# Vocabulary sophistication in first-year composition assignments

Philip Durrant[i], Joseph Moxley[ii], and Lee McCallum[i]

[i]University of Exeter | [ii]University of South Florida

Recently-developed tools which quickly and reliably quantify vocabulary use on a range of measures open up new possibilities for understanding the construct of vocabulary sophistication. To take this work forward, we need to understand how these different measures relate to each other and to human readers' perceptions of texts. This study applied 356 quantitative measures of vocabulary use generated by an automated vocabulary analysis tool (Kyle & Crossley, 2014) to a large corpus of assignments written for First Year Composition courses at a university in the United States. Results suggest that the majority of measures can be reduced to a much smaller set without substantial loss of information. However, distinctions need to be retained between measures based on content vs. function words and on different measures of collocational strength. Overall, correlations with grades are reliable but weak.

**Keywords**: first-year composition, academic writing, vocabulary, vocabulary sophistication, writing assessment

## 1. Introduction

The question of whether quantifiable features of student writing are reliably associated with the grades awarded to that writing goes back several decades (e.g. Golub & Frederick, 1979; Grobe, 1981; Malvern et al., 2004; Olinghouse & Leaird, 2009). Recent years have given a new lease of life to this type of work thanks to the advent of large corpora of student texts and of computational tools capable of identifying key features of those texts (e.g. Crossley et al., 2012; Crossley et al., 2011). One motivation for work of this sort is the potential it offers for creating automated grading and feedback systems (Crossley et al., 2013). Another is its potential for helping us understand readers' subjective reactions to texts. In particular, identifying the objective correlates of judgments about quality enables us to give specific content to our intuitive sense of what constitutes "good writing" (Myhill, 1999). Making explicit usually tacit understandings of quality has at least two clear benefits. First, it enables us to make these understandings accessible to students, demystifying the construct of “good writing”. Second, it gives us the opportunity to reflect on what influences our perceptions of quality and, where appropriate, to critique and revise these influences. Within this area, research on vocabulary use has particularly strong potential. Vocabulary use can be studied computationally with a higher degree of reliability than has been achieved for features such as syntax and cohesion, whose automated analysis (especially in learner texts) can be problematic (Meurers & Dickinson, 2017 describe some of the difficulties). Moreover, as the review below will show, previous research has indicated that quantitative measures of vocabulary use are developmentally significant, being sensitive to linguistic development in both first and second language writing. Finally, patterns of variation in vocabulary use (e.g. amount of repetition; use of lower frequency words; use of register-specific words) are both intuitively meaningful and have clear pedagogical implications.

The potential for development in this area has recently been extended by the availability of Kyle & Crossley’s *Tool for the Automated Analysis of Lexical Sophistication* (*TAALES*) (Kyle & Crossley, 2016). As described below, this allows for the rapid quantification of vocabulary use in large numbers of texts across a wide range of indices of sophistication. While immensely useful, users of this tool are faced with something of an embarrassment of riches. The outputs provided (the latest version gives almost 500 indices) are so plentiful that it is a challenge for users to make sense of them and to decide which measures or sets of measures to use in their analyses. With this in mind, the first aim of this study is to get a better understanding of how *TAALES* data can be used and interpreted by determining how its various frequency-based measures of word-use and phraseology relate to one another and the extent to which each offers distinct information from the others. As well as being of relevance to researchers interested in *TAALES* as a tool, understanding these indices and the relationships between them may also give us insights into the construct of vocabulary sophistication itself. Our second aim is to apply this understanding to uncover the relationships between lexical sophistication and scores in a large corpus of assignments submitted as part of a First Year Composition program at a large public university in the United States. As will be shown below, quantitative measures of vocabulary use have been related to both linguistic development and perceptions of quality in first and second language learners’ writing. The present research aims to extend this to understand how it relates to perceptions of quality of authentic assignments set in a US higher education setting. We will thus address two research questions:

i. How do the various word-frequency and phraseological frequency/association measures provided by TAALES relate to each other?

ii. To what extent are these purported measures of vocabulary sophistication associated with grades awarded to composition assignments at a US university?

## 2. Quantitative measures of vocabulary sophistication

Three main types of measure have been proposed that lend themselves well to automated, quantitative analysis of vocabulary use: measures of density, diversity, and sophistication (Read, 2000). Measures of density quantify the proportion of text made up of content vs. function words. However, while this measure is important for distinguishing text genres (e.g. Biber, 1988), there is little reason to believe that it is of developmental interest (e.g. Berman & Nir, 2010; Golub & Frederick, 1979; Uccelli et al., 2013). Measures of diversity aim to capture the number of different words used by a writer. Clearly, the number of distinct words used is closely related to the total length of a text and the majority of literature in this area has been devoted to identifying ways of accounting for this (Malvern et al., 2004; McCarthy & Jarvis, 2011). Overwhelmingly, the evidence in this area confirms the intuitive claim that more advanced texts employ a more diverse vocabulary in both first (e.g. Berman & Nir, 2010; Crossley et al., 2011; Malvern et al., 2004; Olinghouse & Wilson, 2013; Uccelli et al., 2013) and second language writing (Crossley et al., 2010; Daller et al., 2013; Guo et al., 2013; Hou et al., 2016; Treffers-Daller et al., 2018).

Work on sophistication remains more open. Read (2000: 200) characterizes lexical sophistication as the "selection of low frequency words that are appropriate to the topic and style of writing, rather than just everyday vocabulary". This has been operationalized in several different ways:

*Word length*: perhaps the simplest measure of sophistication has been to quantify the mean length of words in a text. Word length increases with age and quality in L1 writers (Berman & Nir-Sagiv, 2007; Grobe, 1981; Malvern et al., 2004; Massey & Elliott, 1996; Massey et al., 2005; Myhill, 2009; Olinghouse & Leaird, 2009). For L2 writers there is conflicting evidence, with some studies finding that word length is related to proficiency (Cumming et al., 2005; Jarvis et al., 2003) or to longitudinal development (Hou et al., 2016), and others finding no effect (Crossley et al., 2010; Knoch et al., 2015; Verspoor et al., 2012; Vidakovic & Barker, 2010). It is important to note, however, that the construct which these measures tap is rather ambiguous since mean word length is likely to reflect a combination of use of low frequency words and morpho-syntactic complexity (e.g. use of derivations or of participle forms will increase word length).

*Word Frequency*: frequency is most commonly quantified by counting the percentage of words which are not found on a particular list of high-frequency vocabulary. Findings here have also not been consistent. Use of lower-frequency vocabulary in L1 writing does appear to increase with age (Malvern & Richards, 2002; Olinghouse & Leaird, 2009). However, studies looking at correlation with quality have had mixed results, some finding that higher-rated texts use lower-frequency vocabulary (Olinghouse & Leaird, 2009; Roessingh et al., 2015), while others have either found no effect (Malvern et al., 2004) or found the effect to be present only in certain genres (Olinghouse & Wilson, 2013). L2 studies have been similarly inconsistent. Some have reported significant increases in the use of low-frequency words over time (Daller et al., 2013; Knoch et al., 2015; Knoch et al., 2014; Storch, 2009) or with increasing proficiency (Vidakovic & Barker, 2010). Others have reported no increase with time and/or no relationship between use of low-frequency words and text quality (Bulté & Housen, 2014; J.-Y. Kim, 2014) or even a decrease in low-frequency words with time (Hou et al., 2016).

Another approach has been to determine how frequent each word in a text is in a particular reference corpus and average across these to find a mean frequency for the text as a whole. Again, results are ambiguous. Crossley et al. (2011) find no significant difference between L1 ninth and eleventh graders, but ninth-graders used on average significantly more frequent words than college writers. Durrant & Brenchley (in press) separate frequency counts for content words (adjectives, nouns, verbs and adverbs) and function words and find that the mean frequency of the former decreases while that of the latter increases for L1 writers from ages 7 to 14. Results from L2 studies suggest mean word frequency decreases over time (Mazgutova & Kormos, 2015) but this may be task-dependent. Measures calculated using the *Coh-Metrix* tool (Graesser et al., 2014), based on the CELEX[1] database, are found to be weak indicators of text quality in integrated texts written for the Test of English as a Foreign Language (TOEFL) exam but not in independent texts (Guo et al., 2013). Other mean frequency measures using frequency lists from Kucera-Francis (Kucera & Francis, 1967), the Corpus of Contemporary American (COCA)[2] and SUBTLEX-US[3] have yielded negative or weak correlations with quality (Kyle, 2017; Kyle & Crossley, 2016).

*Phraseological sophistication:* while the above research has focused on individual word sophistication, L2 research specifically has moved on to consider sophistication in phraseological units. The most straightforward measures have quantified the proportion of combinations in texts which are attested (i.e. appear at least once) in a reference corpus. Bestgen & Granger (2014), Kyle & Crossley (2016), Bestgen (2017) and M. Kim et al. (2018) all find that a greater proportion of attested forms is significantly associated with a higher score. Strikingly, Crossley et al. (2012) find the opposite tendency – i.e. fewer attested forms associated with higher grades – when looking at the scores awarded to university-level writing, a context similar to our own. This hints at an interesting contrast, whereby the collocational originality (i.e. use of unattested combinations) which is valued in one context is disvalued in the other. Since the evidence for this contrast rests on a single study, however, more work is needed to confirm the pattern.

Another approach has been to quantify collocation use in terms of the mean reference-corpus frequency of items. Kyle & Crossley (2016) find the log frequency of bigrams to correlate positively with scores for the independent writing part of the TOEFL test and to correlate negatively with those for the integrated part. The findings of Crossley et al (2012) again point in the opposite direction, showing negative correlations between log bigram frequency and scores awarded to first-year composition assignments at a US university.

A second main focus of interest has been association measures, and particularly mutual information. One approach has been to look at the proportion of combinations meeting particular thresholds. Granger & Bestgen (2014) find that the proportion of bigrams with MI > 7 is significantly greater in advanced than in intermediate L2 writing, while proportion of bigrams with MI < 3 is significantly greater in intermediate than advanced writing. Paquot (2018), looking at two-word syntactic combinations, does not find any significant differences across proficiency levels for adverb modifier or verb-direct object structures. However, for adjective-noun structures, she finds that the proportion of combinations with MIs of 3-5 and 5-7 significantly increase with proficiency while the proportion of combinations with MI < 3 decrease. Studies looking at the mean MI of combinations have been more consistent. This measure has been found to increase across L2 proficiency levels for both syntactic combinations (Paquot, 2018, 2019) and adjacent bi- and trigrams (Bestgen, 2017; Bestgen & Granger, 2014; Garner et al.; M. Kim et al., 2018).

Two other association measures that have been studied are t-score and Delta-P. Granger & Bestgen (2014) find that advanced L2 users use significantly fewer very high-scoring combinations ($t > 6$), and significantly more moderately-scoring combinations ($6 < t > 2$) than intermediate students. They also find that very low-scoring combinations (t < 2) are more common in intermediate than in advanced learner writing. Studies of mean t-score have been inconsistent, with Garner et al. (2018) reporting a significant positive correlation with L2 proficiency, while Bestgen & Granger (2014) find no such correlation. Delta-P has shown more consistent results, with both M. Kim et al. (2018) and Garner et al. (2018) findings significant positive correlations between Delta-P of bi/tri-grams and scores for L2 writers.

Bringing the above together, research to date has shown clear patterns for lexical density (no relationship with development or writing quality) and diversity (increases with development and increased quality). Results for lexical sophistication are less straightforward. This construct has been operationalised in a wide range of ways and, while some of these measures appear to be related to development/quality in certain contexts, results have been inconsistent and the relationships between the various measures remain unclear. The present study aims to move this work forward by determining how measures of sophistication are related to each other and how they correlate with perceptions of writing quality in the context of a first-year composition programme.

## 3. Methods

This section will first (Section 3.1) introduce the corpus on which our study was based and the context in which it was collected. It will then (Section 3.2) describe the process of analysis, setting out the TAALES indices of interest and our procedures for evaluating their relationships with each other and with essay grades.

### **3.1.** Corpus

This study is based on a corpus of writing produced by students on the First-Year Composition (FYC) program of a large public university in the United States. The program comprises two courses, ENC 1101 and ENC 1102 (hereafter 1101 and 1102). Both courses are required for all first-year students, yet some students qualify for exemption from 1101 and/or 1102, either because they have taken equivalent courses or because they have received scores on exams that have been deemed equivalent. Reflecting the international makeup of the university, these courses are attended by students with a range of first languages. In this research, we do not make a distinction between “native” and “non-native” speakers of English since our focus is on what makes for successful writing in a “mainstream” US university writing program, regardless of the origins of the writer. The FYC program has received the CCC Writing Program Certificate of Excellence Award from the National Council of Teachers of English.

Both 1101 and 1102 focus on student writing. Students write three drafts for each of the three major projects in each course. The projects in 1101 evolve from learning research skills, to identifying arguments in a scholarly conversation, and culminate in an argumentative paper and presentation. 1102 begins where 1101 concludes, by asking students to find a compromise between two opposed stakeholders, analyse a stakeholder's argument through visual rhetoric, and finally call an audience to action through written and multimodal argumentation. Student learning outcomes in 1101 include understanding the writing process, research on current and controversial ideas, and understanding human diversity. 1102 aims to promote understanding of producing knowledge, evaluating information, and becoming a responsible steward of the environment. These learning outcomes ask students to choose and research a topic, and to write about that topic in different genres. While 1102 can function independently for students who do not take 1101, it is designed as a continuation of 1101.

Our study corpus comprised final draft assignments submitted for 1101 and 1102 from January 2016 to January 2017. In their raw form, the majority of texts include reference lists, which are likely to skew vocabulary counts. In the majority of texts, these lists are marked clearly at the end of the text, with headings such as *Works Cited* or *References*. These portions of the texts were automatically removed.[4] Texts which did not mark the reference list with a commonly-used section heading were not included in the analysis. An exception to this was the texts for ENC1101, Project 1. This project asked students to produce an annotated bibliography. These texts did not have a reference list at the end of the text, but rather included references through the length of the text. These texts were included unchanged. Table 1 summarises the contents of the corpus.

The metadata for the corpus included grades assigned by class instructors to each text. 1101 and 1102 are taught primarily by graduate assistants in English with either an emphasis in literature, rhetoric and composition, or creative writing. Some adjunct faculty also teach in the program, and they typically have a Ph.D. in literature or an MFA in Creative Writing. Grade norming and practice are informally conducted at the beginning of each academic year for all composition instructors. Additionally, each fall in a graduate-level Composition Practicum, English department faculty conduct grade-norming exercises to help new graduate assistants grade and comment on undergraduate papers. Furthermore, mentors, who are outstanding graduate students or adjuncts, work with first-time instructors to help improve their feedback and scoring. The Writing Program Administrators also review all instructors' feedback, and they meet with the staff when there are student complaints or when instructors seek mentorship.

Each project uses its own standardized rubric for assessment and grading purposes. These rubrics were developed via peer-production, as documented in Vieregge et al. (2012)

**Table 1:** Corpus makeup

| Class | Project | Number of texts | Words per text | | | Final score distribution | | | | Style score distribution | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Mean | Min | Max | Min | Max | Mean | St.Dev | Min | Max | Mean | St.Dev |
| 1101 | 1 | 1,422 | 1,730 | 234 | 4,304 | 0 | 15 | 11.05 | 2.83 | NA | NA | NA | NA |
| | 2 | 1,140 | 985 | 376 | 2,038 | 0 | 15 | 11.49 | 2.71 | 2 | 8 | 6.61 | 1.23 |
| | 3 | 1,131 | 1,129 | 384 | 2,296 | 0 | 15 | 11.81 | 2.56 | 1 | 8 | 6.87 | 1.17 |
| 1102 | 1 | 1,783 | 1,176 | 551 | 2,497 | 0 | 15 | 11.35 | 2.59 | 1 | 8 | 6.41 | 1.46 |
| | 2 | 1,705 | 1,265 | 25 | 2,377 | 0 | 15 | 11.78 | 2.44 | 1 | 8 | 6.58 | 1.35 |
| | 3 | 1,561 | 1,186 | 513 | 2,361 | 0 | 15 | 11.78 | 2.55 | 1 | 8 | 6.70 | 1.29 |

**Figure 1:** *Style* rubric

| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Low emerging | Emerging:<br>• Frequent diction, grammar and/or punctuation issues<br>• Frequent shifts in point of view<br>• Frequent problems demonstrating accurate usage of standard edited English<br>• Language significantly interferes with the communication of ideas | High emerging | Low developing | Developing:<br>• Some diction, grammar and/or punctuation errors<br>• Inconsistent points of view<br>• Inconsistently demonstrates accurate usage of standard edited English<br>• Language does not interfere with communication of ideas | High developing | Low mastering | Mastering:<br>• Appropriate diction, grammar, and punctuation<br>• Consistent point of view<br>• Consistently polished and appropriate usage of standard edited English<br>• Language consistently complements and facilitates communication of ideas | High mastering |

and have subsequently been used in a number of research studies (Moxley, 2013). The rubric criteria are “analysis”, “evidence”, “organisation”, “format” and “style”. An overall holistic grade is also given, independently of these. All grades are on a scale from zero to eight, while the holistic grade is on a 15-point scale, expressed as letter grades from “A+” to “F”. In Table 1, and in the following analyses, final scores have been converted to a numerical scale, where “1” represents an “F” and “15” represents an “A+”.

The “style” grade is awarded on the basis of language use and so is particularly relevant to the present research. The rubric for this (see Figure 1) is identical for all projects. This scale is worth 15% of the final grade and instructors are told to assess diction, grammar, punctuation, point of view, and Standard English. “Vocabulary” is never explicitly mentioned in the rubric, though diction and Standard English are mentioned. Each assignment is graded by a single instructor. Since composition grading is notoriously unreliable (Meadows & Billington, 2005), this raises the possibility that grades would have been different if marked by different instructors or on a different day. It is important to stress, therefore, that our aim in Research Question 2 is not to compare lexical measures with an objective construct of “writing quality”, but rather with the actual practices of instructors in a specific real-world context where, due to resource limitations, single-instructor grading is the norm.

### **3.2.** Analysis

The aims of this study were to understand how the various components of lexical sophistication measured by the *TAALES* (V2.2) (Kyle & Crossley, 2015) relate, firstly, to each other and, secondly, to the grades awarded to student writing in the FYC context. Our analysis takes these two types of relationship in turn. First, we consider how *TAALES* measures relate to each other and whether these relationships allow us to reduce the large number of measures available to a smaller set. We then consider how each of these smaller sets relates to the grades awarded to texts.

*TAALES* quantifies lexical use on a total of 484 indices, covering the broad measure types of word frequency; word range; n-gram measures; measures of academic vocabulary; and psycholinguistic and semantic measures. Given the nature of the following analysis, it is not possible to look at all of these within the space of a single paper, so we will focus only on those measures which are most relevant to the construct of vocabulary sophistication set out in Section 2. As discussed there, word and phrase frequency and phrase association have been central to this construction of sophistication. By extension, measures of range, which can be seen as another way of approaching frequency – looking at the number of different texts in which a word appears, rather than the raw number of occurrences – are clearly important. Finally, Read's (2000: 200) emphasis on selection of words which are appropriate to a particular "topic and style of writing, rather than just everyday vocabulary" implies that, for the academic context of the current study, measures of academic vocabulary are also relevant. While semantic and some psycholinguistic measures could also be seen as an aspect of sophistication, they do not fit naturally into Read's definition of sophistication, so have been excluded from the current study. These may be a useful focus for future work looking at a broader construction of sophistication (indeed, Crossley et al., 2011 have already taken some steps in this direction). With this in mind, our analysis focuses on seven sets of measures. Specifically:

i. Word frequency
ii. Word range
iii. N-gram proportion
iv. N-gram frequency
v. N-gram range
vi. N-gram association measures
vii. Academic vocabulary

The following paragraphs describe these sets in more detail.

*Frequency measures:* For each analysed text, *TAALES* retrieves the frequency of each constituent word from a specified reference list. It then provides an average word frequency for the text. It offers a number of variations on this basic idea. First, reference frequencies can be taken from any of eleven lists: Kucera-Francis, Thorndike-Lorge (Thorndike & Lorge, 1944), Brown (1984), SUBTLEX-US, British National Corpus (BNC Consortium, 2007) (BNC, with separate counts available for written and spoken parts of the corpus) or Corpus of Contemporary American (COCA, with five lists, reflecting the five component genres of the corpus). Second, counts can be based on all words in the text, only on function words or only on content words. Finally, raw count or logarithmic transformations of those can be used.

*Word Range measures:* Range refers to the number of different texts in a reference corpus in which a word appears and is intended to capture how widely that word is used. This can be found based on the Kucera-Francis, SUBTLEX-US, BNC (spoken and written), and

COCA (again divided into five registers) corpora, and again can be found for all words, content words, or function words, and can be provided in its raw form or as a logarithmic transformation.

*N-gram proportion, frequency and association measures:* N-gram indices break each text down into two- or three-word chunks (e.g. the beginning of the previous paragraph could be broken into the two-word chunks *Range refers*; *refers to*; *to the*; *the number*, etc. or into the three-word chunks *Range refers to; refers to the; to the number*, etc.). *TAALES* determines how frequently each chunk appears in any of the BNC written and spoken corpora or the five register components of COCA and provides the average bi- or trigram frequency for the text as a whole. It also provides a logarithmic transformation of these figures, a range count, and a number of 'proportion' scores, showing what percentage of bi/trigrams in a text are amongst the most frequent N-thousand chunks in the reference corpus.

In addition to frequency, bi/trigrams can also be quantified using association measures. The general logic behind association scores is that some word combinations tend to occur with high frequency simply because their component words are very frequent, rather than because there is a phraseologically-interesting relationship between them (compare the very high-frequency bigrams *in the* and *but it* with the much less frequent *refuse to* and *well off*). Association measures aim to highlight combinations which reflect strong relationships between their component words, rather than simply a high frequency of occurrence. Many different measures have been proposed in the research in phraseology (Gries, 2013 provides a good survey of popular measures). In *TAALES*, five association measures are available: mutual information; mutual information squared; t-score; Delta-P; and collexeme strength.

*Measures of academic vocabulary and phraseology:* To measure the use of academic vocabulary *TAALES* quantifies the percentage of words in a text which can be found on Coxhead's (2000) Academic Word List. Coxhead's list is divided into eight sub-lists of descending frequency. *TAALES* enables both an overall count of AWL words and separate counts for each sub-list. As a measure of academic phraseology, *TAALES* also provides a count of phrases taken from Simpson-Vlach & Ellis's (2010) Academic Formulas List. Separate counts are available for formulas which are characteristic of written and of spoken texts and those which are common to both.

Each of these main headings include measures which are of potential educational interest. A key decision for the analyst is which, of the many options available, should be included in an analysis. A common approach with data of this sort has been to allow algorithms to decide, putting all indices into the initial analysis and seeing which indices make the most statistically robust predictions. The strongest predictors are then usually entered into a regression model to determine the optimal combination of items (Crossley et al., 2010; Guo et al., 2013; Kyle & Crossley, 2016). A recent variation on this approach can be found in M. Kim, et al. (2018), who first enter all *TAALES* variables which are normally-distributed and which do not show multicollinearity into a principal components analysis (PCA) before performing regression analyses with these components as predictors. While analyses of this sort are variable, for the present study we have taken a different approach for two reasons. First, by filtering out all but the strongest predictors, these studies (with the exception of M. Kim et al., 2018) provide information about only a small portion of indices, not giving a sense of how the wider set of measures relate either to each other or to the outcome variables. Second, the atheoretical nature of this approach, in which indices are grouped on purely statistical terms, can result in components and models which are not easily interpretable. When the aim of the research is to inform computational models for the sake of (for example) predicting grades, these issues are not particularly important. However, when the aim – as in the present study – is to increase our understanding of the construct of vocabulary sophistication, it is problematic. The present research therefore takes a different approach.

Our starting point is the seven categories, outlined above, which we take to be conceptually-distinct aspects of the construct of vocabulary sophistication. While there are likely to be overlaps between these seven categories (to which we will return below), each clearly comes at vocabulary from a different angle and so can be usefully studied independently. Within each category, we work on the hypothesis that many of the measures provided by *TAALES* are likely to be variations on the same construct. For example, the word frequency measures provided by different corpora are all measures of word frequency. We therefore start by determining the extent to which the variations on each measure type overlap with each other. In order to do this, we use the following general procedure:

i. Results for each index are converted to z-scores to overcome the issue that different measures often use very different scales.
ii. It is hypothesised that indices operationalizing the same measure with different corpora are essentially measuring the same thing. This hypothesis is both intuitively plausible and has received partial support from M. Kim et al.'s (2018) PCA, in which indices based on different reference corpora tended to load together. Overlaps between these variants are quantified using inter-item correlations, summarised for the group using Cronbach's alpha.

iii. Where the variants of a measure are found to strongly correlate, they are combined into a single composite scale.

iv. Inter-item correlations are determined between the various scales which make up each of the seven main categories described above. Where two scales are strongly correlated, only one is retained for the main analysis. Where scales appear to give different information, both scales are retained.

v. The resulting sets of measures are then correlated with both the final grade and style grade awarded to the text by the class instructor. Because criteria of quality might vary from project to project, this final step is performed separately for each of the six projects in the corpus. Because grades are not always normally distributed, and because it is not clear that they are on an interval scale, spearman correlations are used. Style grades had not been awarded for Project 1 so are included only for the other five projects.

Steps i.-iv. show, for each of the seven distinct measure types described above, how the different variants of each measure relate to each other, which measures overlap and which are distinct (Research Question 1). Step v. shows how they relate to the variables of overall quality and style quality (Research Question 2).

As mentioned above, it is likely that there are correlations between, as well as within, the seven main categories. To understand these correlations, the key representative measures from each category are entered into a principal components analysis. This provides a broader picture of the construct of vocabulary sophistication as a whole, enabling us to elaborate further our answer to Research Question 1.

## 4. Findings

In what follows, we first (Section 4.1) look separately at results for each of the seven main sets of measures described in Section 3.3. In Section 4.2, we consider how the seven sets of measures relate to each other, using a principal components analysis to determine the relationships between the composite variables determined in Section 4.1.

### **4.1** Analysis by category

For each of the categories, we first evaluate the relationships between the various measures within each set and whether/how they can be reduced to a smaller number of composite variables. We then determine how these composite variables relate to essay grades.

### **4.1.1** *Word Frequency*

Analysis of frequency indices started from the hypothesis that indices based on different corpora are likely to be highly correlated with each other, giving slightly different operationalizations of the same construct. As Table 2 shows, this hypothesis was confirmed. Cronbach alphas for sets of indices based on different corpora ranged from .98 to 1. These sets were therefore combined into six composite scales. Table 3 shows the relationships between these scales.

**Table 2:** Reliability of frequency scales combining different reference corpora

| Scale | Alpha |
|---|---|
| Frequency: AW Raw | 1 |
| Frequency: AW Log | .98 |
| Frequency: CW Raw | .99 |
| Frequency: CW Log | .98 |
| Frequency: FW Raw | 1 |
| Frequency: FW Log | .99 |

**Table 3:** Correlations between composite frequency scales

| | Frequency: AW Raw | Frequency: AW Log | Frequency: CW Raw | Frequency: CW Log | Frequency: FW Raw | Frequency: FW Log |
|---|---|---|---|---|---|---|
| Frequency: AW Raw | 1 | | | | | |
| Frequency: AW Log | 0.35 | 1 | | | | |
| Frequency: CW Raw | -0.03 | **0.69** | 1 | | | |
| Frequency: CW Log | -0.19 | **0.79** | **0.78** | 1 | | |
| Frequency: FW Raw | **0.86** | -0.01 | -0.31 | -0.42 | 1 | |
| Frequency: FW Log | **0.75** | -0.02 | -0.3 | -0.4 | **0.91** | 1 |

Table 3 brings out three important points: (i) for both content words (CW) and function words (FW), raw and log-transformed indices are strongly positively correlated. (ii) Both sets of CW indices are moderately negatively correlated with both sets of FW indices. (iii) The all word

(AW) raw index is strongly positively correlated with FW indices; the AW log index is strongly positively correlated with CW indices.

These findings suggest a number of conclusions. Firstly, CW and FW indices offer distinct information from each other and tend to work in opposite directions. Secondly, raw and log-transformed versions of the CW and FW indices offer very similar information to each other. Thirdly, the raw AW index is principally determined by use of function words. This is likely to be because a small number of function words have extremely high token frequencies which will strongly influence any correlations. Fourthly, the log transformed AW index is principally a reflection of content words. On this index, the extreme frequencies of the function words have been neutralized and the higher type frequency of context words enables them to dominate the index. Finally, given the distinctive behaviour of content and function words, and the ambiguous way these are conflated in the AW indices, the latter appear to be of little value.

For these reasons, the following analysis of quality focuses only on CW and FW measures. Since log-transformed indices help neutralize the effects of skew in frequency counts, they are more likely to provide statistically meaningful results, so only these versions of the indices are used. Correlations between these indices and quality ratings are shown in Table 4.

**Table 4:** Spearman correlations between composite log frequency scales and grades

| | | **1101** | | | **1102** | | | **Mean** |
|---|---|---|---|---|---|---|---|---|
| | | **Project 1** | **Project 2** | **Project 3** | **Project 1** | **Project 2** | **Project 3** | ***r*** |
| CW Log | Final grade | -.17*** | -.19 *** | -.20 *** | -.18 *** | -.20 *** | -.10 *** | -.17 |
| | Style grade | NA | -.14*** | -.18*** | -.11*** | -.17*** | -.08** | -.14 |
| FW Log | Final grade | .09** | .03 | .04 | -.01 | .13*** | .01 | .05 |
| | Style grade | NA | -.03 | .04 | -.08** | .14*** | .03 | .02 |

**NOTE:** *** p < .0005; ** p < .005; * p < .05

Mean CW frequency shows a weak but consistent negative correlation with quality ratings: lower-frequency words are found in more highly-rated texts. Mean FW frequency did not show a consistent correlation. Small positive correlations were found for two projects, but the lack of consistency suggests that this relationship is not robust. Strikingly, content word frequency was less strongly related to style grades than to final grades. Given that the style grade is intended to reflect language use only, whereas final grade combines language use with a range of other factors, this is an unexpected finding. These lower correlations are most likely due to the narrower range of scores which were awarded under this category (see Table 1).

### 4.1.2 *Word Range*

As with frequency, the starting assumption was that range counts from different corpora would be highly correlated. This was again confirmed, the sole exception being the KFNCats version of the Raw FW index. Cronbach alphas for sets of indices based on different corpora ranged from .97 to .99 (see Table 5). Composite scales were again created to combine counts from different corpora. Correlations between the six composite scales are shown in Table 6.

**Table 5:** Reliability of range scales combining different reference corpora

| Scale | Alpha |
|---|---|
| Range: AW Raw | .99 |
| Range: AW Log | .97 |
| Range: CW Raw | .99 |
| Range: CW Log | .97 |
| Range: FW Raw | .97* |
| Range: FW Log | .96 |

**NOTE:** * KFNCats FW excluded (corrected item-total correlation = .43)

**Table 6:** Correlations between composite range scales

| | **Range: AW Raw** | **Range: AW Log** | **Range: CW Raw** | **Range: CW Log** | **Range: FW Raw** | **Range: FW Log** |
|---|---|---|---|---|---|---|
| **Range: AW Raw** | 1 | | | | | |
| **Range: AW Log** | **0.91** | 1 | | | | |
| **Range: CW Raw** | **0.89** | **0.89** | 1 | | | |
| **Range: CW Log** | **0.8** | **0.95** | **0.91** | 1 | | |
| **Range: FW Raw** | -0.1 | -0.15 | -0.26 | -0.21 | 1 | |
| **Range: FW Log** | -0.01 | -0.05 | -0.14 | -0.11 | **0.95** | 1 |

As with frequency, the CW and FW indices offer distinct information, while their respective raw and log-transformed versions are highly correlated with each other. The AW range measures are very similar to the CW measures so do not seem to add useful information. As with frequency, and for the same reasons, the log-transformed CW and FW indices were retained for the correlation with quality (see Table 7).

**Table 7:** Spearman correlations between composite range scales and grades

| | | 1101 | | | 1102 | | | Mean |
|---|---|---|---|---|---|---|---|---|
| | | Project 1 | Project 2 | Project 3 | Project 1 | Project 2 | Project 3 | *r* |
| Frequency: CW Log | Final grade | -.16*** | -.17*** | -.19*** | -.18*** | -.18*** | -.11*** | -.16 |
| | Style grade | NA | -.13*** | -.17*** | -.12*** | -.16*** | -.08** | -.13 |
| Frequency: FW Log | Final grade | .01 | -.04 | .00 | -.05* | .03 | .05* | -.02 |
| | Style grade | NA | -.10** | .01 | -.10*** | .03 | -.04 | -.04 |

**NOTE:** *** p < .0005; ** p < .005; * p< .05

As with frequency, the CW measure shows a reliable negative correlation with quality ratings (i.e. words attested in a narrow range of texts in the reference corpora are associated with higher grades), while FW measures show only a weak and inconsistent relationship. For the CW measure, correlations with the final grade were again stronger than those with the style grade. This pattern was reversed for the FW measure, but the weak and inconsistent nature of this relationship suggests we should treat the difference with caution.

### **4.1.3** *N-gram measures*

N-gram frequency measures comprise two rather different types:

i. proportion measures, which show what percentage of n-grams in a text are attested in a reference corpus;
ii. mean frequency measures, which show the mean reference corpus frequency of n-grams.

The latter measure is dependent on, but distinct from, the former in that only n-grams which are found in the reference corpus are included in the mean frequency measure. A text which has a low proportion score (because only a small number of its n-grams are attested in the target corpus) can have either a high or a low frequency score, as the n-grams which are attested could come from any frequency level.

Analysis of proportion measures started from the hypothesis that counts based on different corpora and different frequency bands of those corpora would be highly correlated – both hypotheses being supported by M. Kim et al.'s (2018) PCA. This was confirmed, with both bigram and trigram proportion measures showing a Cronbach alpha of 1. These measures were therefore combined into overall bigram and trigram proportion measures. These two measures in turn were highly correlated (*r* = .88, *p*< .0001). Since it is unclear which, if either, of these two measures should be given priority, both were included in the analysis. Their correlations with final grade and style grade are shown in Table 8. Both measures show consistent negative correlations with grades, implying that higher-scoring papers used fewer n-grams which were found in the reference corpora. As with the previous measures, correlations were lower for the style grade than for the final grade. They were also higher for bigram proportions than for trigram proportions.

**Table 8:** Spearman correlations between composite n-gram proportion scales and grades

| | | **1101** | | | **1102** | | | **Mean** |
|---|---|---|---|---|---|---|---|---|
| | | **Project 1** | **Project 2** | **Project 3** | **Project 1** | **Project 2** | **Project 3** | ***r*** |
| Bigram Proportion | Final grade | -.11*** | -.21*** | -.21*** | -.19*** | -.11*** | -.09*** | -.15 |
| | Style grade | NA | -.15*** | -.18*** | -.13*** | -.07* | -.04 | -.12 |
| Trigram Proportion | Final grade | -.07** | -.17*** | -.17*** | -.13*** | -.08*** | -.07** | -.12 |
| | Style grade | NA | -.09** | -.15*** | -.06** | -.04 | -.03 | -.07 |

**NOTE:** *** p < .0005; ** p < .005; * p< .05

Moving on to the mean frequency of items which were attested in the reference corpora, as with previous categories, indices based on different corpora were highly correlated (see Table 9) and so each combined into composite scales. Again, raw and log-transformed versions of each scale were highly correlated, so only bigram log frequency and trigram log frequency were retained for the comparison with grades (Table 10).

**Table 9:** Reliability of n-gram frequency scales combining different reference corpora

| **Scale** | **Alpha** |
|---|---|
| Bigram/Count | 1 |
| Bigram/Log | .97 |
| Trigram/Count | .95 |
| Trigram/Log | .94 |

**Table 10:** Correlations between composite n-gram frequency scales

| | Bigram Count | Bigram Log | Trigram Count | Trigram Log |
|---|---|---|---|---|
| Bigram Count | 1 | | | |
| Bigram Log | **0.51** | 1 | | |
| Trigram Count | 0.23 | 0.24 | 1 | |
| Trigram Log | 0.25 | 0.41 | **0.79** | 1 |

Correlations here are weak and inconsistent. The strongest and most robust measure is bigram log frequency. As with previous measures, correlations with style grades were lower than those with final grades.

**Table 11:** Spearman correlations between composite n-gram frequency scales and grades

| | | **1101** | | | **1102** | | | **Mean** |
|---|---|---|---|---|---|---|---|---|
| | | **Project 1** | **Project 2** | **Project 3** | **Project 1** | **Project 2** | **Project 3** | ***r*** |
| Bigram log frequency | Final grade | -.08** | -.15*** | -.16*** | -.10*** | -.02 | -.08** | -.10 |
| | Style grade | NA | -.08* | -.14*** | -.09*** | .00 | -.03 | -.07 |
| Trigram log frequency | Final grade | .01 | -.08** | -.07* | .01 | -.04 | -.04 | -.03 |
| | Style grade | NA | -.04 | -.08* | .04 | -.03 | -.01 | -.03 |

**Note:** *** p < .0005; ** p < .005; * p< .05

N-gram range followed a similar pattern to n-gram frequency. As before, measures drawn from different corpora were highly correlated and so were combined into scales (Table 12). Again, raw and log-transformed versions were highly correlated. Only a weak to moderate correlation was found between bigram and trigram counts (Table 13). Bigram and trigram log frequencies were retained for the correlation with grades (Table 14). As with frequency, correlations are weak and inconsistent, with the strongest and most robust correlations being for bigram frequency. Again, correlations with style grades were lower than those with final grades.

**Table 12:** Reliability of n-gram range scales combining different reference corpora

| **Scale** | **alpha** |
|---|---|

| | |
|---|---|
| Bigram/Count | .98 |
| Bigram/Log | .96 |
| Trigram/Count | .94 |
| Trigram/Log | .92 |

**Table 13:** Correlations between composite n-gram range scales

| | **Bigram/Count** | **Bigram/Log** | **Trigram/Count** | **Trigram/Log** |
|---|---|---|---|---|
| Bigram/Count | 1 | | | |
| Bigram/Log | **0.73** | 1 | | |
| Trigram/Count | 0.31 | 0.23 | 1 | |
| Trigram/Log | 0.39 | 0.42 | **0.82** | 1 |

**Table 14:** Spearman correlations between composite n-gram range scales and grades

| | | **1101** | | | **1102** | | | **Mean** |
|---|---|---|---|---|---|---|---|---|
| | | **Project 1** | **Project 2** | **Project 3** | **Project 1** | **Project 2** | **Project 3** | ***r*** |
| CW Log | Final grade | -.10*** | -.17*** | -.17*** | -.11*** | -.04 | -.07** | -.11 |
| | Style grade | NA | -.09** | -.15*** | -.08** | -.03 | -.04 | -.08 |
| FW Log | Final grade | -.01 | -.09** | -.07* | .01 | -.04 | -.04 | -.04 |
| | Style grade | NA | -.05 | -.08* | .05 | -.05 | -.03 | -.03 |

**Note:** *** p < .0005; ** p < .005; * p< .05

As with previous measures, the analysis of association measures combined indices from different reference corpora. It also combined trigram-1 (the association from first word to following bigram) with trigram-2 (the association from initial bigram to following word). Alphas for these composite measures are shown in Table 15. The correlations between these measures are shown in Table 16. MI2 strongly correlates with both MI and t-score, though these latter two measures are relatively independent of each other. For this reason, MI2 is discarded from further analysis. Collexeme strength is also strongly correlated with t-score. Since the latter is more widely used in the literature, it was retained and collexeme strength excluded from further analysis. Bigram-based measures correlate either moderately (t-score) or strongly (MI, DP) with their trigram counterparts. As with previous analyses, both sets of measures will be retained.

**Table 15:** Reliability of n-gram association measure scales combining different reference corpora

| **Description** | **alpha** |
|---|---|
| Bigram/MI | .96 |

| | |
|---|---|
| Bigram/MI2 | .96 |
| Bigram/T | .98 |
| Bigram/DP | .98 |
| Bigram/Coll | .99 |
| Tri/MI | .96 |
| Tri/MI2 | .97 |
| Tri/T | .96 |
| Tri/DP | .93 |
| Tri/Coll | .97 |

**Table 16:** Correlations between composite association measure indices

| | **Bigram** | | | | | **Trigram** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **MI** | **MI2** | **T** | **DP** | **Coll** | **MI** | **MI2** | **T** | **DP** | **Coll** |
| Bigram/MI | 1 | | | | | | | | | |
| Bigram/MI2 | 0.55 | 1 | | | | | | | | |
| Bigram/T | 0.12 | 0.56 | 1 | | | | | | | |
| Bigram/DP | 0.49 | 0.32 | 0.24 | 1 | | | | | | |
| Bigram/Coll | -0.13 | 0.31 | 0.69 | 0.3 | 1 | | | | | |
| Tri/MI | 0.67 | 0.15 | -0.16 | 0.32 | -0.35 | 1 | | | | |
| Tri/MI2 | 0.62 | 0.36 | 0.08 | 0.39 | -0.12 | 0.83 | 1 | | | |
| Tri/T | 0.31 | 0.37 | 0.38 | 0.29 | 0.16 | 0.32 | 0.69 | 1 | | |
| Tri/DP | 0.29 | 0.00 | 0.11 | 0.71 | 0.28 | 0.36 | 0.41 | 0.3 | 1 | |
| Tri/Coll | 0.23 | 0.33 | 0.31 | 0.28 | 0.25 | 0.2 | 0.55 | 0.86 | 0.3 | 1 |

Correlations between grades and MI, t-score and DP measures for both bigrams and trigrams are shown in Table 17. For both bigrams and trigrams, DP has the strongest and most consistent relationship with grades, showing a weak but robust positive correlation. The difference between the correlation with overall and style grades is negligible. MI shows a weaker and somewhat inconsistent positive correlation. T-score is almost entirely independent of grades.

**Table 17:** Spearman correlations between n-gram association measures scales and grades

| | | | **1101** | | | **1102** | | | **Mean** |
|---|---|---|---|---|---|---|---|---|---|
| | | | **Project 1** | **Project 2** | **Project 3** | **Project 1** | **Project 2** | **Project 3** | ***r*** |
| Bigram | MI | Final grade | .13*** | .13*** | .10** | .09*** | -.07* | .02 | .07 |
| | | Style grade | - | .14*** | .09** | .09*** | -.06* | .01 | .05 |
| | t-score | Final grade | .02 | .00 | -.07* | -.04 | .05* | -.06* | -.02 |
| | | Style grade | - | .01 | -.06 | -.04 | .04 | -.06* | -.02 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | DP | Final grade | .20*** | .17*** | .14*** | .13*** | .11*** | .08** | .14 |
| | | Style grade | - | .20*** | .13*** | .13*** | .11*** | .08** | .13 |
| Trigram | MI | Final grade | .15*** | .09** | .11*** | .11*** | -.07* | .05 | .07 |
| | | Style grade | - | .10** | .11** | .12*** | -.06* | .05* | .07 |
| | t-score | Final grade | .07* | -.05 | -.03 | .04 | -.04 | .00 | .00 |
| | | Style grade | - | -.02 | -.03 | .04 | -.04 | .01 | -.01 |
| | DP | Final grade | .17*** | .15*** | .17*** | .15*** | .13*** | .06* | .14 |
| | | Style grade | - | .16*** | .17*** | .14*** | .13*** | .06* | .13 |

**NOTE:** *** p < .0005; ** p < .005; * p< .05

### **4.1.4** *Academic Language*

Table 18 shows the relationships between use of the AWL and the three sub-lists of the AFL. The weak correlations suggest that each index gives distinct information, so the relationships between these and grades were calculated separately. As Table 19 shows, use of the AWL is positively correlated with grades, but there is a striking difference between ENC1101 and ENC1102, with the correlation becoming much weaker (and for project 3, disappearing entirely) in the latter module. Use of formulas from the core and written components of the AFL are unrelated to grades, while use of formulas from the spoken AFL show a weak negative correlation.

**Table 18:** Correlations between academic language indices

| | **All AWL** | **Core AFL** | **Spoken AFL** | **Written AFL** |
|---|---|---|---|---|
| All AWL | 1 | | | |
| Core AFL | 0.13 | 1 | | |
| Spoken AFL | -0.21 | -0.01 | 1 | |
| Written AFL | 0.01 | 0.16 | 0.05 | 1 |

**Table 19:** Spearman correlations between academic language indices and grades

| | | **1101** | | | **1102** | | | **Mean** |
|---|---|---|---|---|---|---|---|---|
| | | **Project 1** | **Project 2** | **Project 3** | **Project 1** | **Project 2** | **Project 3** | ***r*** |
| AWL | Final grade | .18*** | .19*** | .16*** | .06* | .06* | .03 | .11 |
| | Style grade | - | .10** | .15*** | .00 | .00 | .00 | .05 |
| | Final grade | .07* | .01 | .01 | .02 | .04 | -.01 | .02 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| AFL Core | Style grade | - | .03 | .00 | .03 | .00 | -.01 | .01 |
| AFL Written | Final grade | .07* | .02 | .05 | .06* | .01 | .02 | .04 |
| | Style grade | - | .01 | .01 | .07** | .01 | .01 | .02 |
| AFL Spoken | Final grade | -.04 | -.07* | -.13*** | -.03 | -.09*** | -.03 | -07 |
| | Style grade | - | -.04 | -.13*** | -.01 | -.09*** | -.01 | -.06 |

**NOTE:** *** p < .0005; ** p < .005; * p< .05

## **4.2** Principal Components Analysis

While the conceptual distinctions between the above categories means that it is useful to consider each separately, it is also worth considering how they relate to each other. To help understand this, the twenty indices selected in the above analyses were entered into a PCA. To avoid multicollinearity, an initial correlation analysis of all measures was conducted and, where indices correlated at more than *r*=.75, the index which was considered less conceptually fundamental was excluded. "Conceptual fundamentalness" followed the order of presentation of categories above. Specifically: single-word measures were considered more fundamental than n-gram measures; frequency measures were considered more fundamental than range measures and association measures. This led to five deletions being made:

i. CW log range (correlates with CW log frequency, *r*=.97)
ii. Bigram range (correlates with bigram log frequency, *r*=.95)
iii. Trigram range (correlates with trigram log frequency, *r*=.97)
iv. Bigram proportion (correlates with CW log frequency, *r*=.79)
v. Trigram t-score (correlates with trigram log frequency, *r*=.83)

The remaining 15 measures were entered into a PCA. Based on consultation of both eigenvalues (values greater than .94 were retained) and the scree plot, a four-component solution was selected and varimax rotation used to strengthen loadings. The resulting model accounted for 63% of the variance in the data and demonstrated a goodness of fit, based on off-diagonal values of .93. The four components are shown in Table 20. Loadings less than .40 are deleted. Piloting showed this cut-off to provide the most easily-interpretable output. Only the strongest loading of each measure is shown.

**Table 20:** Principal Components Analysis

| | C1 | C2 | C3 | C4 |
|---|---|---|---|---|
| CW log frequency | -.87 | | | |
| Trigram DP | .78 | | | |
| Bigram DP | .69 | | | |
| AWL | .67 | | | |
| AFL spoken | -.45 | | | |
| Bigram t-score | | .77 | | |
| Bigram log frequency | | .73 | | |
| Trigram log frequency | | .64 | | |
| Trigram proportion | | .63 | | |
| Trigram MI | | | .85 | |
| Bigram MI | | | .78 | |
| FW log frequency | | | -.64 | |
| FW range | | | -.57 | |
| AFL written | | | | .71 |
| AFL core | | | | .70 |

**NOTE:** Goodness of fit =.93

Component 1 (which accounts for 23% of the variance) combines positive loadings for both n-gram DP measures and use of AWL with negative loadings for CW log frequency and use of AFL spoken formulas. It will be remembered that the former set of measures correlated positively with scores, while the latter set correlated negatively. This component can be interpreted as contrasting everyday vocabulary (high frequency words, spoken formulas) with vocabulary which is lower in frequency (high DP is associated with one part of the combination being rare) and/or register-specific (AWL items). Of the 15 measures in the analysis, these four have the strongest correlations with quality. Component 2 (accounting for 17% of variance) combines all of the measures associated with the use of attested and high-frequency n-grams. As was discussed in Section 2, use of high-frequency n-grams is associated with lower levels of writing, and measures on this component either show a negligible or negative correlation with quality. Component 3 (accounting for 16% of variance) combines positive loadings for n-gram MI measures with negative loadings for FW frequency and range. The relationship between these measures is perhaps due to higher MI scores being associated with use of lower frequency words. There is little relationship between these and quality. Component 4 (accounting for 9% of variance) combines AFL written and core lists, suggesting that these are a somewhat distinct construct. As we have seen above, these also have little relationship with grades.

## 5. Discussion

The above analysis allows us to draw a number of conclusions. First, indices based on different corpora tend to be very highly correlated. For the purpose of characterizing the lexical sophistication of learner texts using these indices, therefore, choice of corpus does not appear to be an important variable. As noted above, M. Kim et al.'s (2018) analysis seems to point to a similar conclusion. This is perhaps surprising, given previous findings that, for example, correlations between the frequencies/association measures of collocations and learner knowledge of those collocations can differ dramatically depending on the corpus used (Durrant, 2014). Two factors may have rendered the current indices less susceptible to corpus effects. First, the positionally-variable two-word combinations within a span that were studied in Durrant (2014) are likely to be more contextually variable than the measures used here. It is notable that the lowest levels of reliability between corpora were found for the more contextually-specific trigram-based measures. Second, Durrant's findings were based on relatively small numbers of selected collocations (items included in tests). It may be that corpus effects are washed out when measures are built on comprehensive lists of all items in a text.

A second implication of our analysis is that measures of word frequency and range which are based on content words behave rather differently from those based on function words. These appear to represent different constructs, which are not strongly correlated with each other and which differ from each other in their correlations with text quality ratings. Frequency and range counts which are based on all words appear to represent a conflation of these two constructs and are consequently not likely to be a useful measure of sophistication. This conflation is seen most clearly in the fact that raw counts for all word measures correlate with counts for function word measures, whereas log-transformed counts correlate with counts for content word measures.

Turning to phrasal measures, it appears that indices based on the percentage of n-grams in a text which are found within the first N% of a reference corpus do not differ much as N increases. Moreover, bigram and trigram measures tend to be strongly correlated with each other. Both of these findings mirror the findings of M. Kim et al. (2018). The large range of indices offered by TAALES for ngram proportion, frequency and range can therefore be safely reduced to one measure each for each category, with little loss of information.

With regard to association measures, MI and t-score – as has been frequently pointed out (Durrant & Schmitt, 2009) – tap clearly distinct constructs. MI2 appears to be intermediate between these, correlating with both. Collexeme strength is also strongly correlated with t-score, suggesting this does not add useful additional information. DP, on the other hand, is distinct from the other measures and so likely to be of analytical use.

The PCA allowed us to summarise relationships across, as well as within, categories of indices. It suggested that the 20 indices retained in our analyses could be effectively summarised by four components:

i. Use of everyday vocabulary vs. lower frequency/register-specific vocabulary
ii. Use of high-frequency/attested n-grams
iii. Use of strongly-associated n-grams (which combines with low-frequency function words as high-frequency function words do not enter into such combinations)
iv. Use of items from the *Academic Formulas List*

As well as understanding the relationships between measures, we have also attempted to evaluate the extent to which measures genuinely capture something about vocabulary sophistication by correlating them with the scores assigned to texts. Table 21 summarises these relationships, showing the mean correlation across six projects for each of our 20 main measures.

**Table 21:** Summary of correlations between indices and grades

| **Measure** | **Final grade *r*** | **Style grade *r*** |
|---|---|---|
| CW log frequency | -.17 | -.14 |
| FW log frequency | .05 | .02 |
| CW log range | -.16 | -.13 |
| FW log range | -.02 | -.04 |
| Bigram proportion | -.15 | -.12 |
| Trigram proportion | -.12 | -.07 |
| Bigram log frequency | -.10 | -.07 |
| Trigram log frequency | -.03 | -.03 |
| Bigram MI | .07 | .05 |
| Bigram t-score | -.02 | -.02 |
| Bigram DP | .14 | .13 |
| Trigram MI | .07 | .07 |
| Trigram t-score | .00 | -.01 |
| Trigram DP | .14 | .13 |
| AWL | .11 | .05 |
| AFL core | .02 | .01 |
| AFL written | .04 | .02 |
| AFL spoken | -.07 | -.06 |

Overall, indices of lexical sophistication are both weak and highly variable predictors of grades. This is perhaps unsurprising in a context which does not explicitly focus on vocabulary use and in which grading reliability may not be high. In spite of this, however, eight indices consistently correlated with final grades with an *r* of at least ±.1:

*CW log frequency and range*, suggesting that less frequent, less widely-used words are associated with higher scores. This is in line with the intuitive understanding of sophistication, as described by Read (2000) and with the empirical findings of M. Kim et al. (2018). Because of the modest size of the correlations, it is important not to overstate any conclusions that can be drawn from this. However, they suggest that current rubrics and course aims – neither of which refer to vocabulary use – may not fully capture what tutors value in their students' writing.

*Bigram and trigram proportion and bigram frequency*, suggesting that less well-attested n-grams are associated with higher scores. This accords with the negative correlations which Crossley et al. (2012) found between bigram proportion and scores awarded to SAT-like essays written by L1 college students. There is an interesting contrast here with findings from the L2 literature that more proficient second language writers of English use *more* n-grams which are attested in a reference corpus (Bestgen & Granger, 2014; Kyle & Crossley, 2015; 2016; M. Kim et al., 2018). It seems that this apparent indicator of success in second language writing is an indicator of lack of success in the context of mainstream US university classes. A plausible interpretation of these conflicting findings is that the variance in L2 writers' use of unattested n-grams is driven by linguistic inaccuracies or infelicities (which are likely to be disvalued), whereas the variance in writers' use of unattested n-grams in mainstream contexts is driven by originality of thought or expression (which is likely to be valued). In a brief review of the unattested bigrams found in their L2 corpus, Bestgen & Granger (2014) found that around two-thirds were grammatically incorrect pairings. Further research will be required to substantiate this in detail, but it draws attention to the possibility that our relatively abstract indices of vocabulary use may be tapping very different constructs in different contexts.

*Bigram and trigram DP*, suggesting that the use of more closely-associated word pairs is correlated with higher scores. The finding that this outperforms other association measures as a predictor of quality ratings accords with M. Kim et al.'s (2018) finding for L2 writing. DP is a relatively recent addition to the range of association measures used to study phraseology (Gries, 2013). What distinguishes it from other measures is its directionality – i.e. it takes account of the fact that the association from the first word to the second may by stronger or weaker than the association from the second word to the first. Why this measure should be more closely linked to scores than bidirectional measures such as mutual information is unclear and deserves further investigation.

*Academic Word List*, suggesting that use of items from this generic academic vocabulary are associated with higher scores. Given the university context in which these texts were written, it is unsurprising that use of academic words should be associated with higher scores. It is striking, however, that these words were far less predictive of grades in the ENC1102 program than in ENC1101. This suggests that, even within a university context, the value of items from the AWL can be course-specific.

As Table 22 summarizes, there is wide variation across the six projects in the strength of these relationships, especially, as noted above, in the case of AWL use. It is striking that such differing results are found within the boundaries of one pair of classes, offered by a single department to the same group of students. In general, correlations were weaker for the second semester, ENC1102 class, and particularly in the second and third projects of this class. It seems that, whatever these indices of vocabulary use are tapping, its relationship with grades weakened through the course of the academic year. Whether this shift is due to students' development, changes or inconsistencies in grading practices or in the types of writing done, or is merely a random fluctuation in already weak correlations, is unclear. More research will therefore be needed to understand the contextual-dependency of these correlations.

**Table 22:** Summary of consistently correlating indices across projects – final grades only

| | **ENC1101** | | | **ENC1102** | | |
|---|---|---|---|---|---|---|
| | **Project 1** | **Project 2** | **Project 3** | **Project 1** | **Project 2** | **Project 3** |
| CW log frequency | -.17 | -.19 | -.20 | -.18 | -.20 | -.10 |
| CW log range | -.16 | -.17 | -.19 | -.18 | -.18 | -.11 |
| Bigram proportion | -.11 | -.21 | -.21 | -.19 | -.11 | -.09 |
| Trigram proportion | -.07 | -.17 | -.17 | -.13 | -.08 | -.07 |
| Bigram log frequency | -.08 | -.15 | -.16 | -.10 | -.02 | -.08 |
| Bigram DP | .20 | .17 | .14 | .13 | .11 | .08 |
| Trigram DP | .17 | .15 | .17 | .15 | .13 | .06 |
| AWL | .18 | .19 | .16 | .06 | .06 | .03 |
| **Mean absolute *r*** | **.14** | **.18** | **.18** | **.14** | **.11** | **.08** |

## 6. Conclusions

This paper has had two principal aims: (i) to increase our understanding of the various indices of vocabulary sophistication provided by *TAALES* and, by extension, of the various aspects of vocabulary sophistication in general; (ii) to understand the relationship between these indices and achievement in university composition tasks. With regard to the first aim, analysts of vocabulary sophistication should be aware of the following. Firstly, measures of mean word frequency, range, n-gram frequency and n-gram association differ little when different reference corpora are used. Use of multiple indices is therefore likely to be superfluous. Secondly, mean word frequency and range measures which do not distinguish content from function words conflate different constructs and are therefore probably not meaningful. Mean frequency statistics should, at minimum, distinguish content from function words. Further research is needed to determine the extent to which finer distinctions (e.g. individual parts of speech) is important. Thirdly, separate indices of n-grams attested in growing portions of a reference corpus do not offer distinct information from each other. Similarly, bigram and trigram proportion, range and frequency measures do not offer distinct information. Fourthly, to capture a range of distinctive information about n-gram association, three measures are optimal: t-score, mutual information and Delta-P. Finally, four components (described above) offer a good summary of the range of sophistication measures investigated in this study.

With regard to the second aim, we have seen that the lexical measures tested here correlate only weakly with scores, but that correlations are reliable for certain measures. Specifically, higher grades are associated with the use of: lower-frequency content words that are found in a narrow range of texts; unattested n-grams; n-grams with high directional association; use of academic words is also associated with higher grades, but this effect differs markedly between the two courses. Pedagogical conclusions based on these findings can at present only be tentative, and further research is needed to confirm and further clarify the patterns. However, they suggest that tutors' grading may reflect constructs which are not explicitly acknowledged as course aims or in grading rubrics and may shift in unacknowledged ways within programs of study. As set out in Section 3.1, these courses are run with high levels of professionalism and are regarded as excellent examples of their kind. This suggests that such potentially "hidden" aspects of what is rewarded in first-year composition will not be a peculiar feature of courses at this institution. More research is therefore required in similar contexts to understand more fully the relationship between the linguistic correlates of grades and what courses claim to teach and evaluate.

## Notes

**1.** Centre for Lexical Information. The CELEX database provides word-frequency information based on the Cobuild corpus (Burnage, 1990).

**2.** COCA is a large corpus of speech and writing in contemporary American English (Davies, 2008-).

**3.** SUBTLEX-US frequency data are based on a corpus of subtitles from US films and television series (Brysbaert & New, 2009).

**4.** This, and all following corpus and statistical analyses were performed using *R* (R Development Core Team, 2013).

## References

Berman, R. A., & Nir, B. (2010). The lexicon in writing-speech-differentiation. *Written Language and Literacy*, *13*(2), 183-205.

Berman, R. A., & Nir-Sagiv, B. (2007). Comparing narrative and expository text construction across adolescence: A developmental paradox. *Discourse Processes*, *43*(2), 79-120.

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, *69*, 65-78.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28-41.

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

BNC Consortium. (2007). British National Corpus, version 3 (BNC XML ed.). Retrieved from http://www.natcorp.ox.ac.uk (Last accessed February 2019).

Brown, G.D.A. (1984). A frequency count of 190,000 words in the London-Lund corpus of English conversation. *Behavior Research Methods, Instrumentation & Computers*, *16*, 502-532.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977-990.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of second language writing*, *26*, 42-65.

Burnage, G. (1990). *CELEX: A guide for users*. Nijmegen: CELEX - Centre for Lexical Information.

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

Crossley, S. A., Cai, Z., & McNamara, D. (2012). Syntagmatic, Paradigmatic, and Automatic N-gram Approaches to Assessing Essay Quality. In G.M.Youngblood & P.M.McCarthy (Eds.), *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference* (pp. 214-219). Palo-Alto, CA: The AAAI Press.

Crossley, S. A., DeFore, C., Kyle, K., Dai, J., & McNamara, D. (2013). Paragraph specific n-gram approaches to automatically assessing essay quality. In S.K.D'Mello, R.A.Clavo & A.Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining* (pp.216-219). Heidelberg: Springer. Retrieved from http://www.educationaldatamining.org/EDM2013/papers/rn_paper_31.pdf (Last accessed February 2019)

Crossley, S. A., Salsbury, T., McNamara, D., & Jarvis, S. (2010). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561-580.

Crossley, S. A., Weston, J. L., Sullivan, S. T. M., & McNamara, D. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, *28*(3), 282-311.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, *10*(1), 5-43.

Daller, H., Turlik, J., & Weir, I. (2013). Vocabulary acquisition and the learning curve. In S. Jarvis & H. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures* (pp.185-215). Amsterdam/Philadelphia, PA: John Benjamins.

Davies, M. (2008-). *The Corpus of Contemporary American: 450 million words, 1990-present.* Retrieved from http://corpus.byu.edu/coca/ (last accessed February 2019).

Durrant, P. (2014). Corpus frequency and second language learners' knowledge of collocations. *International Journal of Corpus Linguistics*, *19*(4), 443-477.

Durrant, P., & Brenchley, M. (in press). Corpus research on the development of children's writing in L1 English. In A. Glaznieks, A. Abel, V. Lyding, & V. Nicolas (Eds.), *Corpora and Language in Use: Proceedings of the learner corpus research conference, 2017*. Louvain: Presses Universitaires de Louvain.

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International review of applied linguistics*, *47*(2), 157-177.

Garner, J., Crossley, S. A., & Kyle, K. (2018). Beginning and intermediate L2 writers' use of N-grams: An association measures study. *International Review of Applied Linguistics*. Advance online publication. https://doi.org/10.1515/iral-2017-0089

Golub, L. S., & Frederick, W., C. (1979). *Linguistic Structures in the discourse of fourth and sixth graders*. Madison, WI: Center for Cognitive Learning, The University of Wisconsin.

Graesser, A. C., McNamara, D., Louwerse, M. M., & Cai, Z. (2014). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, *36*(2), 193-202.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-ntive writers: A bigram-based study. *International Review of Applied Linguistics*, *52*(3), 229-252.

Gries, S. Th. (2013). 50-something years of work on collocations: What is or should be next... *International Journal of Corpus Linguistics*, *18*(1), 137-165.

Grobe, C. (1981). Syntactic maturity, mechanics, and vocabulary as predictors of quality ratings. *Research in the Teaching of English*, *15*(1), 75-85.

Guo, L., Crossley, S. A., & McNamara, D. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*, 218-238.

Hou, J., Verspoor, M., & Loerts, H. (2016). An exploratory study into the dynamics of Chinese L2 writing development. *Dutch Journal of Applied Linguistics*, *5*(1), 65-96.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, *12*, 377-403.

Kim, J.-Y. (2014). Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study. *English Teaching*, *69*(4), 27-51.

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

Knoch, U., Rouhshad, A., Oon, S. P., & Storch, N. (2015). What happens to ESL students' writing after three years of study at an English medium university? *Journal of second language writing*, *28*, 39-52.

Knoch, U., Rouhshad, A., & Storch, N. (2014). Does the writing of undergraduate ESL students develop after one year of study in an English-medium university? *Assessing Writing*, *21*, 1-17.

Kucera, H. & Francis, W. (1967). *Computational Analysis of Present-day American English*. Providence, RI: Brown University Press.

Kyle, K. (2017). Modelling quality in source-based texts. Retrieved from https://a4li.sri.com/archive/papers/Kyle_2017_Writing_Quality.pdf (last accessed February 2019).

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757-786.

Kyle, K., & Crossley, S. A. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12-24.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, *19*(1), 85-104.

Malvern, D., Richards, B. J., Chipere, N., & Duran, P. (2004). *Lexical Diversity and Language Development*. Basingstoke: Palgrave Macmillan.

Massey, A. J., & Elliott, G. L. (1996). *Aspects of Writing in 16+ English Examinations Between 1980 & 1994*. Cambridge: University of Cambridge Local Examinations Syndicate.

Massey, A. J., Elliott, G. L., & Johnson, N. K. (2005). *Variations in Aspects of Writing in 16+ English Examinations Between 1980 and 2004: Vocabulary, Spelling, Punctuation, Sentence Structure, Non-standard English*. Cambridge: Cambridge Assessment.

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing*, *29*, 3-15.

McCarthy, P. M., & Jarvis, S. (2011). MTLD, voc-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381-392.

Meurers, D., & Dickinson, M. (2017). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, *67:S1*, 66-95.

Moxley, J. (2013). Big data, learning analytics, and social assessment. *Journal of Writing Assessment*, *6*(1), 1-10.

Myhill, D. (1999). Writing matters: Linguistic characteristics of writing in GCSE English examinations. *English in Education*, *33*(3), 70-81.

Myhill, D. (2009). From talking to writing: Linguistic development in writing. *BJEP Monograph Series II, Number 6 - Teaching and Learning Writing*, *27*(44), 27-44.

Olinghouse, N. G., & Leaird, J. T. (2009). The relationship between measures of vocabulary and narrarive writing quality in second- and fourth-grade students. *Reading and Writing*, *22*, 545-565.

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. *Reading and Writing: An Interdisciplinary Journal*, *26*, 45-65.

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.

R Development Core Team. (2013). *R: A language and environment for statisticaly computing* (Version 1.0.136) [Computer software]. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R- project.org/ (last accessed February 2019).

Read, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

Roessingh, H., Elgie, S., & Kover, P. (2015). Using lexical profiling tools to investigage children's written vocabulary in grade 3: An exploratory study. *Language Assessment Quarterly*, *12*(1), 67-86.

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

Storch, N. (2009). The impact of studying in a second language (L2) medium university on the development of L2 writing. *Journal of Second Language Writing*, *18*(2), 103-118.

Thorndike, E.L. & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: NY: Teachers College, Columbia University.

Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics*, *39*(3), 302-327.

Uccelli, P., Dobbs, C. L., & Scott, J. (2013). Mastering academic language: Organization and stance in the persuasive writing of high school students. *Written Communication*, *30*(1), 36-62.

Verspoor, M., Schmid, M. S., & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, *21*(3), 239-263.

Vidakovic, I., & Barker, F. (2010). Use of words and multi-word units in Skills for Life Writing examinations. *University of Cambridge ESOL Examinations Research Notes*, *41*, 7-14.

Vieregge, Q., Stedman, K., Mitchell, T., & Moxley, J. (2012). *Agency in the Age of Peer Production*. Urbana, IL: National Council of Teachers of English.

## Address for correspondence

Philip Durrant
Graduate School of Education
University of Exeter
St Luke's Campus
Heavitree Road
Exeter, EX2 2LU
United Kingdom

p.l.durrant@exeter.ac.uk

## Co-author information

Joseph Moxley
Department of English
University of South Florida

Lee McCallum
Graduate School of Education
University of Exeter