Validating an L2 academic group oral assessment: Insights from a spoken learner corpus

Abstract

This study determines the fine-grained bottom-up linguistic features involved in successful second language (L2) English academic group oral tutorial discussion through the use of a spoken learner corpus comprised of over 20 hours of L2 production. Student performances were graded by teacher-raters using a can-do rating scale which assessed students' ability to participate in a group oral discussion. The performances were transcribed and annotated for linguistic features of group discussion described in the literature such as L2 errors, a range of interactive and interpersonal metadiscourse features, and a range of temporal, prosodic, lexical and syntactic markers (or 'fluencemes') of (dis)fluency. The results of the corpus study suggest that frequent use of metadiscourse is the primary indicator of raters' positive evaluation of student performance in L2 academic tutorial discussion, alongside frequent use of discourse markers, filled pauses and a high speech rate per minute as fluencemes. Most L2 error types as well as syntactic fluencemes did not particularly feature in raters' positive (or negative) evaluations. The detailed cross-sectional data afforded by this corpus analysis serves as quantitative evidence of the linguistic features accompanying each grade awarded across the rating scale, and contributing to the construct validity of the assessment.

Keywords: speaking assessment, group discussion task, corpus analysis, English for Academic Purposes (EAP)

Introduction

Corpora and Language Assessment

The use of language corpora, or "principled collections of language materials, spoken or written, compiled into an electronic database for the purpose of linguistic analysis" (Park, 2014, p. 27; Sinclair, 2004), are now prevalent in the field of language assessment as tools for the validation of assessments (Taylor & Barker, 2008, Barker, 2010). In particular, it has been used to provide evidence to support validity inferences of domain description (i.e., to show that test tasks and performances are relevant to the target language use) and explanation (i.e., to show that test scores represent different levels of proficiency) (Xi, 2017). This approach could involve analysing test materials or test data against 'representative', native-speaker first language (L1) corpora such as the British National Corpus (Oxford University Computing Services, 2007), or comparison against second language (L2) 'learner' corpora such as the International Corpus of Learner English (Granger, Dagneaux, Meunier & Paquot, 2002) and the Cambridge Learner Corpus (Nicholls, 2003). These methods ensure that the input, tasks and outcomes are aligned with the particular register or target language the test is designed to assess (Hawkey & Barker; 2004, Biber, 2006), or to use the corpus data to derive linguistic descriptors for scales of L2 competence/proficiency as seen from the use of the Cambridge Learner Corpus during the *English Profile* project (Hawkins & Buttery, 2010), which sought to compile the linguistic features involved at each competency level of the Common European Framework of Reference Standards (Council of Europe, 2001). These large corpora–in the hundreds of millions of words–are purposefully built to allow for fine-grained bottom-up analyses of (mostly) written assessment data at the syntactic, lexical and discourse level and at different levels of proficiency. However, while excellent resources, such corpora are typically costly and/or unavailable (in their annotated form at least) to the general public.

In addition, there is also currently a gap in the availability of spoken learner corpora for assessment purposes with annotations for L2-specific linguistic features such as errors or features of fluency/disfluency, and (perhaps symptomatic of this) spoken corpora composed of production at different proficiencies (Xi, 2017). Two notable exceptions are the EF-Cambridge Open Language Database (EFCAMDAT, Geertzen, Alexopoulou, & Korhonen, 2013) with error-tagged data across 16 proficiency levels, although at the time of writing, the researchers were yet to add to the corpus the large amount of spoken data they had collected. The multimillion spoken word Trinity Lancaster Corpus (Gablasova, Brezina & McEnery, 2017), has already produced a range of studies focusing on stance (*ibid*), fluency (Götz, forthcoming) and multiword expressions (Coccetta, forthcoming) found at different L2 proficiency levels. The general lack of L2 annotated oral data, however, means that the typical benchmarking approach that corpus-based validation studies have used is not always possible.

An alternative approach to L1/L2 or L2/L2 corpus comparison for test validation purposes is that of corpus analysis of the test data itself, with the corpus used to determine the linguistic features present/absent across the assessment's rating scale. More recently, corpus-based studies into test validation have taken on a multidimensional approach to the investigation of language features salient across rating scales, without the need for benchmarking against an established reference corpus. Notably, a recent special issue of *Language Testing* has featured studies involving the use of natural language processing techniques to explore the linguistic features involved in the assessment of writing (Kyle & Crossley, 2017; Lu, 2017) and also of speaking (LaFlair & Staples, 2017). These studies utilized a bottom-up analysis of the linguistic features involved in the assessment process and may serve to "augment" (Park, 2014, p. 35) ratings of student performance on a given assessment; they involved characterising the linguistic features involved in the students' production that (presumably) influenced raters as they came to

grading decisions across the rating scale. The analysis would potentially provide test stakeholders with a set of linguistic features in each criteria on the rating scale.

In this respect, corpora analysis may be particularly vital for oral assessments, given that working memory constraints of storage and of controlled processing account for significant difficulty in language comprehension over extended periods (Just & Carpenter, 1992). Working memory may vary significantly, leading to differences in ability to recall language used at later times (Unsworth & Engle, 2007). Given these concerns, compiling evidence of the bottom-up linguistic features involved in raters' decisions to award certain grades across the rating scale is (or at least should be) an important aspect of the process of ensuring construct validity of such scales, complementing other methodologies commonly employed in language assessment such as Many-Facets-Rasch Measurement analysis.

Academic Group Discussion Task

In Hong Kong, the vast efforts spent on preparation for high school examinations in a competitive exam-oriented system (Kennedy, 2002) has had freshman students having relatively little opportunity to discuss academic topics with their peers, at least compared to the amount of time spent on writing and rote memorization (Kennedy, 2002; Lee, 2008). Production of argumentative discourse by freshman undergraduates is often criticised as falling short of academic expectations (Matsuda & Jeffery, 2012), with many students arriving at university "use[ing] and respond[ing] to the features of stance and voice differently" to the expectations of their academic tutors (Sancho-Guinda & Hyland, 2012, p. 2) and with Hyland (2016, p. 246) claiming that many students arrive at university "thinking they have landed on Mars". In response, many tertiary institutions offer freshman pre- or in-sessional courses in EAP, which

while primarily focused on writing, also include a variety of elements of oral production including presentations, speeches, and, of interest for the present study, group tutorial discussion.

In Hong Kong, peer-to-peer/group oral L2 assessments have been a feature of the assessment scene since the 1990s, following the addition of a group discussion requirement into the HK secondary English examinations. Group oral assessment have been perceived to create numerous, obligatory situations for 'negotiation for meaning' (Long, 1996) in that the test-takers themselves are responsible for interactionally modifying the available input to increase its 'comprehensibility' wherever breakdowns in communication occur (Krashen, 1987). In negotiating authentic, real-time communication, students are actively involved in 'noticing the gap(s)' (Schmidt, 1992) in their or others' linguistic knowledge as part of general co-operative principles (e.g. Grice, 1975), and seek to repair communication breakdowns through the use of conversation management techniques such as confirmation checks and clarification requests (e.g. 'what did you say?'), which, for the increased benefit of L2 learners, are prompted and received by the learners themselves rather than organised by their interviewers (Foster & Ohta, 2005). This assessment context obligatorily leads to student output (Swain & Lapkin, 1995), again contributing to 'noticing' and subsequent peer-to-peer co-construction and negotiation of knowledge (McNamara, 1997). This communication context is considered particularly fruitful for lower-level L2 learners, as they are more likely to require extended conversational management than that required by higher-level learners (Gan, 2010).

Academic group discussions differ from general group discussions in that a key aspect of this genre is the ability to produce nuanced yet argumentative, persuasive discourse (Biber, 2006). In particular, the task provides candidates opportunities to demonstrate their ability to use metadiscoursal devices such as hedging, boosting, self-mentions or attitude markers used to "stamp their personality or beliefs onto their arguments" (Hyland, 2016, p. 247) in the form of a

stance and *engagement*. Stance refers to the linguistic projection of a language user's views toward the topic under discussion, while engagement involves the dialogic way in which speakers "relate [to their listeners] with respect to the positions advanced" (Hyland, 2016, p. 169). Interlocutors in academic group discussions have to manage the presentation, support and defence of stance, while structuring conversational interaction via the production of a range of interactional metadiscoursal markers include code glosses, evidentials, frame markers (including sequencing devices, label stages, announce goals and topic shifts) and transition markers. The combination of these features leads to organised, structured, and impactful academic production intended to successfully persuade their audience, their peers and the teacher-rater, to support a given point of view.

Assessment of academic production, unlike the assessment of English for general purposes, thus should focus on the presentation, support, and defence of the language user's *position* on the topic under discussion and some studies have demonstrated how raters interpret these metadiscoursal features. Gan (2010) investigated group discussions of secondary school students and found that higher-rated students are more capable of engaging with others' ideas through a higher frequency of register appropriate suggestions, (dis)agreements, explanations and challenges. Gan, Davison and Hamp-Lyons (2009) considered students' ability to both pursue and shift the topic of talk (while still making meaningful individual contributions) as a real measure of success as seen by raters during group assessments. He and Dai (2006) studied performances on the *Chinese College English Test–Spoken English Test* and found that students lacked an appropriate range of interpersonal language functions for the purposes of making claims, presenting their stance, and defending their stance from others, which made it difficult for raters to rate students. Specific to the assessment featured in the present study, Crosthwaite, Boynton and Cole (2017) performed a think-aloud protocol study on six teacher-raters as they

observed and graded students' production of stance and engagement features during a 25-minute-long academic group tutorial test. Considerable variation was noted in how raters perceived successful presentation of stance and engagement during the test as well as variation in how raters arrived at similar grading decisions. Test-related factors such as topic (Van Moere, 2007) and different proficiency levels between interlocutors (Iwashita, 1996) have also been found to impact successful or unsuccessful group oral interaction. Finally, studies on test taker interactions in group oral assessments noted affective factors such as shyness (Bonk & Van Moere, 2004), assertiveness (Ockey, 2009), introversion or extroversion (Berry, 2004), and talkativeness (Van Moere & Kobayashi, 2004) to impact performance.

However, rater judgment on test-taker's performance in an academic discussion is primarily a *top-down* concern, derived from an impression a given rater has drawn (from memory) regarding a student's successful overall presentation of the totality of their arguments, and whether they successfully supported and defended their points in a register-appropriate manner. What is less-known is the extent to which the totality of the *bottom-up* linguistic features present in students' performances can also be considered as constitutive of a raters' opinion of the success or failure of that performance against the scale they are rating against. Extrapolating the features involved in raters' positive evaluations will provide evidence of the construct assessed (i.e., academic English oral proficiency) and backing to support the different levels of the rating scale. It could also reveal a 'hidden curriculum' (Legg, 2016) of linguistic features that contribute to teacher-raters' evaluation of 'successful' academic production, but that are not explicitly taught in current EAP curricula.

Linguistic Features of Academic Group Discussions

The present study is focused on the linguistic features involved in raters' appraisals of 'successful' production on a rating scale used for an English for Academic Purposes (EAP) oral assessment at a leading university in Hong Kong. The group oral tutorial assessment rating scale in the present study is comprised of three main criteria against which students are graded on (see Appendix A): *ability to explain academic concepts*, *ability to interact with others*, and *ability to communicate comprehensibly*. The first two criteria represent the metadiscourse linguistic features of an academic group discussion as found in the literature. In order to receive high ratings in these criteria, descriptions in the rating scale indicate that students need to demonstrate a variety of interactive and interpersonal functions of language which include the following: asking direct or indirect questions to other candidates, rebutting other candidates' claims, deriving counter-arguments to their own arguments so as to sufficiently strengthen their position, and to appropriately present a range of facts, opinions or statistics from academic sources (presented orally in the form of 'spoken citations', such as '…according to an article by Smith in 2009…').

The second language (L2) aspect of the assessment is encapsulated in the criterion *ability to communicate comprehensibly*. Here, the prevalence of errors in student production is likely to impede comprehensibility, while an additional factor is that of fluency. Evans & Green (2007) found in interviews with HK tertiary students that their problems with academic discussion stemmed particularly from issues with grammar, fluency and pronunciation. Gan (2012) in further interviews with HK students found that inadequate vocabulary contributed directly to a lack of fluency in speech; that grammar was a stumbling block for students' oral production (with a knock-on effect on fluency as well); and that improper pronunciation and intonation (e.g. articulation problems) reduced overall fluency. Tsui (2001) suggests L2 learners in such contexts who are prone to errors or disfluencies tend to be more likely to receive negative evaluation in

the language classroom than in other subjects. While fluency is 'an epiphenomenon to which many individual (interrelated) factors contribute' (Götz, 2013:1), fluency can be broken down into individual 'fluencemes', or 'abstract and idealised features of speech that contribute to the production or perception of fluency' (Götz, 2013:8). This approach allows for fluency markers to be annotated and quantified using corpora.

Given the above, when considering these potential linguistic features involved in successful academic group oral discussion, the suggested interaction of the features in question and the assessment criteria against which they may be applicable and (presumably) indicative of proficiency are shown in Figure 1.

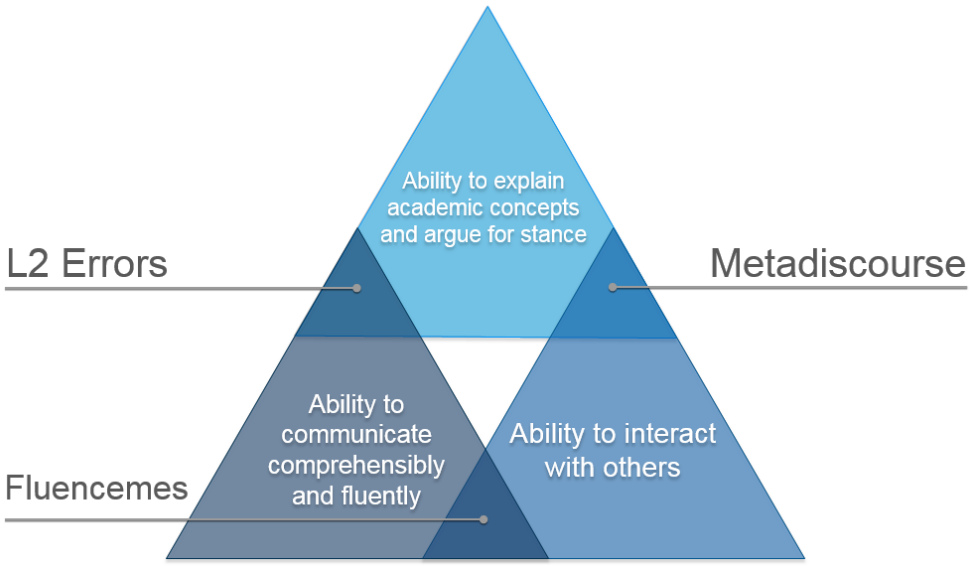

Figure 1. Interaction of annotated features and assessment criteria

The goal of the present study is to build and analyse a learner corpus of group oral academic discussion production in order to determine whether and how these linguistic features (metadiscourse, errors, and fluencemes) affect rater judgments across the rating scale. The following research questions are posed for the present study:

1. Are certain specific linguistic features involved in L2 academic group oral assessment predictors of teacher-raters' grading of (un)successful student performance across an EAP rating scale?
2. Which linguistic features contribute the most to the perceived (lack of) success of L2 learner performance?

Methodology

The data was taken in 2016 at a tertiary institution in Hong Kong, focusing on an undergraduate EAP program taken by approximately 1,500 freshman undergraduates per semester. As part of the range of assessments for this program, students sit a group oral tutorial discussion assessment at the last week of the semester, having had several scaffolded mock tests throughout the semester.

Corpus Sample and Assessment Procedure

The task requires students to take the test in groups of five. Students are given a topic with four academic sources 72 hours before the assessment. They prepare for the assessment by formulating a stance on the topic, brainstorming arguments on a note sheet, and reading four academic sources to support their arguments. During the test, students are given 25 minutes to discuss the topic if there are five participants (with a reduction of 5 minutes per participant if a participant is absent for the test). The discussions are video-recorded and teachers either mark during the live assessment or upon subsequent viewing of recordings.

There were 30 teachers rating the students, with each teacher marking at least 20 students (i.e., one class). Most of the teachers have rated this assessment at least twice a year for more than five years. Teachers use a 12-point scale analytic marking rubric (A+ to F) to mark students

with three domains representing students' group discussion skills ability (Appendix A). Students are given a mark across the three domains and an aggregate mark is produced. To compute the aggregate grade, the letter grades are converted to a raw score (0-100) and then transformed to percentages, which are then added up to a total raw score. This final raw score is converted back to a letter grade and given to students as the final aggregate score.

Reliability of raters and scores are ensured through several methods. Every year, all teachers go through a standardisation procedure prior to the exam. Standardisation involves all teachers individually marking three sample videos and the marks are compared against a standard grade. During the assessment, students are marked by another course teacher to ensure fairness. Inter-rater reliability of scores are ensured by random double-marking of sample scripts by the course coordinator. If there is more than a full letter grade difference between the coordinator's grade and the teacher's grade, they have a discussion and come to an agreement on the final grade to be assigned to the student.

For this study, the final aggregate grade awarded is used to determine the predictors of grade rather than the individual criteria scores. The aggregate grades are used in this study for two reasons. Firstly, one concern with this assessment flagged in previous think-aloud protocol research involving raters of this assessment (Crosthwaite, Boynton & Cole, 2017) has shown that failure for the criteria '*ability to communicate comprehensibly*' has an additional negative impact on grades for the other two criteria. Secondly, the linguistic features involved in the analysis cut across each of the three criteria rather than being independent to any one criterion, justifying the model shown previously in Figure 1.

All students attending the course that semester were approached by the researcher to give consent for their exam data and for their grades to be made available for the study. Approximately half of them agreed. As agreement from *all* participants in a video had to be

given before we could analyse the data, we aimed to collect a final sample of 250 participants (across 50 video recordings with 5 participants each), which would span 20 hours of tutorial discussion. Due to the need to secure consent from all participants in a given video, many students at the lowest grades declined to give consent. 'A-' was selected as the highest graded performance in our corpus sample, as the number of 'A+' and 'A' grade performances where all participants agreed to give consent for their video to be used was quite low. C+ was the lowest graded performance in our sample, with only 6 graded performances giving consent at C grade, and 1 graded performance giving consent at C-, so data at 'C' and 'C-' grades were not included. No participants with D or F grades gave consent to provide data for the current study. However, the subcorpora are still broadly representative of the actual distribution of grades awarded to students on the assessment (Table 1).

Table 1

Corpus word counts and participants by grade

Corpus feature	A-	B+	B	B-	C+
Word count	37010	48672	31001	17585	9019
Participants	55	71	57	37	22

The complete data for the corpus was finally drawn from a total of 59 videos, between 20-25 minutes long, spanning approximately 20 hours 20 minutes of group oral tutorial discussion. The total word count was 143,287 words, across 242 participants.

Annotated Features

Rather than using a standard transcription protocol for representing oral data in text as this would affect a number of automated analyses, the data was transcribed in plain, unmarked text by a bilingual English/Cantonese-speaking research assistant, to which we then added layers of annotation using corpus software after subsequent re-viewings (see below regarding the annotation method). Three main linguistic features, i.e., the errors, metadiscourse, and (dis)fluency features, would be both readily annotatable using a corpus-based approach.

Errors. Our error coding taxonomy was adopted from a similar scheme found in Dahlmeier, Ng and Wu (2013), with some modifications for errors that rarely feature in our data and with some errors from their scheme merged with other errors to avoid redundancy (Figure 2)

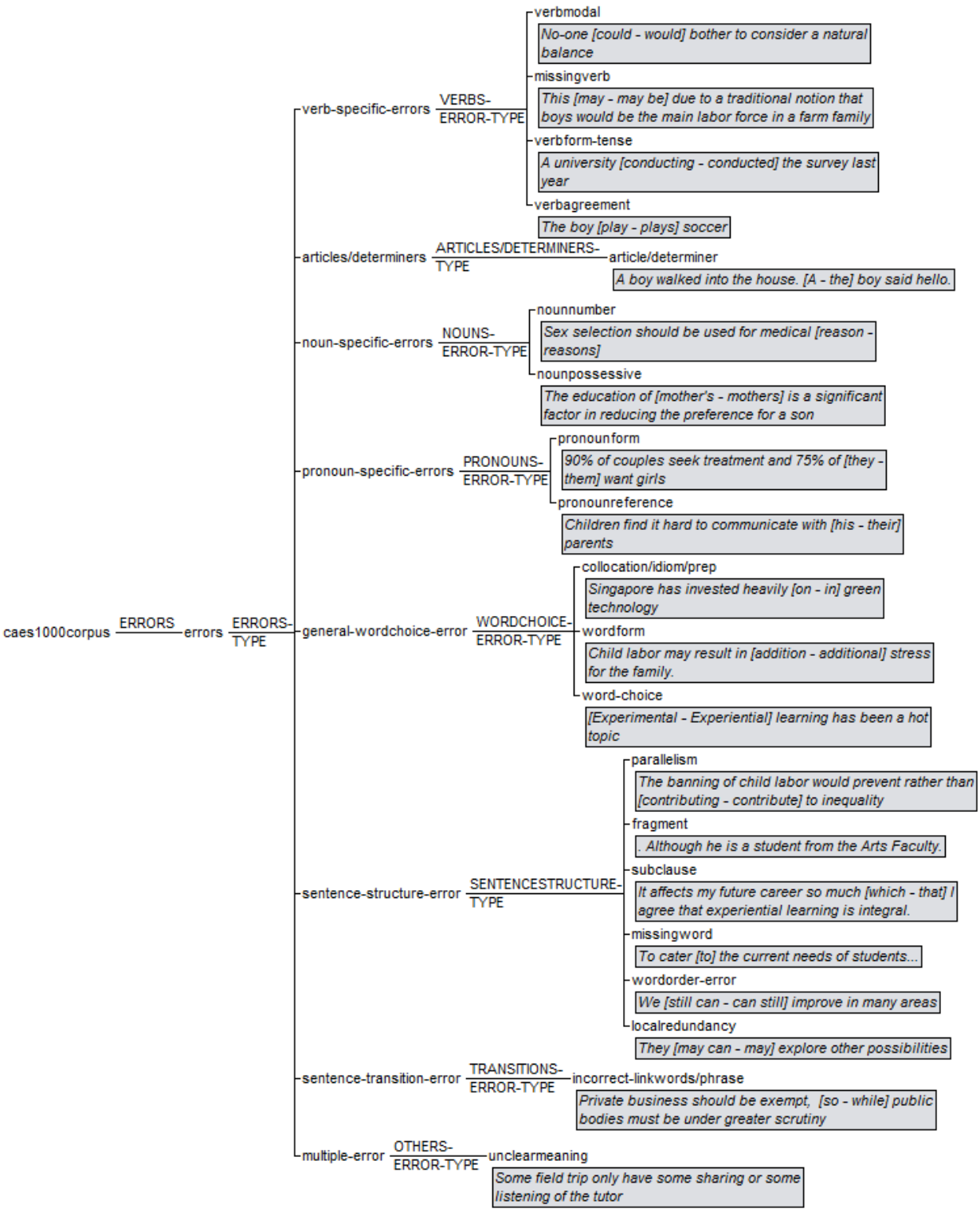

Figure 2. Error coding scheme with glosses [error – correction]

These categories follow the guidelines of Chuang and Nesi (2006) in that "error categories should not overlap, should have precise definitions, and should describe, not explain" (p. 252). The final categories selected in the present study include each of Chuang and Nesi's (2006) error categories of misformation, omission, overinclusion, misordering, and misselection. These categories were trialled on previous corpus research in EAP using written corpus data in Crosthwaite (2017).

Metadiscourse features. For the criteria 'ability to explain academic concepts and argue for a stance' and 'ability to interact with others', we annotated for interactional and interpersonal metadiscourse features from the list of such features found in an appendix in Hyland (2005). Each feature was checked by a bilingual English/Cantonese research assistant to ensure the items in the list were being used as a metadiscoursal feature within the corpus data.

(Dis)fluency features. We annotated for a number of features of (dis)fluency from a modified list of 'fluencemes' Götz (2013) as identified in Crible, Dumont, Grosman & Notarrigo (2016). Our final selection of fluencemes for analysis are divided into prosodic markers, lexical markers, and syntactic markers. Among these, certain prosodic, lexical and temporal features are considered as features of productive fluency on the one hand, with other lexical and syntactic features representative of perceptive fluency[1] on the other.

For prosodic markers, we incorporate measures of *speech rate* as a temporal productive fluency marker as well as *unfilled* and *filled pauses*, the former a temporal productive marker and the latter as a fluency-enhancement strategy, with sentence-final *falling, neutral* and *rising*

[1] At least, in terms of what an abstract, conceptual 'native speaker' would consider as a feature facilitating the perception of fluency following Götz's (2013) approach.

intonation annotated as a perceptive prosodic fluenceme. Our lexical markers include the list of *discourse markers* from Aijmer (2004) (Appendix B). We also include the frequency of *identical repeats* (the problem…the problem), and an umbrella category coined *reformulations*, including false starts (they learn ne...it's as if they discover…), modified repetitions, and substitutions. Our syntactic markers include a measure of *dependent clauses* as a perceptive fluency marker (see the following section for how a measure of dependent clauses was derived), as well as an umbrella category coined *interrupted structures*, including truncation at the lexical level, and incomplete utterances. Non-verbal markers (e.g. hands, gaze) were not annotated in the present study.

Annotation Procedure

The data for the corpus was transcribed in the form of plain text files with header information added to each turn regarding the participant ID, turn number, time turn began / ended, and overall grade awarded to that participant. For annotation of all errors, all metadiscourse, prosodic/lexical (dis)fluency markers and interrupted structures, the transcribed files were converted into a searchable, annotatable corpus using UAM Corpustool (O'Donnell, 2008). Errors, prosodic fluencemes, repeats, reformulations, and interrupted structures were manually annotated by a bilingual English/Cantonese speaking research assistant of near-native L2 proficiency and by the researcher (a native speaker of English). As the vast majority of the assessment candidates are L1 Cantonese, our approach follows the suggestion of Dagneaux, Denness and Granger (1998) in that "efficiency is increased" if the annotators have both a high knowledge of English grammar and the L1 mother tongue of the interlanguage variety to be analysed (p. 165).

Metadiscourse and discourse markers were annotated automatically in UAM Corpustool using Corpus Query Language (Christ, 1994) to bring up concordance lists of target words/phrases for automated annotation. For annotation of dependent clauses as a disfluency marker, we used Nini's (2015) Multidimensional Analysis Tagger (MAT) to tag the texts for past/present participle clauses, pied-piping relative clauses, 'that' relative clause on subject/object position, sentence relatives, split infinitives, WH-clauses, WH-relatives in subject/object position, and past/present participle deletion relatives, working out the raw and normalised values for each turn and total by graded performance. Due to differences in sample sizes across subcorpora, the raw frequencies of the annotated items were converted into a normalised frequency per 1,000 tokens in UAM Corpustool and MAT (1,000 tokens is the output generated by both UAM Corpustool and MAT, even though many students' production was less than 1,000 tokens). Due to the large number of errors and error types involved, we were unable to provide statistics for occurrences of L2-appropriate usage (such as Target Language Use, Pica, 1983) as recommended under a true Computer-aided Error Analysis approach. Thus, only the normalised frequencies of errors per 1,000 tokens are analysed for comparison in the present study.

After annotation of the three main annotation categories (errors, fluency markers, metadiscourse), the researcher and research assistant checked each files' annotations for accuracy, changing the annotation where there was disagreement. As the texts were produced in English, the researcher had the final decision on any disagreement after consultation with the research assistant. Two other native speakers of English then analysed a random sample of 20 of the complete annotated video transcripts to check for inter-rater reliability (just over 33% of the total corpus data), marking each annotation as correct/incorrect. Rater agreement on metadiscourse was measured by Intraclass Correlation Coefficient with a final statistic of .823,

of which a value greater than .750 is considered 'excellent' in the literature (Fleiss, 1981). Agreement on fluenceme annotations (not including dependent clauses, which was derived using the MAT Tagger) and errors was .686 and .735 respectively, with these values considered ‘good’ under Fleiss (1981).

Statistical Procedures

In order to determine whether the three main annotated groups of linguistic features (metadiscourse, errors, and fluencemes) were real predictors of grading decisions across the rating scale, of each relevant subcategory was transformed to a standardized z-score in SPSS. This step allows for linguistic features that are relatively infrequent, but potentially quite salient for raters (e.g. errors of idiom) to be placed on equal footing with features that far more frequent in the data (e.g. filled pauses). These z-scores were then transformed into a unified z-score for their relevant superordinate category (e.g. the z-score for ‘interactive metadiscourse’ as a superordinate feature was made by adding the z-scores for each sub-component of this feature before dividing this sum by the number of sub-components). Once this transformation process was complete, we then performed ordinal regression analysis with grade as the dependent variable (given that ‘grade’ is an ordinal variable) to determine the significant predictors of grade. Each time more than one variable was included in a model, we checked for multicollinearity of the included variables (using the option to do so in SPSS) using a cut-off for correlation of .75 for removal/combination of any variables. This cut-off was never close to being reached in our data, ensuring no issues with multicollinearity were present.

Results

The normalised frequencies of each annotated category are shown in full in the tables in Appendix C; this results section will thus summarise these results for the reader before concentrating on the results of the regression analyses. To remind the reader, the annotated categories were:

A) *Metadiscourse* features, including *interactive* metadiscourse for structuring and organising text (e.g. *code glosses* – 'as a matter of fact', s*equencing markers* – 'first', 'second', *topic shifts* – 'with regards to', etc.) and *interpersonal* metadiscourse for attitude and stance (e.g. *hedges* – 'possibly', *boosters* – 'obviously', etc.);
B) *Fluencemes*, including *prosodic* (e.g. *words per minute*), *lexical* (e.g. *discourse markers* – 'and so on…') and *syntactic* fluencemes (e.g. *interrupted structures*);
C) Errors, including *verb-specific* (e.g. *missing verb, agreement*), *word choice*, *sentence structure* (e.g. *parallelism, local redundancy*) and other error types.

We begin by presenting the results of the regressions analyses for each individual annotated category first, before we present the results of a regression analysis of the three annotated categories combined.

Metadiscourse Features and grade

Table 2 shows the normalised median frequencies of each metadiscourse feature in our corpus by grade (The raw figures can be found in Appendix C).

Table 4

Normalised metadiscourse features by grade assigned per 1,000 words

Feature	A- Median (Median Absolute Deviation)	B+	B	B-	C+
Interactive metadiscourse					
Code glosses	1.06 (0.66)	0.74 (0.45)	0.62 (0.32)	0.45 (0.45)	0.36 (0.36)
Evidentials	0.28 (0.28)	0.31 (0.25)	0.26 (0.26)	0 (0)	0 (0)
Frame markers					
a) Sequencing	0.61 (0.33)	0.61 (0.61)	0.30 (0.30)	0 (0)	0 (0)
b) Label stages	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
c) Announce goals	1.3 (0.68)	1.27 (0.72)	0.76 (0.43)	0.45 (0.45)	0.52 (0.23)
d) Topic shifts	3.01 (1.59)	4.57 (1.79)	2.36 (1.25)	1.65 (0.84)	0.87 (0.53)
Transition markers	14.05 (4.78)	19.65 (5.65)	11.1 (4.05)	9.22 (3.39)	7.02 (3.31)
Interpersonal metadiscourse					
Attitude markers	1.18 (0.88)	2.06 (0.78)	1.15 (0.64)	0.85 (0.64)	0.71 (0.58)
Boosters	5.84 (3.10)	8.41 (2.92)	5.59 (2.78)	3.35 (1.60)	2.71 (0.69)
Self-mention	6.78 (2.89)	10.31(3.13)	5.99 (3.05)	4.62 (2.15)	3.60 (1.40)
Engagement markers	10.45 (4.97)	11.2(3.46)	7.32 (3.26)	5.45 (2.95)	3.46 (1.45)
Hedges	7.57 (3.25)	7.74(2.89)	4.71 (2.10)	3.57 (2.20)	2.84 (0.93)

Ordinal regression was performed on the unified z-scores for interactive and interpersonal metadiscourse separately. The model for interactive metadiscourse (-2LL[2]=707.6, $\chi 2$=30.2, p<.001) suggested that this feature is a significant positive predictor of grade (β =.636, sig<.001, Exp β=1.88[3]). The model passed Pearson goodness of fit (p=.165[4]) with a Nagelkerk pseudo r^2 value of .123[5]. Another model (-2LL=704.5, $\chi 2$=34.1, p<.001) suggested interpersonal metadiscourse was also a significant positive predictor of grade (β =.698, sig<.001, Exp β=2.01), with Pearson sig.= .149 and Nagelkerk r^2 = .138.

Looking now at the subordinate features of interactive metadiscourse, the results of the ordinal regression of these features and grade (-2LL=705.2, $\chi 2$=41.2, p<.001, Pearson sig= .147, Nagelkerk r^2=.164) suggests *topic shifts* (β =.548, sig<.001, Exp β=1.76) and *label stages* (β

[2] -2LL = 2 log-likelihood statistic, $\chi 2$ = chi squared statistic
[3] β=expected regression value, Exp(B)=odds ratio
[4] The Pearson goodness-of-fit test should be non-significant if the model is to be useful.
[5] The Negelkerk r^2 value is a measure of the model's predictive power.

=.279, sig=.033, Exp β=1.32) were significant positive predictors of grade (Figure 3). At A-level, the most frequent topic shifts included *'so', 'well', 'now', and 'back to'*, while the most frequent label stages included *'now', 'overall', 'so far',* and *'on the whole'*.

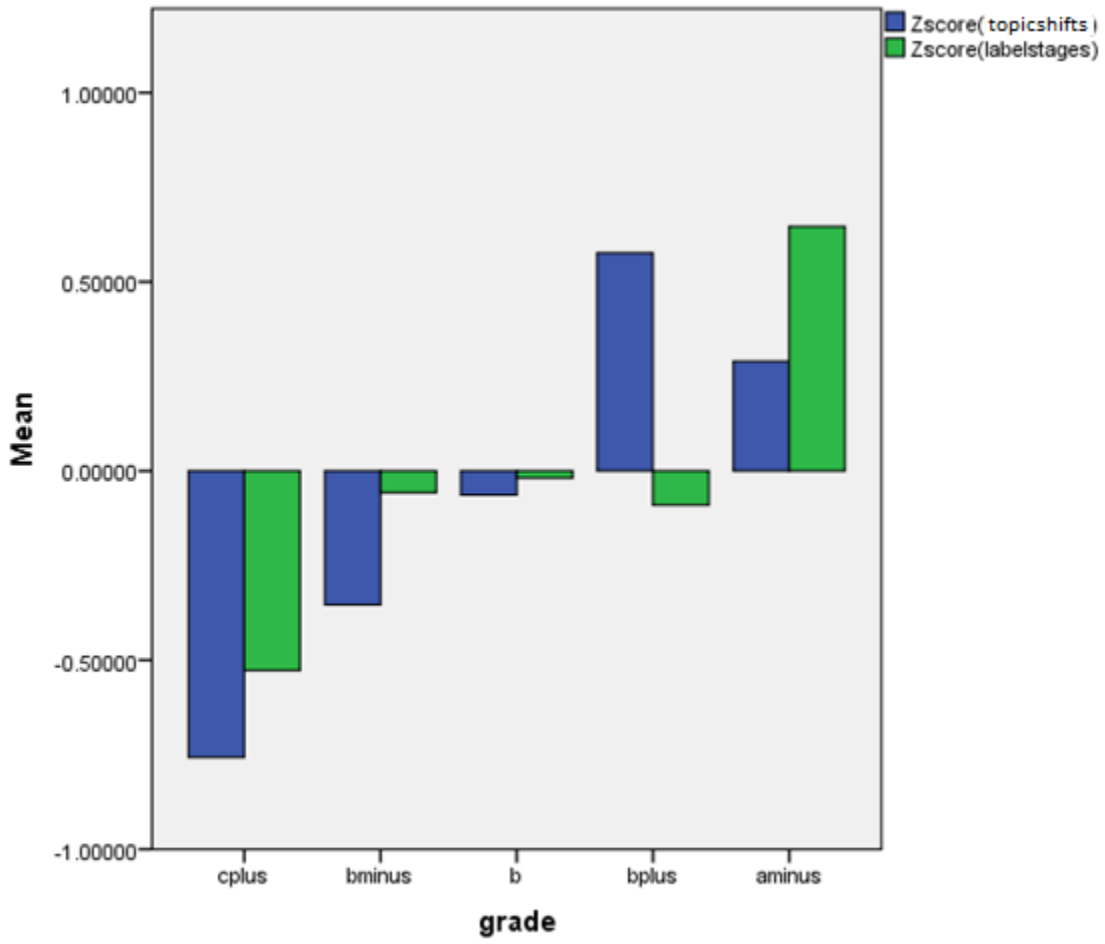

Figure 3. Topic shifts and label stages as predictors of grade

Of the subordinate categories of interpersonal metadiscourse, an ordinal regression model involving these categories and grade (-2LL=705.3, $\chi 2$=41.1, p<.001, Pearson sig= .152, Nagelkerk r^2=.164) suggested that *engagement markers* were also significant positive predictors of grade (β =.625, sig=.023, Exp β=1.86) (Figure 4). At A- level, the most frequently used

engagement markers included *'we', 'you', 'should', 'have to', 'our', 'do not'* and *'need to'*, as speakers referred to their co-participants directly and made numerous recommendations.

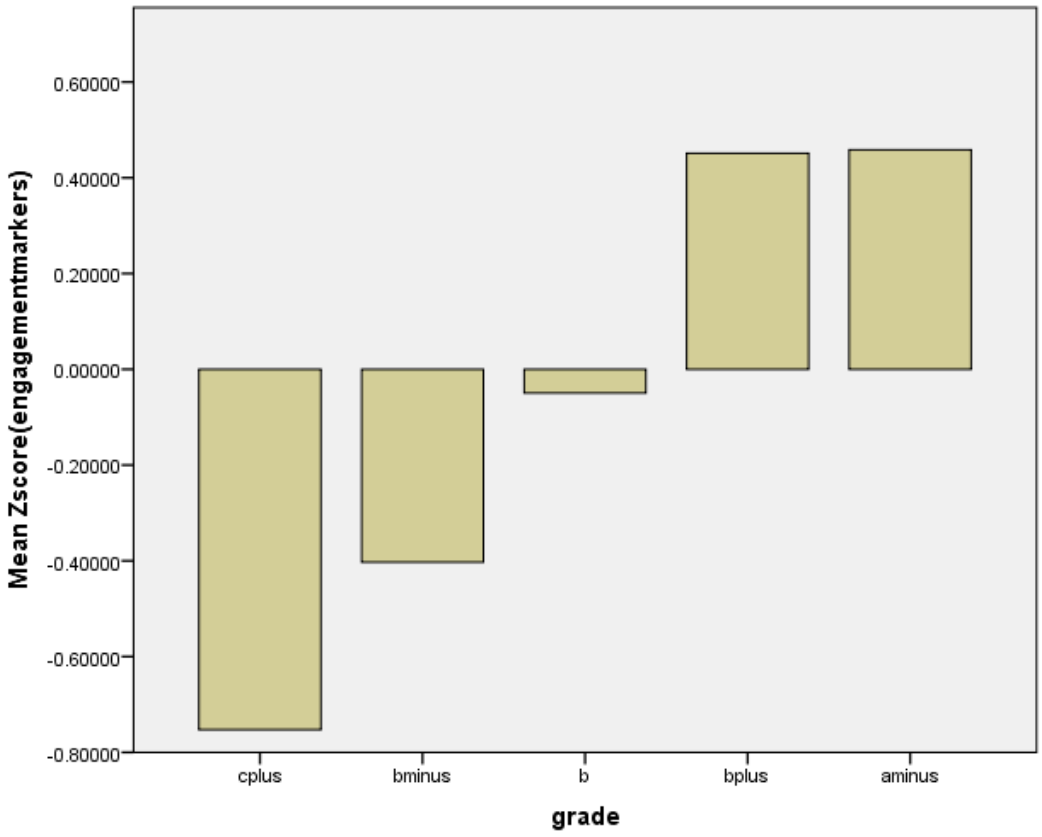

Figure 4. Engagement markers as predictors of grade

Error type, frequency and grade

Table 3 shows a summarised version of the normalised median frequencies of errors in our corpus by error type (the raw and full table including all error subcategories are shown in Appendix C).

Table 3

Normalised error features by grade assigned (Median/Median Absolute Deviation, per 1,000 words)

Feature	A-	B+	B	B-	C+
Errors (all categories)	11.2 (5.09)	16.2(6.39)	12.6 (5.65)	11.7 (4.96)	13.8 (6.86)
Verb-specific errors (e.g. tense, missing modal)	1.47 (0.68)	2.66 (1.43)	2.28 (1.2)	1.83 (0.97)	2.87 (1.09)
Article errors	1.19 (0.80)	1.40 (0.80)	0.94 (0.49)	0.91 (0.53)	0.92 (0.35)
Noun-specific errors (e.g. number, possessive)	1.13 (0.67)	1.30 (0.58)	1.44 (0.91)	0.74 (0.48)	1.04 (0.52)
Pronoun-specific errors (e.g. form, reference)	0.28 (0.28)	0.36 (0.36)	0.31 (0.31)	0.28 (0.28)	0.37 (0.16)
General word choice errors (e.g. collocation, word form)	2.85 (1.43)	3.97 (1.86)	3.39 (1.78)	2.65 (1.15)	3.46 (1.83)
Sentence structure errors (e.g. parallelism, word order)	1.63 (0.74)	2.05 (0.85)	1.74 (0.94)	1.16 (0.57)	1.56 (0.84)
Sentence transition errors	0.81 (0.51)	1.03 (0.55)	0.38 (0.38)	0.39 (0.27)	0.63 (0.50)
Multiple errors - unclear	0.30 (0.30)	0.35 (0.28)	0.36 (0.36)	0.26 (0.26)	0.36 (0.36)

The data suggest that there does not appear to be a linear relationship between errors produced overall and grade, with the median error frequency at its lowest at A-, highest at B+, with the number of errors at B- equivalent to that of A- before rising again at C+. Ordinal regression of errors as a superordinate category was of poor fit (-2LL=746.2, $\chi 2$=.170, p=.680) and insignificant (β = -.081, sig=.691, Exp β=0.92). Ordinal regression for the subordinate categories (Figure 5) was acceptable, if weaker than those for metadiscourse (-2LL=705.1, $\chi 2$=41.3, p=.005, Pearson sig=.024, Nagelkerk r^2=.165) with errors of *verb tense* (β =-.472, sig=.002, Exp β=0.62), *multiple errors leading to unclear meaning* (β =-.296, sig=.046, Exp

β=0.74) and *pronoun form* (β =-.298, sig=.025, Exp β=0.74) suggested to be significant negative predictors of grade.

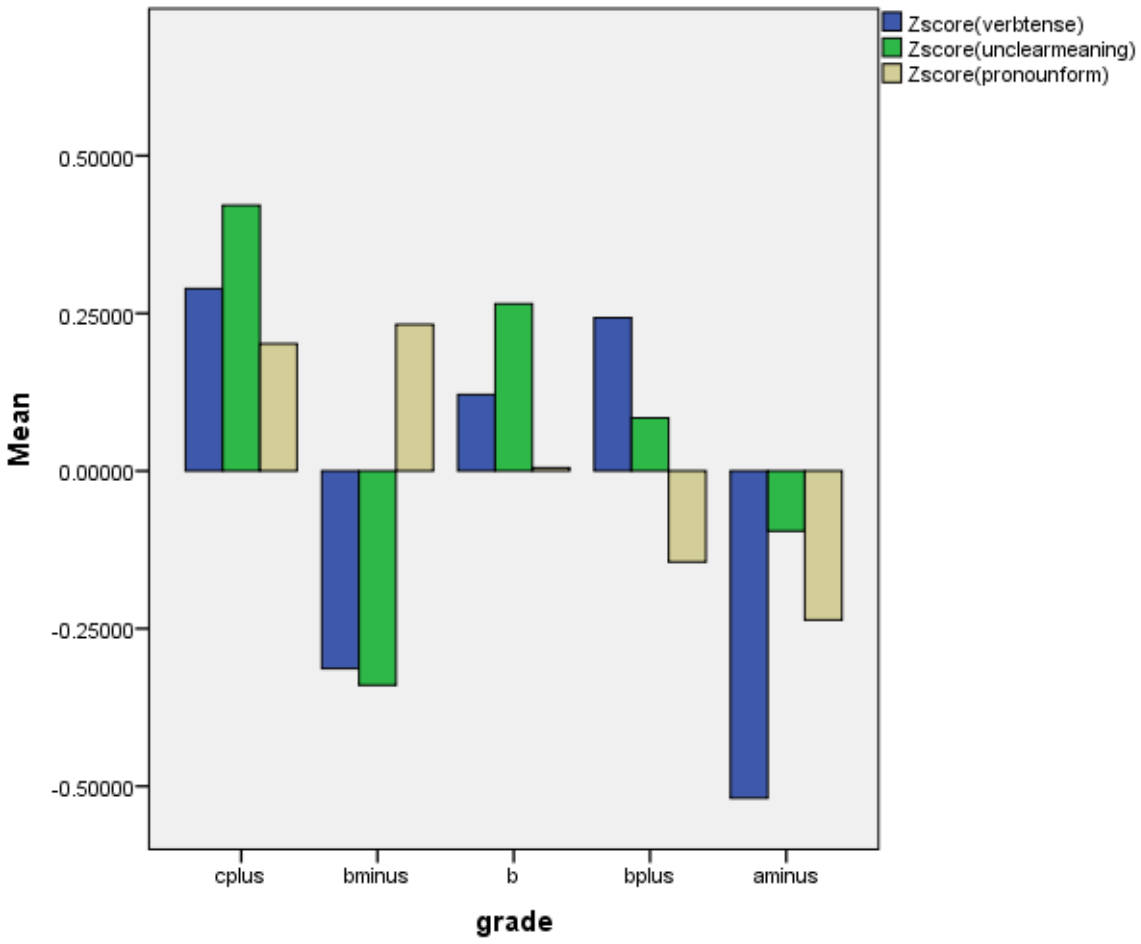

Figure 5. Errors as significant negative predictors of grade

Errors of verb tense were found in a variety of contexts, from the use of the progressive form in infinitive clauses ('Students who finish their normal schooling need to ***having [have]** a private tutor after school), to using the present simple tense when reporting on previous findings in spoken citations ('According to a journal article in Asia in 2015, the Korean Education Panel ***employ [employed]** a study…'). Instances of the code *Multiple errors leading to unclear*

meaning being used were numerous in the data, e.g. '…for the poorer students who are living in rural area, and they receive fewer tutoring, and it is ***wanna poor**', and apparently represent serious breakdowns in coherence for raters. *Pronoun form* errors also appeared to be particular salient for raters, e.g. 'He said that when ***he [his]** studies started to involve algebra…'

Fluencemes

Table 4 shows the normalised median frequencies of each fluenceme category (the raw frequencies are shown in Appendix C).

Table 4

Normalised (dis)fluency features by grade assigned (Median/Median Absolute Deviation, per 1,000 words)

Feature	A-	B+	B	B-	C+
Syntactic fluencemes					
Interrupted structures	1.95(1.26)	2.75(1.48)	2.30(1.72)	2.00(0.94)	1.80(0.71)
Dependent clauses[6]	4.35(0.71)	4.28(0.67)	4.15(0.45)	4.03(0.80)	4.37(0.85)
Lexical fluencemes					
Discourse markers	5.52(2.91)	5.44(1.94)	3.95(1.52)	2.74(1.72)	1.82(0.60)
Repeats	0.36(0.36)	0.37(0.37)	0.47(0.34)	0.38(0.38)	0.32(0.31)
Reformulations	0.69(0.42)	0.96(0.59)	0.87(0.64)	0.72(0.43)	0.80(0.39)
Prosodic markers					
Unfilled pauses	0(0)	0(0)	0(0)	0(0)	0(0)
Filled pauses	7.58(3.56)	10.77(7.31)	8.11(5.02)	6.72(2.55)	4.41(3.35)
Stressing	0(0)	0(0)	0(0)	0(0)	0(0)
Falling intonation	5.88(2.61)	6.00(2.69)	3.89(1.45)	3.22(1.32)	2.54(0.95)
Neutral tone	3.5(1.88)	3.12(1.42)	2.44(1.12)	3.6(1.71)	1.62(0.60)
Rising intonation	0.26(0.21)	0.27(0.27)	0(0)	0(0)	0(0)
Word lengthening	0.31(0.31)	0.48(0.48)	0.24(0.24)	0.51(0.51)	0(0)

[6] Tables 6 & 7, Calculated via the sum total of infinitive clauses, *that* relative clauses in object position, *that* relative clauses in subject position, past participle clauses, pied-piping relative clauses, present participle clauses, sentence relatives, split infinitives, subordinator *that* deletion, wh-clauses, wh-relatives in object position, wh-relatives in subject position, past participle deletion relatives and present participle deletion relatives, as tagged by the Nini (2015) MAT tagger

In addition to the categories shown in the appendix, we also calculated the average speech rate of each participant at each grade as prosodic fluencemes, by dividing the total number of words by the total time in seconds a participant spoke for, shown here in Table 5.

Table 5

Temporal prosodic fluencemes by grade

Fluenceme Category	A-	B+	B	B-	C+
Avg. word count	1208	780	780	657	462
Avg. speaking time	483	420	420	364	282
Words per second	2.1	1.9	1.8	1.7	1.6
Words per minute	127	114	110	105	99

The z-scores for the superordinate fluencemes categories (*syntactic, lexical and prosodic)* were each entered into a separate ordinal regression analysis. The model for syntactic fluencemes as a superordinate category was inconclusive (-2LL=744.0, $\chi 2$=2.36, p=.124, Pearson sig=.021, Nagelkerk r^2=.010), as it was for the subordinate categories (-2LL=741.8, $\chi 2$=4.58, p=.101, Pearson sig=.037, Nagelkerk r^2=.020). The model for lexical fluencemes as a superordinate category was acceptable, albeit weak (-2LL=735.4, $\chi 2$=10.91, p=.001, Pearson sig=.020, Nagelkerk r^2=.047) with these fluencemes suggested to be significant positive predictors of grade (β =.600, sig=.001, Exp β=1.82). Looking at the subordinate categories, the model (-2LL=705.2, $\chi 2$=41.15, p<.001, Pearson sig=.229, Nagelkerk r^2=.164) suggested that *discourse markers* were strong significant positive predictors of grade (β=.786, sig=<.001, Exp β=2.19) (Figure 6). The most frequent of these at A- grade were *'I think', 'actually', 'like', 'really', 'well', 'kind of', 'now', 'yeah'* and *'you know'.*

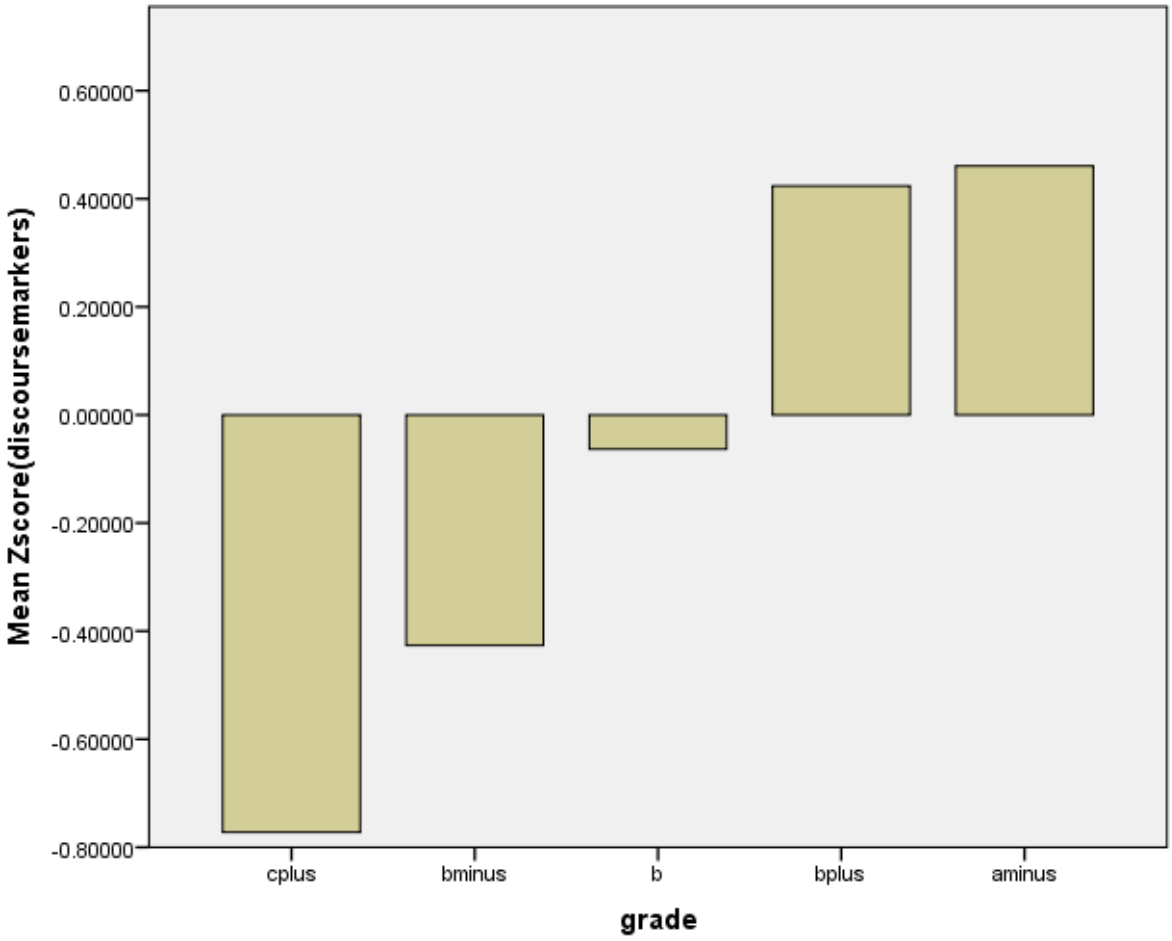

Figure 6. Discourse markers as predictive of grade

For prosodic fluencemes, the model for the superordinate category was valid (-2LL=702.9, $\chi 2$=39.3, p=.001, Pearson sig=.051, Nagelkerk r^2=.157), with this category a significant positive predictor of grade (β=.814, sig<.001, Exp(β)=2.56). The model for the subordinate categories (-2LL=678.1, $\chi 2$=68.3, p=.001, Pearson sig=.000, Nagelkerk r^2=.258) suggested that *speech rate per minute* (β=.997, sig<.001, Exp(β)=2.71) and *filled pauses* (β=.353, sig=.007, Exp(β)=1.42) were significant positive predictors of grade (Figure 7), with filled pauses a potentially controversial positive predictor (and apparently in contrast with the

normalised results). Such pauses may be seen as a disfluency marker on the one hand, or as a method of maintaining the turn of talk on the other.

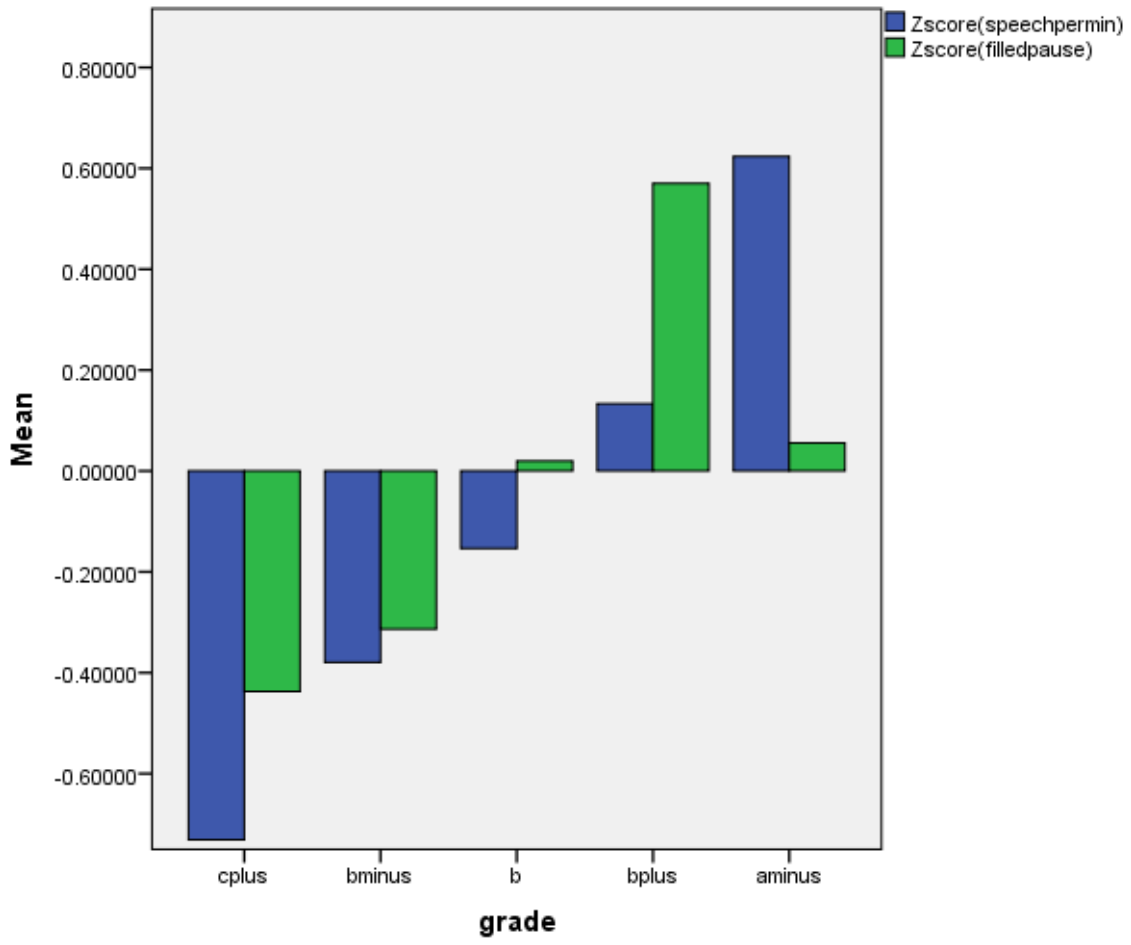

Figure 7. Prosodic fluencemes predictive of grade

Combined Features

In order to determine the most significant predictors of grade among all subordinate categories (Figures 8 and 9), a final ordinal regression analysis was performed, resulting in a reliable model (-2LL=578.7, $\chi 2$=167.6, p<.001, Pearson sig=.089, Nagelkerk r^2=.524). The most significant positive predictors, in order of significance, were *discourse markers* (β=1.03,

sig<.001, Exp(β)=2.80), *speech rate per minute* (β=1.02, sig<.001, Exp(β)=2.78), *code glosses* (β=.367, sig=.026, Exp(β)=1.43) and *label stages* (β=0.35, sig=.036, Exp(β)=1.39). The most significant negative predictors, in order of significance, were errors of *collocation/idiom* (β=-.393, sig=.016, Exp(β)=0.67) e.g '…because it really reflects several problems ***behind [within]** society', *multiple errors leading to unclear meaning* (β=-.470, sig=.007, Exp(β)=0.62), and *reformulations* (β=-.731, sig=.001, Exp(β)=0.48), e.g. '*…and I feel like* ***we could*** *–* ***we should*** *use Shanghai as a* ***representation*** *–* ***representative*** *for Asian society*'.

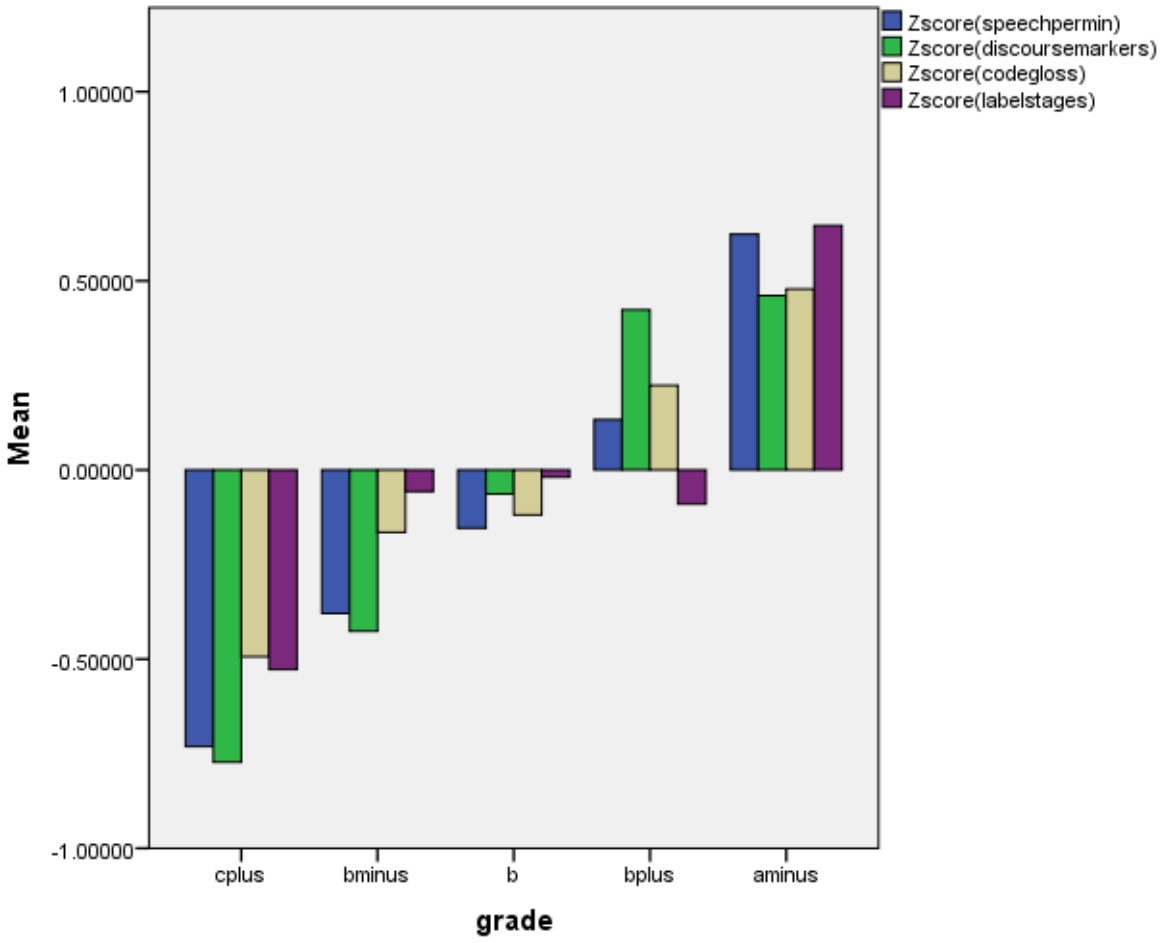

Figure 8. Combined positive predictors of grade

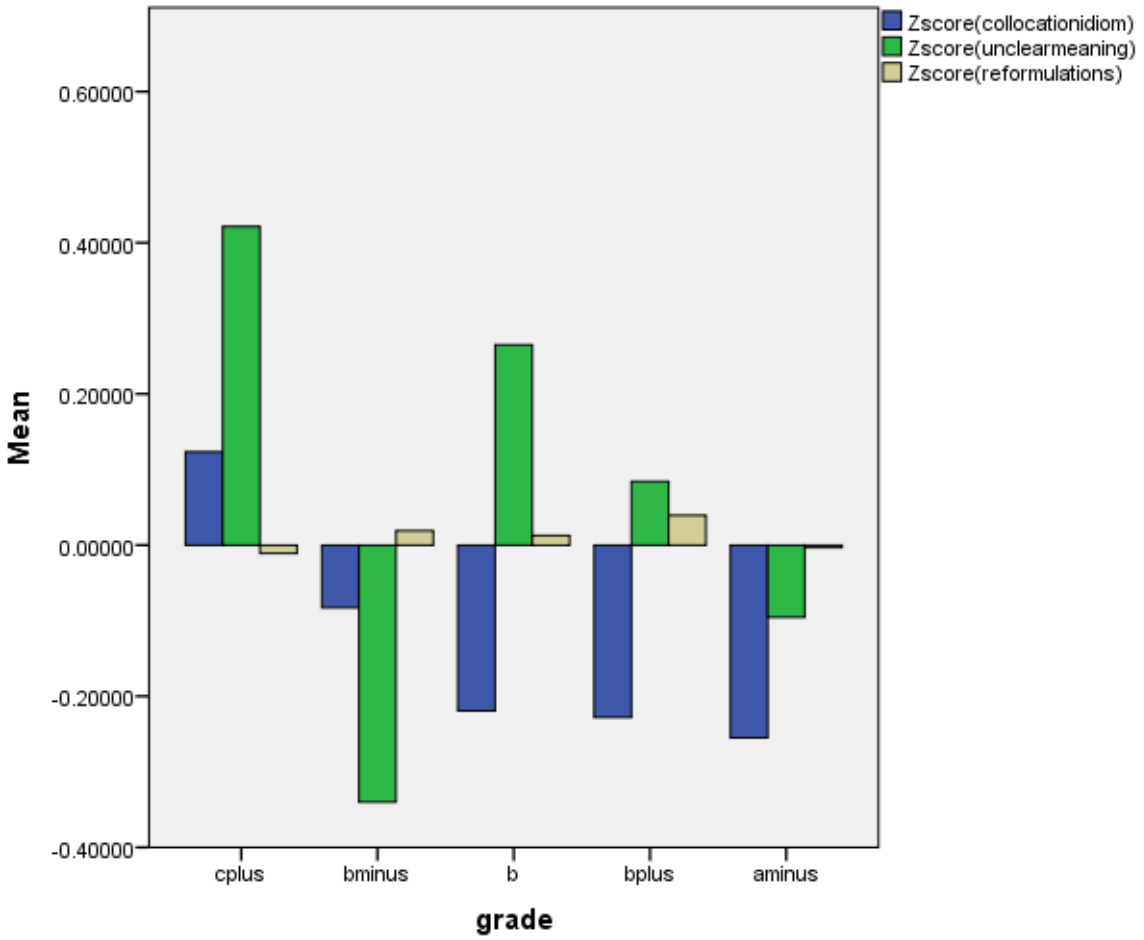

Figure 9. Combined negative predictors of grade

Discussion

This study has presented a fine-grained corpus-based analysis of a range of linguistic features potentially indicative of successful and unsuccessful performance on an L2 academic group oral assessment across the rating scale. The analysis and findings serve to highlight the overall usefulness of consulting learner corpora in defining the construct assessed by academic group discussion tasks. In addition, the detailed cross-sectional data afforded via this kind of corpus analysis serves as quantitative evidence of the linguistic features accompanying each grade awarded across the rating scale, which can be used in discussions of standardisation and

moderation for the raters involved. Finally, the data will, in future, serve as a critical source of information from which to re-write or at least refine the assessment criteria so as to provide both raters and test-takers with greater accountability and reliability of such criteria in terms of the expectations of successful academic discourse.

In summary, it was found that participants' frequent use of interactive and interpersonal metadiscourse during their assessment has a significant positive effect on raters' perception of success in academic discussion. It is certainly the case that metadiscourse plays a main role in two of the three assessment criteria for this particular assessment (*ability to support a stance*, *ability to interact with others*). As the data was analysed against the aggregate grade of the three assessment criteria rather than the individual subgrades, it is possible that the importance of metadiscourse to grading decisions was essentially predetermined by the nature of the assessment criteria. This is supported by the finding that most L2 errors or certain features of (dis)fluency were not necessarily primary indicators of student success or failure in this particular assessment context. However, other studies on this same oral assessment (Crosthwaite, Boynton & Cole, 2016, 2017) and other studies on L2 writing (Cumming 1990; Gebril & Plakans, 2014) suggest that raters are often unable to offer high grades for stance and interaction if comprehensibility (including errors and disfluency) is a significant concern, with students being penalised across all three assessment (stance, interaction, comprehensibility) criteria if this the case. This has strong implications in rater scale development in that instead of a separate *comprehensibility* criterion, descriptions of errors could be embedded with the stance criteria and fluency embedded with the interaction criteria.

This study also shows that metadiscoursal linguistic features are important in grading decisions across extended, interactional, academic L2 production. It appears that students' mastery of this range of structural and rhetorical linguistic features–many of which are explicitly

taught on the EAP course in question–is a *prerequisite* to being able to produce and defend a coherent and persuasive stance over 20-25 minutes, and to suitably impress the raters in doing so. In particular, when taking the analysis of the total combined standardised linguistic features into account, the interpersonal *engagement markers* and interactive *topic shifts* appear to be the primary positive indicators of successful student performance among the other metadiscoursal features analysed. Engagement markers "explicitly refer to or build a relationship with the reader [or listener]" (Hyland, 2005, p. 3), focusing the audience's attention or "including them as participants" (p. 4), including second person-pronouns, imperatives, and question forms. Topic shifts, or metadiscourse that "refers to discourse acts, sequences, or text stages" (Hyland, 2005, p. 3) are commonly used in lectures and seminars, both in L1 for L1 (Mauranen, 2001) and by L1 speakers addressing L2 audiences (Mauranen, 2010) to organise speech and draw attention to specific content. Our findings for the tertiary group academic oral discussion in this regard appears to be similar to those of Gan (2010) and Gan, Davison and Hamp-Lyons (2009) in that students who were more capable of engaging with others' ideas through a higher frequency of register appropriate suggestions, (dis)agreements, explanations and challenges and who could pursue and shift the topic of talk are seen by raters as more successful than those who lack the means or ability to lead the discussion in this manner. This ability is termed *confluence* by McCarthy (2010), and is considered to be "more crucial than strict grammatical accuracy" (p. 11) when considering successful L2 production. This corpus study thus provides tentative evidence to support the explanation inference in this assessment's validity argument (i.e., features of performances vary across the rating scale and concurs with the theoretical construct). Future corpus studies should collect more data, particularly at the lower end of the scale, to determine if this is variation of features is systematic across different levels of the scale).

One surprising finding in this study was that for the combined standardised features analysis, results suggest that errors of collocation/idiom are significant negative predictors of grade. Here, it is likely that malformed standard English idioms or fixed expressions (e.g. 'on the another hand') are particularly salient to raters amongst other types of errors commonly produced by L1 Cantonese or L1 Mandarin speakers such as verb agreement, plural marking or article errors. However, we have no current way of confirming whether this is the case without online experimental data. This finding though has implications again in rater training in that teachers should be made aware of the impact of these errors in overall rating, and to consider harsh rating of errors of this type in parallel to rating decisions for other error types. It could be the case here that raters are unfairly rating students for attempting idiomatic language, although the use of idioms generally correlates to a high degree of proficiency.

The analysis of the fluencemes involved in raters' appraisal of (dis)fluency suggests that it is not so much fluency at the linguistic level that is important to raters' grading decisions, but rather it is fluency at the temporal level, as evidenced by the significant positive linear relationship between grade, the number of words produced, the amount of time spent talking, and the concurrent speech rate per second/minute. The use of discourse markers and filled pauses to maintain fluency were also significant positive predictors of grade. When interviewing raters about this assessment, Crosthwaite, Boynton and Cole (2016) found that for students who did not talk for long enough or frequently made short turns, raters found the grading of student performance considered difficult:

> Student 2 spoke less and I think he was the hardest to rate for it too. Comparatively speaking, Student 4 is quite difficult for me to give a grade to because she didn't have lots of turns (Crosthwaite, Boynton & Cole, 2016, p. 23).

It seems the findings in this study reveal the importance of the quantity of this feature of academic group discussions generally taken for granted by raters and test takers.

Syntactic fluencemes did not appear to factor in grading decisions, which is perhaps not surprising given that raters are involved in top-down decision making, and lack the working memory and attention needed to recall syntactic information over extended periods. In other words, raters appeared concerned with the overall impact of what was said, rather than the frequencies of the syntactic structures involved in what was said. However, it may be possible that the normalising of these fluency features into instances per 1,000 words is not providing the full picture, in that in a 25-minute discussion, students may go through periods of relative fluency, along with other periods of relative disfluency. Raters alluded to this possibility when interviewed about this particular assessment:

> Like fluency, I think that particular feature [is hard], because some of the students, he or she may be quite fluent in certain point of time, but towards the end of the assessment or discussion, the level of fluency is not very strong, so I found [that] very difficult (Crosthwaite, Boynton & Cole, 2016, p. 25).

Conclusion

Although the present study focuses on only one assessment in only one context, primarily on one linguistic variety (L1 Cantonese/Mandarin speakers of L2 English), and uses a local rather than standardized rubric, the findings generated from the corpus should be of considerable

interest to those developing or using academic group oral assessments. In terms of the value of the present study, the findings have already had a significant impact on rater training for the assessment in question, in terms of raising awareness of the linguistic features that may be indicative of rating decisions across the scale, as well as their potential impact on the perception of successful group academic discussion more generally. The present study has also again shown the usefulness of learner corpora as a resource that serves to 'augment' human judgement across rating scales. The corpus has revealed insights into L2 production - and how that production is assessed - that could not have been derived with reference to the grades or raters alone. In addition, through the analysis of extralinguistic features as exhibited in spoken data, the study has also contributed to corpus-based studies on language assessment that have hitherto primarily focused on the written mode. A limitation of this study is that the selection of linguistic features for annotation is not exhaustive and based primarily on the assessment rating scale used in our context, and other features such as formulaic language, visual aspects of performance including gesture, multidimensional comparisons with other established values of register/genre suitability (e.g. Biber, 2006), positioning of features in the utterance, or existing automated L2 natural language processing techniques (e.g. Lu, 2010, 2017) were not taken into account. We have not even attempted to take into account the propositional content of what was produced by our test takers, whether this was both well-formed and fluent, or not. It is, of course, impractical or even inadvisable to make analyses across too many dimensions on the same dataset, and three superordinate linguistic features analysed here (metadiscourse, errors, and fluency markers) represent a broad enough range of features whose function can be applied to all instances of natural spoken language, either L1 or L2.

References

Aijmer, K. (2004). Pragmatic markers in spoken interlanguage. *Nordic Journal of English Studies*, *3*(*1*), 173-190.

Barker, F. (2010). How can corpora be used in language testing? In A. O'Keeffe & M. McCarthy (Eds.) *The Routledge Handbook of Corpus Linguistics* (pp. 633–645). New York: Routledge.

Berry, V. (2004). *A study of the interaction between individual personality differences and oral performance test facets*. Unpublished doctoral dissertation. King's College, University of London, UK.

Biber, D. (2006). *University Language: A Corpus-Based Study of Spoken and Written Registers*. Amsterdam, the Netherlands: John Benjamins.

Bonk, W. J., & Van Moere, A. (2004). L2 group oral testing: The influence of shyness/outgoingness, match of interlocutors' proficiency level, and gender on individual scores. In *Annual Meeting of the Language Testing Research Colloquium*, Temecula, California.

Christ, O. (1994). A Modular and Flexible Architecture for an Integrated Corpus Query System. In *Proceedings of the 3rd Conference on Computational Lexicography and Text Research (COMPLEX '94)*, Budapest, 1994. http://xxx.lanl.gov/abs/cs.CL/9408005.

Chuang, F. Y., & Nesi, H. (2006). An analysis of formal errors in a corpus of L2 English produced by Chinese students. *Corpora*, *1*(*2*), 251-271.

Coccetta, F. (forthcoming). Formulaic expressions in learner speech: New insights from the Trinity Lancaster Corpus.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching and assessment*. Cambridge, UK: Cambridge University Press

Crible L., Dumont A., Grosman I., Notarrigo I. (2016). *Annotation manual of fluency and disfluency markers in multilingual, multimodal, native and learner corpora. Version 2.0*. Technical report. Belgium: Université Catholique de Louvain and Université de Namur.

Crosthwaite, P. (2017). Does EAP writing instruction reduce L2 errors? Evidence from a longitudinal corpus of L2 EAP essays and reports. *International Review of Applied Linguistics in Language Teaching (IRAL)*, *56*(*3*), 315-344. DOI: 10.1515/iral-2016-0129.

Crosthwaite, P., Boynton, S. & Cole, S. (2016). Validating an academic group tutorial discussion speaking test, *International Journal of English Linguistics*, *6*(*4*), 12-30.

Crosthwaite, P., Boynton, S. & Cole, S. (2017). Exploring rater conceptions of academic stance and engagement during group tutorial discussion assessment. *Journal of English for Academic Purposes*, *28*, 1-13. DOI: 10.1016/j.jeap.2017.04.004.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing,* 7(*1*), 31-51.

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System*, *26*(*2*), 163–174.

Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013, June). Building a large annotated corpus of Learner English: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 22-31). Association for Computational Linguistics.

Evans, S., & Green, C. (2007). Why EAP is necessary: A survey of Hong Kong tertiary students. *Journal of English for Academic Purposes*, *6*(*1*), 3-17.

Fleiss, J. L. (1981). Balanced incomplete block designs for inter-rater reliability studies. *Applied Psychological Measurement*, *5*(*1*), 105-112.

Foster, P., & Ohta, A. S. (2005). Negotiation for meaning and peer assistance in second language classrooms. *Applied Linguistics*, *26*(*3*), 402-430.

Gablasova, D., Brezina, V., Mcenery, T., & Boyd, E. (2017). Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics*, *38*(*5*), 613-637.

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher-and lower-scoring students. *Language Testing*, 27(*4*), 585-602.

Gan, Z. (2012). Understanding L2 speaking problems: Implications for ESL curriculum development in a teacher training institution in Hong Kong. *Australian Journal of Teacher Education*, *37*(*1*), 43-59.

Gan, Z., Davison, C., & Hamp-Lyons, L. (2009). Topic negotiation in peer group oral assessment situations: A conversation analytic approach. *Applied Linguistics*, *30*(*3*), 315-334.

Gebril, A. & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing, 21*, 56-73, DOI: 10.1016/j.asw.2014.03.002.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum*. Somerville, MA: Cascadilla Proceedings Project.

Götz, S. (2013). *Fluency in Native and Nonnative English Speech* (Studies in Corpus Linguistics, Vol. 53). Amsterdam: John Benjamins Publishing.

Götz, S. (forthcoming). Describing fluency across proficiency levels: From 'can-do- statements' towards learner-corpus-informed descriptions of proficiency.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.). (2002). *International Corpus of Learner English*. Belgium: Presses Universitaires de Louvain.

Grice, P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.) *Syntax and Semantics: Speech Acts* (Vol. 3, pp. 41–58). New York, NY: Academic Press.

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, *9*(*2*), 122–159.

Hawkins, J. A., & Buttery, P. (2010). Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, *1*, *e5*. DOI:10.1017/s2041536210000103

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing*, *23*(*3*), 370-401.

Hyland, K. (2005). *Metadiscourse*. New York, NY: John Wiley & Sons, Inc.

Hyland, K. (2016). Writing with attitude: Conveying a stance in academic texts. In E. Hinkel (Ed.) *Teaching English Grammar to Speakers of Other Languages* (pp.246-265). London, UK: Routledge.

Iwashita, N. (1996). The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing*, 5(*2*), 51-66.

Just, M. A. & Carpenter, P.A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99* (*1*), 122-149.

Kennedy, P. (2002). Learning cultures and learning styles: Myth-understandings about adult (Hong Kong) Chinese learners. *International Journal of Lifelong Education*, *21*(*5*), 430-445.

Krashen, S. D. (1987). *Principles and Practices in Second Language Acquisition.* New York, NY: Prentice-Hall.

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

LaFlair, G.T. & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing*, *34*(*4*), 451-475.

Lee, I. (2008). Understanding teachers' written feedback practices in Hong Kong secondary classrooms. *Journal of Second Language Writing*, *17*(*2*), 69-85.

Legg, M. (2016). *An Exploration of the Voices of a New University Curriculum in Hong Kong: Implications for the Teaching of English for Academic Purposes*. (Unpublished doctoral thesis). Macquarie University, Sydney, Australia.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.) *Handbook of Research on Language Acquisition. Vol. 2: Second Language Acquisition.* (pp. 413-468). New York, NY: Academic Press.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(*4*), 474–496.

Lu, X. (2017). Automated measurement of syntactic complexity in corpus-based L2 writing research and implications for writing assessment. *Language Testing*, *34*(*4*), 493-511.

Matsuda, P.K & Jeffery, J.V. (2012). Voice in student essays. In K. Hyland & C. Sancho-Guinda (eds.) *Stance and Academic Voice in Written Academic Genres* (pp. 151-166). Hampshire: UK.

Mauranen, A. (2001). Reflexive academic talk: Observations from MICASE. In R. Simpson & J. Swales (eds.), *Corpus Linguistics in North America: Selections from the 1999 Symposium* (pp. 165–178). Ann Arbor, MI: University of Michigan Press.

Mauranen, A. (2010). Features of English as a lingua franca in academia. *Helsinki English Studies*, *6*, 6-28.

McCarthy, M. (2010). Spoken fluency revisited. *English Profile Journal*, *1*(*1*), 1-15.

McNamara, T. F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics*, *18*(*4*), 446-466.

Nicholls, D. (2003, March). The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference* (Vol. 16, pp. 572-581).

Nini, A. (2015). *Multidimensional Analysis Tagger (Version 1.3)*. [Software]. Available at: http://sites.google.com/site/multidimensionaltagger

O'Donnell, M. (2008, April). The UAM CorpusTool: Software for corpus annotation and exploration. In *Proceedings of the XXVI Congreso de AESLA, Almeria, Spain* (pp. 3-5).

Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing*, *26*(*2*), 161-186.

Oxford University Computing Services (2007). The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Park, K. (2014). Corpora and language assessment: The state of the art. *Language Assessment Quarterly*, *11*(*1*), 27-44, DOI: 10.1080/15434303.2013.872647

Pica, T. (1983). Adult acquisition of English as a second language under different conditions of exposure. *Language Learning*, *33*(*4*), 465-497.

Sancho-Guinda, C. & Hyland, K. (2012). Introduction: a context-sensitive approach to stance and voice. In K. Hyland & C. Sancho-Guinda (eds.) *Stance and Academic Voice in Written Academic Genres* (pp. 1-15). Hampshire: UK.

Schmidt, R. (1992). Awareness and second language acquisition. *Annual Review of Applied Linguistics*, *13*, 206-226.

Sinclair, J. (2004). *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins.

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*, *16*(*3*), 371-391.

Taylor, L., & Barker, F. (2008). Using corpora for language assessment. In E. Shohamy & N. H. Hornberger (eds.), *Encyclopedia of Language and Education, 2nd Edition, Volume 7: Language Testing and Assessment* (pp. 241–254). New York, NY: Springer.

Tsui, A., (2001). Classroom interaction. In R Carter and D. Nunan (eds.), *The Cambridge Guide to Teaching English to Speakers of Other Languages* (pp. 120-126). Cambridge: Cambridge University Press.

Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, *114*(*1*), 104-132.

Van Moere, A. (2007). *Group oral tests: how does task affect candidate performance and test scores?* Unpublished doctoral dissertation, Lancaster University, Lancaster, UK.

Van Moere, A. and Kobayashi, M. (2004). Group oral testing: does amount of output affect scores? Paper presented at the *Language Testing Forum*, Lancaster University, Lancaster, UK.

Xi, X. (2017). What does corpus linguistics have to offer to language assessment? *Language Testing*, *34*(*4*), 565-577.

Appendix A – Table 6, Speaking Test Assessment Criteria *(2015-16 Revised)*

	A+, A, A-	B+, B, B-	C+, C, C-	D+, D, D-	F
Ability to explain academic concepts and argue for a stance supported by sources 40% of grade	You can **always** clearly explain academic concepts. You are **always** able to argue for a critical stance with the support of valid academic sources where appropriate. You show an **excellent** ability to critically respond to / question other students' stance.	You can **almost always** clearly explain academic concepts. You are **usually** able to argue for a critical stance with the support of valid academic sources where appropriate. You show a **good** ability to critically respond to / question other students' stance.	You are **usually able to** explain academic concepts but **sometimes not** clearly. While you can **usually** argue for a stance, it is **not very** detailed or supported by any academic sources and is **usually** simplistic rather than critical. You show a **limited** ability to critically respond to / question other students' stance.	There is **only some evidence** of an ability to explain academic concepts and these are **usually** unclear. There is **only some evidence** of an ability to argue for a stance and when you do, it is **almost always** simplistic rather than critical and not supported by any academic sources. You show a **limited** ability to critically respond to / question other students' stance and when you do attempt to, the meaning is unclear. You are **mostly silent** throughout the discussion.	What you say is **almost always** unclear. You are **unable** to express a stance. You **never** critically respond to other students' stance. You have **no notesheet /** You have **plagiarized** your notesheet from another student. You **never** use any sources.
Ability to interact with others 30% of grade	You **never** dominate the discussion. You **never** read from your notes when expressing your stance. Your contributions to the discussion are **always** naturally linked to what has been said before. You **always** use active listening skills (nodding, eye contact etc.) when appropriate.	You **never** dominate the discussion. You **never** read from your notes when expressing your stance. Your contributions to the discussion are **almost always** naturally linked to what has been said before. You **almost always** use active listening skills (nodding, eye contact etc.) when appropriate.	You dominate the discussion in **one or two places**. You **sometimes** read from your notes when expressing your stance. Your contributions to the discussion are **usually** naturally linked to what has been said before. You **usually** use active listening skills (nodding, eye contact etc.) when appropriate.	You **often** dominate the discussion. You **often** read from your notes when expressing your stance. Your contributions to the discussion are **only sometimes** naturally linked to what has been said before. You **only sometimes** use active listening skills (nodding, eye contact etc.) when appropriate.	Your interaction skills are **too limited** to be able to successfully take an active role in the tutorial discussion.

Ability to communicate comprehensibly and fluently 30% of grade	You are **always** comprehensible. Mistakes with grammar / vocabulary are **infrequent** and **never** interfere with understanding. You are **always** fluent.	You are **nearly always** comprehensible. Mistakes with grammar / vocabulary are **infrequent** and **rarely** interfere with understanding. You are **usually** fluent.	You are **generally** comprehensible. Mistakes with grammar / vocabulary **occur throughout** but **rarely** interfere with understanding. You are **generally** fluent.	You are **only sometimes** comprehensible. Mistakes with grammar / vocabulary **occur throughout** and interfere with understanding **in multiple places.** You are **only sometimes** fluent.	Your spoken language causes **repeated and sustained** strain on the listener.

Appendix B

Table 7

Discourse Markers analysed

I think	You know
Sort of	Well,
Really,	I mean
You see	And so on
…or something	I suppose…
Actually,	Or anything
Like	And that sort of thing
Or anything of that sort	And such like
And things	Well actually
Well I think	Oh you know
You know…and things	I think, you know
I think actually	I think really
I think, you see	Okay?
, So,	Now,
By the way	That is,
I guess	oh
But anyway	

Taken from:

Aijmer, K. (2004). Pragmatic markers in spoken interlanguage. *Nordic Journal of English Studies*, 3(1), 173-190.

Redeker, G. (2006). Discourse markers as attentional cues at discourse transitions. In Fischer, Kerstin (Ed.), *Approaches to Discourse Particles*. Elsevier, Amsterdam, pp. 339-358.

Appendix C – Raw and normalised (per 1,000 words) frequency of linguistic features by grade

Table 7 – Raw frequencies of coded metadiscourse features by grade assigned

Feature	A-	B+	B	B-	C+
Interactive metadiscourse					
Code glosses	126	123	85	57	27
Evidentials	46	54	38	21	16
Frame markers					
a) Sequencing	82	105	68	26	9
b) Label stages	39	27	21	9	1
c) Announce goals	172	208	153	78	37
d) Topic shifts	474	650	382	176	77
Transition markers	2092	3015	1804	1050	539
Interpersonal metadiscourse					
Attitude markers	192	305	226	98	62
Boosters	844	1151	774	367	201
Self-mention	1106	1500	959	461	243
Engagement markers	1419	1647	1072	562	252
Hedges	927	1188	707	421	178

Table 8 – Raw frequencies of coded error features by grade assigned

Feature	A-	B+	B	B-	C+
Errors	1554	2375	1931	1200	711
Verb-specific errors	223	400	355	228	146
a) Modal	35	90	87	60	22
b) Missing verb	73	104	78	57	36
c) Verb form/tense	59	125	110	58	52
d) Verb agreement	56	81	80	53	36
Article errors (missing/extra/wrong form)	148	226	163	104	51
Noun-specific errors	134	220	198	115	62
a) Noun number	128	214	189	108	58
b) Noun possessive	6	6	9	7	4
Pronoun-specific errors	50	70	65	46	27
a) Form	4	5	8	4	7
b) Reference	46	65	57	42	20
General word choice errors	423	610	488	280	184
a) Collocation/idiom	144	195	135	90	63
b) Word form	60	82	96	55	20
c) Word choice	219	333	257	137	101
Sentence structure errors	230	325	260	175	79
a) Parallelism	24	24	15	21	7
b) Fragment	38	71	48	31	15
c) Subordinate/relative clause	38	50	44	24	16
d) Missing word	110	149	127	87	38
e) Word order	20	31	26	12	3
f) Local redundancy	204	316	268	151	86
Sentence transition errors	90	138	75	62	47
Multiple errors - unclear	41	60	57	34	27

Table 9 – Full normalised error features by grade assigned (Median/Median Absolute Deviation, per 1,000 words)

Feature	A-	B+	B	B-	C+
Errors	11.2 (5.09)	16.2(6.39)	12.6 (5.65)	11.7 (4.96)	13.8 (6.86)

Verb-specific errors	1.47 (0.68)	2.66 (1.43)	2.28 (1.2)	1.83 (0.97)	2.87 (1.09)
a) Modal	0.25 (0.25)	0.60 (0.35)	0.35 (0.35)	0.36 (0.36)	0.31 (0.31)
b) Missing verb	0.40 (0.40)	0.67 (0.40)	0.55 (0.31)	0.33 (0.33)	0.55 (0.24)
c) Verb form/tense	0.30 (0.30)	0.80 (0.40)	0.74 (0.48)	0.39 (0.39)	0.76 (0.24)
d) Verb agreement	0.36 (0.36)	0.55 (0.25)	0.61 (0.35)	0.45 (0.45)	0.68 (0.63)
Article errors	1.19 (0.80)	1.40 (0.80)	0.94 (0.49)	0.91 (0.53)	0.92 (0.35)
Noun-specific errors	1.13 (0.67)	1.30 (0.58)	1.44 (0.91)	0.74 (0.48)	1.04 (0.52)
a) Noun number	1.13 (0.75)	1.30 (0.58)	1.29 (0.94)	0.74 (0.46)	0.95 (0.50)
b) Noun possessive	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Pronoun-specific errors	0.28 (0.28)	0.36 (0.36)	0.31 (0.31)	0.28 (0.28)	0.37 (0.16)
a) Form	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
b) Reference	0.28 (0.28)	0.36 (0.36)	0.31 (0.31)	0.27 (0.27)	0.36 (0.36)
General word choice errors	2.85 (1.43)	3.97 (1.86)	3.39 (1.78)	2.65 (1.15)	3.46 (1.83)
a) Collocation/idiom	0.98 (0.41)	1.16 (0.61)	0.78 (0.54)	0.91 (0.38)	0.88 (0.33)
b) Word form	0.40 (0.39)	0.57 (0.34)	0.66 (0.45)	0.36 (0.27)	0.31 (0.31)
c) Word choice	1.46 (0.90)	2.28 (1.06)	1.59 (1.08)	1.17 (0.59)	1.87 (1.23)
Sentence structure errors	1.63 (0.74)	2.05 (0.85)	1.74 (0.94)	1.16 (0.57)	1.56 (0.84)
a) Parallelism	0.25 (0.25)	0 (0)	0 (0)	0 (0)	0 (0)
c) Fragment	0.27 (0.27)	0.47 (0.36)	0.35 (0.21)	0.33 (0.33)	0.13 (0.13)
d) Subordinate/relative clause	0 (0)	0.28 (0.28)	0.27 (0.27)	0.12 (0.12)	0.35 (0.26)
e) Missing word	0.90 (0.34)	0.85 (0.46)	0.73 (0.40)	0.71 (0.48)	0.94 (0.57)
f) Word order	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
g) Local redundancy	1.51(0.73)	2.05(1.06)	1.52(0.81)	1.36(0.80)	1.23(1.03)
Sentence transition errors	0.81 (0.51)	1.03 (0.55)	0.38 (0.38)	0.39 (0.27)	0.63 (0.50)
Multiple errors - unclear	0.30 (0.30)	0.35 (0.28)	0.36 (0.36)	0.26 (0.26)	0.36 (0.36)

Table 10 – Raw frequency of coded fluencemes by grade assigned

Feature	A-	B+	B	B-	C+
Syntactic fluencemes					
Interrupted structures	352	473	341	267	120
	1557	1849	1293	742	413

Dependent clauses[7]					
Lexical fluencemes	918	1087	719	336	159
Discourse markers	68	91	77	50	31
Repeats	111	145	124	84	49
Reformulations					
Prosodic markers	5	12	5	8	3
Unfilled pauses	1195	2025	1257	749	397
Filled pauses	8	4	11	2	0
Stressing	835	1084	699	365	195
Falling intonation	595	664	492	328	152
Neutral tone	54	99	50	36	19
Rising intonation	56	79	53	74	13
Word lengthening					

[7] Tables 6 & 7, Calculated via the sum total of infinitive clauses, *that* relative clauses in object position, *that* relative clauses in subject position, past participle clauses, pied-piping relative clauses, present participle clauses, sentence relatives, split infinitives, subordinator *that* deletion, wh-clauses, wh-relatives in object position, wh-relatives in subject position, past participle deletion relatives and present participle deletion relatives, as tagged by the Nini (2015) MAT tagger