Exploring complexity at the lexis-grammar interface: diversity and sophistication of verb-argument structures in L2 Dutch writing

Abstract

While the construct of phraseological complexity has received increasing attention within the field of L2 complexity research, few measures have targeted more abstract phraseological units such as verb-argument structures (VAS) that have been found to play a substantial role in L2 proficiency, performance, and development. The present study therefore aims to expand upon the units targeted in existing phraseological complexity research by investigating measures of VAS diversity and sophistication, respectively operationalized as the type-token ratio (TTR) and mean mutual information (MI) score of these units, alongside previously investigated diversity and sophistication measures for verb + direct object dependency units. The complexity measures are computed for a corpus of L2 Dutch writing collected from the B1, B2, and C1 levels of the Certificaat Nederlands als Vreemde Taal (CNaVT) exams, and are entered as predictors in multiple regression models predicting the numeric overall exam scores. Results indicate that the measure of VAS sophistication, in particular, demonstrates strong potential as a quantitative measure to include alongside more established measures of dependency-based phraseological units. Not only does this measure exhibit a clear relationship with learner proficiency at the more advanced stages, but it also appears to supplement the predictive ability of existing phraseological complexity measures.

Keywords: Phraseological complexity; argument structure; L2 Dutch; proficiency; lexis-grammar interface

1. INTRODUCTION

Increased interest in linguistic complexity has enriched L2 research with a deeper understanding of the core components underpinning learner proficiency, performance, and development, in addition to a battery of quantitative measures and tools to investigate these constructs. However, despite the undeniable contributions towards theory and methods in L2 research, there have also been growing concerns about the validity of measures of linguistic complexity (Bulté & Housen, 2012; Pallotti, 2015), as well as the relevance and usefulness of complexity measures for tracking L2 development and/or evaluating L2 proficiency (Biber et al., 2020; Pallotti, 2023). Recent investigations into L2 complexity have also attempted to address ‘reductionist’ trends by expanding the domains of language, as well as the range of features within each domain of language, that are operationalized within the construct of linguistic complexity (Biber et al., 2020; Housen et al., 2019). The field has thereby largely progressed from the broad measures of syntactic complexity that were once representative of the majority of studies, in favor of measures targeting more narrow linguistic features across various domains of language (Kyle & Crossley, 2018; Housen et al., 2019; Crossley, 2020). This progression has helped to broaden the scope of the complexity construct, to increase the specificity of the resulting analyses, and to better capture L2 proficiency and development. For example, fine-grained phrasal measures of syntactic complexity provide greater linguistic interpretability than omnibus measures of sentence/T-unit/clause length, and have been reported in several studies to better predict measures of L2 proficiency, in particular L2 writing proficiency, than more traditional, and less-interpretable, measures (Biber et al., 2020; Kyle & Crossley, 2018).

Recently, the construct of phraseological complexity and measures thereof have increasingly featured in L2 complexity research (Paquot, 2018, 2019). The phraseological dimension of language deals with various linguistic phenomena that straddle the lexis-grammar interface, and are therefore not adequately gauged by either syntactic or lexical complexity measures. The increasing interest in phraseological complexity measures follows from the important role that phraseological competence has been found to play in L2 proficiency and development, specifically at the more advanced levels of learner proficiency (Durrant & Schmitt, 2009; Granger & Bestgen, 2014). Paquot (2019) defined the construct of phraseological complexity as “the range of phraseological units that surface in language production and the degree of sophistication of such phraseological units” (p. 124), and put forward diversity and sophistication measures for several phraseological units. Particularly noteworthy is the development of indices of diversity and sophistication for phraseological units defined by their grammatical relation, such as verb + direct object relations (e.g., ‘consider’ + ‘idea’), which have been found to be strong predictors of L2 writing quality and L2 proficiency across several L2s and learning contexts (L2 English: Kyle & Eguchi, 2021; Paquot 2018, 2019; L2 Dutch: Rubin et al., 2021; L2 French: Vandeweerd et al., 2021).

Thus far, the scope of these phraseological measures targeting grammatical relations has been restricted to only a few grammatical relations, and only one type of verb-argument relation, namely verb + direct object co-occurrences. The co-selection of verb lemmas and argument structures (i.e., the number and nature of verbal arguments) therefore offers an interesting new direction for phraseological complexity, allowing for the definition and extraction of a broader set of phraseological units that are not entirely dependent upon lexical representations. As L2 development progresses, learners have been found to use more structurally complex argument structures in combination with increasingly sophisticated main verb lemmas (Römer, 2019). Whereas existing research has taken a more qualitative approach to evaluating the argument structure and verb lemma repertoires of learners, quantitative measures of verb-argument structure (VAS) diversity and sophistication can capture this increase in the complexity of VAS systematically for a variety of applied purposes, such as the assessment of learner proficiency and performance. The main aim of the present study is to incorporate VAS units (e.g., 'consider' + transitive) into the battery of measures tapping into the phraseological dimension of linguistic complexity. To this end, we propose a number of quantitative measures and apply them to a learner corpus of L2 Dutch. The focus on L2 Dutch constitutes another important contribution to L2 research by allowing us to expand the scope of L2 complexity research in general, as most L2 research thus far focuses on English as an L2. More specifically, the acquisition of VAS in Dutch has yet to be addressed in L2 research.

2. PHRASEOLOGICAL COMPLEXITY

2.1 Collocation complexity

Phraseological complexity measures currently implemented in L2 complexity research tap into both the diversity and sophistication of various phraseological units, including statistical collocations (positional and dependency-based) and n-grams. Positional collocations consist of lexical items that occur together within a specified window span more frequently than would be expected from their respective distributions and are therefore entirely defined by their frequency of (co-)occurrence; consequently, these statistical collocations are situated primarily in the lexical domain. To include phraseological units situated at the interface of lexis and grammar, Paquot (2018, 2019) made use of automatic grammatical dependency annotations to extract statistical collocations that constitute a grammatical relation (specifically verb + direct object, adjectival modifier, and adverbial modifier relations). Being both lexically and structurally defined, these phraseological units exemplify the overlap between phraseology and the lexis-grammar interface.

Frequency of co-occurrence information is used to compute statistical association measures, which gauge the likelihood of co-occurrence given the respective distributions of the individual components of the phraseological units. The degree of association between the components of phraseological units has been operationalized as an indicator of sophistication within existing L2 complexity research. Pointwise mutual information (MI), in particular, has been found to be a good indicator of association strength (Gries, 2022). Whereas sophistication measures are grounded in frequency of (co-)occurrence and are computed on the basis of frequencies in a reference corpus, diversity measures gauge the range of phraseological units produced by a learner within a text. Phraseological diversity measures typically consist of a variation of the type-token ratio (TTR).

Several studies have reported promising findings in terms of the relationship between diversity and sophistication measures computed for statistical collocations and learner proficiency and development, specifically when targeting the intermediate-to-advanced stages of learner proficiency. Paquot (2018, 2019) identified no significant differences in syntactic or lexical complexity (with the exception of a small difference between B2/C1 and C2 for one lexical measure, lexical density) between L2 English written productions at the B2, C1, and C2 levels of the Common European Framework of Reference for Languages (CEFR), while significant differences were identified for MI-based sophistication measures of three dependency-based units: verb + direct object, adjectival modifier, and adverbial modifier dependencies (though no significant differences for diversity measures for these units were identified). Similarly, Kyle & Eguchi (2021) found that dependency-based sophistication measures (as opposed to bigram-based measures), in combination with word-level complexity measures, serve as significant predictors of holistic writing scores for L2 English across a range of learner backgrounds. For L2 French, measures of phraseological sophistication (mean MI of verb + direct object units) and diversity (of adjectival modifiers) were found to contribute to the classification of learner proficiency in a random forest model, in combination with measures of lexical, syntactic, and morphological complexity (Vandeweerd et al., 2021). Phraseological diversity and sophistication measures for dependency-based units have recently been applied to L2 Dutch exam data, where these measures contributed more to the prediction of the pass/fail exam outcome at the B1 and B2 exam levels than traditional measures of syntactic and lexical complexity (Rubin et al., 2021).

2.2 VAS complexity

Studies investigating measures of verb + direct object diversity and sophistication have identified the VAS as a promising locus for operationalizations of phraseological complexity suitable for predicting learner performance across a variety of L2s and learning contexts (Paquot, 2018, 2019; Rubin et al., 2021; Vandeweerd et al., 2021). Usage-based approaches have also found VAS, in the form of verb-argument constructions (henceforth VACs[1]), to be a central element in the process of language acquisition, and various findings have confirmed an increase in the complexity of these structures as language proficiency develops (Ellis & Ferreira-Junior, 2009; Ellis et al., 2014). This line of research originates in child language acquisition (e.g., Goldberg,

[1] "Verb-argument structure" is used here as a more general term for the relationships between a verbal head and its arguments. "Verb-argument construction", on the other hand, is a term coined in the constructionist literature to describe the same type of grammatical relations, and is therefore subject to more theory-specific criteria. The differences between VAS and VACs are presented in more detail in Section 3.3.

2004), but more recent research has focused on the acquisition of VACs in L2 learners. Ellis and Ferreira-Junior (2009) investigated the factors driving L2 VAC acquisition and found that both the frequency of verbs and the association strength between verbs and the VACs they occur in as measured in the input determine the first-learned verbs within a particular VAC. Similar sensitivity to verb frequency and verb-VAC association strength was found for advanced L2 English learners in a later study, where an experimental design was implemented to determine whether the most readily accessible verbs for a particular VAC are influenced by the frequency, association strength, and semantic prototypicality of those verbs in a reference corpus (Ellis et al., 2014). These findings indicate that phraseological sophistication, operationalized as association strength in existing phraseological complexity research, might be expected to capture some of the developmental trends observed in L2 VAC acquisition. However, because both studies focused on rather homogeneous learner groups in terms of proficiency, development of VAC sophistication did not factor into these investigations. In a comparison of VAC types produced by L2 English learners across proficiency levels, Römer (2019) found that learners at higher proficiency levels possessed a larger repertoire of VAC types and also made use of lower frequency verbs within VACs more often than lower proficiency learners. Moreover, the lower proficiency learners were found to rarely, if at all, produce more complex VACs. While the findings reported from these usage-based investigations emphasize VAC frequency and association strength as mechanisms driving L2 VAC acquisition, little research has implemented quantitative measures targeting verbs and their argument structures to describe learner proficiency and development.

The co-selection of verbs and argument structures can be operationalized and measured in a similar way to the verb + direct object phraseological units that have already been incorporated into the construct of phraseological complexity. Although they target similar grammatical structures, verb + direct object units and VAS units differ in their proximity to the lexis-grammar interface. Verb + direct object units consist of co-occurring lexical items that form a grammatical relation, more specifically a transitive verb and a noun, and are influenced by different lexical and grammatical constraints than the combination between an argument structure (without lexical representation) and a lexical verb. The complexity measures that target these different phraseological units therefore tap into different aspects of a learner's knowledge of the language. Increased access to VAS of varying degrees of structural complexity and a growing repertoire of verb lemmas leads to a wider range of VAS observed in learner language production as proficiency develops (Römer, 2019), which is captured by measures of diversity. In terms of sophistication, both frequency and association have been reported to impact the acquisition of VAS (Ellis & Ferreira-Junior, 2009). Argument-structures tend to occur first with verbs that occur more frequently and exclusively in those structures, subsequently expanding to less frequent verbs that exhibit less predictable patterns of co-occurrence with argument structures. Sophistication measures operationalizing MI are suited for identifying the degree to which patterns of co-occurrence are predictable, and in this way capture some of the primary factors influencing the acquisition of VAS.

Similar indices of VAS sophistication have been explored from a usage-based perspective by Kyle and Crossley (2017), who reported findings that indicate a significant relationship between L2 writing quality and both frequency and association strength of VACs. Their approach adopts a broader definition of syntax, and they conceptualize these indices within the construct of syntactic sophistication as opposed to phraseological sophistication. As a result, their analysis considers indices of VAC sophistication in comparison to traditional measures of syntactic complexity rather than complexity measures targeting other phraseological units, such as statistical collocations, leaving unanswered questions about the relationship between the complexity of different types of phraseological units. Moreover, VACs as defined in a traditional usage-based framework are highly lexicalized, e.g., incorporating function words, and are therefore more restrictive in terms of possible combinations with lexical verbs in comparison to more schematic approaches to argument structure like the one adopted here.

In the present study, we propose measures of VAS diversity and sophistication to expand the range of units included within the operationalization of phraseological complexity by including units situated at the interface of lexis and grammar. The research is guided by the following research questions:

1. To what extent do measures of VAS complexity contribute to the prediction of the assessment of L2 Dutch proficiency?

2. Do measures of VAS complexity exhibit similar relationships with L2 Dutch proficiency as measures of verb + direct object complexity?

3. METHOD

In this section, we describe the following: (i) the learner data, (ii) the methods used to extract the target phraseological units, (iii) the definition of the target units, (iv) all complexity measures computed in this study, and (v) the statistical analyses used to answer the research questions. All scripts used in the processing of the learner and reference corpora in this study, in addition to the processed data used to carry out the statistical analyses, have been made available through OSF (https://osf.io/hp7mu/). The learner texts have not been made available, as these texts are not open access. The reference corpus used here (Lassy Groot corpus; van Noord et al., 2006) is available upon request from the Instituut voor de Nederlandse taal [Institute for the Dutch language][2].

3.1 Learner data

The data analyzed in the present study were collected from the Certificaat Nederlands als Vreemde Taal (CNaVT) exams administered by the Centrum voor Taal en Onderwijs (CTO) at KU Leuven, from 2016-2018. These exams are offered at four levels of the CEFR, i.e., A2, B1, B2, and C1, and serve to fulfill various Dutch language requirements (e.g., educational, professional, and residency). In total, over 3,000 written exam productions have been collected from three exam levels: the B1, B2, and C1 levels. As the exams are administered locally worldwide and evaluated centrally in Belgium, the learners vary in L1 background and learning context, although the majority consists of L1 French speakers from Belgium (52%). The exams consist of several tasks each, of which two writing tasks per exam year/level have been collected, with the exception of the B1 level exams, for which only one writing task has been collected per exam year. Only one text per exam was included in this study, meaning that no learner produced more than one text in the sample used here. Each exam excerpt is written in response to a prompt that is unique to the exam year and level, though the tasks can be grouped into three genres: informative, argumentative, and scientific. The full CNaVT dataset has been reduced here to include only the exams for which exam scores have been recorded, as well as to ensure that the texts at each exam level belong to the same text genre (B1: informative; B2: argumentative; and C1: scientific). The exam topics and genres for each exam level that have been included in the present study can be consulted in Table 1.

[2] https://taalmaterialen.ivdnt.org/download/tstc-lassy-groot-corpus/

Table 1. Distribution of L2 Dutch data across learners, topics, genres, and CEFR levels

CEFR	Genre	Topic	Learners	Tokens	Mean(SD)
B1	Informative	Information evening: traffic in Utrecht	171	34,920	204.21(56.32)
B2	Argumentative	Radio segment: newspaper under pressure	425	100,631	236.78(45.64)
	Argumentative	Admissions exams	386	100,491	261.02(55.10)
C1	Scientific	"The reading brain"	79	21,345	270.19(59.10)
	Scientific	Medical research	111	34,400	309.91(59.91)

The exams are evaluated according to formal and content assessment grids for each task, which together are converted to a numeric overall score for each separate exam level, and ultimately a binary pass/fail decision. Because of the discrepancies in task genre, topic, and assessment grids across the exam levels, in combination with the uneven distribution of learners across exam levels and the fact that the exams are intended to function as isolated entities, the exam data are analyzed independently within each CEFR exam level in the present investigation. This is thought to be the most ecologically valid way of handling the data. The numeric overall exam score achieved by each learner is used as the indicator of proficiency/performance in the present study.

3.2 Corpus pre-processing

Automatic linguistic annotation was provided for the learner corpus texts using the Dutch-language dependency parser Alpino (van Noord, 2006). The Lassy Groot corpus (van Noord et al., 2006), the Dutch reference corpus used here in the computation of phraseological sophistication measures, is annotated with the same tool. The linguistic annotations include lemmas, POS (part-of-speech) tags, dependency relations, lexical categories, and subcategorization frames, among others. The texts were not corrected for orthographic or other errors. The parser used in this study is not intended for the annotation of learner language, but a thorough investigation of the accuracy of the tool on uncorrected L2 Dutch writing found an overall accuracy of 83.7%, compared to the original reported accuracy of 90.9%, with some variation across dependency relation and learner proficiency level (Rubin, 2021). Both the dependency-based and VAS phraseological units operationalized in the current study are extracted using these linguistic annotations.

The verb + direct object dependency-based units were extracted from both the learner and reference corpora using POS tags and dependency relation annotations by means of in-house R scripts. A verb + direct object dependency consists of a verbal head with a nominal dependent that has been annotated with the 'obj1' dependency label. The argument structures (defined below) were extracted from both the learner and reference corpora, in combination with the main verb lemma that occurs in each finite verb phrase, using a combination of in-house XQuery and R scripts. The lexical category annotations were used to distinguish finite and non-finite verb phrases, subcategorization frames were used to distinguish between passive and active verb phrases, and dependency labels were used to reconstruct argument structures for each finite verb phrase.

3.3 VAS in Dutch

There are various approaches to the definition and classification of the relationship between a verb and its arguments. While similar analyses have approached the definition of argument structures from a constructionist perspective (Kyle & Crossley, 2017; Römer, 2019), the approach taken here aims at a broader classification of argument structures that can be applied more generally across theoretical frameworks. Although there is substantial overlap between VAS (as they are treated in the present article) and VACs as defined in usage-based approaches (e.g., Kyle & Crossley, 2017), they differ in several important respects: (i) VACs often include modal verbs and copulas as a main verb lemma, (ii) VACs are often highly lexicalized, whereas the definition adopted here is purely schematic, and (iii) VACs often include arguments with a wider range of lexical categories, compared to the present definition which accounts for nominal and verbal complements only.

An annotation scheme for Dutch argument structures was devised according to the types of arguments allowed in Dutch and the number of those arguments taken by a verb, following several Dutch reference grammars (Balk-Smit Duyzentkunst, 2000; Klooster, 2001). A major departure from constructionist approaches is the treatment of tense, aspect, and modality (TAM) marking in the verb phrase, as well as the treatment of copular verbs. Argument structures across various tenses, aspects, and moods were collapsed into single argument structure categories, as these TAM markers do not influence the number or nature of the arguments in the verb phrase. Copular verbs are not included among the argument structures as predicates of copular verbs are subject complements (as opposed to object complements in the other argument structures considered in this study). Because word order is less rigid in Dutch than in English (for which previous automatic extraction protocols for VAS have been developed), variable word order was not considered an instance of additional argument structures. The treatment of prepositional arguments in Dutch is a complex issue. The so-called *meewerkend* ('cooperative') and *belanghebbend* ('interested') indirect objects are the most widely accepted prepositional arguments, while the *voorzetselvoorwerp* ('prepositional object'), which behaves more like a direct object, has been the subject of much debate (Duinhoven, 1989). Because of the complex set of semantic and syntactic criteria for identifying these arguments, which additionally causes substantial difficulty for automatic identification, as well as the debated status of certain prepositional arguments, only nominal and verbal complements are considered as part of the argument structure in the present analysis.

We distinguish between five different argument structures, exemplified in (1), forming the basis of the annotation criteria used in the automatic extraction protocol.

(1)

(a) Intransitive:

'…dat	mijn	toetsenbord	niet	meer	goed functioneer-de.'
…that	1SG.GEN	keyboard	NEG	longer	good function-PAST.SG

'…that my keyboard didn't function well anymore.'

(Lassy groot corpus, sentence id: WS-U-T-B-0000000892.p.5.s.3)

(b) Transitive NP:

'We	gaan	dat	direct	proberen.'
1PL.NOM	go.1PL	that	directly	try.INF

'We will try that right away.'

(sentence id: WS-U-T-B-0000000887.p.8.s.6)

(c) Transitive VC:

'…dat	men	pas	na	de	oorlog	bemerk-te	hoe
…that	PRO	only	after	DEF.SG	war	notice-PAST.SG	WH

slecht	de	kwaliteit	was	geweest.'
bad	DEF.SG	quality	be.PAST.SG	be.PTCP

'…that people only noticed how bad the quality had been after the war.'

(sentence id: WR-P-E-J-0000028945.p.4.s.1)

(d) Ditransitive NP NP:

'Veel	mens-en	stel-den	mij	regelmatig
Many	person-PL	pose-PAST.PL	1SG.ACC	regularly

de	vraag…'
DEF.SG	question…'

'Many people asked me the question regularly.'

(sentence id: WS-U-T-B-0000000901.p.4.s.4)

(e) Ditransitive VC NP:

'…sommige	internetproviders	schrijv-en	je	voor	wat
…some	internet.providers	prescribe-PL	2SG.ACC	VP	WH

hier	in	moet	staan.'
here	in	must:AUX.3.SG	stand.INF.'

'…some internet providers tell you what needs to be included here.'

(sentence id: WS-U-T-B-0000000937.p.7.s.5)

A sample of 166 sentences was extracted from the Lassy Groot reference corpus, and each finite verb phrase was annotated by the first author and a second annotator for argument structure. The percentage agreement between the two annotators reached 94.1%. The annotations provided by the first author were then used to assess the extraction accuracy of the automatic extraction procedure. It is also worth noting that the automatic extraction procedure made use of several layers of linguistic annotation provided by the Dutch-language dependency parser Alpino (van Noord, 2006), and the annotation of argument structures is therefore dependent on the way that arguments are annotated by the tool. The argument structure categories included in the present analysis, their extraction criteria, and their automatic extraction accuracy rates are presented in Table 3. The extraction accuracy overall suffers from erroneous automatic linguistic annotations, but some additional patterns can be observed that contribute to the suboptimal precision and recall values: prepositional phrases labeled as complements in the automatic annotations and as adjuncts in the manual annotations; lexical category of infinitive verbs labeled as 'noun phrase' instead of 'infinitive' in the automatic annotations; and auxiliary verbs labeled as passive auxiliary in one annotation set and tense auxiliary in the other annotation set.

Table 3. Automatic extraction accuracy for VAS

Argument structure	**Extraction criteria**	**Manual N**	**F1-score**	**Precision**	**Recall**
Intransitive	Zero object complements	79	0.63	0.59	0.67

Transitive NP	Nominal direct object complement	166	0.82	0.87	0.78
Transitive VC	Verbal direct object complement	9	0.70	0.64	0.78
Ditransitive	Direct and indirect object complements	9	0.57	0.5	0.67

Note: Manual N = token count for each structure in manual annotations provided by the first author. The precision and recall measures for the two types of ditransitive structures are merged due to an overall low occurrence of these structures in the manually annotated sample (only nine occurrences combined).

3.4 *Phraseological complexity measures*

The same measures of phraseological diversity and sophistication were computed for the verb + direct object dependency-based units and for the VAS units. As a measure of diversity, the TTR is computed from the learner texts as the number of unique phraseological unit types over the total number of unit tokens per text. Because of the overall low number of phraseological units per text, the bare TTR was used as a measure of phraseological diversity as opposed to a diversity measure less sensitive to text length (e.g., McCarthy & Jarvis, 2010).

To compute MI-based measures of phraseological sophistication, frequency counts are recorded in a contingency table for all phraseological units extracted from the reference corpus and for the individual components of each unit. Based on these frequency counts, MI is computed as a measure of association strength for each phraseological unit, using the Ngram Statistics Package (Banerjee & Pedersen, 2003). The phraseological units extracted from the learner corpus (following the same procedure outlined for the reference corpus) are then searched in the lists of units extracted from the reference corpus, and are assigned the frequency and association strength values recorded for the same units in the reference corpus. The mean MI score for each of the two types of phraseological units targeted in this study is computed per text as a measure of phraseological sophistication. Additionally, the average frequency of each of the individual components of the VAS units (i.e. the argument structure and the main verb lemma) is computed per text to account for the effects of the average frequency of argument structures and main verb lemmas on the proficiency assessment of the learners. By considering frequency and association as separate variables, it is possible to distinguish between their effects on the proficiency exam outcome. A comprehensive list of complexity indices computed in the framework of the present analysis can be consulted in Table 4.

Table 4. Phraseological complexity measures computed in current study

Phraseological unit	Diversity	Sophistication
Verb + direct object dependency	Type-token ratio	Mean MI
VAS	Type-token ratio	*Association* Mean MI *Frequency* Mean frequency argument structures

3.5 *Statistical analysis*

The relationship between the measures of phraseological complexity and the assessment of L2 Dutch proficiency (RQ1) is first explored by means of bivariate Pearson correlation coefficients, followed by multiple linear regression modeling. For each proficiency level, the diversity and sophistication measures were entered as predictors into a linear regression model predicting the numeric overall exam score. For the B2 and C1 levels, for which two different writing tasks are included among the data analyzed, interactions were included between all complexity measures and the topic of the prompt to control for variable relationships between the measures and the outcome variable for the two writing tasks. Sum contrast coding (or deviation coding) was used to code the categorical topic variable so that the main effects in the model reflect the average of the two topics in each of the B2 and C1 models, and the interaction effects indicate the deviation of each topic from the main effect for each predictor.

The extent to which the dependency-based and VAS measures capture overlapping or complementary information (RQ2) is explored both via the bivariate Pearson correlation coefficients between all pairs of phraseological complexity measures as well as through the significance and effect size (i.e. standardized coefficients) of the measures when modeled together in the linear regression models.

4. RESULTS

Before considering the relationship between the measures of phraseological complexity and the exam results, we first look at the evidence for argument-structure use in L2 Dutch across the three CEFR levels. As is typical for corpus data (Gries, 2014), the distribution across argument structures is very uneven, with two of the five structures accounting for the majority of occurrences in our L2 Dutch corpus (see Table 5). For each of the three CEFR levels included in the current study, transitive NP structures account for the largest percentage of the five argument structures extracted from the corpus, followed by intransitive, transitive VC, ditransitive NP NP, and finally ditransitive NP VC structures.

Turning to the bivariate correlations between the diversity and sophistication measures and the exam scores, we can observe certain trends that hold across the three CEFR levels and others that are more divergent (see Table 6). The measures exhibiting the strongest and most stable correlation with the exam score across the three levels are the mean MI of verb + direct object dependencies, the mean MI of VAS, and one frequency variable for sophistication of VAS: mean verb frequency for main verbs in VAS units.

Table 5. Distribution of extracted VAS across CEFR levels

	B1	**B2**		**C1**	
	Information evening: traffic in Utrecht	*Radio segment: newspaper under pressure*	*Admissions exams*	*“The reading brain”*	*Medical research*
Intransitive	1,774 (42.7%)	3,599 (32.6%)	2,888 (26.7%)	967 (33.0%)	888 (27.6%)
Transitive NP	1,984 (47.8%)	5,889 (53.3%)	6,260 (57.9%)	1,511 (51.6%)	1,907 (59.2%)
Transitive VC	250 (6.0%)	1,227 (11.1%)	1,412 (13.1%)	328 (11.2%)	310 (9.6%)

Ditransitive NP NP	114 (2.7%)	271 (2.5%)	175 (1.6%)	74 (2.5%)	92 (2.9%)
Ditransitive VC NP	31 (0.7%)	53 (0.5%)	71 (0.7%)	46 (1.6%)	22 (0.7%)
Total	4,153	11,039	10,806	2,926	3,219

Correlations between the complexity measures are low to moderate (see supplementary materials on OSF), with the strongest correlation reaching -.52. Diversity measures (TTR) for both types of phraseological units demonstrate a low to moderate correlation (.24, .39, .43) across the CEFR levels. Sophistication measures for the two types of phraseological units exhibit slightly lower correlations (.26, .06, .21) that are more variable across the proficiency levels. None of the correlations between complexity measures warrant exclusion in the modeling of learner proficiency undertaken in the next step of the analysis.

The following sections review the output of three linear regression analyses that model the numeric exam score for each CEFR exam level as a function of the phraseological complexity measures. These analyses evaluate whether the trends observed in the bivariate correlations hold when all complexity measures are allowed to covary.

Table 6. Bivariate Pearson correlations between complexity measures and exam scores for each exam level

CEFR	*MI mean dobj*	*TTR dobj*	*MI mean VAS*	*TTR VAS*	*Mean AS freq*	*Mean verb freq*
B1	.18	-.13	.29	-.17	-.19	-.15

B2	.27	.02	.29	.00	.05	-.14
C1	.24	.06	.32	.14	-.05	-.22

	B1 Model			B2 Model			C1 Model		
Main effects	*B*	*S.E.*	*β*	*B*	*S.E.*	*β*	*B*	*S.E.*	*β*
MI mean dobj	9.1	5.1	13.5	20.2**	3.4	22.5	15.9*	6.6	15.6
TTR dobj	-22.9	141.1	-1.3	129.6*	58.6	9.1	185.6	140.4	9.4
Mean MI VAS	24.9	19.2	11.8	98.9**	10.9	40.3	63.3**	18.4	25.4
TTR VAS	-248.6**	79.4	-29.7	-42.1	41.1	-4.9	142.0	79.5	13.9
Mean AS freq	$-1.9e^{-5}$*	$9.3e^{-6}$	-17.2	$1.6e^{-5}$**	$4.0e^{-6}$	15.2	$1.7e^{-6}$	$8.0e^{-6}$	1.6
Mean verb freq	$-5.1e^{-5}$**	$1.9e^{-5}$	-28.8	$-8.7e^{-6}$	$1.1e^{-5}$	-4.1	$-3.6e^{-6}$	$1.9e^{-5}$	-1.5
Topic1				-22.7	72.6	-31.9	-328.8	168.1	-12.7
Interactions									
Topic1:Mean MI dobj				-3.0	3.4	-3.3	4.4	6.6	4.3
Topic1:TTR dobj				-26.1	58.6	-1.8	-3.1	140.4	-0.2
Topic1:Mean MI VAS				-32.6**	10.9	-13.3	10.8	18.4	4.3
Topic1:TTR VAS				67.9	41.1	7.9	293.5**	79.5	28.8
Topic1:Mean AS freq				$4.5e^{-8}$	$4.0e^{-6}$	$4.0e^{-2}$	$3.1e^{-6}$	$8.0e^{-6}$	2,8
Topic1:Mean verb freq				$-6.0e^{-6}$	$1.1e^{-5}$	-2.8	$4.4e^{-5}$*	$1.9e^{-5}$	18.3

Table 7. Regression model output across CEFR levels

B = unstandardized coefficients; S.E. = standard error of unstandardized coefficients; β = standardized coefficients; Topic1 refers to the topic “newspaper under pressure” at the B2 level and “The reading brain” at the C1 level
* significant at $p < .05$
** significant at $p < .01$

4.1 *B1 model*

Three of the six predictors are significant in the linear regression model for the B1 texts: the TTR of VAS, the mean argument structure frequency, and the mean verb frequency. These three variables demonstrated moderate bivariate correlations with the exam score (-.17, -.19, and -.15, respectively), though the variable correlating strongest with the exam score, the mean MI of VAS, was not significant in the regression model. The standardized regression coefficients (represented by β in Table 7) indicate that the three significant predictors also have the largest effects in the prediction of the exam score. The overall R^2 for the model is .18 (adjusted R^2: .15). When comparing the sophistication and diversity measures for the dependency-based units and the VAS units, the significant and important predictors in the model are all measures relating to the VAS units.

4.2 *B2 model*

In the regression model for the B2 texts, there are four significant main effects and one significant interaction among the seven predictors (six measures plus topic) and six interaction terms entered into the model. The topic "newspaper under pressure" is coded 1 and the topic "admissions exams" is coded -1, and coefficients for interactions with topic therefore represent the deviation of each topic from the slope reported under main effects (for each complexity measure, the slope for "newspaper under pressure" is the main effect coefficient plus the interaction effect coefficient, while the slope for "admissions exams" is the main effect coefficient minus the interaction effect coefficient). The significant interaction between topic and mean MI of VAS indicates that the positive effect of an increased average MI score for texts written in response to the topic "newspaper under pressure" on the exam score is nearly doubled for the second prompt topic at the B2 level, “admissions exams”. The four significant main effects in the model are the mean MI of verb + direct object dependencies, TTR of verb + direct object dependencies, mean MI of VAS, and the mean argument-structure frequency. The significant predictors in the model largely correspond to the bivariate correlations reported in Table 6, with the exception of TTR of verb + direct object dependencies and the measure for mean argument-structure frequency, which did not exhibit strong bivariate correlations with the exam score for the B2 exam data. Looking at the standardized coefficients, the largest effects in the model are associated with the sophistication measures for both types of phraseological units. The overall R^2 for the model is .26 (adjusted R^2: .24). In this model, sophistication measures for both types of phraseological units (dependency-based and VAS units) are prominent, whereas diversity measures do not appear to contribute as substantially to the prediction of the exam score.

4.3 *C1 model*

Finally, two significant main effects and two significant interaction effects are identified in the C1 regression model. The topic “The reading brain” is coded 1, and the topic “medical research” is coded -1. The significant interaction between topic and TTR of VAS indicates a strong positive effect of an increase in the TTR of VAS on the exam score for texts written in response to the topic “The reading brain”, but a smaller negative effect of an increase in the TTR of VAS on the exam score for texts written in response to the second topic, “medical research”. The significant interaction between topic and mean verb frequency indicates a positive effect of an increase in the mean verb frequency on the exam score for texts written in response to the topic “The reading brain”, but a negative effect of an increase in the mean verb frequency on the exam score for texts written in response to the second topic, “medical research”. The two significant main effects are mean MI of verb + direct object dependencies and mean MI of VAS, both measures of phraseological sophistication. The significant predictors in the model reflect the bivariate correlations in Table 6, including the moderate negative correlation (-.22) between mean verb frequency and exam score, which is reflected in the significant interaction effect. The largest effects in the model, determined by the standardized coefficients, are for the two significant main effects and the mean verb frequency. The overall R^2 for the model is .24 (adjusted R^2: .19). In the C1 model, sophistication measures for both types of phraseological units emerge as significant and important predictors, as well as a frequency measure related to the verb lemmas occurring within VAS units in the learner corpus.

5. DISCUSSION

The present study aimed to include diversity and sophistication measures for VAS within the battery of measures that serve to investigate phraseological complexity and to investigate the relationship between these variables and an indicator of learner proficiency. Additionally, the relationship between these measures and complexity measures targeting verb + direct object dependencies was investigated to compare the behavior of complexity measures computed for different types of phraseological units.

5.1 *VAS complexity and proficiency*

One of the primary goals of L2 complexity research has been to establish links between quantitative complexity measures and learner proficiency (Bulté & Housen, 2012; Housen et al., 2019). The results of the present analysis show that the proposed diversity and sophistication measures for VAS contribute significantly to the prediction of L2 Dutch exam scores across three CEFR levels. The measure of VAS sophistication included in this study, mean MI for VAS, emerges as a significant and important predictor for the B2 and C1 level exams. The measure of VAS diversity on the other hand, TTR of VAS, is a significant and important predictor at the B1 level, and additionally at the C1 level (though the direction of the effect is reversed for the two topics). While the sophistication measure for these phraseological units demonstrates a positive linear relationship with exam score across the B2 and C1 levels and for all writing topics, the relationship between the diversity measure for VAS and the exam score is less consistent across the models. At the B1 level, an increase in the TTR for VAS is associated with a decrease in the exam score, and at the C1 level the direction of the effect depends on the topic of the writing prompt.

The pattern of increasing verb-argument sophistication with increasing exam scores observed in the present study reflects findings from usage-based approaches to the study of language acquisition that highlight the role of frequency and contingency in the acquisition of language (Ellis & O'Donnell, 2012; Wulff & Ellis, 2018). More specifically, L2 acquisition of VACs has been found to be heavily influenced by their frequency in a reference corpus and their structural complexity (i.e., highly frequent VACs with lower structural complexity are among the earliest acquired VACs) (Römer, 2019). The prototypicality of main verb lemmas, meaning the extent to which a verb is representative of a particular class of argument structures in terms of contingency as well as lexical semantics, is also considered to be one of the key factors driving L2 VAC acquisition (Ellis & Ferreira-Junior, 2009b). In this way, frequency and association appear to be two of the driving factors behind VAC acquisition, echoing common sentiments in usage-based theory regarding the processes underlying language acquisition (Ellis, 2002).

To understand the trends observed in this study more comprehensively, we take a closer look at the VAS extracted from the learner texts across the three proficiency levels and their corresponding frequency statistics in the Lassy Groot reference corpus. Table 8 presents the 10 VAS with the highest MI scores extracted from the learner data at each of the CEFR levels. All of the VAS in Table 8 are ditransitive structures (either with two nominal arguments or a nominal and a verbal argument), and have been assigned an MI score above 4.68. When compared with the 10 most frequently occurring VAS at each of the CEFR levels (see supplemental materials on OSF), relationships between the structural complexity, frequency, and association strength of the argument structure emerge. The structural complexity of the argument structure is directly related to its frequency of occurrence, in that more complex structures occur much less frequently (see Table 5 where ditransitive structures account for less than 5% of all argument structures). The results of this study show that ditransitive VAS units are both relatively rare and also fall in the higher MI range, as compared to the more frequent and less structurally complex transitive NP structure. Previous VAC research (e.g., Römer, 2019) has found that learners tend to acquire more structurally complex argument structures later, which corresponds to the positive effect of verb-argument sophistication on L2 Dutch exam scores reported in the present analysis.

In terms of verb frequency, the verbs occurring in the VAS with the highest MI scores tend to be much less frequent overall in the reference corpus, compared with the verbs extracted from the VAS with lower MI scores. When a verb with a higher frequency is encountered in a VAS with a high MI value, as is the case with the verb ‘geven’ in the ditransitive NP NP structure, the argument structure typically accounts for a large proportion of occurrences of the verb (41% for ditransitive_NP_NP + geef). These verbs that exhibit an overall high frequency of occurrence in combination with a high degree of association, such as ‘geven’ in the ditransitive NP NP argument structure, adhere to the principle of ‘prototypicality’ in that they tend to be representative of the class of argument structures that they most often occur in (Fenk-Oczlon, 2001). Argument structures have been found to be acquired and used first and foremost with such prototypical verb lemmas (Ellis et al., 2014; Goldberg et al., 2004). Based on the L2 VAC acquisition literature, this is the type of VAS unit we would expect to see as the learners start to acquire more structurally complex units. Interestingly, this particular example of a prototypical ditransitive NP NP occurs among the VAS with the highest MI scores at the B1 level, but is below the range for the highest MI scores at the B2 and C1 levels (see Table 8).

While the comparison of the VAS that occur most frequently (see OSF), and that have the highest MI scores, across the exam levels reveals similarities in the types of argument structures produced by learners, the VAS presented in Table 9 illustrate the variation within one level, namely C1. The left side of the table presents all VAS units extracted from a (randomly selected) high-scoring C1 level text and their corresponding MI scores, and the right side presents all VAS units extracted from a low-scoring C1 level text. The VAS extracted from the low-scoring C1 text exhibit an overall lower range of MI scores, and do not include any ditransitive argument structures. On the other hand, the high-scoring C1 text includes two ditransitive argument structures, both of which have an MI score more than twice that of the VAS with the highest MI score in the low-scoring text. Within the class of transitive NP argument structures, the most frequent argument structure class at all exam levels, the mean MI score for the high-scoring text is also 36% higher than that of the low-scoring text. This qualitative comparison demonstrates that, as learner performance increases (determined by the numeric exam score at each exam level), learner texts will exhibit: (i) more structurally complex

VAS, which exhibit lower frequency in the reference corpus compared to their less complex counterparts in addition to higher MI scores and (ii) more sophisticated, i.e., less frequent and more strongly associated, verb choices within each class of argument structure. This is directly reflected in the positive correlation between VAS sophistication and exam score identified at the C1 level in the quantitative analysis.

Both main verb frequency and argument structure frequency have also been explored independently as influential factors in L2 acquisition. Kyle, Crossley, and Verspoor (2021) found the average main verb lemma to be a stronger indicator of L2 development than both the frequency and association strength of the VAC unit. While their analysis did not find any significant development in the construction lemma frequency in their learners, Römer (2019) reported an overall increase in VAC types as proficiency level progressed from A1 to C1. To account for the potential impact of these frequency variables on L2 Dutch proficiency, we included measures of main verb frequency and argument structure frequency in the regression analyses to account for the potential impact of these frequency variables on L2 Dutch proficiency. These frequency measures were found to be significant predictors for the B1 and B2 models (both measures at B1 and only mean argument structure frequency at B2). Mean verb lemma frequency and mean argument structure frequency were the only sophistication measures that emerged as significant predictors at the B1 level, indicating that learners at this exam level might make more progress in expanding their repertoire of verb and argument structure lemmas than identifying co-occurrence patterns between the two. This could also be indicative of an intermediary stage of VAS development where learners are exploring the use of new verb lemmas in the argument structures with which they have already become more acquainted (Ellis & Ferreira-Junior, 2009). While mean argument structure frequency is a significant predictor in the B2 model, mean MI of VAS is a significant predictor with a larger effect size. At the C1 level, mean MI of VAS is the only significant sophistication measure in the model (leaving aside verb + direct object sophistication for the moment). These results, taken together, point towards a shift that occurs in the development of VAS around the B2 level, where learners have successfully expanded their argument structure and verb lemma repertoires to include the most structurally complex and sophisticated exemplars and are now developing a more acute familiarity with the co-occurrence patterns exhibited by VAS units in written Dutch.

Table 8. VAS with highest MI values across CEFR levels

Rank	VAS	MI	Freq-ref	V-freq	AS-freq
	B1				
1	ditransitive_NP_VC + dwing ‘force’	6.68	8,358	17,882	188,593
2	ditransitive_NP_VC + verplicht ‘require‘	6.57	10,853	25,021	188,593
3	ditransitive_NP_VC + raad_aan ‘recommend’	6.44	3,006	7,589	188,593
4	ditransitive_NP_VC + moedig_aan ‘encourage’	6.07	1,735	5,684	188,593
5	ditransitive_NP_VC + roep_op ‘call on’	5.72	6,339	26,463	188,593
6	ditransitive_NP_VC + vraag ‘ask’	5.72	35,244	147,402	188,593
7	ditransitive_NP_VC + nodig_uit ‘invite’	5.65	4,518	19,812	188,593

8	ditransitive_NP_NP + geef ‘give’	5.14	101,445	247,381	481,048
9	ditransitive_NP_VC + vertel ‘tell’	4.80	13,469	105,989	188,593
10	ditransitive_NP_NP + stuur_op ‘send’	4.68	852	2,872	481,048
	B2				
1	ditransitive_NP_VC + verzoek ‘request’	6.84	4,615	8,801	188,593
2	ditransitive_NP_VC + dwing ‘force’	6.68	8,358	17,882	188,593
3	ditransitive_NP_VC + verplicht ‘require’	6.57	10,853	25,021	188,593
4	ditransitive_NP_VC + raad_aan ‘recommend’	6.44	3,006	7,589	188,593
5	ditransitive_NP_VC + moedig_aan ‘encourage’	6.07	1,735	5,684	188,593
6	ditransitive_NP_NP + bezorg ‘deliver’	5.91	8,457	12,076	481,048
7	ditransitive_NP_VC + vraag ‘ask’	5.72	35,244	147,402	188,593
8	ditransitive_NP_VC + nodig_uit ‘invite’	5.65	4,518	19,812	188,593
9	ditransitive_NP_NP + wijs_toe ‘allocate’	5.61	3,882	6,842	481,048
10	ditransitive_NP_NP + verschaf ‘provide’	5.45	2,379	4,687	481,048

	C1				
1	ditransitive_NP_VC + spoor_aan ‘urge’	7.03	1,697	2,859	188,593
2	ditransitive_NP_VC + dwing ‘force’	6.68	8,358	17,882	188,593
3	ditransitive_NP_VC + verplicht ‘require’	6.57	10,853	25,021	188,593
4	ditransitive_NP_NP + wens_toe ‘wish’	6.28	1,228	1,361	481,048
5	ditransitive_NP_VC + moedig_aan ‘encourage’	6.07	1,735	5,684	188,593
6	ditransitive_NP_VC + adviseer ‘advise’	5.90	1,760	6,498	188,593
7	ditransitive_NP_NP + veroorloof ‘afford’	5.87	2,662	3,915	481,048
8	ditransitive_NP_NP + schrijf_toe ‘attribute’	5.74	3,308	5,319	481,048
9	ditransitive_NP_VC + vraag ‘ask’	5.72	35,244	147,402	188,593
10	ditransitive_NP_VC + nodig_uit ‘invite’	5.65	4,518	19,812	188,593

Table 9. VAS extracted from high- and low-scoring C1 texts

High score		Low Score	
VAS	MI	VAS	MI

ditransitive_NP_NP + schrijf_toe	5.74	transitive_VC + stel_voor ‘imagine’	2.23
‘attribute’	5.15	transitive_VC + onderzoek	2.07
ditransitive_NP_NP + geef_door	2.38	‘research’	1.70
‘pass on’	2.20	transitive_NP + beleef ‘experience’	1.65
intransitive + evolueer ‘evolve’	2.18	transitive_VC + vind (x2) ‘find’	1.56
intransitive + duur ‘last’	1.72	intransitive + reageer ‘react’	1.48
intransitive + hoest ‘cough’	1.56	transitive_NP + neem ‘take’	1.37
transitive_NP + bevat ‘contain’	1.44	transitive_NP + benader (x2)	1.23
transitive_NP + krijg (x2) ‘get’	1.26	‘approach’	1.16
transitive_NP + organiseer (x2)	1.24	transitive_NP + pas_aan ‘adapt’	1.15
‘organize’	1.24	transitive_NP + neem_waar	1.14
transitive_NP + publiceer (x5)	1.11	‘observe’	1.09
‘publish’	1.08	transitive_NP + behandel ‘treat’	1.08
transitive_NP + implementeer	0.80	transitive_NP + analyseer ‘analyze’	1.07
‘implement’	0.60	transitive_NP + vind ‘find’	0.86
transitive_NP + haal_aan ‘quote’	0.36	transitive_NP + lees (x3) ‘read’	0.66
transitive_NP + ontwikkel	0.32	transitive_NP + stimuleer	0.41
‘develop’	-0.54	‘stimulate’	0.36
transitive_NP + lees ‘read’		intransitive + verander ‘change’	-0.04
intransitive + ga ‘go’		transitive_NP + vestig (x2)	-0.90
transitive_NP + geef ‘give’		‘establish’	-1.25
transitive_NP + heb (x4) ‘have’		transitive_NP + sluit_aan ‘join’	
intransitive + blijk ‘appear’		transitive_NP + heb (x2) ‘have’	

intransitive + ben ‘be’	ditransitive_NP_NP + trek ‘pull’
	transitive_NP + leg_uit ‘explain’
	transitive_NP + wil ‘want’

5.2 *VAS and phraseological complexity*

Various findings have contributed towards establishing the validity of measures of phraseological complexity as indicators of L2 proficiency and writing quality within the framework of L2 complexity research. The results reported in the present study are in line with those reported by Kyle and Crossley (2017) for their measures of VAC frequency and association strength, which served as strong predictors of L2 English writing quality, demonstrating a certain degree of cross-linguistic validity for measures of VAS complexity. The results of the present study additionally demonstrate the complex interplay between various indices of phraseological complexity across a range of proficiency levels and text types. Measures of verb + direct object diversity and sophistication were included in the models at each level to determine how or whether the complexity measures of these two types of phraseological units complement one another in the prediction of learner proficiency. Diversity of verb + direct object dependency units only contributed significantly to the B2 regression models, whereas diversity of VAS only contributed significantly to the B1 regression model. However, the mean MI for verb + direct object dependencies was a significant predictor at both the B2 and C1 levels, alongside the mean MI of VAS. This consistency in the results for the phraseological sophistication measures suggests that the sophistication of the two different types of phraseological units investigated in this study contributes novel information for the prediction of learner proficiency. The stability of the effect of the complexity measures across text topic also factors into the analysis at the B2 and C1 levels, where verb + direct object sophistication does not appear to interact with text topic. Though the interaction between topic and sophistication of VAS is significant at the B2 level, the direction of the effect is consistent across both topics. The direction of the effect of diversity of VAS, however, does differ between the two topics at the C1 level, indicating that this measure may be less consistent across proficiency level and task types.

For the construct of phraseological complexity, the measure of VAS sophistication, in particular, demonstrates strong potential as a quantitative measure to include alongside more established measures of dependency-based phraseological units. Not only does this measure exhibit a clear relationship with learner proficiency at the more advanced stages, but it also appears to supplement the predictive ability of existing phraseological complexity measures. From a more theoretical perspective, the results reported in the present analysis for the measures of phraseological sophistication also provide support for the predictive validity of phraseological complexity, by establishing that similar measures targeting distinct phraseological units exhibit similar behavior in the prediction of L2 proficiency.

6. CONCLUSION

The battery of phraseological complexity measures is continuously expanding with the addition of measures operationalizing various phraseological units that play a demonstrable role in L2 acquisition. In the present study, L2 Dutch exam scores across three CEFR levels were modeled as a function of both verb + direct object dependency and VAS complexity measures. This approach has explored whether phraseological units situated at the lexis-grammar interface demonstrate similar relationships with learner proficiency and performance, and whether these measures are complementary to measures of phraseological complexity that target more lexically-defined phraseological units. The findings highlight measures of VAS complexity, and in particular VAS sophistication, as significant predictors of learner proficiency, exhibiting variable relationships with the L2 Dutch exam scores across the three CEFR exam levels.

The relationships observed between the measures investigated here and learner proficiency provide support for calls to widen the scope of L2 complexity research through the operationalization of a wider range of linguistic features across more domains of the language system. While the sophistication measure for VAS emerges as a significant and important predictor for the B2 and C1 models, with a relatively stable effect across the multiple text topics, measures of the frequency of main verb lemmas and argument structures appear to pattern more clearly with exam scores at the B1 level. Moreover, the measure of verb + direct object dependency sophistication also contributes significantly to the B2 and C1 models. These trends signal a potential development of the role that phraseological complexity plays in the assessment of L2 Dutch performance as proficiency increases. Together, these trends also emphasize the need for complexity measures capturing a range of phraseological phenomena, including those situated at the interface of lexis and grammar, in order to trace L2 Dutch (writing) development across the full spectrum of learner proficiency.

The analyses of phraseological complexity carried out here are limited by the nature of the dataset, which required separate statistical analyses for each exam level. The extraction of the phraseological units measured in this study is also heavily dependent on automatic linguistic annotation, which has been shown to introduce a certain degree of bias into the analysis of L2 data (Rubin, 2021), in addition to a moderate degree of erroneous annotations overall. This is evident in the suboptimal automatic extraction accuracy rates for the target structures in the present analysis. Despite the limitations, this study has contributed to establishing the cross-linguistic validity of phraseological complexity measures by focusing on L2 Dutch, for which the body of L2 complexity research is still quite limited. Nevertheless, the measures introduced in the present study should be implemented across a wider range of data types and L2s in order to further support the validity of these measures.

References

Balk-Smit Duyzentkunst, F. (2000). *Grammatica van het Nederlands*. The Hague: Sdu.

Banerjee, S., & Pedersen, T. (2003). The design, implementation, and use of the Ngram Statistics Package. In Proceedings of the fourth international conference on intelligent text processing and computational linguistics.

Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes, 46*, 100869.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), Dimensions of L2 performance and proficiency (pp. 21−46). Philadelphia, Amsterdam: John Benjamins.

Crossley, S. (2020). Linguistic features in writing quality and development: an overview. Journal of Writing Research, 11(3), 415-443.

Duinhoven, M. A. (1989). Het voorzetselvoorwerp. een zinspatroon in wording. De Nieuwe Taalgids, 82, 40−55.

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.

Ellis, N. C. (2002). Frequency effects in language processing: a review with implications for theories of implicit and explicit language acquisition. Studies in Second Language Acquisition, 24, 143−188.

Ellis, N. C., & Ferreira-Junior (2009). Constructions and their acquisition: islands and the distinctiveness of their occupancy. Annual Review of Cognitive Linguistics, 7, 405-431.

Ellis, N. C., & O'Donnell, M. B. (2012). Statistical construction learning: Does a Zipfian problem space ensure robust language learning. In P. Rebuschat & J. N. Williams (Eds.), Statistical learning and language acquisition (pp. 265−304). Boston, Berlin: De Gruyter.

Ellis, N. C., O'Donnell, M. B., Römer, U. (2014). Second Language Verb-Argument Constructions are Sensitive to Form, Function, Frequency, Contingency, and Prototypicality.

Fenk-Oczlon, G. (2001). Familiarity, information flow, and linguistic form. In J. Bybee & P. Hopper (Eds.), Frequency and the emergence of linguistic structure (p. 431−448). Philadelphia, Amsterdam: John Benjamins.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. Cognitive Linguistics, 15, 289−316.

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.

Gries, S. T. (2014). Quantitative corpus approaches to linguistic analysis: seven or eight levels of resolution and the lessons they teach us. In I. Taavitsainen, M. Kytö, C. Claridge, & J. Smith (Eds.), Developments in English: expanding electronic evidence (pp. 29−47). Cambridge, UK: Cambridge University Press.

Gries, S. T. (2022). What do (some of) our association measures measure (most)? Association? *Journal of Second Language Studies*, *5*(1), 1-33.

Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (2019). Multiple approaches to complexity in second language research. Second Language Research, 35(1), 3−21.

Klooster, W. (2001). Grammatica van het Hedendaags Nederlands. The Hague: Sdu.

Kyle, K., & Crossley, S. A. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. Language Testing, 34(4), 513−535.

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. The Modern Language Journal, 102(2), 333−349.

Kyle, K., Crossley, S. A., & Verspoor (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition*, *43*, 781-812.

Kyle, K., & Eguchi, M. (2021). Automatically assessing lexical sophistication using words, n-gram, dependency, and verb argument construction indices. In S. Granger (Ed.), Perspectives on the L2 phrasicon: the view from learner corpora (pp. 126−151). Bristol: Multilingual Matters.

McCarthy, P. M. & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods, 42(2), 381−392.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. Applied Linguistics, 30(4), 555−578.

Pallotti, G. (2015). A simple view of linguistic complexity. Second Language Research, *31*(1), 117−134.

Pallotti, G. (2023). Appropriate complexity. In C. Granget, I. Repiso, & G. Fon Sing (Eds.), *Language, creoles, varieties: From emergence to transmission*. Language Science Press.

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.

Römer, U. (2019). A corpus perspective on the development of verb constructions in second language learners. International Journal of Corpus Linguistics, 24, 268−290.

Rubin, R. (2021). Assessing the impact of automatic dependency annotation on the measurement of phraseological complexity in L2 Dutch. International Journal of Learner Corpus Research, 7(1), 131−162.

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: the view from learner corpora (pp. 101−125). Bristol: Multilingual Matters.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.

van Noord, G. (2006). At last parsing is now operational. In TALN 2006 (pp. 20−42).

van Noord, G., Schuurman, I., & Vandeghinste, V. (2006). Syntactic annotation of large corpora in STEVIN. In *Proceedings of LREC 2006*. ELRA.

Wulff, S., & Ellis, N. C. (2018). Usage-based approaches to second language acquisition. In D. Miller, F. Bayram, J. Rothman, & L. Serratrice (Eds.), *Bilingual cognition and language:*

The state of the science across its subfields (pp. 37−56). Philadelphia, Amsterdam: John Benjamins.