1 Introduction

Holistic marking is often the norm for assessing essays in academic contexts. An on-going question is what makes a rater give higher or lower marks to different essays. This Element aims to gain an understanding of the linguistic and non-linguistic variables that may play a role in shaping the writing quality scores awarded to student essays by raters. Our work answers one central question: what linguistic and non-linguistic variables may play a role in shaping the writing quality scores awarded to student essays in a first-year composition (FYC) writing context?

The underlying premise is that collocation may play a role in understanding what shapes writing quality scores. We also assume that non-linguistic factors such as variation in individual writers, the writing task, and the language status of the writers may play a role in shaping scores.

To understand these variables, our work engages with the following methodologically driven sub-questions:

(i) How can we choose appropriate measures of collocation?
(ii) How can we measure and understand the potential role of different linguistic and non-linguistic variables involved in shaping writing quality scores in an appropriate way?

By answering these questions, we hope to illuminate the complexity of the linguistic and non-linguistic variables themselves, as well as how these variables operate with a degree of nuance in the rating process overall. We hope to illuminate, and to some extent demystify, different aspects of the rating process in a relatively underexplored FYC writing context in the United States (see Section 2 for details of the writing context).

Three empirical studies are carried out to answer these questions. The first question is answered by engaging with past literature and through the study of collocation measures in a cluster analysis. The second question is answered via the use of a cumulative-link mixed effects regression model. This model can accommodate different variable types to appreciate how they may play a role in shaping writing scores. A follow-up qualitative study provides a deeper understanding of how writers use collocations in their writing and also helps answer the second question.

The introductory section of this Element presents the rationale for the focus on these particular questions and our methods for answering them.

1.1 Understanding Writing Quality via Quantitative Linguistic Features

There has been a long-standing interest in understanding rater judgements of writing quality in first and second language research. This interest has adopted several theoretical and methodological lenses (e.g., see the overview in Durrant et al., 2021). One popular lens used to tap into these judgements and the inferences we can (and cannot) make from them has been the quantitative study of the relationship between linguistic features and writing quality grades. Under this lens, linguistic features are identified (normally by adhering to a specific theoretical framework that governs how to identify the features) and counted (manually or automatically with corpus software), and then relationships between these frequencies and writing quality grades are established numerically using statistical techniques such as correlation and regression analyses. Writing quality grades represent subjective ratings made by text evaluators who largely make their judgements from a predetermined set of criteria which to an extent presupposes what 'good' writing involves. These criteria therefore guide evaluators in their judgements (e.g., see the IELTS and TOEFL grade bandings mentioned in Durrant et al., 2021). This quantitative approach has enjoyed sustained popularity in the literature and is currently experiencing something of a 'boom', thanks to the increasing creation of corpus software tools which make the counting and analyses of such features increasingly user-friendly. This boom is well documented across overviews provided in Durrant et al. (2021) and Crossley (2020).

In their studies, researchers make two key assumptions. First, there is an assumption of, or perhaps appreciation for, the role that linguistic features themselves may play in our understandings of writing quality judgements as a measurable construct. This means there is an underlying belief that by counting linguistic features and looking at their relationships with writing quality via statistical methods, we can learn something about how these features may be being judged/perceived by raters. Second, and linking back to the first assumption, is the belief that the linguistic features chosen are (a) worth counting (because they have an established linguistic history/history in models of writing proficiency/quality), and (b) that they can indeed be reliably counted.

Findings of past feature-writing quality work have gone on to inform two often connected areas of research: the development of writing proficiency scales/rubrics by referring to differences in linguistic feature use across bandscales (e.g., Hawkins & Filipovic, 2012), and/or the training of large-scale feedback and grading systems (e.g., Chen et al., 2017). Researchers have most persistently studied features of grammar (e.g., clauses (see Bulté & Housen, 2014))

and vocabulary (e.g., percentage of words appearing in the Academic Word List (Daller et al., 2013)), with features of cohesion occupying an inconsistent position of interest for researchers (e.g., see the review in Durrant et al. (2021)). Features of phraseology, for example lexical bundles (e.g., see Appel & Wood, 2016), occupy an increasingly prominent position, especially in second language literature (e.g., see reference made to this emerging importance in Durrant et al. (2021) and Paquot (2018, 2019)).

It is this latter linguistic area that this Element focusses on. The following sub-sections make an explicit case for the study of one specific area of phraseology: that of collocation. The sub-sections present the rationale for such a focus and explain how this Element contributes to understanding the role collocation may play alongside several non-linguistic writing assessment variables in shaping writing quality judgements.

1.2 The Rationale for Studying Collocations and Writing Quality

The use of appropriate language is viewed as a key component of success for meeting programme outcomes in the FYC programme our Element focusses on (CWPA, 2014; CWPA et al., 2011). In this sense, using appropriate academic language is therefore a requirement of fitting into students' respective academic disciplines/communities. Wray (2006, p. 593) notes on this matter that 'when we speak, we select particular turns of phrase that we perceive to be associated with certain values, styles and groups', with the learning of these phrases or word combinations acting as a badge of identity and this badge is linked to particular academic communities.

Later, Wray (2019, p. 267) emphasises that the status of a word combination as a formula lies in the decision-making or perceptions of the agent. She states that a formulaic sequence is 'any multiword string that is perceived by the agent (i.e., learner, researcher, etc.) to have an identity or usefulness as a single lexical unit'. Siyanova-Chanturia and Pellicer-Sánchez (2019, p. 6) highlight that there are several frameworks or reasons that guide the perceiver's decision-making. It may entail high frequency of occurrence (since frequently produced strings, other than being useful by virtue of being frequent in language, may also benefit from being treated as a single unit), a teacher's perceived value of a string (no matter how frequent), some sort of basic holistic storage and processing, a specific pragmatic function, or, indeed, something altogether different.

Although these definitions allow researchers flexibility in their theoretical and methodological approaches to capturing formulas, two particular approaches have dominated the literature: phraseological and frequency-based approaches (see Nesselhauf (2005) for an in-depth overview of the differences

between these lenses). This Element grounds its theoretical and methodological conceptualisation of collocation in the frequency-based lens. Under a frequency-based lens, the tenets of collocation rest on understandings from Firth (1968, p. 181), who believed that part of a word's meaning is the 'habitual collocations' in which it appears. Meaning here is said to include both the concept with which the word is associated and the ways in which it is used. Firth (1968) gives the example of '*dark night*', resting his understanding on the belief that collocating words are part of each other's meaning. Thus, because *dark* appears frequently alongside *night*, collocability with *night* is one of the meanings of *dark* (Firth, 1957, p. 196). Later, Firth (1968) articulates these thoughts further to indicate that collocation is a type of mutual expectancy between words. Collocating words are said to predict each other, in the sense that the presence of one word makes the presence of the other more likely.

Corpus linguists have unpacked this mutual expectancy further by stating that collocation is 'the relationship a lexical item has with items that appear with greater than random probability in its (textual) context' (Hoey, 1991, p. 7), with Jones and Sinclair (1974) simply stating that words are collocates if they appear together more frequently than their individual word frequencies would predict. Under these views, there is also a psycholinguistic nature to collocation to consider, with Sinclair's (1987, p. 391) idiom principle setting out this mental association where 'a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments'. Further still, Hoey's (2005, pp. 3–5) theory of lexical priming also draws attention to the psycholinguistic nature of collocation in that it is seen as the 'psychological association between words . . . evidenced by their occurrence together in corpora more often than is explicable in terms of random distribution'. Under these guiding thoughts then, collocation is bound up in the idea that word combinations are more frequent than their individual word frequencies would explain and that there is a degree of non-random use to these pairings.

These thoughts have led researchers to develop multiple taxonomies and dictionaries of collocations (e.g., Benson et al., 2009) with studies presenting word combinations such as (i) '*heavy rain*', (ii) '*rancid butter*', and (iii) '*apologise profusely*', as collocations (e.g., Paquot, 2018). Under the frequency-based school of thought, researchers have commonly captured collocations like these by focussing on the belief that collocations are pairs of words which regularly co-occur within a given span or window of text, for example two to four words either side of the node/search word. This approach has been criticised as capturing syntactically unrelated or uninteresting collocations (e.g., Evert, 2009). More recently a smaller group of studies have identified

collocations using syntactic parsers which capture collocations according to a particular syntactic dependency relationship. Dependency pairings are decided by parsing a text using an automated parser. Dependency grammar operates on the notion that in every sentence each word is dependent on another, apart from the root of the sentence which is independent (Debusmann, 2000). A word depends on another if it is a complement or a modifier of the latter. For example, dependency pairings which might be collocations include an adjective modifying a noun.

1.2.1 The Study of Collocation in Student Writing

After using a span or syntactic approach, many researchers have focussed on studying the often-arbitrary partnering and the degree of exclusivity in combinations extracted. 'Arbitrariness' here refers to the way that collocational preferences can sometimes appear to defy logical explanation. For example, the fact that an idea can be 'utterly ridiculous' but not 'utterly sensible'. 'Exclusivity' refers to the way that some words are found almost exclusively in combination with another particular word, or group of words. For example, the fact that few things other than 'rain' can be described as 'torrential' and that few things can be 'shrugged' other than our 'shoulders'. It is these often arbitrary, complex, and exclusive relationships that researchers have attempted to study in learner writing.

Many of the foundations for studying collocation in learner writing stem from the belief that second language learner writers struggle to use collocations appropriately. This is because the combinations are assumed to be stored mentally as single units and must be used appropriately with an understanding of their arbitrary combinatory nature and intended, expected meaning. However, there is growing evidence that first language learners also struggle with collocation. They struggle to navigate the expected writing of university genres and disciplines for the first time. Both groups therefore encounter barriers in using the language expected and ultimately gaining acceptance, through that language use, into their academic communities (Durrant, 2019; Wray, 2002).

Under a frequency-based approach, scholars have used frequency information to statistically show two pieces of information about learner collocation: (i) how confident we can be that the word combination does or does not occur because of random chance and (ii) the degree of exclusivity the words in the combination have with each other; in other words – the degree to which they may in fact have other possible combinatory partners. The formulae used to capture these are known as association measures (Evert, 2004).

Evert (2004, p. 75) defines an association measure as 'a formula that computes an association score from the frequency information in a pair's contingency table'. A contingency table is a 2 × 2 table that lays out a word combination's frequency information. The table contains frequency information relating to the frequency of the combination, the frequency of word 1 and word 2 in the pair, the frequency of other possible word combinations using either word 1 or 2, and the size of the reference corpus being used. An illustration of a contingency table is provided in Brezina et al. (2015, pp. 144–5). Evert (2004) groups measures able to capture confidence as (i) significance measures, and those able to capture exclusivity as measures of (ii) association strength. These types of information have been interpreted as the higher the score, the more confident we can be that the combination is a collocation (i.e., not occurring because of random chance) in (i), and in (ii), the higher the score, the more exclusive the pairing and the less likely it is to have multiple other word partners that it pairs with naturally.

Those researching learner writing have mostly relied on two representative measures from the significance and the degree of strength groups. In the former, this has been the t-score, and in the latter, the mutual information (MI) score. The t-score, as a measure of confidence, has been found to flag up word combinations that comprise high-frequency words (e.g., 'little bit', 'other hand') (Granger & Bestgen, 2014). In contrast, the MI has been found to flag up word combinations that comprise low-frequency words (which make them more exclusive to each other). For example, 'tectonic plate' and 'juvenile delinquency' (Durrant & Schmitt, 2009; Granger & Bestgen, 2014).

Several studies have used these measures to inform understandings of second language learner writing (e.g., Bestgen, 2017; Bestgen & Granger, 2014; Chen, 2019; Durrant & Schmitt, 2009; Garner et al., 2019, 2020; Granger & Bestgen, 2014; Kim et al., 2018), with few studies of first language learner writing (e.g., Durrant & Brenchley, 2021; Kyle et al., 2018). In their English for Academic Purposes (EAP) study, Durrant and Schmitt (2009) found that second language writers used more high-scoring t-score combinations, while first language writers used more high-scoring MI combinations (more exclusive pairings found in discipline- and genre-specific writing). To some extent, this finding has been corroborated in other second language contexts (e.g., Bestgen, 2017; Bestgen & Granger, 2014; Garner et al., 2019, 2020; Granger & Bestgen, 2014); however, across these individual contexts, increases in MI combination use have not always been linear across year groups of learners or proficiency levels (e.g., Durrant & Brenchley, 2021; Paquot, 2018, 2019).

1.3 Emerging Questions from Current Studies

1.3.1 How Can We Choose Appropriate Measures of Collocation?

The first question that this Element engages with is 'How can we choose appropriate measures of collocation'? Engaging with this important question is warranted because, as previous sections of the Element have noted, past studies have raised several issues relating to the use of association measures. Scholars have relied on a narrow set of measures that have been restricted to association measures used in the language learning/assessment literature with the t-score and MI featuring prominently. There has been sparse mention of alternatives or an awareness of how the hundreds of other association measures touted in the literature align with the MI or t-score or may be able to illuminate different collocation properties to those highlighted by the MI and t-score (e.g., see criticisms in Öksuz et al. (2021) and acknowledgement of the hundreds of measures in Pecina (2005, 2010), Wiechmann (2008), Gries and Ellis (2015), and more recently Kyle et al. (2018) and Kyle and Eguchi (2021)). This Element's starting position is that the use of these measures needs to be understood against the wider bank of association measures that researchers have access to. The measures need to be understood in terms of their ability to illuminate different types of collocation properties. There is also a need to bring together the fragmented association measure literature. This fragmented picture means measurement choice is often underexplored and/or undertheorised because measures are spread out across different disciplines and scholars (Öksuz et al., 2021).

1.3.2 How Can We Measure and Understand the Potential Role of Different Linguistic and Non-linguistic Variables Involved in Shaping Writing Quality Scores in an Appropriate Way?

The second question that this Element engages with is 'How can we measure and understand the potential role of different linguistic and non-linguistic variables involved in shaping writing quality scores in an appropriate way'? Studies in this research area have started to use a wider range of statistical methods, such as recently mixed/multi-effects models (e.g., Garner et al., 2019, 2020; Paquot, 2018, 2019), to measure relationships between collocations and writing quality. Thus, a key goal of the Element is to explore how these types of models offer an appropriate method of studying writing quality scoring.

1.4 The Organisation of the Element

This Element proceeds by providing an overview of the FYC context in Section 2. Section 3 then guides readers through the current collocation-grade

landscape and emphasises how the empirical work in the Element adds to this landscape. Section 4 sets out the methodological steps taken in the three individual studies. Then, Section 5 describes the results of the cluster analysis carried out to answer the first question, while Section 6 describes the results of the mixed effects modelling, carried out to partially answer the second question. Section 7 also helps answer the second question by qualitatively unpacking the possible reasons for the statistical relationships between the measures of collocation and writing quality by looking at text samples from the FYC corpus itself. Section 8 concludes the Element by summing up the key findings and limitations, and importantly how we reflect on our methodological approach and its promise in future work.

2 FYC Programmes and the Writing Context

2.1 Overview of the Section

This section will explain the rationale for focussing on a FYC programme in the United States. The section will explain how these programmes may benefit from closer engagement with language instruction. We chose to base our study on the programme at the University of South Florida (USF) because of its focus on different writing tasks.

2.2 The Nature of the FYC Programme at USF

The University of South Florida is a large public university with a diverse student population. Of its 50,000 students, as many as 41 per cent identify as African American, Black, Asian American, Hispanic, Native American, or multiracial (USF, 2018). The university provides degrees in business, engineering, arts and social sciences, and interdisciplinary sciences (USF, 2018).

As a state requirement, students who enter a Florida College or University State system have been required since 2015–16 to complete thirty-six hours of general education coursework from a list of courses in communication, mathematics, social sciences, humanities, and natural science, among others. This requirement means students develop the academic and numeracy skills needed for the demands of university study.

In the FYC programme, students complete writing as a ‘process’. They develop strategies in pre-writing, co-authoring, revising, and editing, as well as learning to follow academic/disciplinary conventions for different genres. They must achieve a minimum C-grade to continue their studies.

The programme’s learning objectives are set out across two modules: ENC (English Composition) 1101 and ENC 1102. Some of these objectives include the following:

- Learning and applying strategies to facilitate a range of skills, including critical reading, the stages of process writing, and giving peer feedback.
- Composing academic genres and adhering to academic conventions (structure, citation, and linguistic features).
- Synthesising disparate or conflicting thoughts when evaluating questions/problems to form cohesive and collaborative solutions.

2.2.1 Individual Project Information

Students complete six projects: three on each module. They produce drafts, carry out peer review activities, and develop a revision plan from this feedback. Across the two modules, students choose a controversial topic that they explore from different stakeholder perspectives. In ENC 1101, the three projects are: producing an annotated bibliography, analysing a stakeholder's platform, and synthesising multiple perspectives in the form of a literature review. In ENC 1102, the three projects are: developing a Rogerian argument on common ground between stakeholders and how they can compromise, analysing a visual rhetoric, and finally composing a multimodal argument in the final project.

Course ENC 1101 focusses on solidifying writing practices by introducing and practising paraphrasing, citing sources, drafting and editing work, peer review, and collaboration. Course ENC 1102 focusses more on developing students' argumentation and reasoning skills as well as their agency. Project 1 from ENC 1102 requires students to develop arguments that look at differences in stakeholder views for their chosen topics and explore how these stakeholders may reach a compromise. This project builds on Project 3 from ENC 1101, which sets out the key arguments for each stakeholder.

2.3 Teaching, Evaluation, and Feedback at USF

Modules are taught by permanent staff, adjunct instructors, and Graduate Teaching Assistants (GTAs). The ethos on the programme is that writing is constructed at a community level. This means peer review and teacher-led writing conferences feature heavily. Writing is commented on and evaluated using 'My Reviewers', a bespoke learning management system (LMS) which allows instructors and students to view programme material, draft and final projects, and to give peer and instructor feedback via PDF annotation tools.

Each of the six projects is worth between 20 and 30 per cent of students' overall grade. Students are awarded the remaining percentage of their grade for

Table 1 Holistic grades awarded for ENC 1101 and ENC 1102

Grade Types	Grade Breakdown for ENC 1101 and ENC 1102		
A	A+ (97–100) GPA: 4.00	A (94–96.9) GPA: 4.00	A– (90–93.9) GPA: 3.67
B	B+ (87–89.9) GPA: 3.33	B (84–86.9) GPA: 3.00	B– (80–83.9) GPA: 2.67
C	C+ (77–79.9) GPA: 2.33	C (74–76.9) GPA: 2.00	C– (70–73.9) GPA: 1.67
D	D+ (67–69.9) GPA: 1.33	D (64–66.9) GPA: 1.00	D– (60–63.9) GPA: 0.67
F		F (59.99 or below) 0.00	

homework tasks and class participation, equalling 100 per cent. Projects are evaluated using custom-made rubrics which instructors are trained to use. These rubrics evaluate projects according to analysis, use of evidence, organisation, focus, and style. An overall holistic grade out of fifteen points is awarded, expressed by the letters A–F. These bandings are shown in Table 1.

2.4 Language Instruction in FYC Programmes

2.4.1 The Focus on Language Instruction in FYC Programmes

Although the CWPA Outcomes Statement (2014) helps standardise FYC programmes across universities, Eckstein and Ferris (2018) highlight how the statement and the Conference on College Composition and Communication's fluid guidance on students' linguistic needs means there is the potential for explicit language input to be overlooked in favour of a focus on traditional composition processes. Indeed, several scholars have started to draw attention to the lack of language focus on FYC programmes (e.g., Matsuda et al., 2013). They acknowledge that this lack of language instruction is common despite many raters downgrading coursework because of language problems.

The CWPA Outcomes Statement (2014) makes most specific reference to language instruction on FYC programmes under its 'Rhetorical Knowledge' and 'Knowledge of Conventions' sections. When developing rhetorical knowledge, students are expected to develop the ability to respond to a variety of different contexts, that is, they must be able to shift tone, level of formality, medium, and/or structure. Instructors are expected to guide students towards learning about the main features of genres.

Despite these connections to language use, Aull (2015) emphasises that most FYC programmes focus on process pedagogies and neglect focussing on how