverb + adjective, adverb or verb) and verb + direct object (*dobj*: verb + noun). We first made use of *spaCy* (Honnibal & Montani, 2017) with the pipeline en_core_web_lg 3.6.0 to lemmatize, part-of-speech (POS) tag and parse each VESPA learner text. While we did not check the accuracy of the *spaCy* parser on our own corpora, reports indicate that the tool achieves an accuracy score of 90,27% for dependency labelling and 97,35% for POS-tagging.[2] Recent studies have demonstrated its effectiveness with L2 texts. In particular, Kyle and Eguchi (2021) reported F1 scores ranging from .960 to .986 for the three dependency relations. The next steps consisted in extracting from each VESPA text all the pairs of words found in the *amod*, *advmod* and *dobj* Stanford typed dependency relations in the form of lemmas and simplified part-of-speech tags. As illustrated in (1), a Stanford typed dependency is a binary grammatical relation between a governor and a dependent (see De Marneffe & Manning, 2016).

(1) a. *amod* adjective modifier

She has black hair. *Amod* (hair + NN, black + JJ)

b. *advmod* adverbial modifier

She has very black hair. *Advmod* (black + JJ, very + RB)

Repeat less quickly. *Advmod* (quickly + RB, less + RB)

She eats slowly. *Advmod* (eat + VBZ, slowly + RB)

c. *dobj* direct object

He won the lottery. *Dobj* (win + VV, lottery + NN)

We used the same corpus processing pipeline to extract dependencies from reference corpora employed to compute external measures of phraseological sophistication. We implemented and

[2] https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.6.0

tested two methods that tap into different dimensions of phraseological sophistication. First, to operationalize the register specificity dimension of phraseological sophistication, sophisticated word combinations were defined and operationalized as academic collocations. Secondly, we also measured phraseological sophistication in relation to frequency, whereby more sophisticated word combinations are characterized by lower occurrence rates. These two approaches required the development of new resources as explained below.

3.2.1. Phraseological sophistication by register specificity

The most salient register-specific characteristic of dependency relations such as adjective + noun, adverb + verb, or verb + direct object structures in academic texts is the prevalence of academic collocations. Academic vocabulary is used "to refer to those activities that characterize academic work, organize scientific discourse and build the rhetoric of academic texts" (Paquot, 2010, p. 28). As such, academic collocations are particularly frequent and well distributed in academic texts (e.g., *limited ability*, *potential benefit*, *provide + evidence*, *test + hypothesis*) but they are less common in other registers.

To explore phraseological sophistication by register specificity, we initially considered using the *Academic Collocation List* (ACL), a list of the 2,469 most frequent and pedagogically relevant cross-disciplinary lexical collocations compiled from the written curricular component of the *Pearson International Corpus of Academic English* (Ackermann & Chen, 2013). However, as reported in Paquot (2019), a first attempt at using this list to explore another dimension of phraseological sophistication proved unsuccessful, most likely because the ACL is too small for research purposes. The ACL was primarily compiled to help EFL learners increase their collocational competence and support English for Academic Purposes (EAP) teachers in their lesson planning. As a result, the number of items was limited to a manageable

dataset. The direct consequence, however, is that its text coverage is relatively limited: Ackermann and Chen (2013) report an overall coverage of the ACL in the source corpus of 1.4%. This is not much compared to the coverage of word frequency lists that are commonly used to assess lexical sophistication in learner texts. The Academic Word List (Coxhead, 2000), for example, has been shown to cover ca. 10% of academic texts.

Consequently, we had to create a new database of academic collocations to explore the register specificity of dependency relations in learner texts. For this purpose, we used an extended version of the *Louvain Corpus of Research Articles* (LOCRA) that includes research articles from prominent journals in nine disciplines within the Social Sciences and Humanities (SSH; as detailed in Table 2). As our goal was to identify academic collocations and not discipline-specific collocations, it was important to include several disciplines in the reference corpus.

Table 2. LOCRA composition

	Number of texts	Number of words
Anthropology	250	2,040,272
Business	200	1,979,903
Education	228	1,994,118
Law	226	2,104,869
Linguistics	219	2,152,744
Literature	252	1,930,173
Political sciences and international relations	247	1,959,513
Psychology	313	2,651,290
Sociology	261	2,024,834

To be considered an academic collocation, a word co-occurrence (in the form of a relational dependency) had to meet the following criteria:

1. Relative frequency: A word co-occurrence had to appear with a relative frequency > 1 per million words (pmw) (Durrant, 2009; Rogers et al., 2021). This criterion removed

relatively infrequent co-occurrences such as *abandon + assumption*, *accept + assistance*, *ask + informant*, *political alliance*, *strong attachment*.

2. Range: A word co-occurrence had to appear in a minimum of seven out of nine disciplines (see Liu (2012) for a similar use of the criterion of range to create a list of multi-word constructions in academic written English). This criterion further removed discipline-specific word combinations such as *entrepreneurial culture*, *metalinguistic feedback*, *proficient learner*, *adopt + law*, *acquire + language*, *ask + teacher*, and *amend + constitution*.
3. Keyness: A word co-occurrence had to be identified as a positive keyword in LOCRA against the ENCOW16 web corpus (Schäfer, 2015) (log-likelihood < 0.01 with Holm correction for multiple testing). This criterion further removed collocations that are more frequent in general English and, as such, less typical of academic writing (e.g., *dramatic change*, *personal choice*, *high expectation*, *serious consequence*, *build + relationship*, *change + nature*, *draw + inspiration*, *express + feeling*, *give + direction*).
4. Association: A word co-occurrence had to occur with a pointwise mutual information score above 0.5 (this criterion removes all the word co-occurrences that belong to the first quartile of the MI score distribution). The MI-based criterion allowed us to remove all word co-occurrences with negative MIs (typically word combinations with at least one high frequency word such as *other analysis*, *other relation*, *other effect*, *have + measure*, as well as some less idiomatic word combinations (e.g., *use + difference*, *include + relationship*).

The final list consists of 6,529 academic collocations (Table 3). Compared to the ACL, it is increased by 62% and includes a much larger set (and a much larger proportion) of *advmod* and *dobj* dependencies.

Table 3. Distribution of academic collocations and comparison with the ACL (Ackermann & Chen, 2013)

	Academic collocation dataset used in this study	ACL
Amod	2,669 (40.9%)	1,773 (71.8%)
Advmod	2,102 (32.2%)	294 (11.9%)
Dobj	1,758 (26.9%)	310 (12.6%)
Other	-	92
Total	6,529	2,469

The second step consisted in the identification of academic collocations in each learner text. Table 4 lists the six measures of phraseological sophistication by specificity used in this study. The six measures are ratios of the number of academic collocations to the total number of co-occurrences for each type of relational dependencies, both in the form of types and tokens. Ratios based on types provide insights into the diversity of academic collocations across CEFR levels, while ratios based on tokens primarily reflect repetition and text coverage rather than variety.

Table 4. Measures of phraseological sophistication based on the academic collocation dataset

	Formula
PS_ACAD_AMOD_TYPES	$acad\_amod_{types}/amod_{types}$
PS_ACAD_AMOD_TOKENS	$acad\_amod_{tokens}/amod_{tokens}$
PS_ACAD_ADVMOD_TYPES	$acad\_advmod_{types}/advmod_{types}$
PS_ACAD_ADVMOD_TOKENS	$acad\_advmod_{tokens}/advmod_{tokens}$
PS_ACAD_DOBJ_TYPES	$acad\_dobj_{types}/dobj_{types}$
PS_ACAD_DOBJ_TOKENS	$acad\_dobj_{tokens}/dobj_{tokens}$

3.2.2. Phraseological sophistication by frequency

To explore phraseological sophistication by (general) frequency, we extracted frequency lists of *amod*, *advmod* and *dobj* dependencies from the Web corpus ENCOW16AX (Schäfer, 2015). We opted for the ENCOW16AX corpus as reference corpus because it is one of the largest corpora available for English (9,538,719,919 tokens in the form of randomised sentences) and it has the advantage of being distributed with Stanford-typed dependencies. However, we were

quickly confronted with a problem in our endeavour to operationalize phraseological sophistication in relation to frequency: what are frequent, less frequent or rare co-occurrences? When should word combinations be considered sophisticated in the sense of "relatively unusual or advanced" (Read, 2000: 203) in a learner text? For lack of a principled approach to guide the methodological decision regarding the selection of a critical size for a list of the most frequently occurring word combinations, TAALES, for example, provides a wide range of metrics computed based on the proportion of *n*-grams in a text that rank within the top 10,000, 20,000, 30,000 up to 100,000 most frequent *n*-grams in various COCA sub-corpora. However, the use of all these measures combined poses a notable challenge when it comes to interpreting the scores.

For lack of research in this area, we drew on L2 vocabulary research to provide a practical solution to this currently unresolved problem. In L2 vocabulary studies, sophisticated words are generally defined as those beyond the first 2,000 most frequent words that constitute approximately 80% of an average written text (Wolfe-Quintero et al., 1998; Schmitt, 2010: 69). Although empirical research has not yet provided insights into the frequency at which word combinations should be considered frequent, one possible approach to addressing this question is via text coverage analysis. As a first exploration into this, we decided to operationalize sophisticated (here, in the sense of infrequent or rare) dependencies as co-occurrences above an 80% mean coverage threshold in VESPA. Given that the proportion of off-list dependencies in VESPA was systematically different from the 5% typically reported in vocabulary research for single words, however, we had to adapt this threshold for each dependency as follows:

Threshold above which dependencies are considered less frequent or rare =

(proportion of dependencies represented in ENCOW16/95) * 80

Table 5 lists the metrics used to explore the frequency dimension of phraseological sophistication for each dependency (both as types and tokens) and reports the target coverage threshold used based on the cumulative proportions for each measure.

Table 5: measures of phraseological sophistication by frequency

	Formula
PS_FREQ_AMOD_TYPES	% of *amod* dependencies (types) above 76.6% coverage threshold
PS_FREQ_AMOD_TOKENS	% of *amod* dependencies (tokens) above 76.6% coverage threshold
PS_FREQ_ADVMOD_TYPES	% of *advmod* dependencies (types) above 81.6% coverage threshold
PS_FREQ_ADVMOD_TOKENS	% of *advmod* dependencies (tokens) above 81.3% coverage threshold
PS_FREQ_DOBJ_TYPES	% of *dobj* dependencies (types) above 79% coverage threshold
PS_FREQ_DOBJ_TOKENS	% of *dobj* dependencies (tokens) above 79.6% coverage threshold

3.3. Statistics

To test Hypothesis 1 and Hypothesis 2, we used the newly developed measures of phraseological sophistication to compare texts categorized as B2, C1, and C2 in the VESPA corpus. We systematically assessed the distributions for normality using the Shapiro–Wilk test. Given that at least some of the metrics were not normally distributed, we report all descriptive statistics in the form of median and inter-quartile range in the main text. Other descriptive statistics are available in the supplementary material (see below). Frequency counts demonstrating normal distribution underwent ANOVA, followed by Tukey contrasts for post-hoc analysis. For non-normally distributed data, we applied Kruskal–Wallis rank sum tests, followed by Dunn tests with Holm corrections. The significance threshold for all statistical tests was set at 0.05.

To test Hypothesis 3, we conducted a comparison of phraseological sophistication measures pertaining to the dimensions of register specificity and frequency, as proposed in this study, along with the phraseological sophistication measures used in Paquot (2019) to tap into association. Correlations were computed using the Spearman method and their interpretations follow the recommendations outlined in Gries (2013: 147).

All analyses were conducted with R (R Core Team, 2022) and the *rstatix* and *corrplot* packages. The code and data are available on the Open Science Framework page dedicated to this study: https://osf.io/jxbn8/?view_only=ce748f05c1c5494ba0cea6eebe483336.

4. Results

4.1. Phraseological sophistication by register specificity

As shown in Table 6, no statistically significant difference is detected between CEFR groups when phraseological sophistication is conceptualized in terms of register specificity and operationalized with measures based on academic collocations in the form of adjective-noun dependencies. By contrast, when exploring phraseological sophistication with *advmod* relations, a statistically significant difference emerges. This difference is noticeable both when academic collocations are investigated as tokens ($F(2,95) = 3.21$, $p = 0.04$*, $\eta 2 = 0.04$) and types ($H(2,95) = 6.37$, $p = 0.04$*, $\eta 2 = 0.06$). The observed effect size is small for types, and a pairwise post-hoc Dunn test with Holm adjustments showed significance only for the comparison between B2 and C1 ($p = .04$). For tokens, the observed effect is moderate, and a pairwise post-hoc Dunn test with Holm adjustments similarly showed significance only for the comparison between B2 and C1 ($p = 0.05$).

Table 6: Measures of phraseological sophistication by register specificity (academic collocations) in VESPA

	B2		C1		C2		Between-group comparisons
	Median	IQR	Median	IQR	Median	IQR	
PS_ACAD_AMOD_TYPES	0.13	0.03	0.13	0.05	0.14	0.04	$H(2,95) = 3.05$, $p = .22$
PS_ACAD_AMOD_TOKENS	0.12	0.07	0.12	0.07	0.14	0.03	$F(2,95) = 0.8$, $p = .43$

PS_ACAD_ADVMOD_TYPES	0.23	0.1	0.27	0.08	0.26	0.09	H(2,95) = 6.37, p = .04*
PS_ACAD_ADVMOD_TOKENS	0.24	0.1	0.31	0.1	0.26	0.08	F(2,95) = 3.21, p = .04*
PS_ACAD_DOBJ_TYPES	0.14	0.05	0.14	0.05	0.19	0.07	H(2,95) = 9.54, p = .008*
PS_ACAD_DOBJ_TOKENS	0.14	0.05	0.15	0.06	0.2	0.1	H(2,95) = 6.76, p = .03*

Measures of phraseological sophistication targeting academic collocations in the form of direct object dependencies also exhibit a statistically significant difference, both when investigated as tokens ($H(2,95) = 6.76$, $p = 0.03$*, $\eta 2 = 0.05$) and types ($H(2,95) = 9.54$, $p = 0.008$*, $\eta 2 = 0.08$). The observed effect size is moderate for types, and a pairwise post-hoc Dunn test with Holm adjustments revealed significant differences for two comparisons: B2 vs. C2 ($p = .01$) and C1 vs. C2 ($p = 0.02$). For tokens, however, the observed effect is small, and a pairwise post-hoc Dunn test with Holm adjustments showed significance only for the comparison between B2 and C2 ($p = .03$).

4.2. Phraseological sophistication by frequency

Table 7 shows that there is no statistically significant difference detected among CEFR groups when conceptualizing phraseological sophistication as low frequency. Although some linear trends can be observed for PS_FREQ_ADVMOD_TYPES, PS_FREQ_DOBJ_TYPES, and PS_FREQ_DOBJ_TOKENS, there is a notable degree of intra-group variability.

Table 7: Measures of phraseological sophistication in relation to frequency in VESPA

	B2		C1		C2		Between-group comparisons
	Median	IQR	Median	IQR	Median	IQR	
PS_FREQ_AMOD_TYPES	24.41	6.08	23.35	10.24	27.27	6.03	F(2,95) = .54, p = .59
PS_FREQ_AMOD_TOKENS	13.79	9.13	12.07	8.67	13.26	7.10	H(2,95) = 1.83, p = .4

PS_FREQ_ADVMOD_TYPES	12.68	9.43	15.3	8.11	16.99	5.25	H(2,95) = 1.84, p = .4
PS_FREQ_ADVMOD_TOKENS	14.19	7.41	15.96	7	15.45	4.78	H(2,95) = 2.22, p = .33
PS_FREQ_DOBJ_TYPES	21.74	9.28	20.81	8.17	18.09	4.15	H(2,95) = 3.47, p = .18
PS_FREQ_DOBJ_TOKENS	21.11	8.31	20.11	7.97	17.31	2.33	H(2,95) = 5.48, p = .06

4.3. Shared patterns of variation

Figure 1 presents a correlation matrix for all the new measures of phraseological sophistication proposed in this study as well as the MI-based metrics employed in Paquot (2019). For visualization purposes, measures are organized by dependency type (i.e. all measures that were computed on *dobj* dependencies are clustered together, etc.). The correlation matrix highlights several important results:

1. For *dobj* and *advmod* dependencies, measures of phraseological sophistication operationalized in relation to association (MI-based measures) correlate positively with measures of phraseological sophistication measured in relation to register specificity (academic collocations). There are intermediate correlations between PS_ASSOC_DOBJ and PS_ACAD_DOBJ_TYPES [$\tau = 0.33$], PS_ASSOC_DOBJ and PS_ACAD_DOBJ_TOKENS [$\tau = 0.24$], PS_ASSOC_ADVMOD and PS_ACAD_ADVMOD_TYPES [$\tau = 0.29$] and PS_ASSOC_ADVMOD and PS_ACAD_ADVOMOD_TOKENS [$\tau = 0.24$]. By contrast, correlations are lower for PS_ASSOC_AMOD and PS_ACAD_AMOD_TYPES [$\tau = 0.12$] and almost null for PS_ASSOC_AMOD and PS_ACAD_AMOD_TOKENS [$\tau = 0.04$]. In general, this means that a learner text with more highly associated collocations will also contain more academic collocations.
2. For each dependency, measures of phraseological sophistication measured in relation to register specificity (academic collocations) correlate negatively with measures of

phraseological sophistication by low frequency. Correlations are intermediate (e.g. PS_ACAD_DOBJ_TYPES – PS_FREQ_DOBJ_TYPES [τ = - 0.36], PS_ACAD_AMOD_TYPES – PS_FREQ_AMOD_TYPES [τ = - 0.51], PS_ACAD_ADVMOD_TYPES – PS_FREQ_ADVMOD_TYPES [τ = - 0.23]. This suggests that learner texts with more academic collocations will contain fewer low frequency co-occurrences.

3. Correlations between measures of phraseological sophistication operationalized in relation to association and measures of phraseological sophistication by low frequency are very low except for *advmod* (PS_ASSOC_ADVMOD – PS_FREQ_ADVMOD_TYPES [τ = 0.21], PS_ASSOC_ADVMOD – PS_FREQ_ADVMOD_TOKENS [τ = 0.21]). This suggests that there is not a strong relationship between the presence of highly associated collocations and low co-occurrences in VESPA learner texts.
4. Measures of phraseological sophistication operationalized in relation to association correlate positively across dependency types. All correlations are intermediate (PS_ASSOC_DOBJ – PS_ASSOC_AMOD [τ = 0.23], PS_ASSOC_DOBJ – PS_ASSOC_ADVMOD [τ = 0.21], PS_ASSOC_AMOD – PS_ASSOC_ADVMOD [τ = 0.27]). This means that a learner text with more highly associated *dobj* dependencies will also contain more highly associated *amod* and *advmod* dependencies.
5. Similarly, measures of phraseological sophistication operationalized in relation to register specificity also correlate positively across dependency types. For example, correlations between the type-based measures are as follows: PS_ACAD_DOBJ_TYPES – PS_ACAD_AMOD_TYPES [τ = 0.25], PS_ACAD_DOBJ_TYPES – PS_ACAD_ADVMOD_TYPES [τ =0.21], and PS_ACAD_AMOD_TYPES – PS_ACAD_ADVMOD_TYPES [0.13]. This means that a learner text with more academic collocations of the *dobj* type will also contain more academic collocations of the *amod* and *advmod* types.

6. For *dobj* dependencies, measures of phraseological sophistication operationalized in relation to low frequency correlate positively with the same metrics computed for *amod* and *advmod* dependencies. For example, correlations between the type-based metrics are as follows: PS_FREQ_DOBJ_TYPES – PS_FREQ_AMOD_TYPES [τ = 0.33], and PS_FREQ_DOBJ_TYPES – PS_FREQ_ADVMOD_TYPES [τ = 0.15]. By contrast, there are very low if not negative correlations between measures involving *amod* and *advmod* dependencies (PS_FREQ_AMOD_TYPES – PS_FREQ_ADVMOD_TYPES [τ = 0.03], PS_FREQ_AMOD_TOKENS – PS_FREQ_ADVMOD_TOKENS [τ = - 0.12]). This suggests that there is not a clear relationship between the presence of low frequency co-occurrences across dependency types.

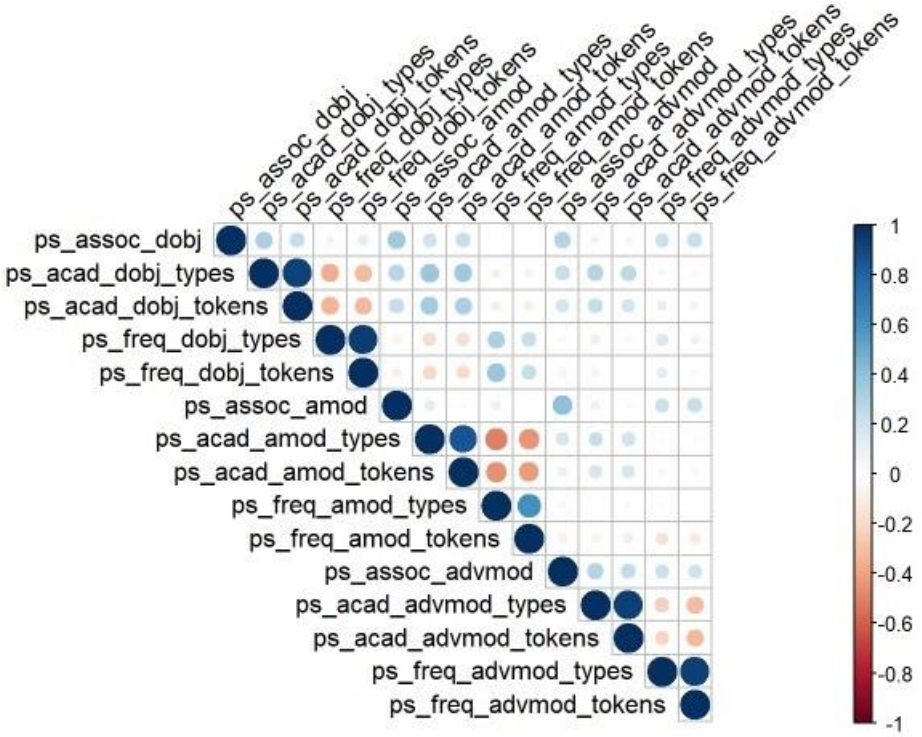

Figure 1: Correlation matrix for measures of phraseological sophistication in VESPA

5. Discussion

Results seem to indicate that at least some measures of phraseological sophistication by register specificity (here operationalized as the proportion of academic collocations present in learner texts), correlate with L2 proficiency. Hypothesis 1, however, is only partially confirmed, as different types of academic collocations exhibit varying behaviours. Measures relying on *amod* relations did not show correlation with CEFR levels. By contrast, those based on *advmod* and *dobj* relations exhibited a relationship with proficiency, albeit in distinct ways. In the VESPA corpus, *advmod* relations distinguish between upper-intermediate (B2) and advanced (C1) learners. By contrast, they fail to separate C2 learners from the other two groups, as C2 learners use fewer *advmod* collocations compared to C1 learners. Academic collocations in the form of direct object (*dobj*) dependencies stand out as the only category capable of distinguishing C2 learners from the other two groups, with a higher frequency observed in C2 texts compared to the others. This observation aligns with Paquot's (2019) finding that when phraseological sophistication was conceptualized in relation to association, MI scores for direct object relations were likewise the most distinguishing metric for C2 learners in comparison to other MI-based measures. Together, these results show that, for the dimensions of association and register specificity, dependency types correlate in learner writing (i.e., the more highly associated *dobj* relations, the more highly associated *amod* and *advmod;* the more academic collocations in the form of *dobj* dependencies, the more academic collocations in the form of *amod* and *advmod* dependencies). However, they also further underscore the specificity of verb + object collocations as a lexicogrammatical context posing particular difficulties to learners in academic settings, and where, unlike the other dependencies perhaps, there is scope for L2 phraseological competence to develop until the more advanced stages.

The findings presented in this study also diverge from those of Paquot (2019), who did not detect statistically significant differences between CEFR levels in VESPA using a similar measure of phraseological sophistication by register specificity (the proportion of academic collocations from the *Academic Collocation List*, Ackerman & Chen 2013). An important distinction between the two studies lies in the size of the two academic collocation lists used (6,529 academic collocations in the current study compared to 2,469 academic collocations in Paquot (2019)), resulting in a significantly higher proportion of collocations identified in learner texts across proficiency levels in the current study. For example, in the case of PS_ACAD_DOBJ_TYPES, mean values have shown a substantial increase from 0.009 – 0.02 to 0.14 – 0.18.[3] This substantiates our claim that the absence of significant results in Paquot (2019) most likely originated from the choice of the resource used. It certainly did not warrant the conclusion that measures of phraseological sophistication in relation to register specificity (here, academic collocations) were ineffective and not worthy of further exploration.

Unlike Hypothesis 1, Hypothesis 2 is not confirmed: different trends are noticed for the distinct metrics and no measure is statistically significant. For lack of theoretical work on co-occurrence frequency distribution, we decided to adopt a criterion from vocabulary research specializing in the development of word frequency lists and their coverage in learner texts. We operationalized low frequency co-occurrences as co-occurrences that appear in the ENCOW16AX web corpus but fall outside the set of co-occurrences necessary to achieve an 80% mean coverage threshold in VESPA. However, none of the computed measures demonstrated statistical significance and effectively differentiated between CEFR levels in the learner corpus. As happened with the first operationalization of phraseological sophistication by register specificity presented in Paquot (2019), we are tempted to believe that the results should not be taken as evidence that frequency is not a useful dimension of phraseological

[3] LS1*dobj* in Paquot (2019).

sophistication. We believe instead that there is some work ahead of us to develop theoretically sound and effective measures of phraseological sophistication in relation to frequency. It is very likely that the measures used in this study consider too few co-occurrences as low frequency co-occurrences. The 80% coverage threshold for each dependency is probably quite high given that high frequency words can be used to produce many different co-occurrences that can become quite rare, even in a large web corpus such as ENCOW16AX. Thus, to reach 80% coverage, *dobj* dependencies with a frequency as low as 0.13 occurrences per million words were categorized as frequent. Examples include co-occurrences such as *yield + argument*, *write + proposition*, *wreak + violence*, *witness + altercation*, *warrant + caution*, and *trigger + curiosity*. This classification may warrant reconsideration in future studies. Subsequent research efforts, however, should first address the pivotal question asked earlier – What are frequent, less frequent or rare co-occurrences? - before we can develop meaningful measures of phraseological sophistication in relation to frequency.

Finally, we explored patterns of variation among the distinct sets of metrics of phraseological sophistication by means of a correlation matrix. Hypothesis 3 is confirmed to the extent that measures that tap into the dimensions of association and register specificity (academic collocations) generally display intermediate correlations (except for *amod* dependencies). Thus, a learner text that has a higher mean MI value (which is often due to the fact that it includes more highly associated collocations and fewer co-occurrences with low or negative MI score) is also likely to feature more academic collocations. These results provide evidence that the two dimensions of association and register specificity are related. These dimensions can be related because some highly associated collocations are also considered academic collocations (e.g. *place + emphasis*, *play + role*, *adopt + stance*, *conduct + study*, *follow + procedure*, *bridge + gap*). However, this is not always the case (otherwise, the measures would correlate more strongly): many highly associated collocations are not academic

collocations (e.g., *pursue + career*, *paint + picture*, *do + justice*, *scratch + surface*, *win + election*, *arouse + curiosity*, *lose + sight*) and many academic collocations are not highly associated collocations (e.g., *note + difference*, *have + origin*, *have + effect*, *include + information*, *have + advantage*, *illustrate + relationship*, *see + figure*, *examine + use*). The fact that the correlation between the two measures is small to moderate is encouraging, as it confirms that they represent different facets of the sophistication of word combinations.

Hypothesis 3 is also disconfirmed to the extent that, unlike what we hypothesized, measures of phraseological sophistication by frequency do not correlate with measures of phraseological sophistication by association; contrary to our expectations, they are also negatively correlated with measures tapping into register specificity. Building upon the observations made in the previous paragraph, the negative correlations may come as further evidence that the set 80% coverage threshold may have been substantially too high to distinguish between frequent and infrequent co-occurrences. The correlation matrix revealed that a higher presence of academic collocations in a learner text within VESPA coincides with a decrease of low-frequency co-occurrences. This raises an important question: What precisely constitute these low-frequency co-occurrences? Why do they seem not to be indicative of enhanced phraseological competence? Among the low-frequency co-occurrences found in VESPA, some are discipline-specific word combinations that can be said to match our definition of sophisticated co-occurrences (*analyse + adjective*, *compare + wordlist*, *teach + collocation*, *translate + adverb*, *zap + instance*). Others, however, appear less idiomatic or more creative, and as such do not align well with our definition of phraseological sophistication. Examples include *articulate + war*, *challenge + dictionary*, *possess + pronunciation*, *present + vagueness*, *proffer + translation*, *rattle + hypothesis*, *reorganise + interaction*, and *trust + grammar*. Further research is warranted to develop measures of phraseological sophistication that effectively capture the dimension of frequency and are able to distinguish between these

various types of low frequency word combinations. In future studies, it may also be more effective to either reduce the set coverage threshold or explore EFL learners' use of mid-frequency co-occurrences.

6. Conclusion

The construct of phraseological complexity has already proved useful to describe L2 performance, assess L2 proficiency and trace L2 development across modes and tasks in a variety of L2s. However, as noted by Paquot (2021), scant attention has been given to construct definition, and there has also been a lack of investigation into the reliability and the validity of the metrics used to operationalize its two main dimensions: phraseological sophistication and phraseological diversity. As advocated by several leading experts in the field, including Norris and Ortega (2003), both the reliability and validity of our methods and measurements should be the focus of more theoretical thinking and empirical evaluation in L2 complexity research, as well as in SLA quantitative research more broadly.

In this study, we focused more specifically on the construct definition of phraseological sophistication and its measurement. A majority of the studies that have employed Paquot's (2019) framework to investigate learners' phraseological complexity in learner language have only used MI scores to measure phraseological sophistication in relation to association. In her initial proposal, however, Paquot (2019) already introduced a second type of measurement of phraseological sophistication that tapped into the specificity of word combinations (i.e., ratios of academic collocations) in learner texts. The limited uptake of these alternative measures in subsequent research can be ascribed to two main factors. First, there is currently a scarcity of essential resources such as lists of register-specific collocations, especially in languages other than English. Secondly, the apparent lack of significant results observed for measures of register

specificity in Paquot's (2019) original work has been used by several authors as grounds for their non-utilization. The consequence, however, is that the metrics used in the existing body of research, including my own work, have not fully represented the construct of phraseological sophistication that they are meant to measure, thereby raising issues of construct under-representation. Drawing on L2 vocabulary research, in this study, we have developed a conceptualization of phraseological sophistication that encompasses not only association but also considers register specificity and frequency – akin to the *n*-gram indices that were developed as part of measures of lexical sophistication in TAALES, Kyle and Crossley (2015).

Our findings provide robust empirical evidence for incorporating register specificity as a dimension of phraseological sophistication in a multidimensional framework. In terms of concurrent validity, like the MI-based measures used in Paquot (2019) and follow-up studies, measures of phraseological sophistication in relation to register specificity (academic collocations) have demonstrated their ability to distinguish at least between some of the CEFR groups that they should theoretically be able to distinguish and correlate with proficiency scores (cf. Crossley et al., 2013). They have also exhibited correlations with related association-based measures of phraseological sophistication (convergent validity), although these correlation values fall within the low-intermediate range. By contrast, we were not able to make an argument for the concurrent validity and convergent validity of frequency-based measures of phraseological sophistication. We believe this may at least be explained by the metrics implemented, which distinguished between a large proportion of frequent co-occurrences and a very limited set of low-frequency co-occurrences that include less idiomatic and more creative combinations. Further research will need to explore the frequency of relational co-occurrences and how it relates to phraseological sophistication in learner language.

With this study, we also want to emphasize the imperative of rigorously testing and validating all measures of phraseological complexity, including the ones first proposed in

Paquot (2019) and the more recently developed metrics. Ideally, empirical support for the reliability and validity of these measures should be drawn on diverse learner corpora representing various L1s and L2s, as well as different modes and tasks. Only through systematic evaluation can we ascertain the extent to which our conclusions can be confidently generalized. To further advance these research directions, the development of appropriate context-sensitive resources is required. To this end, we share the raw list of dependencies extracted from the LOCRA together with a variety of metrics related to frequency, range, keyness, and association (see the OSF page of this project; also Naets et al., 2024). In that way, researchers will be able to tailor the list to their specific needs. This database, however, may have limited applicability in contexts where learner language samples originated outside of academic settings. In such cases, operationalizing phraseological sophistication in relation to register specificity may require an alternative approach. For example, if a study investigates the phraseological complexity of narrative texts written by L2 learners, it may be more appropriate to use a collocation dataset compiled from a large reference corpus of fiction writing to operationalize phraseological sophistication in relation to register specificity. Similarly, future research should also explore the distribution of relational co-occurrences across a range of reference corpora with the aim of developing resources that could be used to investigate phraseological sophistication in relation to frequency.

Open Data badge

This article has been awarded an Open Data badge. The data is available from https://osf.io/jxbn8/?view_only=ce748f05c1c5494ba0cea6eebe483336. Learn more about the Open Practices badges from the Center for Open Science: https://osf.io/tvyxz/wiki.

Acknowledgments

We would like to thank the reviewers and the editors of the special issue for their constructive comments.

References

Ackermann, K., & Chen, Y. -H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247. https://doi.org/10.1016/j.jeap.2013.08.002

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*(2), 81-92. https://doi.org/10.1016/j.esp.2011.08.004

Bartsch, S. (2004). *Structural and functional properties of collocations in English*. Gunter Narr Verlag.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28–41. https://doi.org/10.1016/j.jslw.2014.09.004

Bestgen, Y., & Granger, S. (2017). Using collgrams to assess L2 phraseological development: A replication study. In P. de Haan, R. de Vries, & S. van Vuuren (Eds.), *Language, learners and levels: Progression and variation* (pp. 385–408). Presses Universitaires de Louvain.

Bestgen, Y., & Granger, S. (2018). Tracking L2 writers' phraseological development using collgrams: Evidence from a longitudinal EFL corpus. In S. Hoffman & A. Sand (Eds.), *Corpora and lexis* (pp. 277–301). Brill. https://doi.org/10.1163/9789004361133_011

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman grammar of spoken and written English*. Pearson Education Limited.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at …: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Bybee, J., & Beckner, C. (2012). Usage-based theory. In H. Narrog & B. Heine (Eds.), *The Oxford handbook of linguistic analysis* (pp. 827–855). Oxford University Press.

Chen, Y. -H., & Baker, P. (2016). Investigating criterial discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics*, *37*(6), 849–880.

Cobb, T. (2023). Web Vocab Profile [computer software]. Montréal: UQAM. Accessed 15 September 2023. http://www.lextutor.ca/vp.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Crossley, S. A., Cai, Z., & McNamara, D. S. (2012). Syntagmatic, paradigmatic, and automatic *n*-gram approaches to assessing essay quality. In P. M. McCarthy & G. M. Youngblood (Eds.), *Proceedings of the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 214–219). AAAI Press.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2013). Validating lexical measures using human scores of lexical proficiency. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 105–134). John Benjamins. https://doi.org/10.1075/sibil.47.06ch4

De Marneffe, M.-C., & Manning, C. (2016). *Stanford typed dependencies manual*. http://nlp.stanford.edu/software/dependencies_manual.pdf.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, *28*(3), 157–169. https://doi.org/10.1016/j.esp.2009.02.002

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL – International Review of Applied Linguistics in Language Teaching*, *47*(2), 157–177. https://doi.org/10.1515/iral.2009.007

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking and points of order. *Studies in Second Language Acquisition*, *18*(1), 91–126. https://doi.org/10.1017/S0272263100014698

Evert, S. (2004). *The statistics of word cooccurrences: Word pairs and collocations.* Unpublished PhD thesis. Universität Stuttgart. https://www.stephanie-evert.de/PUB/Evert2004phd.pdf

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, *67*(S1), 155–179. https://doi.org/10.1111/lang.12225

Garner, J., Crossley, S., & Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System*, *80*, 176–187. https://doi.org/10.1016/j.system.2018.12.001

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

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

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers : A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, *52*(3), 229–252. https://doi.org/10.1515/iral-2014-0011

Gries, S. Th. (2013). *Statistics for linguistics with R. A practical introduction* (2nd ed.). De Gruyter Mouton.

Gries, S. Th. (2022). What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies*, *5*(1), 1–33. https://doi.org/10.1075/jsls.21028.gri

Hoey, M. (1991). *Patterns of lexis in text*. Cambridge University Press.

Honnibal, M., & Montani, I. (2017*). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*. https://sentometrics-research.com/publication/72/

Hu, R., Wu, J., & Lu, X. (2022). Word-combination-based measures of phraseological diversity, sophistication, and complexity and their relationship to second language Chinese proficiency and writing quality. *Language Learning*, *72*(4), 1128–1169. https://doi.org/10.1111/lang.12511

Jiang, J., Bi, P., Xie, N., & Liu, H. (2023). Phraseological complexity and low- and intermediate-level L2 learners' writing quality. *International Review of Applied*

Linguistics in Language Teaching, *61*(3), 765–790. https://doi.org/10.1515/iral-2019-0147

Kim, M., & Crossley, S. A. (2023). Lexical and phraseological differences between second language written and spoken opinion responses. *Frontiers in Psychology*, *14*, 1068685. https://doi.org/10.3389/fpsyg.2023.1068685

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon : Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, *102*(1), 120–141.

Kim, S., & Kessler, M. (2022). Examining L2 English university students' uses of lexical bundles and their relationship to writing quality. *Assessing Writing*, *51*, 100589. https://doi.org/10.1016/j.asw.2021.100589

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., & Eguchi, M. (2021). Automatically assessing lexical sophistication using word, bigram, and dependency indices. In S. Granger (Ed.), *Perspectives on the L2 phrasicon: The view from learner corpora* (pp. 126–151). Multilingual Matters.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*(2), 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes*, *31*(1), 25-35. https://doi.org/10.1016/j.esp.2011.07.002

McCallum, L. (2020). Relationships between measures of phraseological complexity and writing quality in a CEFR assessment context. *Arab Journal of Applied Linguistics*, *5*(1), 63-99. https://doi.org/10.1234/ajal.v5i1.197

Naets, H., Shen, J., & Paquot, M. (2024). *A database of English dependencies with measures of frequency, association, range and keyness*. [Data set]. https://doi.org/10.14428/DVN/Y6Y4G8. Open Data @ UCLouvain, V1.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Norris, J., & Ortega, L. (2003). Defining and measuring SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 717–761). Wiley. https://doi.org/10.1002/9780470756492

Paquot, M. (2010). *Academic vocabulary in learner writing: From extraction to analysis*. Continuum.

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. DOI: 10.1080/15434303.2017.140542

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145. https://doi.org/10.1177/0267658317694221

Paquot, M. (2021, August 15-21). *Measures of phraseological complexity: reliability and validity* [paper presentation]. Investigating complexity in L2 phraseology: methods and

applications symposium, AILA World Congress 2021, University of Groningen, The Netherlands.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130–149. https://doi.org/10.1017/S0267190512000098

Paquot, M., Larsson, T., Hasselgård, H., Ebeling, S. O., De Meyere, D., Valentin, L., Laso, N. J., Verdaguer, I., & van Vuuren, S. (2022b). 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. https://doi.org/10.32714/ricl.10.02.02

Paquot, M., Naets, H., & Gries, S. Th. (2021). Using syntactic co-occurrences to trace phraseological complexity development in learner writing: verb + object structures in LONGDALE. In B. Le Bruyn & M. Paquot (Eds.), *Learner corpus research meets second language acquisition* (pp. 122–147). Cambridge University Press.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Read, J. (2000). *Assessing vocabulary*. Oxford University Press.

Rogers, J., Müller, A., Daulton, F. E., 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. https://doi.org/10.1016/j.esp.2021.01.001

Römer, U. (2009). The inseparability of lexis and grammar. Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, *7*, 141–163.

Rubin, R., Housen, A., & Paquot, M. (2021). Phraseological complexity as an index of L2 Dutch writing proficiency: A partial replication study. In S. Granger (Ed.), *Perspectives on the L2 phrasicon* (pp. 101–125). Multilingual Matters. https://doi.org/10.21832/9781788924863-006

Schäfer, R. (2015). Processing and querying large Web corpora with the COW14 architecture. In P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, & W. Andreas (Eds.), *Proceedings of the 3rd workshop on challenges in the management of large corpora* (CMLC-3) (pp. 28–34). Leibnitz-Institut für Deutsche Sprache.

Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing and use*. John Benjamins.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes*, *12*(3), 214–225. https://doi.org/10.1016/j.jeap.2013.05.002

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. https://doi.org/10.1075/ijlcr.20015.van

Vandeweerd, N., Housen, A., & Paquot, M. (2022). Comparing the longitudinal development of phraseological complexity across oral and written tasks. *Studies in Second Language Acquisition*, *45*(4), 1–25. https://doi.org/10.1017/S0272263122000389

Vandeweerd, N., Housen, A., & Paquot, M. (2023). Proficiency at the lexis–grammar interface: Comparing oral versus written French exam tasks. *Language Testing*, *40*(3), 658–683. https://doi.org/10.1177/02655322231153543

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. University of Hawaii Press.

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

Xia, D., Chen, Y., & Pae, H. K. (2022a). Lexical and grammatical collocations in beginning and intermediate L2 argumentative essays: A bigram study. *IRAL – International Review of Applied Linguistics in Language Teaching*, *61*(4), 1421–1453. https://doi.org/10.1515/iral-2021-0188

Xia, D., Ai, H., & Pae, H. K. (2022b). Please let me know: Lexical bundles in business emails by business English learners and working professionals. *International Journal of Learner Corpus Research*, *8*(1), 1–30. https://doi.org/10.1075/ijlcr.20019.xia

Yoo, I. W., & Shin, Y. K. (2022). English lexical bundles in a learner corpus of argumentative essays written by Korean university students. *Corpora*, *17*(Supplement), 23–42. https://doi.org/10.3366/cor.2022.0245

Zhang, Y., & Ouyang, J. (2023). Linguistic complexity as the predictor of EFL independent and integrated writing quality. *Assessing Writing*, *56*, 100727. https://doi.org/10.1016/j.asw.2023.100727

Address for correspondence

Magali Paquot

UCLouvain

ILC

Place Cardinal Mercier 31/L3.03.33

1348 Louvain-la-Neuve

Belgium

Magali.paquot@uclouvain.be

Publication history

Date received: 29 September 2023

Date accepted: 26 July 2024