Abstract

This study investigates whether re-thinking the separation of lexis and grammar in language testing could lead to more valid inferences about proficiency across modes. As argued by Römer (2017), typical scoring rubrics ignore important information about proficiency encoded at the lexis-grammar interface, in particular how the co-selection of lexical and grammatical features is mediated by communicative function. This is especially evident when assessing oral versus written exam tasks, where the modality of a task may intersect with register-induced variation in linguistic output (e.g. Biber et al., 2016). This article presents the results of an empirical study in which we measured the diversity and sophistication of four-word lexical bundles extracted from a corpus of French proficiency exams. Analysis revealed that the diversity of noun-based bundles was a significant predictor of written proficiency scores and the sophistication of verb-based bundles was a significant predictor of proficiency scores across both modes, suggesting that communicative function (Biber et al., 2004) as well as the constraints of online planning (Skehan, 1998) mediated the effect of lexicogrammatical phenomena on proficiency scores. Importantly, lexicogrammatical measures were better predictors of proficiency than solely lexical-based measures, which speaks to the potential utility of considering lexicogrammatical competence on scoring rubrics.

Keywords: phraseology, lexicogrammar, complexity, modality, register, French

Proficiency at the lexis-grammar interface: Comparing oral versus written French exam tasks

1 Introduction

Recently, there have been calls to re-think traditional divisions between "vocabulary" and "grammar" that have been typical in models of language ability used in language testing (e.g. Alderson & Kremmel, 2013; Kremmel et al., 2017; Römer, 2017). As argued by Römer (2017), language proficiency rating scales often fail to acknowledge the fact that lexis and grammar interact in meaningful ways. This is evident in corpus linguistic data showing that certain words and grammatical structures demonstrate patterns of co-occurrence and recurrence (Biber, 2009; Römer, 2009; Sinclair, 1991; Stefanowitsch & Gries, 2003), that is they appear together more often than would be "expected on the basis of chance" (Gries, 2008, p. 6) or they are very often repeated together (Biber et al., 1999; Paquot & Granger, 2012). Importantly, these studies also show that associations between lexical items and grammatical structures often serve specific communicative functions (see Biber et al., 2004), which means that associations between lexis and grammar also need to be considered within the specific communicative context in which they occur. The problem with strictly separating lexis and grammar on scoring rubrics is therefore two-fold.

First, it ignores important information about linguistic competence encoded at the lexis-grammar interface, information which has shown to contribute to holistic evaluations of learner productions. For example, Paquot (2018) found that the association between words in specific grammatical structures (e.g. arouse + curiosity in a direct object structure) was a significant predictor of the proficiency scores assigned to advanced learner texts in a corpus of research

papers. Importantly, no significant differences were found regarding the lexical or syntactic complexity of the texts, suggesting that the most important information about proficiency was not lexical or grammatical but rather lexicogrammatical. Similarly, Kyle and Crossley (2017) found that the association strength of verbs and the constructions in which they appear (e.g. subject-**require**-direct object-direct object, 'sports require teamwork and the ability to cooperate') was likewise a predictor of holistic proficiency scores and that models based on such lexicogrammatical phenomena were better predictors of proficiency than a model based solely on syntactic complexity. These results show that rating scales that do not account for lexicogrammatical phenomena may not be fully representative of the larger construct of proficiency, which constitutes a threat to the *explanation inference* in an argument-based approach to test validation (Chapelle et al., 2008; Kane, 2006).

The second problem with separating lexis from grammar on scoring rubrics is that it ignores the way that both lexis and grammar interact together with communicative function. That language varies according to context of use is certainly not news to language assessment specialists (see Bachman, 1990). What is less clear, however, is how functional differences between tasks relate specifically to the lexicogrammatical features exhibited in exam performances and how those lexicogrammatical features, in turn, relate to the proficiency scores attributed to such exams. For example, in a study comparing oral and written TOEFL exam tasks, Biber and Gray (2013) observed that collocational patterns differed not only with respect to task type, but that such patterns were reflected in the scores attributed to the exams. As Paquot et al. (2021, p. 229) point out, "there is a need for more research into how lexicogrammatical competence interacts with foreign language proficiency and development across modes, tasks, registers and other linguistic, contextual and or situational characteristics." In particular, relatively little research has been done into the linguistic differences between oral and written

production tasks on language exams (Biber et al., 2016). The current study therefore represents an attempt to complement the discussion about the division between lexis and grammar in language testing by investigating the extent to which lexicogrammatical characteristics are predictive of proficiency scores in oral versus written exam tasks.

2 Background

One way to capture complexity at the lexis-grammar interface is through phraseology. Phraseology looks beyond single words to examine the way that words combine together. Psycholinguistic research has shown that both learners and native speakers are sensitive to patterns of co-occurrence and recurrence but in different ways. Ellis, Simpson-Vlach and Maynard (2008), for example, found that for learners of English, the processing of three-, four- and five-word phrases is primarily a function of the frequency of the phrases in large reference corpora. For native speakers however, it is not frequency but rather association strength of the phrases that has the strongest effect on processing. Phrases containing words that occur more exclusively together (and rarely with other words) were the fastest to process. These receptive differences are also found in production data. Durrant and Schmitt (2009) compared written texts produced by native and non-native writers of English and found that compared to the native writers, the non-native writers used more highly frequent collocations and fewer infrequent but strongly associated collocations. Similarly, as learners increase in proficiency, they tend to decrease their use of highly frequent bigrams (adjacent word pairs) and increase their use of infrequent but highly associated bigrams, both when learners are compared cross-sectionally across proficiency levels (Granger & Bestgen, 2014) as well as when individual learners are followed longitudinally (Bestgen & Granger, 2018).

Paquot (2019) termed this aspect of linguistic competence *phraseological complexity*, which she defined as the diversity and sophistication of phraseological units in learner production (following Ortega, 2003). Phraseological diversity refers to the extent to which a learner uses a large number of different phraseological units. Phraseological sophistication on the other hand, refers to the extent to which a learner is able to select more specific or informative phraseological units (i.e., those which may be more appropriate to a specific topic or register). In a series of studies focusing on L2 English, French and Dutch, measures of phraseological complexity have been found to be predictive of L2 proficiency, even when lexical and syntactic complexity measures were not (Paquot, 2018, 2019; Rubin et al., 2021; Vandeweerd et al., 2021). While these studies speak to the relevance of phraseological complexity to L2 proficiency in general, they are limited by the fact that they have focused exclusively on written production.

Writing and speaking differ in a number of important ways including the presence (or absence) of an audience, the degree of online planning required and control over output (Ravid & Tolchinsky, 2002). The fact that speaking occurs online and in real-time also means that learners have less time to conceptualize, formulate and monitor linguistic output as compared to writing (Skehan, 1998). Even the most pressured writing conditions (e.g. timed language exams) still offer more opportunities for planning and revising compared to speech. As a result, L2 speech generally exhibits less lexical complexity as compared to writing (e.g. Granfeldt, 2007). The constraints of real-time production and the lack of online planning may also lead to the use of more frequent (and therefore more accessible) phraseological units, as suggested by the results of Biber and Gray's (2013) study of the collocations used in oral and written TOEFL exam tasks.

In addition to the overall differences outlined above, speech and writing often serve different communicative purposes. As a result, oral and written production tend to be associated

with the use of different linguistic features (Biber, 2019). For instance, the fact that speech is often interactive means that it tends to have a more personal or involved focus. Linguistically, this is associated with a greater use of pronouns, modal verbs, lexical verbs, adverbs and discourse markers. In contrast, writing tends to be more informational in purpose and is marked by the use of nouns, nominalizations, attributive adjectives and prepositional phrases. These differences have been observed across several languages and different text types (Biber, 2014) and even extend to the level of phraseology, with speech being associated with the use of verb-based phraseological units and writing being associated with noun-based phraseological units (Biber et al., 2004).

From a developmental perspective, linguistic features associated with speech are also hypothesized to be acquired early because they are necessary for almost all types of conversation. Linguistic features common to writing (e.g. complex phrasal embedding) on the other hand, are hypothesized to be acquired later, and only in the context of formal education (Biber et al., 2011; Halliday & Matthiessen, 2006). This suggests that although tasks may require different linguistic features, there may be developmental patterns within tasks whereby more advanced language users incorporate more nominal complexity features, regardless of task type. For example, in their study comparing oral and written TOEFL exam tasks, Biber, Gray and Staples (2016) found that although written tasks were indeed associated with the use of phrase-level linguistic features and oral exams were associated with clause-level linguistic features, within each task, there was a positive relationship between the use of phrasal features and holistic proficiency scores given by raters.

With regards to French, the target language of the learners in our study, phraseological complexity has also been found to increase with proficiency both in oral production (Forsberg,

2010) as well as written production (Vandeweerd et al., 2021). Only one study thus far has directly compared phraseological complexity across modes however. Vandeweerd et al. (2022) compared the diversity and sophistication of direct objects (verb + noun) and adjectival modifiers (noun + adjective) across oral and written tasks collected from university-level learners of French over a period of 21 months. While the design of the corpus made it difficult to isolate the precise effects of task characteristics (e.g. modality, interactivity) from register, comparison of the learner group and an L1 group seemed to indicate that the extra planning time afforded by the written task had a positive effect on the phraseological complexity in learner productions but that this effect was somewhat mediated by the functional requirements of the task, as seen by the fact that the complexity of the noun-based units (adjectival modifiers) and verb-based units (direct objects) differed across tasks. However, the applicability of these findings to the language testing context is somewhat limited given that the study involved a relatively small and quite advanced sample learners (n = 29) and the fact that the learners were not directly impacted by the results of their linguistic performance.

The current study therefore aims to determine the extent to which measures of phraseological complexity are predictive of oral versus written French proficiency scores in the context of a high-stakes language assessment exam. In line with previous research, we expect that measures of phraseological complexity will be predictive of proficiency even when controlling for the lexical complexity of a text (Vandeweerd et al., 2021). This would suggest that raters are sensitive to complexity not just at the at the level of lexis but also at the lexis-grammar interface (i.e. that phraseological measures are not simply a subset of lexical complexity measures) as argued by Paquot et al. (2021). We also aim to determine the extent to which various measures of phraseological complexity (e.g. those targeting diversity versus sophistication and noun-based versus verb-based units) differ in their ability to predict proficiency scores in the two modes. We

expect to find that verb-based measures will be more predictive of proficiency in the oral task and that noun-based measures will be more predictive of proficiency in the written task due to the register-induced differences between the two tasks (in line with Biber et al., 2016) but that the constraints of online production will have a mediating effect on the complexity of speech (Skehan, 1998).

3 Methodology

3.1 Data

This study uses a sample of 545 oral and written productions from the *Test d'évaluation de français* (TEF, Chambre de commerce et d'industrie de Paris, 2010), a French proficiency exam recognized by the French Ministry of Education for the admittance of non-native speakers to higher level education in France. Because the exam is also recognized by the federal government of Canada for immigration purposes, many of the candidates who take the exam are themselves native speakers of French. The distinction between native and non-native speakers, however, often has much more to do with identity and national politics than linguistic performance (Davies, 2004; Holliday, 2017). We therefore decided to include both candidates who self-identified as native-speakers of French and those who identified their first language as a language other than French and used L1 background as a control variable in the statistical analysis, in light of previous research showing an effect of L1 background on phraseological competence (e.g., Durrant & Schmitt, 2009; Ellis et al., 2008).

As discussed previously, the communicative function of a task tends to be associated with the use of specific linguistic features. For this reason, we selected two tasks with a similar function:

- Written Production Task B asks candidates to write a 'letter to an editor' (min. 200 words) in which they respond giving their opinion on a controversial topic. The prompts for this task are randomly distributed to candidates prior to starting the exam and consist of sentences extracted from newspaper articles (e.g. 'Television advertisements should be abolished'). Candidates are asked to develop at least three arguments to defend their point of view. They have one hour to complete this task as well as another written production task. Both tasks are completed via computer without access to reference materials.

- Oral Production Task B asks candidates to make a telephone call to a close friend. They are provided with a visual prompt (e.g. an advertisement promoting a vacation to Quebec), which they must present to the examiner and convince them to participate. Candidates have ten minutes to complete the task.

While letters and telephone calls differ in more than just modality (e.g., stylistics, structural conventions, formality), they both require candidates to express their personal attitudes and their certainty about propositions in order to persuade the addressee (i.e. the newspaper editor in the case of the written task and the 'friend' in the case of the oral task). Functionally, this tends to promote the use of linguistic elements such as first-person pronouns and dependent clauses (e.g., *À mon avis, c'est vrai que…*, 'In my opinion, it's true that…'), especially relative to more narrative- or descriptive-type tasks (e.g. Lu, 2011; Michel et al., 2019). This functional similarity makes these two tasks more comparable than the other TEF tasks, which are more narrative or descriptive in nature (e.g., newspaper article, telephone inquiry to request information).

The evaluation of the exams was done externally by raters at the *Chambre de commerce et d'industrie de Paris*. Pairs of raters judged each production on a scale of 0-6 for each of the following criteria: presentation of ideas, quality of argumentation, grammatical precision,

vocabulary control, elocution (in the case of the oral task) and orthography (in the case of the written task). Although inter-rater reliability scores are not available for these texts specifically, reliability has previously been reported to be moderate (r = 0.652) for the written production section of the TEF exam (Casanova & Demeuse, 2016) and high (r = 0.94) for the oral production section (Artus & Demeuse, 2008). A proprietary formula was used to transform the individual scores into a normed score out of 699 for the entire section of the exam. According to Artus and Demeuse (2008), this was done so that raters remain unaware of how the global score for each competence is calculated (to prevent them from simply assigning scores based on an *a priori* estimation of a candidate's global proficiency). The normed scores out of 699 are indexed to the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001), with 100 point increments corresponding to the six CEFR levels.[1]

[INSERT TABLE 1]

The sample of exams used in the study (Table 1) was made according to several criteria. First, we made an initial selection of exams where the written section was at least 100 words in length given that complexity measures have been shown to be unreliable for texts shorter than that. Second, given that topic can have a significant effect on phraseological complexity (Paquot, Naets, et al., 2021) we wanted to ensure that differences observed between oral and written sections of the exam were not simply due to topic differences. We therefore manually categorized the prompts into overarching themes (e.g. environment, travel, language, etc.) and then made an initial selection of exams with a match between the theme of the oral and written sections. The audio from the oral exams was then transcribed and the written exams were manually spell-corrected in order to ensure comparability with the oral transcriptions (which by definition did not have any orthographic errors). Letter headings (e.g. addresses, signature lines) were also

removed from the written productions. R scripts were then used to clean the texts (e.g. remove stylized apostrophes, standardize spelling) and part-of-speech (POS)-tag them using TreeTagger (Schmid, 1994). TreeTagger has previously been shown to have a high level of accuracy (96.12%) on L1 data (Denis & Sagot, 2012) and has also been shown to perform well (89.67%-93.52% accuracy) across various different task types (e.g., oral conversations, online reviews, literature and law texts, Mattiuzzi, 2021). We did not carry out an evaluation of the reliability on the POS annotation directly given that POS-taggers have been shown to perform generally well on learner data, provided that texts are manually spell-corrected before being tagged (De Haan, 2000; Rooy & Schäfer, 2002), which was the case here. Jarvis and Hashimoto (2021), for example, report F-scores above 0.90 for all part of speech categories in a corpus L2 English narrative texts, indicating a high level of reliability when compared to human annotation.

Following pre-processing, we used topic-modeling as an additional method of controlling for topic differences. Topic-modelling is a bottom-up method whereby texts are grouped together into computer-defined topics according to their common vocabulary (see Murakami et al., 2017). For each text in the corpus, the model outputs a probability distribution for each topic. This allows the differences between texts to be quantified in terms of their overlapping probability distributions using similarity metrics such as Jensen-Shannon Divergence (JSD) (Wang & Dong, 2020). In this study, we used the JSD scores between the oral and written sections of the exams to control for topic differences in the statistical analysis. The smaller the JSD between two sections of the exam, the more similar they are in terms of shared vocabulary. For example, in one exam, where the written prompt was *Il faut modifier nos comportements alimentaires pour préserver la planète* ('It is necessary to modify our dietary behaviour to preserve the planet') and the oral prompt related to an advertisement for installing a backyard windmill, both tasks were classified

as belonging to the same theme by the model and the JSD value for the two texts was found to be 0.004, indicating a high level of overlap in vocabulary.

3.2 Complexity measures

This study focuses on the complexity of four-word lexical bundles. This diverges somewhat from previous studies of phraseological complexity that have primarily focused on two-word co-occurrence phenomena (e.g. adjective + noun, verb + noun, adverb + verb collocations). While such phenomena have been shown to be relevant to L2 development (see Altenberg & Granger, 2001; Durrant & Schmitt, 2009; Laufer & Waldman, 2011; Nesselhauf, 2005), they represent only a small part of a learner's possible phraseological repertoire. Equally important to L2 phraseology are recurrence phenomena, such as lexical bundles, defined as "simple sequences of words that commonly go together in natural discourse" (Biber et al., 1999, p. 990ff). Like co-occurrence phenomena, lexical bundles also exist at the interface between lexis and grammar (Römer, 2009) and have been shown to be relevant to the development of L2 proficiency (e.g., Biber & Gray, 2013; Zheng, 2016)

Unlike collocations however, which by definition contain only two linguistic items, lexical bundles can be of any length. This is especially important when it comes to French because as noted by Vinay and Darbelnet (1995, p. 126), French has a tendency to modify nouns and verbs with prepositional phrases rather than adjectives or adverbs. For example, the French equivalent of *comment ironically* would be *s'exprimer en termes ironiques* (lit. 'express oneself in ironic terms'). Looking only at two-word collocations would ignore this type of verbal modification (as argued in Vandeweerd et al., 2021). In addition, the number and type of lexical bundles have both been shown to be different across registers and text types (Biber et al., 1999, 2004), a characteristic which makes them particularly suited to comparing oral and written tasks.

As to the length of the lexical bundles, while previous research has investigated bundles ranging in length from 3 to 7 words in French (see e.g., Granger, 2014), there is an upper limit constraining the length of units, given the computational power required to calculate association measures. Initial experimentation revealed that four word-bundles represented a fair balance between expanding construct representation (e.g., to capture important aspects of French as described above) on the one hand and computational feasibility on the other.[2]

We did however make some modifications to the way that lexical bundles are traditionally extracted. Whereas lexical bundles are typically composed of inflected surface forms, in our approach we used POS-tagged and lemmatized lexical bundles. Using POS-tagged lexical bundles allowed us to filter the units we extracted in order to account for possible cross-modal register differences (Biber, 2019). We focused on two types of lexical bundles in particular: those starting with verbs (verb-bundles) to tap into phraseological complexity at the clause-level and those starting with nouns (noun-bundles) to tap into phraseological complexity at the phrase-level. Admittedly, this is a rather crude operationalization but given that this is the first study to investigate phraseological complexity in this way, we felt that it represented a fair balance between a strictly bottom-up approach and imposing too much of our own intuition on the units extracted. The decision to use the lemmas instead of the surface forms was due to the highly inflected nature of French and follows the suggestion of Treffers-Daller (2013) to use lemmas in the calculation of complexity measures in French. In addition, this protects against a possible confound of accuracy in the calculation of our phraseological complexity measures because all inflected forms are treated equally. This is especially important when comparing oral and written French because many grammatical inflections are only encoded orthographically but not phonologically (see Blanche-Benveniste & Adam, 1999). On a practical level, collapsing inflected forms together also makes it possible to calculate association measures in cases where

an exact inflected form may not be attested in a reference corpus but another inflected form, representing a different tense or mode *is* attested, thereby increasing the coverage of our phraseological complexity measures. Another difference from traditional lexical bundle approaches is that we included bundles that crossed punctuation (but not sentence) boundaries (cf. Biber & Barbieri, 2007). This was done because the oral transcriptions contain no punctuation (other than the segmentation into C-units) so we wanted to ensure comparability across modes.

All lexical bundles were extracted using R scripts. Following extraction, the list of units was filtered to remove any verb-bundles starting with the auxiliary verbs *être* and *avoir* (so that the focus was on lexical verbs) and any bundles containing lemmas unknown to TreeTagger. For example, the abbreviation *H24* ('24/7') was not recognized by TreeTagger, resulting in the following bundle: *ouverture_NOM unknown_ADJ du_PRP magasin_NOM* ('unknown opening of the store'). Bundles containing unknown lemmas such as these only accounted for 2.073% of all bundle tokens and 2.196% of all bundle types. Following filtering, we then collapsed all proper nouns and determiners within the bundles. All proper nouns were re-coded as 'NAM' (the TreeTagger POS tag for proper nouns) and all determiners as 'DET'. For example, the surface form, “Monsieur Dupont, mes sincères...” ('Monsieur Dupont, my sincere...') was re-coded as “monsieur_NOM NAM DET sincère_ADJ”. Without collapsing proper nouns, many of the units are not found in the reference corpus (simply because the specific names are not attested). Collapsing determiners was done out of practical necessity because otherwise indexing all variations of units in the reference corpus was not feasible.

The frequency of each bundle was then checked in a large reference corpus and those that appeared fewer than ten times were excluded from analysis. The use of a frequency threshold

ensures that all bundles that are included in the analysis have at least some degree of conventionalization as phraseological units in L1 usage and is in line with previous studies of lexical bundles (e.g., Biber et al., 2004; Biber & Barbieri, 2007). The reference corpus used was the FRCOW16 corpus (Schäfer, 2015; Schäfer & Bildhauer, 2012): a ten billion word web-scraped corpus of French. An obvious disadvantage of this corpus is that it only contains written texts. As such, it may not be exactly representative of the tasks completed by TEF candidates. That being said, the size of the corpus means that a majority of the lexical bundles extracted from the TEF exams are attested, in comparison to smaller oral-specific corpora in which many of the extracted lexical bundles were simply not found (for a similar argument see Paquot et al. 2021). As a precaution, we checked whether there was a significant difference in coverage of lexical bundles from the oral and written tasks and this was not found to be the case. Table 2 shows the most frequent bundles extracted from the learner corpus.

[INSERT TABLE 2]

Following Paquot (2019), we defined phraseological complexity in terms of the diversity and sophistication of the bundles. Whereas previous phraseological complexity research has operationalized diversity using transformations of type-token ratios, the length of the phraseological units in the current study mean that very few of them are repeated verbatim more than once in a given text. As a result, type-token based measures are not very informative. Instead, we operationalized diversity as the ratio of bundle types to the number of different POS structures. For example, imagine a hypothetical text containing the following four verb-bundles:

- faire_VER part_NOM de_PRP DET ('share the')
- prendre_VER effet_NOM avant_PRP DET ('take effect before the')
- pouvoir_VER te_PRO aider_VER à_PRP ('can help you to')

- trouver_VER que_KON ce_PRO être_VER ('find that it is')

In this text, there are four unique verb-bundles but only three unique verb-bundle POS structures (verb + noun + preposition + determiner, verb + pronoun + verb + preposition, verb + conjunction + pronoun + verb).

$$\frac{\text{No. unique verb-bundles}}{\text{No. unique POS structures}} = \frac{4}{3} = 1.33$$

As shown in the above formula, this text would receive a value of 1.33, which reflects the fact that this learner used more than one surface realization of the POS structure verb + noun + preposition + determiner.

The logic behind this measure is that a learner who has a relatively small lexical repertoire will consistently use the same words in the same grammatical structures (resulting in a value of 1), whereas a learner with a larger lexical repertoire will show more variation in the words they use to fill grammatical structures (i.e., each part-of-speech structure will be realized with multiple different lexical realizations). The main difference with this measure and previous operationalizations of phraseological diversity is that instead of focusing on one single structure at a time (e.g. the type-token ratio of verb + noun collocations), this measure takes into account the diversity exhibited within all grammatical structures (i.e., verb + noun + preposition + determiner, verb + pronoun + verb + preposition, etc.).

Phraseological sophistication was operationalized as the pointwise mutual information (PMI, Church & Hanks, 1990) between the first word in the bundle (e.g. *faire_VER*) and the rest of the sequence (e.g. *part_NOM de_PRP DET*). This represents the extent to which the first word (the noun or the verb) *selects* the rest of the sequence (Dunn, 2018). PMI highlights associations that are highly exclusive. Because of this exclusivity, units with a higher PMI tend to serve

specific discursive functions (Ellis et al., 2008). For example, *salutation_NOM DET plus_ADV distingué_ADJ* ('most distinguished salutation'; PMI = 18.5), is used as a greeting in formal letters. In contrast, bundles with a lower PMI tend to be more general and part of everyday use by virtue of the fact that the noun or verb is often found in other structures (e.g. *trouver_VER que_KON ce_PRO suivre|être|sommer_VER*; 'find that it's'; PMI = 4.5).

Both phraseological complexity measures were calculated using a random sampling method. This was because simply calculating measures based on entire texts would introduce a confound of text length. This is particularly problematic given that the median text length of oral productions is over three times longer than that of written productions (see Table 1). To avoid this, measures of phraseological complexity were calculated as the average over ten random samples of five bundles.[3] This method is similar to one that is used to control for text length when calculating morphological complexity (see Brezina & Pallotti, 2019). Visual inspection of the data revealed that the diversity measures exhibited positive-skew, which was addressed by applying a Box-Cox transformation to both measures.

In addition to the phraseological complexity measures, we also calculated two measures of lexical complexity. This was done to determine the extent to which measures of phraseological complexity provide an additional benefit over and above measures that target single words. If phraseological complexity measures are predictive of proficiency, even when controlling for lexical complexity, it would suggest that lexicogrammatical knowledge indeed represents "a different construct" (as suggested by Paquot et al. 2021). Lexical diversity was operationalized using MTLD (McCarthy & Jarvis, 2010), and lexical sophistication was operationalized as the number of words on the 3-, 4- and 5-thousand frequency bands of the FRCOW16 corpus (i.e.,

those that are not within the two thousand most frequent words) (as in Lu, 2012). To control for text length, this measure was computed as the average over 100-word moving windows.

3.3 Analysis

We built a mixed effects regression model using the *nlme* package (Pinheiro et al., 2020) in R in order to predict the proficiency scores of the exams as a function of the lexical and phraseological complexity measures. Texts with fewer than five noun-bundle or verb-bundle types (n = 15) were excluded because they did have enough bundles to calculate the phraseological complexity measures. Exams where the topic could not be conclusively assigned by the model (i.e., where there was a tie between the most highly predicted topics) were also excluded to ensure that the categorical variable 'topic' was truly characteristic of the vocabulary in a given text (n = 37). Data were split into a training set (67%) for model building and a test set (33%) to evaluate the generalizability of the model on unseen data.

The following variables were included as controls in the model: L1 of the candidate, JSD (the topic similarity between the oral and written sections of exams) and the total number of words (tokens) to control for text length. Because each candidate contributed two productions to the data set, the model included random intercepts per candidate ID. We initially began with a fully-specified model with all levels of L1 and topic (the 7 model-generated topics) and where each of the lexical and phraseological complexity measures were allowed to interact with both mode (written/oral) and topic. We then used backwards step-wise model selection in order to determine if all levels L1 and topic were necessary to the model and whether the interactions between mode and topic were necessary for each complexity measure (i.e. whether they significantly improved model fit). After simplifying interactions, non-significant predictors were retained in the model if they either acted as a control variable (i.e. JSD) or if they were necessary

to answer the research questions (in which case a non significant result would indicate that the null hypothesis that the given measure is not predictive of proficiency could not be rejected). All numeric predictors were standardized and centered prior to modelling. Initial models revealed problems with homoscedasticity which were resolved by weighting variance according to mode.

Following the model selection process, only five levels of L1 (French, Arabic, Haitian Creole, Kinyarwanda and other) and two levels of topic (travel and other) were retained in the model. We also tested whether second and third degree polynomial transformations of the complexity measures would improve model fit in order to determine if any of the measures had non-linear relationships with proficiency. This was only found to be the case for MTLD, for which a second degree polynomial significantly improved model fit.

The supplementary materials for this article are provided on an OSF repository.[4] These include the transcription guidelines, topic-modelling procedure, the full model-selection process as well as all R scripts mentioned in the preceding sections. The descriptive statistics for all measures are provided in Appendix 1.

4 Results

The summary of the multiple regression model is provided in Table 3. The model was able to explain 47.21% of variance in proficiency scores overall (conditional R^2) and 33.66% of the variance was explained by the fixed effects in the model (marginal R^2). On unseen data, the model was able to explain 36.81% of the variance in proficiency scores. Both the residual mean squared error (73.23) and mean absolute error (58.90) indicated that when the model was off, it was usually off by less than one hundred points (i.e. less than one CEFR level). Model diagnostics did reveal some non-linearity of residuals but this was largely due to data scarcity at

the lower end of the proficiency scale. For this reason, the accuracy of the model for proficiency scores lower than 400 (CEFR B2) should be interpreted with caution. In the two sections that follow, we present the results of the model concerning the lexical and phraseological complexity measures. As a reminder, these results show the effects of each measure, while controlling for L1 background, text length and topic similarity between the oral and written components of exams.

[INSERT TABLE 3]

4.1 Lexical complexity measures

As shown in Figure 1, there was a significant curvilinear effect of lexical diversity (MTLD) on exam scores whereby an increase in 1 standard deviation of MTLD was found to be associated with 27.26 (CI = 19.5, 35.01) extra points on the exam but this plateaued at $z > 1$, after which an increase in lexical diversity was no longer associated with an increase in test scores. Lexical sophistication not found to be positively correlated with test scores except for texts classified as belonging to the topic of travel (as shown by the significant interaction between lexical sophistication and topic). For texts on the topic of travel, an increase in one standard deviation was associated with an increase of 21.64 (CI = 1.09, 42.18) points on the exam. The lexical sophistication of texts not classified as belonging to the topic of travel was found to be associated with an increase in 0.97 (CI = -7.23, 9.18) points (n.s.).

[INSERT FIGURE 1]

4.2 Phraseological complexity measures

The effects plots for all phraseological complexity measures are shown in Figure 2. None of the phraseological complexity measures were found to interact significantly with topic and only one of the phraseological complexity measures was found to interact significantly with

mode: the diversity of noun-bundles. In the written task, a one standard deviation increase in the diversity of noun-bundles was associated with 9.36 (CI = -0.16, 18.89) extra points on the exam but in the oral mode, this was associated with 4.69 (CI = -12.86, 3.47) *fewer* points on the exam. The diversity of verb-bundles was found to be slightly positively correlated with proficiency scores but this was not significant (EMM trend = 0.04, CI = -5.49, 5.57). In terms of sophistication, positive correlations were found with proficiency for both noun-bundles (EMM trend = 5.29, CI = -0.65, 11.24) and verb-bundles (EMM trend = 7.77, CI = 1.22, 14.32) in both modes but only the correlation with the sophistication of verb-bundles was found to be significant at an alpha level of 5%.

[INSERT FIGURE 2]

5 Discussion

This study aimed to determine: (1) the extent to which measures of phraseological complexity are predictive of oral versus written proficiency scores and (2) the extent to which various measures of phraseological complexity differ in their ability to predict proficiency scores in the two modes.

5.1 Phraseological complexity and proficiency scores

A mixed effects regression model revealed that two of the four phraseological complexity measures were predictive of proficiency scores, namely the diversity of noun-bundles (in the case of the written productions) and the sophistication of verb-bundles (in both the written and oral productions). Importantly, these correlations held even when controlling for the lexical diversity and sophistication of texts. As a comparison, we also built a model that included only lexical complexity measures. This model performed significantly worse than the model with the

phraseological complexity measures (AIC = 8,318.00 vs. 8,310.25, $p < 0.001$) and explained slightly less variance in the proficiency scores (Marginal $R^2 = 0.31$ vs. 0.34), indicating that the lexical complexity measures did not increase their explanatory power once the phraseological measures were removed from the model.

In addition, phraseological measures were found to exhibit two characteristics that made them especially useful as indicators of proficiency. First, unlike with lexical diversity, there was no plateau in the relationship between phraseological complexity and proficiency scores. Increases in phraseological complexity continued to contribute to proficiency scores even at the highest levels of proficiency. Second, none of the phraseological measures were found to interact significantly with topic (cf. Paquot, Naets, et al., 2021 who found a significant effect of topic on phraseological sophistication). This was not the case for lexical sophistication, for which a significant positive relationship with proficiency was found only for texts on the topic of travel. Looking at the texts themselves, it seems that relative to other topics, travel seems to have allowed candidates to mobilize a broader range of semantic categories from very general, frequent vocabulary describing common vacation activities (e.g., *ski* = 25.38/million words; *nager* , 'to swim' = 9.20/million words) to more specific and thus relatively infrequent vocabulary (e.g., *moustiquaire*, 'mosquito net'; 0.97/million words). The fact that such a topic effect was not found for phraseological complexity measures suggests that measures based on the PMI of 4-word lexical bundles may be more robust to variations in topic than measures based on the frequency of single words.

It is important to highlight, however, the fact that the contribution of phraseological complexity to exam scores amounted to fewer than 10 extra points on the exam for every one standard deviation increase in phraseological complexity (holding all other variables constant).

While this is a rather modest contribution, it is still much larger than that of lexical sophistication, which for texts not on the topic of travel, only amounted to an increase of about 1 extra point (n.s.) for every one standard deviation increase in lexical sophistication. This, despite the fact that examiners are explicitly asked to evaluate vocabulary use in the scoring rubric for the TEF exam.

According to Kane (2006, p. 144), when scoring keys include inappropriate criteria or omit some important relevant criteria, this constitutes a threat to the inferences that can be made on the basis of such scores. Specifically, the results here suggest that intentionally or not, raters were sensitive to the phraseological complexity of the TEF productions, as evidenced by the significant relationship between test scores and phraseological complexity measures. It is not clear from these results precisely how raters took this into account however, given that there is no mention of phraseology or idiomaticity in the scoring rubric. In the case of the TEF, this is especially problematic because the total score is calculated using a formula that takes into account the differences in scores between each sub-criterion (e.g., syntax, lexis, coherence/cohesion). It is therefore crucial that the scores attributed to each sub-criterion actually reflect what they are intended to reflect. If some raters considered phraseological complexity a part of lexis and but others considered it a component of syntax, this could lead to differences in the final score given to exams that are objectively similar in terms of global proficiency.

On the basis of the results in this study alone, it is not possible to elaborate further on exactly how lexicogrammatical/phraseological phenomena should be taken into account on scoring rubrics (e.g., as part of existing sub-criterion or as a completely new one?) but future work would do well to explore this question more systematically. What does seem to be important though is that raters receive more explicit instructions as to how such aspects of linguistic competence should be evaluated, in light of the fact that lexico-grammatical

competence constitutes an important aspect of proficiency from a theoretical level (Biber, 2009; Römer, 2009; Sinclair, 1991; Stefanowitsch & Gries, 2003) and that raters appear to be sensitive to such phenomena even when they are not included in scoring rubrics.

5.2 Differences between phraseological complexity measures

The results of this study also showed that measures of diversity and sophistication differed in their ability to predict proficiency scores for the two types of phraseological units that were extracted. For noun-bundles, diversity was more predictive of proficiency than sophistication. When it comes to verb-bundles however, it was sophistication rather than diversity that was most predictive of proficiency scores. These two findings are each discussed in more detail below.

5.2.1 The diversity of noun-bundles**.**

In the case of written productions, using a diverse set of noun-bundles was associated with higher proficiency scores. This is likely because noun-bundles tap into the phrasal complexity features typical of academic, written discourse (Biber, 2019; Halliday & Matthiessen, 2006). A candidate who uses a variety of noun-bundles in written production therefore demonstrates both knowledge of the functional requirements of the register (e.g. the use of nominalizations and post-modification of nouns) and shows evidence of a broad lexicogrammatical repertoire. In oral productions on the other hand, greater noun-bundle diversity was associated with lower proficiency scores. This may be because phrase-level complexity features are less important in speech, which tends to have a more personal or involved focus.

Another possibility is that the "diversity" of oral productions is actually due to the repetition of a small set of semi-fixed units, in line with Römer's (2017) finding that spoken corpora are made up of highly frequent but variable expressions (e.g., I

think/mean/thought/know/suppose it) and Leech's claim (2000, p. 697) that speech relies on a "restricted and repetitive lexicogrammatical repertoire" due to the constraints of real-time processing. For example, in one low-scoring oral production (CEFR B1), a candidate used the following two noun-bundles within the same utterance: *temps avec tes enfants* ('time with your children'; PMI = 4.5) and *sport avec tes enfants* ('sport with your children'; PMI = 3.05). The two bundles are technically unique and thus contribute to a higher value for noun-bundle diversity, despite only differing by one word. To check if this might the case more broadly, we calculated the number of noun-bundles in each text that differed by only one word from another noun-bundle in the same text. As suspected, noun-bundles were repeated with only a one-word difference 0.55 (SD = 0.94) times on average in oral productions, compared to 0.18 (SD = 0.45) times in written productions. Of course, this is only a very crude estimate (and there is a large amount of variation), but it provides more evidence that speech may indeed be less lexicogrammatically diverse than writing overall and suggest that when it comes to lexicogrammatical diversity, simplistic descriptors such as *the more…the better* are not necessarily accurate for oral production. These results also hint at the potential usefulness of looking at patterns of lexicogrammatical "fixedness" in French, that is, the use of fixed, invariable expressions compared to "slot-and-frame" expressions with variable internal components (along the lines of Biber's 2009 study of such patterns English).

5.2.2 The sophistication of verb-bundles**.** Although we had expected that verb-based measures would be more predictive of proficiency in the oral productions than in the written productions, we found no significant interaction effect of mode. The sophistication of verb-bundles was positively associated with proficiency to the same extent in both written and oral productions. In retrospect, this finding is actually quite logical because both tasks require argumentation: an opinion about a controversial topic in the case of the written task and reasons

why a friend should participate in an activity in the oral task. As shown by Biber, Conrad and Cortes (2004), lexical bundles that are used to express attitudes or assessments of certainty (what the authors refer to as 'stance bundles') tend to be composed of dependent clauses and verb phrases (roughly equivalent to the verb-bundles in the current study). This means that the requirements of both tasks (persuading the addressee) require the use of clause-level syntactic features and so the major factor differentiating candidates is therefore the *sophistication* of the verb-bundles. Candidates who use more sophisticated verb-bundles show evidence of having a larger lexicogrammatical repertoire (all other things being equal) because they are able to draw on more specific or appropriate phraseological units.

This finding is consistent with the longitudinal results of Vandeweerd et al. (2021), who found that the sophistication of verb + noun collocations increased significantly over a 21-month period in both written and oral tasks produced by university-level learners of French but that no significant change was observed for adjective + noun collocations. Together with the findings of Vandeweerd et al. (2021), our results suggest that because verb-bundles tap into communicative functions that are common to both oral and written registers, they may be more useful as broad measures of lexicogrammatical competence when comparing across modes.

Taken together, the results of this study suggest that lexicogrammatical complexity represents a separate construct from that of lexical complexity and provide more evidence in favour of including lexicogrammatical criteria within language testing rubrics (Paquot et al., 2021). At the same time, it is important to be cautious of descriptors that are too simplistic and ignore the functional differences between tasks. In Europe, proficiency tests such as the TEF are indexed to the Common European Framework of Reference (CEFR, Council of Europe, 2001). As Paquot (2018) points out, the CEFR promotes a rather traditional understanding of

phraseology and does not do justice to the ways that both lexis and grammar interact with communicative function, especially at the higher proficiency levels. As an example, in both the original descriptors and the revised version (Council of Europe, 2018), the use of "prefabricated expressions" is listed as an indication of lower proficiency. This contrasts somewhat with the results here, which found that certain lexical bundles are actually quite sophisticated by virtue of their exclusivity (PMI) and the fact that they demonstrate knowledge of the characteristics and linguistic requirements for certain registers (e.g. the use of *je me permets de réagir* in formal written letters). To be fair, the most recent version of the CEFR *does* acknowledge the importance of co-occurrence and recurrence phenomena but this is only within the description of *vocabulary control* in general, not within the descriptor scales: "as competence increases, such ability is driven increasingly by association in the form of collocations and lexical chunks, with one expression triggering another." (Council of Europe, 2018, p. 134). However, as shown by the results of this study, reducing the relationship between lexicogrammatical competence and proficiency to a simple linear association, as has previously been done in phraseological complexity studies, does not do justice to the complex ways that lexis and grammar both interact with register, modality and proficiency.

6 Conclusion

In the introduction to this article, we argued (in line with Römer, 2017) that conceptualizing lexis and grammar as solely separate phenomena may lead to invalid inferences about a candidate's linguistic competence given evidence from target language use. We also discussed the ways in which phraseological phenomena are likely to differ between written and spoken exam tasks. Specifically, we hypothesized that the association between phraseological complexity and proficiency would be mediated by constraints of online processing (Skehan,

1998) and register (Biber, 2019). In order to investigate this, we carried out a study in which we compared the complexity (diversity and sophistication) of two types of phraseological units (noun-bundles and verb-bundles) between oral and written productions from the TEF exam. The results of this study showed that phraseological complexity measures were significant predictors of the proficiency scores assigned to texts by raters, even when controlling for lexical complexity. As it turned out, these measures were even better predictors of proficiency than measures of lexical diversity or sophistication. The results also spoke to the interplay between phraseological complexity, modality and register when it comes to evaluating proficiency.

The advantage of the measures presented here is that they allow vague descriptors of linguistic competence (e.g., "as competence increases, such ability is driven increasingly by association…") to be objectively quantified and measured more accurately than may be possible with human raters (Clauser et al., 2010; Williamson et al., 1999). However, it is important to point out that just because the complexity measures employed in this study are correlated with proficiency scores attributed by professional raters, it does not automatically follow that they perfectly represent the construct they are intended to measure (see Pallotti, 2015). Moving forward, it will be important to further evaluate the extent to which these measures really capture the construct of phraseological complexity (e.g., by grounding them in human judgements of phraseological competence and not just proficiency). Such evidence would contribute to understanding the validity of phraseological complexity measures as indices of linguistic proficiency more broadly (see Chapelle & Chung, 2010) as well as their potential usefulness to language assessment (e.g., in automatic scoring).

It is also important to highlight the limitations of the current study. For one, in sampling the TEF exams, we made a decision to prioritize cross-modal topic similarity but this came at the

expense of a less diverse sample of proficiency levels. This meant that the data that was used to build the model was skewed towards higher proficiency levels so the applicability of these findings to lower-level learners (below CEFR B2) will need to be investigated in future studies. In addition, our rather broad focus on all noun- and verb-bundles in this study may be criticized for lack of linguistic precision. While it is certainly the case that not all units that were extracted were necessarily equally as informative about proficiency, this study does represent a good first step in broadening the scope of phraseological complexity measures beyond co-occurrence phenomena to include recurrence phenomena such as lexical bundles. In future studies, it may be fruitful explore the differences between different noun or verb-bundle structures and to evaluate the usefulness of other aspects of recurrence (e.g. fixedness as in Biber, 2009).

Finally, it is important to acknowledge that although lexicogrammatical complexity was shown to be relevant to the evaluations of proficiency in this study, there are of course many other pragmatic and linguistic factors that raters take into account when scoring an exam. Future studies would do well to explore the extent to which lexicogrammatical complexity works in combination with other aspects of proficiency (e.g. accuracy, fluency) as well as how both factors contribute to the ability to accomplish the exam tasks.

While we realize that a complete description of the complicated relationship between phraseological complexity, register and modality is probably beyond the scope of most scoring rubrics, the results of this study show that if descriptors are too simplistic, they risk painting an inaccurate picture of linguistic competence, which constitutes a threat to inferences about a candidate's proficiency made on the basis of exam results. At best, ignoring lexicogrammatical competence may under- or over-estimate a candidate's proficiency. At worst, scoring rubrics may suggest associations between linguistic features that are actually negatively correlated with

proficiency. Considering the results of this study, it seems clear that scoring rubrics should, at the minimum, make reference to lexicogrammatical competence, in addition to lexis and grammar because as argued by Paquot et al. (2021, p. 229), lexicogrammatical knowledge is not simply the "sum of lexical and grammatical knowledge". That being said, we agree with Römer (2017) that as corpus linguists, we do not have the relevant expertise to provide advice about the precise way that these findings should be applied. We do hope however that the results of the current study *can* help to inform a more empirically-grounded approach to the lexis-grammar interface in language testing.

Notes

1. See Demeuse et al. (2010) for more information about the procedure by which the TEF exam was indexed to the CEFR.
2. Our script was designed for maximum efficiency and used up to 10 cores on a dedicated server. Despite this, it still took more than two weeks of uninterrupted run time to extract all four-word bundles from the reference corpus.
3. The sample size of five was chosen because only 13 exams had fewer than five noun-bundles or verb-bundles in either production. These exams were excluded from analysis.
4. https://osf.io/zn26k/

Acknowledgements

We are extremely grateful to the *Chambre de commerce et d'industrie de Paris* for their participation in this project, and in particular, to Dominique Casanova. We would also like to thank Héloïse Copin, Anthonya Delfosse and Thibaut Mareschal for their work transcribing and annotating the exam data. This work was supported by the Fonds de la Recherche Scientifique (FNRS) under Grant n°T0086.18.

References

Alderson, J. C., & Kremmel, B. (2013). Re-examining the content validation of a grammar test: The (im)possibility of distinguishing vocabulary and structural knowledge. *Language Testing*, *30*(4), 535–556. https://doi.org/10.1177/0265532213489568

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, *22*(2), 173–194. https://doi.org/10.1093/applin/22.2.173

Artus, F., & Demeuse, M. (2008). Evaluer les productions orales en français langue étrangère (FLE) en situation de test. Etude de la fidélité inter-juges de l'épreuve d'expression orale du Test d'Evaluation du Français (TEF) de la Chambre de Commerce et d'Industrie de Paris (CCIP). *Les Cahiers Des Sciences de l'éducation*, *25*, 131–151.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bestgen, Y., & Granger, S. (2018). Tracking L2 writers' phraseological development using collgrams: Evidence from a longitudinal EFL corpus. In S. Hoffmann, A. Sand, S. Arndt-Lappe, & L. M. Dillmann (Eds.), *Corpora and lexis* (pp. 277–301). Brill Rodopi.

Biber, D. (2009). A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics*, *14*(3), 275–311. https://doi.org/10.1075/ijcl.14.3.08bib

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, *14*(1), 7–34. https://doi.org/10.1075/lic.14.1.02bib

Biber, D. (2019). Text-linguistic approaches to register variation. *Register Studies*, *1*(1), 42–75. https://doi.org/10.1075/rs.18007.bib

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, *26*(3), 263–286. https://doi.org/10.1016/j.esp.2006.08.003

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

Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the Toefl iBT test: A lexico-grammatical analysis*. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02311.x

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*(1), 5–35. https://doi.org/10.5054/tq.2011.244483

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics*, *37*(5), 639–668. https://doi.org/10.1093/applin/amu059

Biber, D., Leech, G., Johansson, S., & Conrad, S. (1999). *Longman grammar of spoken and written English*. Longman.

Blanche-Benveniste, C., & Adam, J.-P. (1999). La conjugaison des verbes: Virtuelle, attestée, defective. *Recherches sur le français parlé*, *15*, 87–112.

Brezina, V., & Pallotti, G. (2019). Morphological complexity in written L2 texts. *Second Language Research*, *35*(1), 99–119. https://doi.org/10.1177/0267658316643125

Casanova, D., & Demeuse, M. (2016). Évaluateurs évalués : Évaluation diagnostique des compétences en évaluation des correcteurs d'une épreuve d'expression écrite à forts enjeux. *Mesure et évaluation en éducation*, *39*(3), 59–94. https://doi.org/10.7202/1040137ar

Chambre de commerce et d'industrie de Paris. (2010). *TEF: Le Test d'évaluation de français de la Chambre de commerce et de l'industrie de Paris*.

Chapelle, C. A., & Chung, Y. R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, *27*(3), 301–315. https://doi.org/10.1177/0265532210364405

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). Test score interpretation and use. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the test of english as a foreign language* (pp. 1–25). Routledge. https://doi.org/10.4324/9780203937891

Church, K. W., & Hanks, P. (1990). *Word association norms, mutual information, and lexicography*. *16*, 22–29. https://doi.org/10.3115/981623.981633

Clauser, B. E., Kane, M. T., & Swanson, D. B. (2010). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education*, *15*(4), 413–432. https://doi.org/10.1207/S15324818AME1504_05

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

Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume with new descriptors*. Council of Europe Publishing.

Davies, A. (2004). The Native Speaker in Applied Linguistics. In A. Davies & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 431–450). Blackwell. https://doi.org/10.1002/9780470757000.ch17

De Haan, P. (2000). Tagging non-native English with the TOSCA-ICLE tagger. In C. Mair & M. Hundt (Eds.), *Corpus linguistics and linguistic theory*. Rodopi.

Demeuse, M., Desroches, F., Casanova, D., Crendal, A., & Holle, A. (2010). *Validation empirique d'un test de français langue étrangère en regard du Cadre européen commun de référence pour les langues*. Paper presented at the 22nd Colloque international de l'Association pour le développement des méthodologies d'évaluation en éducation (ADMEE–Europe), Braga, Portugal.

Denis, P., & Sagot, B. (2012). Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, *46*, 721–736. https://doi.org/10.1007/s10579-012-9193-0

Dunn, J. (2018). Multi-unit association measures Moving beyond pairs of words. *International Journal of Corpus Linguistics*, *23*(2), 183–215. https://doi.org/10.1075/ijcl.23.2

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

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics and TESOL. *TESOL Quarterly*, *5*(1), 375–396. https://doi.org/10.1515/CLLT.2009.003

Forsberg, F. (2010). Using conventional sequences in L2 French. *International Review of Applied Linguistics in Language Teaching*, *48*(1), 25–51. https://doi.org/10.1515/iral.2010.002

Granfeldt, J. (2007). Speaking and writing in French L2: Exploring effects on fluency, complexity and accuracy. In S. van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning & teaching* (pp. 87–98). Koninklijk Vlaamse Academie van Belgie voor Wetenschappen en Kunsten.

Granger, S. (2014). A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast*, *14*(1), 58–72. https://doi.org/10.1075/lic.14.1.04gra

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. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 3–25). John Benjamins. https://doi.org/10.1075/z.139.06gri

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2006). *Construing experience through meaning*. Continuum.

Holliday, A. (2017). Native-Speakerism. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1–9). John Wiley & Sons, Inc.

Jarvis, S., & Hashimoto, B. J. (2021). How operationalizations of word types affect measures of lexical diversity. *International Journal of Learner Corpus Research*, *7*(1), 163–194. https://doi.org/10.1075/ijlcr.20004.jar

Kane, M. (2006). Content-related validity evidence in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 131–153). Lawrence Erlbaum. https://doi.org/10.4324/9780203874776.ch7

Kremmel, B., Brunfaut, T., & Alderson, J. C. (2017). Exploring the role of phraseological knowledge in foreign language reading. *Applied Linguistics*, *38*(6), 848–870. https://doi.org/10.1093/applin/amv070

Kyle, K., & Crossley, S. A. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*(4), 513–535. https://doi.org/10.1177/0265532217712554

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

Leech, G. (2000). Grammars of spoken english: New outcomes of corpus-oriented research. *Language Learning*, *50*(4), 675–724. https://doi.org/10.1111/0023-8333.00143

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36–62. https://doi.org/10.5054/tq.2011.240859

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–206. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Mattiuzzi, S. (2021). *An evaluation of part-of-speech taggers for French* [PhD thesis]. Université de Genève.

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. https://doi.org/10.3758/BRM.42.2.381

Michel, M., Murakami, A., Alexopoulou, T., & Meurers, D. (2019). Effects of task type on morphosyntactic complexity across proficiency. *Instructed Second Language Acquisition*, *3*(2), 124–152. https://doi.org/10.1558/isla.38248

Murakami, A., Thompson, P., Hunston, S., & Vajn, D. (2017). 'What is this corpus about?': Using topic modelling to explore a specialised corpus. *Corpora*, *12*(2), 243–277. https://doi.org/10.3366/cor.2017.0118

Nesselhauf, N. (2005). *Collocations in a learner corpus*. John Benjamins.

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

Pallotti, G. (2015). A simple view of linguistic complexity. *Second Language Research*, *31*(11), 117–134. https://doi.org/10.1177/0267658314536435

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. https://doi.org/10.1080/15434303.2017.1405421

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

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

Paquot, M., Gries, S. Th., & Yoder, M. (2021). Measuring Lexicogrammar. In P. Winke & T. Brunfaut (Eds.), *The Routledge handbook of second language acquisition and language testing* (pp. 223–232). Routledge. https://doi.org/10.4324/9781351034784-25

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.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2020). *nlme: Linear and Nonlinear Mixed Effects Models*. https://cran.r-project.org/package=nlme

Ravid, D., & Tolchinsky, L. (2002). Developing linguistic literacy: a comprehensive model. *Journal of Child Language*, *29*(2), 417–447. https://doi.org/10.1017/s0305000902005111

Römer, U. (2009). The inseparability of lexis and grammar: Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics*, *7*(2009), 140–162. https://doi.org/10.1075/arcl.7.06rom

Römer, U. (2017). Language assessment and the inseparability of lexis and grammar: Focus on the construct of speaking. *Language Testing*, *34*(4), 477–492. https://doi.org/10.1177/0265532217711431

Rooy, B. van, & Schäfer, L. (2002). The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies*, *20*(4), 325–335. https://doi.org/10.2989/16073610209486319

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 second language phrasicon: The view from learner corpora* (pp. 101–125). Multilingual Matters.

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

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)* (pp. 486–493). European Language Resources Association (ELRA).

Schmid, H. (1994). *TreeTagger - a part-of-speech tagger for many languages*. [Computer software]. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

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

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Stefanowitsch, A., & Gries, S. Th. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8*(2), 209–243. https://doi.org/10.1075/ijcl.8.2.03ste

Treffers-Daller, J. (2013). Measuring lexical diversity among L2 learners of French: an exploration of the validity of D, MTLD and HD-D as measures of language ability. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp. 79–104). John Benjamins.

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 [Advance online publication]. *Studies in Second Language Acquisition*. https://doi.org/10.1017/S0272263122000389

Vinay, J.-P., & Darbelnet, J. (1995). *Comparative stylistics of French and English* (J. Sager & M.-J. Hamel, Trans.). John Benjamins.

Wang, J., & Dong, Y. (2020). Measurement of text similarity: A survey. *Information (Switzerland)*, *11*(9), 1–17. https://doi.org/10.3390/info11090421

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). 'Mental model' comparison of automated and human scoring. *Journal of Educational Measurement*, *36*(2), 158–184.

Zheng, Y. (2016). The complex, dynamic development of L2 lexical use: A longitudinal study on Chinese learners of English. *System*, *56*, 40–53. https://doi.org/10.1016/j.system.2015.11.007

Figures and Tables

Table 1: TEF Sample

	Oral	Written
Number of exams by L1 (%)		
Arabic	226 (41%)	
French	173 (32%)	
Haitian Creole	11 (2.0%)	
Kinyarwanda	12 (2.2%)	
other	123 (23%)	
Median number of tokens (range)	966 (145, 2,006)	271 (104, 700)
Median score (range)	551 (263, 699)	475 (165, 699)

Table 2: Most frequent bundles extracted from the TEF corpus

Bundle[1]	Example	Type	PMI	Written	Oral	Total	FRCOW
penser_VER que_KON ce_PRO suivre|être|sommer_VER	*je pense que c'est positive*	verb-bundle	6.27	5	368	373	91,426
monsieur_NOM DET rédacteur_NOM en_PRP	*monsieur, le rédacteur en chef*	noun-bundle	5.62	135	0	135	137
lecteur_NOM de_PRP DET journal|journau_NOM	*lecteur de votre journal*	noun-bundle	8.48	114	0	114	187
faire_VER part_NOM de_PRP DET	*faire part de mon point de vue*	verb-bundle	5.87	100	0	100	59,530
dire_VER que_KON ce_PRO suivre|être|sommer_VER	*Je tiens à vous dire que c'est absurde*	verb-bundle	5.53	3	84	87	138,227

Bundle = Extracted bundle, Example = Example in context, Type = Category of bundle, PMI = Pointwise Mutual Information calculated on the basis of the FRCOW16 corpus, Written = Frequency in written TEF texts, Oral = Frequency in oral TEF texts, Total = Total frequency in TEF corpus, FRCOW = Frequency in the FRCOW16 Corpus

[1]Vertical bars indicate lemmas that could not be disambiguated by TreeTagger and were therefore collapsed in the analysis.

Table 3: Model summary

SCORE ~ L1 + JSD + TOKENS + poly(MTLD, 2) + FRCOWOFF2K * TOPIC + DIV.PMI.N + DIV.PMI.V + DIVERSITY.N.bcn * MODE + DIVERSITY.V.bcn						
Fixed Effects	b	SE	df	t	p	
(Intercept)	559.658	7.28	359	76.879	<0.001	***
L1ara	-42.35	6.977	359	-6.069	<0.001	***
L1hat	-6.573	21.764	359	-0.302	0.763	
L1kin	-16.252	20.722	359	-0.784	0.433	
L1other	-59.504	8.335	359	-7.139	<0.001	***
JSD	-4.352	2.972	359	-1.464	0.144	
TOKENS	55.705	4.12	353	13.521	<0.001	***
poly(MTLD, 2)1	690.68	109.983	353	6.28	<0.001	***
poly(MTLD, 2)2	-157.539	70.679	353	-2.229	0.026	*
LEX.SOPHISTICATION	0.972	3.411	353	0.285	0.776	
TOPIC.travel	-2.456	6.835	353	-0.359	0.720	
PHRAS.SOPHISTICATION.N	5.291	3.023	353	1.751	0.081	.
PHRAS.SOPHISTICATION.V	7.768	3.331	353	2.332	0.020	*
PHRAS.DIVERSITY.N	-4.694	3.394	353	-1.383	0.168	
MODEwritten	-28.372	10.978	353	-2.584	0.010	*
PHRAS.DIVERSITY.V	0.036	2.811	353	0.013	0.990	
LEX.SOPHISTICATION:TOPIC.travel	20.664	8.88	353	2.327	0.021	*
PHRAS.DIVERSITY.N:MODEwritten	14.059	5.184	353	2.712	0.007	**
Random Effects	Variance	SD				
(Intercept)	1603.769	40.047				
Residual	6253.346	79.078				

Marginal R^2: 0.337
Conditional R^2: 0.472
*** p< 0.001, ** p < 0.01, *, p < 0.05, . p < 0.1

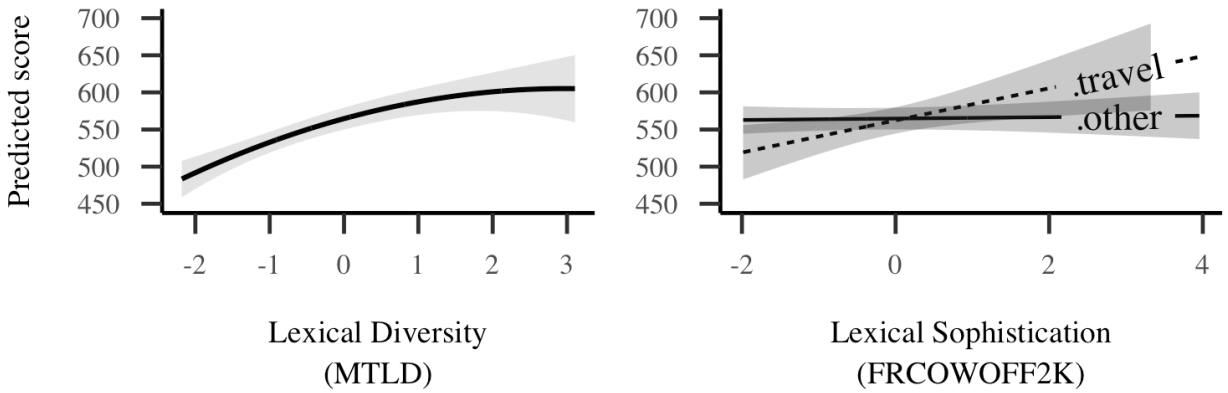

Figure 1. Effects plots for lexical complexity measures

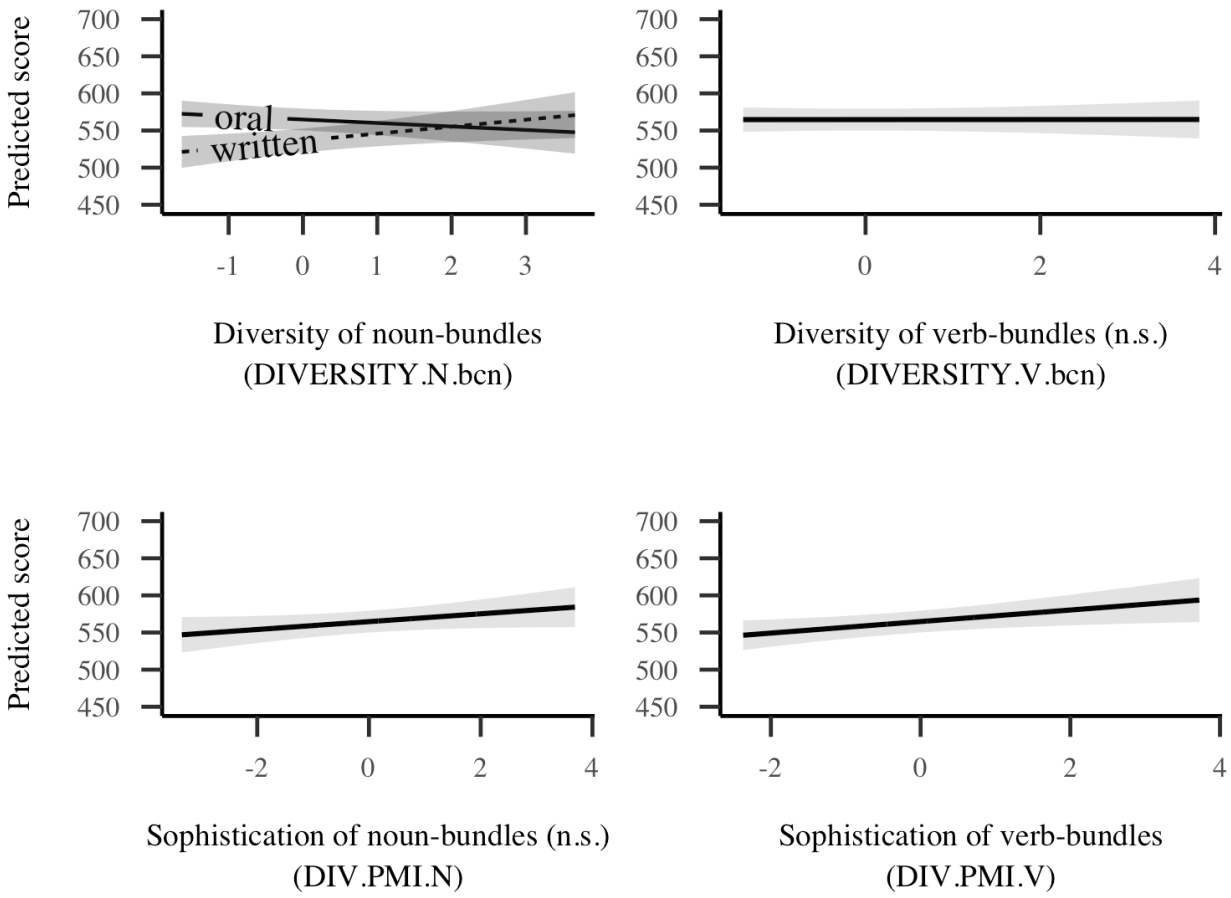

Figure 2. Effects plots for phraseological complexity measures

Appendix 1

Table 4: Descriptive statistics for data used in model building

		oral, N = 545	written, N = 545
Exam characteristics	Exams per L1 (%)		
	Arabic	226 (41%)	226 (41%)
	French	173 (32%)	173 (32%)
	Haitian Creole	11 (2.0%)	11 (2.0%)
	Kinyarwanda	12 (2.2%)	12 (2.2%)
	other	123 (23%)	123 (23%)
	Median JSD (IQR)	0.34 (0.17, 0.50)	0.34 (0.17, 0.50)
	Exams per topic (%)		
	.other	422 (77%)	487 (89%)
	.travel	123 (23%)	58 (11%)
	Median number of tokens (IQR)	966 (771, 1,131)	271 (223, 325)
	Median score (IQR)	551 (499, 598)	475 (404, 549)
Median value for complexity measures (IQR)	LEX.DIVERSITY (MTLD)	40 (35, 46)	58 (50, 66)
	LEX.SOPHISTICATION (FRCOWOFF2K)	3.18 (2.49, 3.95)	4.81 (3.55, 6.06)
	PHRAS.DIVERSITY (noun-bundles)	0.08 (0.05, 0.12)	0.08 (0.05, 0.13)
	PHRAS.DIVERSITY (verb-bundles)	0.07 (0.05, 0.08)	0.07 (0.03, 0.12)
	PHRAS.SOPHISTICATION (noun-bundles)	4.68 (4.04, 5.41)	5.91 (5.07, 6.64)
	PHRAS.SOPHISTICATION (verb-bundles)	4.81 (4.38, 5.35)	6.22 (5.48, 6.91)