Journal of Research in Applied Linguistics
ISSN: 2345-3303 – E-ISSN: 2588-3887 – https://rals.scu.ac.ir

Please cite this paper as follows:

Martín-Monje, E., & Barcena, E. (2024). Tutor vs. automatic focused feedback and grading of student ESP compositions in an online learning environment. *Journal of Research in Applied Linguistics*, *15*(2), 22-42. https://doi.org/10.22055/rals.2024.45636.3198

Research Paper

Tutor vs. Automatic Focused Feedback and Grading of Student ESP Compositions in an Online Learning Environment

Elena Martín-Monje[1] & Elena Barcena[2]

[1]Corresponding author, Foreign Philologies and Their Linguistics, Philology, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain; *emartin@flog.uned.es*

[2]Foreign Philologies and Their Linguistics, Philology, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain; mbarcena@*flog.uned.es*

Received: 25/12/2023 Accepted: 29/06/2024

Abstract

This article discusses the affordances and limitations of an automatic text evaluator in the context of the online teaching/learning of composition writing skills within a specialized linguistic domain, namely, English for Tourism. The system, named G-Rubric, was designed and built by an interdisciplinary team of linguists, psychologists, educationalists, and computer engineers to explore the applicability of data-driven language learning in education, for which it subsequently obtained several awards and distinctions. This article describes the adaptation process of G-Rubric to English for Tourism, contextualized in a distance learning university degree, and analyses its potential to substitute or complement frontline tutors in the task of revising and assessing student compositions. Two types of textual evaluation are provided by G-Rubric: numerical grading and focused feedback on form (writing) and function (content). Content evaluation is based on pattern-matching and machine reasoning against a specialized corpus and associated knowledge previously inserted in the tool as appropriate. The paper compares the performance of both tutors and system and proposes specific lines of research to gain insights into their optimal integration.

Keywords: Focused Feedback; Automatic Feedback; ESP; LSP; Writing; Online Learning; Data-Driven Learning.

1. Introduction

This article discusses the use of automated grading and feedback in text writing instruction. To this end, the potential of a multilingual tool called G-Rubric (Jorge-Botana, Olmos & Luzón, 2020; Martin-Monje & Castrillo, 2021) is analysed in the context of Tourism students' work on their essay drafts. This research is contextualized in a real massive distance learning environment, where L2 teachers encounter difficulties for providing individualised guidance in student writing. The task is not only time-consuming for the teacher but also frustrating for students because the tool interface and space constraints make the feedback provided difficult to understand and interpret. More importantly, from a pedagogical perspective, corrections are always executed on the final product, not the process, all of which limits the didactic value of the teacher's input.

G-Rubric was built as part of an award-winning UNED (the Spanish National Distance Learning University) project (https://eng.grubric.com/), as an automatic multilingual evaluator of written compositions. It was initially applied effectively for the assessment of general Spanish texts (Hernández Benítez & Santamaría Lancho, 2016). Subsequently, G-Rubric was adapted in the context of a didactic innovation project called 'Data-Driven Learning (DDL) for the improvement of the written competence in English: a pilot experience in MOOCs, degrees, and master courses' (UNED, 2021, https://www.uned.es/universidad/inicio/en/institucional/IUED/innovacion-docente/grupos-innovacion/grupo-9/proyectos.html), to be used in online ESP courses following the latest developments in DDL. Therefore, the aim here was two-fold: firstly, as its title suggests, to explore the potential of DDL (Boulton, 2017) for the improvement of free writing competence, taking into account the specificities and divergences from standard language that characterize each

specialized linguistic domain; and secondly, to gain insights on the affordances and limitations, integration and other practicalities, of the use of automatic text evaluators, both for teachers and students from mass distance learning courses.

After compiling and implementing an English subject-specific corpus in G-Rubric and incorporating related specialized knowledge, the system was piloted with students from the 1st year "English for Professional Purposes" subject of the degree in Tourism at UNED (Barcena, Martin-Monje & Jordano, 2016) during the academic year 2020-21. The students were required to submit a report on one of the topics that made up the syllabus, for which they were provided with preliminary guidelines and relevant web links. The tool matched compositions submitted by the students against an embedded corpus-based model answer and associated knowledge and produced both a numerical score and a more qualitative feedback report. The G-Rubric assisted students in the learning process because it presented them with data and information on their performance of both form and content of each composition draft. The students were expected to improve their writing in an iterative manner, by working cyclically on composition drafts and using each feedback report to produce further refined versions, according to their autonomous judgement and expectations. This article delves on the findings of the quasi-experiment undertaken with the G-rubric tool in the abovementioned case scenario and undertakes a contrastive analysis of human vs. machine assessment, raising new research goals on the improvement of writing quality by machine, given state-of-the-art technology, and the optimal integration of its capabilities with those of professional teachers.

2. Literature Review

2.1. LSP Teaching

Languages for specific purposes (henceforth, LSPs) can be defined to deal with the communicative needs and practices of particular social groups. LSPs share all the universal properties of standard languages, such as unlimited generative capacity and completeness, in relation to the corresponding subfield of reality that they cover (Kittredge, 1982). Furthermore, they are generally characterized by features that include discursive systematicity and patterns of usage, a tendency towards syntactic closure or restrictiveness, a focus on content, a drastic reduction of lexical and structural ambiguity, and a prominent role of terminology for semantic reference (Barcena, Martin-Monje, & Jordano, 2016). However, as these authors claim, there are noticeable differences among the plethora of specialized (professional, academic) domains that correspond to LSPs, in terms of the degree of closedness of the practice community, main channel of communication, register, and so on. Tourism as an LSP is comparatively rather open and heterogeneous in the sense that it covers a vast number of subfields (gastronomy—culinary tourism; health and wellbeing—medical tourism; art and history—cultural tourism; etc.) with fuzzy delimitations with respect to standard language and other related domains. In this sense, unlike other subgenres, the linguistic domain of Tourism can be said to be on the opposite side of Harris' (1968) mathematical conception of sublanguages as computationally tractable subsystems.

Although LSP teaching has often been taken as a unitary methodology (Dudley-Evans & St. John, 1998; Hutchinson & Waters, 1987), there is a repertoire of ways in which any given specialized linguistic domain can be taught (Ghanbar & Rezvani, 2024). As in general language courses, LSP designs include topic-based, structural/situational, functional-notional, discourse (rhetorical), and skill-based. They are learner-based to various degrees, and attention is paid to the concepts of relevance and appropriateness. At the lowest level, the students are trained by giving high priority to the language forms they would normally require in order to communicate in the given field and, in turn, low priority to rare forms. At the highest level, explicit needs analyses are carried out and ad hoc syllabi are built for intensive tailor-made courses. LSP students are typically adults (either in tertiary education or in professional contexts) with an intermediate to advanced knowledge of the corresponding standard language (Barrantes Montero, 2009) and the world in general. Beyond their comparably more mature psychological profile than that of younger students following the standard academic path, as learners they have been recognised to exhibit comparatively high motivation and low anxiety levels, all of which the teacher can rely on (Mitchell, Myles, & Mardsen, 2019). Materials, contents, and activities are taken or adapted from the corresponding field of reference, and evaluation goes beyond standardised tests and certification, into capability enhancement related to real world usage.

Given the role played by English in international communication, it is not surprising that there is an outstanding number of courses and contents for the teaching of this language when used in specialized spheres of reality, and a substantial body of research literature around them (Dağ Akbaş, 2021; Flowerdew, 2016). Therefore, just as many

students of English view the need to learn a rather standardized culture-free form of the language in order to meet the demands and practicalities of global communication, they are also expected to command the jargon that pertains to their present or future occupation. Experts (Hyland, 2022; Swales, 2000; Trace, Hudson & Brown, 2015) note that just as students with the same level of English as a second language may be more proficient at some processes than others (e.g., speaking and writing), they may also perform differently across the spectrum of genres. The authors claim that the regulations and constraints that underlie the most specialized domains often lead to better levels of accomplishment. Conversely, the eclectic range of texts and discourses that fall into domains like Tourism and the rather open nature of most of them render such fields far from such an inherent advantage found in the teaching/learning of other LSPs.

A relevant modality in many spheres of communication is composition. This is taken here generically to include various text-types, such as essays, articles, reports, and plans; and genres like expository writing, argumentative writing, narrative writing, descriptive writing, analytical writing, and persuasive writing. They have a tripartite structure (introduction-body-conclusion) in common and a double axis of subject matter and purpose that leads the structuring of sequences of data, facts, and/or opinions. The updated Companion Volume of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2018) singles out reports and essays as a separate category within the production of written texts, within each specific scale. According to the CEFR, these types of compositions have distinctive features in terms of content and extension and complexity of discourse. Content can range from familiar subjects and routine factual information to complex academic and professional topics, and the writer must distinguish their viewpoints from those in the sources. The same applies to complexity of discourse, from linking sentences with simple connectors to "smoothly flowing expositions with effective logical discourse" (p. 68). The scale provided by the CEFR for B1 (the level required at university entrance in Spain) is shown in Figure 1:

	Reports and Essays
B1	Can produce short, simple essays on topics of interest. Can produce a text on a topical subject of personal interest, using simple language to list advantages and disadvantages, and give and justify their opinion. Can summarise, report and give their opinion about accumulated factual information on familiar routine and nonroutine matters within their field with some confidence.
	Can produce very brief reports in a standard conventionalised format, which pass on routine factual information and state reasons for actions. Can present a topic in a short report or poster, using photographs and short blocks of text.

Figure 1. *Descriptors for CEFR B1 Level in Reports and Essays*

Following the CEFR's notional-functional approach, two types of competence criteria can be distinguished, the majority of which are related to content and the rest to form. Prominence is given to the level of discourse, communicative functions like exposition and argumentation, and rhetorical aspects such as relevance and adequacy (to domain/genre/text-type). Structural criteria considered are complexity and length. Lumley (2002) distinguishes the following four writing performance criteria: task fulfilment and appropriacy, conventions of presentation, cohesion and organisation, and grammatical control. Rubrics for written production include that of Fleckenstein, Keller, Krüger, Tannenbaum, and Köller (2020), which is based on the B2 CEFR level and covers form (correctness, adequacy, and variety of lexical choice and grammatical structure) and content (addressing, organising and developing topic and task with effectiveness, coherence, and appropriateness). Fox and Arteneva (2017) defend the need for the adoption of an ESP-based approach for diagnosing the need for academic support in entrance writing. The variables considered by these authors are: fluency (defined by criteria relating to organisation, paragraphing, length, and general academic style), content (defined by criteria relating to data interpretation and analysis), and form (defined by criteria relating to grammar, vocabulary, and spelling).

2.1. Providing Focused Feedback in LSP Writing

According to many authors (Bird, Downs, McCracken & Rieman, 2019; Harris, 1984; Smelstor, 1978), the teaching of composition writing typically involves guidance through four steps, namely: prewriting, writing, revising, and proofreading. The first step, when students gather information and begin to organize it into a cohesive unit, is generally considered as the most important one. Writing is complex in that it requires attention to be paid to various levels in parallel: orthography and punctuation, layout, lexis and semantics, morphosyntax, and discourse. Revision deals with

organization, audience, and focus. The main aspects that need to be reviewed are intelligibility, lexical accuracy, grammatical correlates, coherence, and cohesion. Proofreading deals with surface errors and trying to view the writing as a whole message and as readers eventually will.

There are many ways in which a teacher may guide their students through the first two steps. Strategies for prewriting, like brainstorming and clustering, aim at helping the students activate and forward their mental structures on the topic in hand, and formulate preliminary ideas, while boosting their motivation and confidence in their significance. The general principles of textual composition can be presented to the student group with the help of dedicated materials with illustrative purposes. Once individual drafts have been produced, the number and heterogeneity of possible errors—particularly those of a linguistic and communicative nature—render group revision insufficient because personalized feedback is hampered by the unbalanced ratio between teachers and students. Furthermore, because all the errors and mistakes can hardly be tackled in one attempt, students may require the revision of several drafts. The authors claim that it is reasonable to predict that at this point the instructional process will particularly benefit from the incorporation of a feedback tool.

Feedback in L2 teaching consists of the transmission of expert formative judgment on student performance. A question that has received considerable interest from language teachers and researchers is what type of feedback is optimal for second language learners. Two major types are distinguished in the literature: focused feedback and unfocused feedback. The former targets a subset of error types, whereas in the latter any type of error can be the object of identification. Characterizations of these two options based on theoretical, methodological, and pedagogical criteria are ultimately relativistic and continue to be controversial even to this day (Kao & Reynolds, 2016; van Beuningen, 2021). However, most studies value the effect of prioritising a small subset of goals, depending on the level of knowledge, starting from sheer intelligibility upwards.

There are four major arguments in favor of focused feedback, according to Fletcher-Wood (2021). The first one is cognitive load reduction. Students can only process a handful of ideas at once. After all, as Sadler (2010) says, feedback is new information for them to parse and process, so it is not advisable to introduce too many new unsequenced concepts together at the risk of a subset of them being inevitably filtered out. The second argument is the promotion of behavior change. Making a lasting habit takes effort and support over a period of time. If a student is given too many items to work on, they will struggle to practise all of them, and the teacher will have similar difficulties recalling them in future sessions. Third, confidence and self-efficacy. A reduced list of issues feels like an achievable task. The individual's belief in their executive capacity to produce specific performance attainments is, according to Bandura (1997), a fundamental aspect for potential improvement. The fourth and last argument in favour of focused feedback is workload reduction. Providing written feedback is extremely time consuming. Assuming the student will only assimilate a percentage of the feedback provided at a time and, subsequently, at least part of it will need to be periodically repeated, which renders the process unsustainable. The less feedback offered, the quicker the teacher can undertake the task, and the more frequently they can do so.

In the last decades, consensus has been reached in the research community that error correction is a fundamental part of the composition teaching process and that the key lies more in the selection of error types than in the way of delivering them. However, regarding the selection of error types for focused feedback, its effects on students' writing quality are inconclusive, partly because of the spurious nature of short-term effects. It is still uncertain which feedback content is beneficial to students, even though numerous types of feedback have been contrasted and analysed to date (Ellis, 2009; Hyland & Hyland, 2006).

Automated feedback can play an important role in teaching/learning ESP writing in the following senses: Firstly, at the most fundamental level, it compensates the unbalanced ratio between teachers and students in many institutions and, in that sense, can be said to lift a burden off the teacher, who can focus on the most creative and least computationally tractable aspects of their job. Secondly, once error types are identifiable by the tool, reliability is maximum, and all instances will be marked. Thirdly, if presubmission versions are revised by machine, the working time and effort underlying the task in hand is expanded exponentially. Furthermore, leaving the teacher aside during this phase helps students' motivation and sense of autonomy, and lowers the anxiety filters that often emerge with teacher intervention (Kurdi, 2018).

Many computer-based systems have been developed so far to support writing teaching and assessment in various ways (Allen, Jacovina, & McNamara, 2015; Limpo, Nunes, & Coelho, 2020). Three main categories are usually identified: grammar and vocabulary checkers, intelligent tutoring systems, and automated evaluators. Despite there being a targeted user profile for most tools, their functionality is often flexible and overlap considerably. Furthermore, all three types have in common that they act upon a completed writing product, not the process. Checkers, like Grammarly and Ginger (to mention two of the most popular ones), pretend to offer interactivity and activate the students' high-order metal skills, for example, reflection through oriented questions. As can be expected, their rate of success is related to the degree of formal restriction of the domain (e.g., organic chemistry). Two examples of intelligent tutoring systems targeted at supporting writing are eWritingPal (Roscoe, Allen, Weston, Crossley, & McNamara, 2014) and ThesisWriter (Rapp & Kauf, 2018). Some automated evaluators provide scoring (e.g., e-rater; Attali & Burstein, 2006), but others are formative and provide suggestions for correction, such as *Criterion* (Link, Dursun, Karakaya, & Hegelheimer, 2014) and AWA/AcaWriter (Knight, Martinez-Maldonado, Gibson, & Buckingham Shum, 2017; Knight, Shibani, Abel, Gibson, & Ryan, 2020). After analysing over a hundred feedback tools, authors like Keuning, Jeuring, and Heeren (2019) note that they have diversified greatly in the 21st century, and they keep on doing so. Yet, they are still designed mostly to identify mistakes, rather than correct the underlying knowledge/skill gap and scaffold the subsequent building process. These authors also point out that there is a noticeable mismatch between tool performance and teacher needs, and scope for improving integration in a blended didactic context.

There is a rather fragmented and inconclusive body of research exploring the most reliable configurations of automated error identification (diversity, types, sequences) and which of them have the highest impact on the quality of results. These working areas cover generalist studies of the long-term value of error correction with mostly positive results (e.g., Miao, Badger & Zhen, 2006; Sachs & Polio, 2007) and equally positive local studies on the impact of specific configurations in single contexts of application (Stevenson & Phakiti, 2019). Other longitudinal research projects of a more critical nature have focused on several related methodological and contextual issues (Bitchener & Knoch, 2008; Bruton, 2009; Ellis, 2009). These authors claim that the most common type of automatic feedback is still superficial in nature and aims at improving the accuracy of the final product, not its readability or providing an understanding of the students' composition process (Crossley, 2020). Comparative studies between teacher feedback and automated feedback suggest that the former has a more positive effect on the psychological aspect of writing (i.e., perceived usefulness and ease of use), while the latter may be more effective in helping develop writing proficiency in the long run (i.e., cognitive aspects) (Wang & Han, 2022). While authors like these claim that, at present, the main scope for improvement of automated feedback lies in bettering the psychological student experience, others focus their attention on teachers' perceptions because of their direct implications for institutional stakeholders and policy makers (Wilson, Fudge, Ahrendt, & Raiche, 2021).

3. Methodology

G-Rubric[1] is a multilingual generic learner-centred tool that offers open writing training with multiple assessment opportunities. Its development was coordinated by Jorge-Botana, Olmos and Luzón (2020) at UNED in order to formally analyse and evaluate discursive text in different quantitative and qualitative ways. According to its authors, G-Rubric offers automatically assessed practice that can be adapted to any subject, educational level, degree of difficulty, length and complexity. It follows a consolidated instructional model based on academic writing production and formative assessment. Each academic task is developed following an instructional sequence that the student can go through up to six times (cycles) so that, based on the results of the automatic evaluation of a given response/version, they can improve the subsequent performance with the guidance and assistance provided. G-Rubric is online, universally accessible and permanently available. The results of each automatic assessment are instant and systematic: they include a score for the adequacy and correctness of the content of a given student answer, an assessment of the appropriateness of its written expression, and a diagnosis on what has been analysed with guidance for improvement.

Partly due to the capability of G-Rubric to be customised for the evaluation of specialised texts, it was subsequently adapted and piloted by the authors in a real educational context in order to explore the potential of DDL for providing automatic assistance for student ESP essays, incorporating the latest research in this field. At the same time, there was an evident need to improve the students' attitude towards composition writing and, hence, raise their participation rates in such a task, particularly low-level students. There was another research goal of an applied nature

related to finding support to teachers' highly time-consuming tasks in massive online educational institutions, like free writing revision, with a view to replacing/ complementing them. The present article specifically addresses the following question in relation to the authors' research work of application and adaptation of G-Rubric for English student compositions in the field of Tourism: How do human and automated focused feedback and grading compare in terms of academic effectiveness and usability?

3.1. Scenario and Participants

The project designed and implemented around this and other related research questions was carried out over two academic years at a distance learning tertiary institution. Apart from the inherent complexities of such a methodology, there were some added challenges, namely the heterogeneous academic level of students in the English for professional purposes subject from the first year of their degree in Tourism, and also the massive nature of the course, with an unbalanced teacher/student ratio with over a thousand registered students and only three teachers. The authors' hypothesis was that this semantic evaluator of written texts drafted by the students would benefit from the closure and systematicity of LSPs and be adapted to help both teachers and students of this type of disciplines overcome the abovementioned challenges. If so, the teachers would be able to use the tool as an aid/substitute to provide personalised focused feedback, and students would be encouraged to focus extensively on the work of redrafting, which could result in both a cost-effective teaching strategy and a more effective learning experience.

The criteria for grading written assignments in the target subject of the project were incorporated by the teachers into a rubric, since compositions were to be evaluated by a considerable number of tutor-teachers of this subject (henceforth referred to simply as tutors to distinguish them from the teachers at UNED's headquarters), who are distributed around more than 50 regional centers over Spain (and assigned a group of students at the beginning of each academic year on the basis of geographical proximity). The more balanced student-tutor ratio (in comparison to student-teacher ratio) enables teachers of English for professional purposes to be supported by the corresponding tutors in individual didactic tasks, as part of the task division established autonomously by the teaching teams of each subject at the central headquarters of the university. Following institutional methodological recommendations, in English for professional purposes, the following evaluation schema was established: 60% of the grade would correspond to the final closed written exam, and 40% would be divided into two blocks of tasks (20% was equally divided between two free writing compositions and the rest consisted of oral comprehension and production tasks), because tutors can assist with formative evaluation tasks during the term. With this schema, not only was the grade distributed across different language-communicative competence, but also students could be provided with formative evaluation at different points in the course as a didactic strategy.

The project was designed around the second of the two written tasks of the course, which corresponded to 10% of the final grade, and was undertaken using the tool of the institution's official platform, which centralizes most of the academic activity of the university and where the virtual courses are. The rubric developed by the teaching team served the double purpose of granting sufficient homogeneity to tutors' evaluations and providing students with a similar level of formative feedback on their composition writing skills. See Figure 2 with the criteria grid for grading written assignments:

CRITERIA FOR ESSAY EVALUATION	LEVEL OF PERFORMANCE			
	D (0-0.5)	C (0.5-1)	B (1-1.5)	A (1.5-2)
• Successful accomplishment of the contents of the units				
• Adequate use of the vocabulary and expressions studied in the units				
• Adequate use of the grammar rules learned in the units				
• Fluency				
• Discourse cohesion (use of connectors and other cohesive devices)				
TOTAL SCORE				

Figure 2. *Rubric Proposed for Compositions in Tourism* (Own Source)

As can be seen, the criteria were clustered in five items, and this, together with decimal assessment, facilitated the usability of the rubric by tutors and its understanding by students. The scoring was structured as follows: 40% was

devoted to content (the first two criteria) and 60% to the use of English (the last three). Therefore, there was considerable balance between content and form and, while not completely explicit in the rubric, the full descriptors in the tutors' guide uploaded to the virtual course at the beginning of each academic year showed their direct connection with the ESP approach, so the evaluation could be adapted to the features of the specialized linguistic domain in hand.

In contrast with the above, the type of evaluation performed by G-Rubric consisted of an overall assessment on the quality and appropriateness of the writing, and a numerical score (between 0 and 10 points) referred to its content, including a diagnosis of the ideas developed in the submitted work and the vocabulary used, and suggestions on how both ideas and vocabulary could be further developed. Both sets of criteria unfolded in such a way that a correspondence could be established with the tutors' rubric above:

- Writing (Form): Its evaluation was provided as a percentage and the maximum that could be obtained is 100%. Within Writing: Three aspects were considered (which were already in the initial version of the tool):
    - Correctness: It corresponded with the criterion "adequate use of the grammar rules learned in the units" of the IFP rubric.
    - Clarity: It corresponded to the criterion "discourse cohesion" in the IFP rubric.
    - Lexical amplitude: Although it does not correspond exactly, this was roughly related to the "fluency" criterion of the IFP rubric.
- Content. This part was subject of domain adaptation. Within Content, two aspects were considered:
    - Ideas: This corresponded to the criterion "successful accomplishment of the contents of the units" of the IFP rubric.
    - Vocabulary: This corresponded to the criterion "adequate use of vocabulary and expressions studied in the units" of the IFP rubric.

When comparing the grade distributions of assignments assessed by tutors and by the computer tool, it should be remarked that there was a considerable parallelism between both sets of assessment criteria: 40% of G-Rubric's grade was devoted to content (the first two criteria in the tutors' rubric) and 60% was devoted to use of form (the other three criteria).

Figure 3 shows the G-Rubric interface. It should be emphasized that G-Rubric was intended as a tool for formative assessment, that is, to help students improve their written production in successive attempts, taking into account both the numerical scoring and the feedback provided by the tool. Based on the information provided, students could decide to redraft their assignment and submit a revised version up to six times:

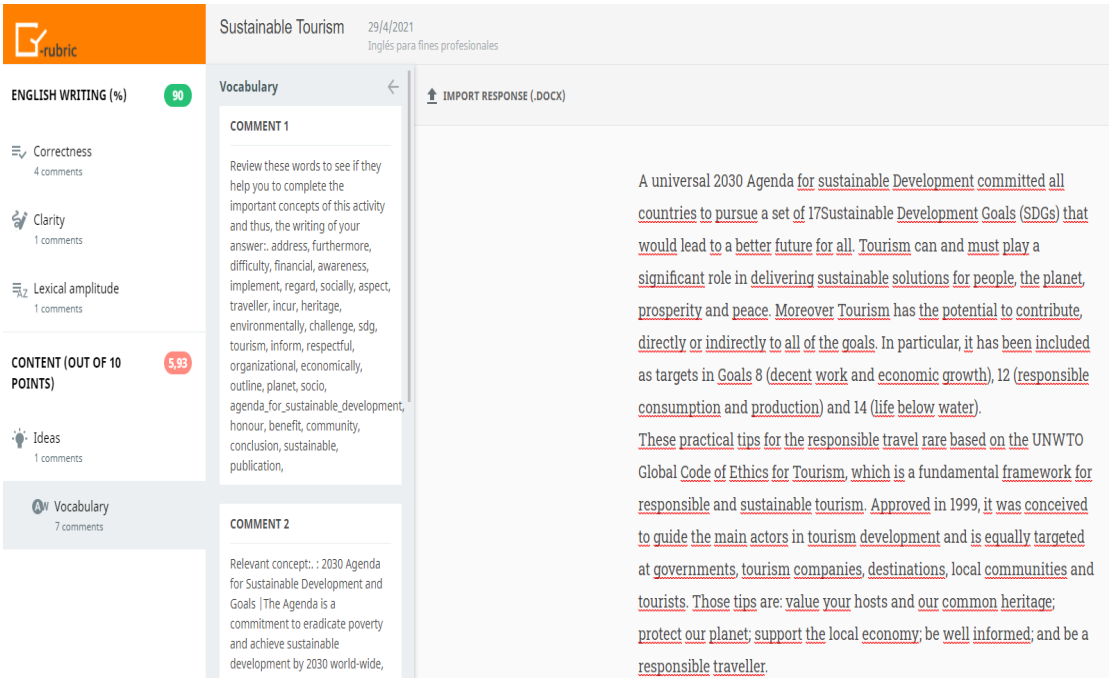


Figure 3. *Screenshot of the Feedback Provided by G-Rubric*

As explained above, the assignment under study was equivalent to 10% of the final grade. An average of less than 20% of the students do the assignment on a yearly basis, despite being aware that the mark for it will count as zero if they fail to hand it in. One of the assumptions that was implicitly tested in this research work was that online students, the majority of whom are mature and attracted by the anonymity granted by the institution, might resent tutors' continuous evaluation of their work, which has aspects in common with traditional classroom-based education. However, the edition when G-Rubric was made available to the students, of the 1,025 students that conformed the group, only 238 did the assignment (10.92%) and 26 used the tool to prepare it (2.54%), which supports the widely observed tendency of UNED students to simplify learning paths to focus on those activities that are most relevant for their final grades. Appendix B shows the table with the scores obtained by those students who decided to do the proposed written assignment and use G-Rubric, which supports the authors' interpretation of the existence of certain self-consciousness on the part of low-level students that prevented them from undertaking the assignment, at the expense of lowering their global grade for that subject.

3.2. 1st Phase: Adaptating the G-Rubric System to the Development of ESP Composition Writing

During the first year of the project, the theoretical framework on DDL was established. In its most basic form, DDL is an approach to language learning which propounds that in order to learn the target language effectively, students should access as much authentic linguistic data as possible (Warren, 2015). In line with this, the use of linguistic data, corpora, and DDL have led to the identification of lexico-grammatical patterns that support more effective language learning by improving written production and enhancing learner-centredness and autonomy (Boulton, 2017; Curado, 2017; Boulton & Cobb, 2017; Karpenko-Seccombe, 2020). These findings were expected to be accentuated in ESP due to the closure, restriction, regularity, and systematicity that characterize specialized linguistic domains to various extents (Díez-Arcón & Martín-Monje, 2021). Based on this conceptual framework, the existing G-Rubric tool was adapted to English following the principles of DDL, to be used for ESP teaching and learning.

Subsequently, as shown in Figure 1 below, the new version of the G-Rubric tool was developed. The process involved seven steps. Firstly, the different layers of the software were adapted to English, since the tool only existed in Spanish beforehand. Secondly, the semantic space in English -where the corpus would be hosted- was created. Thirdly, an ESP corpus was compiled focusing on "Sustainable Tourism", which was the chosen topic for this project. The corpus comprised of 52 web pages related to sustainable tourism, 29 digitized documents on the same topic and a selection of the digital content of the subject coursebook which covered areas related to sustainable tourism. The fourth step was related to computer programming (AIML coding) and translating the instructions that the interface would show. Then, the student task was designed. The students had to draft a written report of around 300 words on the topic of Sustainable tourism. They were provided with scaffolding: detailed guidelines, websites with relevant information and access to the G-Rubric tool, which gave them focused feedback on the successive versions of the report which they submitted. Next, the new version of the G-Rubric tool was assessed by an expert in Corpus Linguistics and finally it was tested by the students.

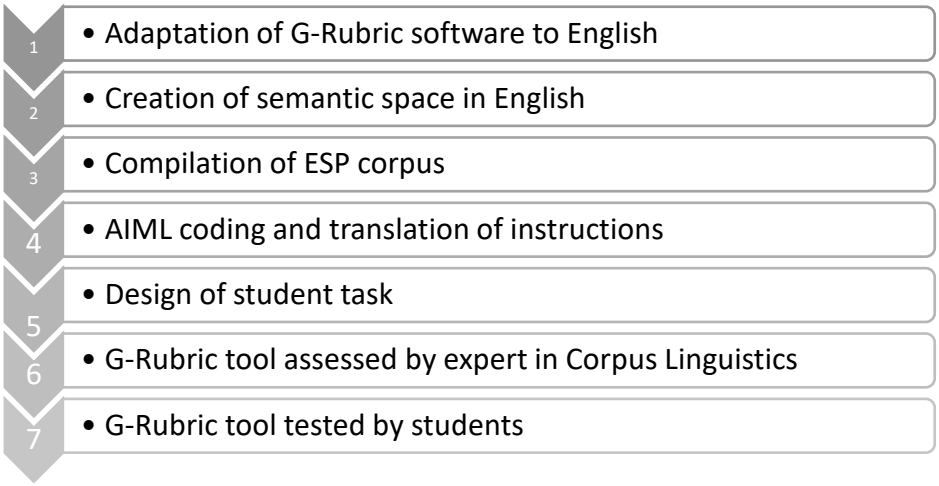


Figure 4. *G-Rubric Adaptation Process*

3.2. 2nd Phase: Implementing the G-Rubric System

Once the ESP-based version of G-Rubric had been built in the first phase of the project, it was time to get it tested with the students. During the academic year 2020-21, those who registered in the subject English for professional purposes, a compulsory course in the first year of the Degree in Tourism at UNED, were offered the possibility of using the G-Rubric tool. This evolved around a course assignment that consisted of the production of a professional proposal to adapt a given company in the Tourism sector to the Sustainable Development Goals (SDGs) due to the relevance of the 2030 Agenda for Sustainable Development and the SDGs in the tourist industry. Therefore, the topic chosen for the compilation of the corpus was one of high practical relevance for them: Sustainable tourism.

Appendix A contains the complete document of instructions for the assignment. The report was required to be relatively short (between 200 and 300 words) and the students were required to build a logical structure, with an opening paragraph explaining the relationship between Tourism and SDGs; a second paragraph providing tips to encourage responsible travelling; a third one which listed the difficulties and challenges in implementing the SDGs in a company in the tourist sector; and a final one with some concluding remarks on the above. Prior to the task, given the intermediate level of this first-year subject, the students were taught general principles of ESP writing, including the use of short sentences, simple connectors, and verbal cohesion. They were also offered weblinks where they would find the information they had to include in the report. A so-called golden answer or model written assignment was created by the authors and entered into the G-Rubric tool, so that the software could compare the contents of each composition against that "ideal" answer (Díez-Arcón & Martín-Monje, 2021). The model answer consisted of five sections:

- A composition on the topic provided (up to 400 words)
- 2-5 key concepts or ideas related to the topic
- 5-10 keywords for each concept
- Up to 10 pieces of content that must appear in the composition (characters, events, etc.)
- 6 multiple-choice questions on the model composition with one only correct answer and three distractors.

The quasi-experiment described in this article involved both an experimental group and a control group. The former was formed by the 26 students who voluntarily used the G-Rubric tool when drafting their written compositions, whereas the control group consisted of 26 randomly selected compositions from the rest of the cohort, who did not make use of the tool. As for data collection, the institution's virtual learning environment provided tracking of student progress, including all their intermediate and final grades and tutors' feedback that made up the students' formative evaluation. Besides, some (but not all) data were collected from the G-Rubric tool, namely the numerical grades and the feedback text on the students' gradual versions of each composition (with additional evaluation graphs). Once the robustness of G-Rubric was tested and confirmed by different members from the teaching team, it was piloted. The compositions were generated by the students with or without the quantitative and qualitative input from the tool, and then uploaded on the virtual course, where they would all be corrected and scored by the corresponding tutors.

4. Results and Discussion

In order to provide evidence to compare human and automated grading, Table 1 shows the results executed by the tutors and G-Rubric on the same compositions clustered in the standard way:

Table 1. *Distribution of the Grading Undertaken by Tutors and G-Rubric*

No. of Students	Grade Distribution (Eval. by Tutors)	No. of Students	Grade Distribution (Eval. by G-Rubric)
0	0-4.9	0	0-4.9
3	5-6	2	5-6
2	6-6.9	13	6-6.9
4	7-7.9	10	7-7.9
3	8-8.9	1	8-8.9
14	9-10	0	9-10

It is to be noted that there was no interference between the use of G-Rubric and tutor scoring because the latter were not informed of the ongoing project to prevent performance monitorization. As can be seen, they coincide in that

there are no fails, but tutors' grades are noticeably higher across the scale. This tendency is complemented with the data in Table 2:

Table 2. *Statistical Data to Compare the Grading Undertaken by Tutors and G-Rubric*

	Eval. by Tutors	Eval. by G-Rubric
Mean	8.62	6.89
Median	9.00	6.75
Mode	9.00, 9.50	--
Span (Highest and Lowest)	10-5.00	8.14-5.86

A qualitative analysis reveals inconsistencies between the feedback provided by the tutor and the numerical score. Table 3 shows two compositions with a striking lack of coherence between the feedback text and the numerical score:

Table 3. *A sample of Two Composition Evaluations Performed by Tutor*

Student No. 21427	**General comment**: I value the effort, but you really need to improve your writing skills. See the attached document for feedback on your work. **Comments provided in edited document**: First paragraph highlighted. Teacher's comment: "I don't get what you mean." Second paragraph highlighted. Teacher's comment: "This means nothing." Third paragraph highlighted. Teacher's comment: "Incomplete sentence." **Comment at the end of the document**: "No se entiende nada- parece sacado de un traductor- unas cuantas frases lo parecen, de hecho" [*trans.: "Unintelligible - it seems that the text has been generated by a machine translation system – particularly certain sentences."*] **SCORE: 5**
Student No. 11959	**General comment**: Excellent job. A few comments in the attached file only **Comments provided in edited document**: "the nature" changed to "nature" "avoiding" changed to "avoid" "focus into" changed to "focus on" "it is to bet" changed to "it means betting" **SCORE: 10**

As can be noted, even compositions with mistakes obtain a straight 10 out of 10 points. The issue of overgrading has only been addressed internally within each teaching teams, with measurements such as random check-ups, but not at institutional level. In order to increase the sample, twice as many tutor-corrected compositions were examined, which confirmed the remarkably high scores. Thus, in the randomly selected sample of 52 student essays: 28.84% obtained 10/10, 53.84% obtain 9-10/10, 78.85% obtained 7 or more/10, and only 7.69% scored less than 5/10 (failed). These results contrast with the over 30% fails in the final exam the same year, corrected at the university headquarters. There is long term evidence of a general tendency for overgrading on the part of tutors who know and have certain degree of personal contact with their reduced group of students, which can lead them to lose impartiality. It is also reasonable to assume that UNED tutors operate under some pressure as they have temporary contracts that depend each year on the results and evaluations of their students.

Another feature of tutor evaluation is the fact that it is undertaken with little or no consideration to the rubric elaborated by the teaching team (and included as part of the instructions for both tutors and students at the beginning of each academic year). Only two out of the 26 evaluations provided explicit reference to the different sections of the rubric. Generally, the tutors tended to establish a score largely based on grammatical accuracy, rather than on content or other formal aspects. Grammar also had a predominant role in feedback texts, as part of the individual diagnosis, error correction, and suggestions for improvement. See the example of student no. 21123 (G-Rubric's developers provided

researchers with a numerical code to comply with personal data protection regulations while enabling traceability across tutor and machine evaluation):

- G-Rubric: Writing 93%, Content 1.57
- Teacher's grade: 9

As part of the quality control of the quasi-experiment, 10 assignments were sampled which had achieved the highest grades when assessed by the corresponding tutors. A contrastive analysis between the score given by human and machine revealed considerable discrepancies, so all the assignments were sent to four external expert evaluators for discernment. Interestingly, expert evaluation of the compositions (see Table 3 in Díez-Arcón & Martín-Monje, 2021) proved to be numerically closer to that of G-Rubric. Specifically, of the 36 evaluations undertaken in total (4 were cancelled due to plagiarism), only 7 were closer to the tutor's (19.44%).

Unlike holistic tutor grading, it was possible to retrieve separately the scores of the two sets of criteria performed by the system: form and content. This revealed abrupt differences, as can be seen in Table 4 (these data are specified in Appendix II):

Table 4. *Distribution of G-Rubric's Grading of Form and Content in the Compositions*

No. of Students	Grade Distribution (Form)	No. of Students	Grade Distribution (Content)
0	0-4.9	20	0-4.9
0	5-6	1	5-6
0	6-6.9	4	6-6.9
1	7-7.9	0	7-7.9
14	8-8.9	0	8-8.9
11	9-10	0	9-10

The right-hand columns show the figures that correspond to 60% weighting of the Writing grade and 40% of the Content grade and have been converted into out-of-ten scores in order to enable comparability with tutor grading. Here, each of the five criteria is assigned a percentage of 20%, which is more balanced than the strong grammar bias in tutor evaluation commented above. As can be seen in Table 3, G-Rubric is very demanding, and there are no maximum scores (10/10) in Content in the sample. The two best scores in Content have been found in the following submissions, which reflect the disparity with form (Writing):

- 11959:
  - G-Rubric: Writing 82%, Content **6.48**
  - Tutor's Grade: 10
- 21188:
  - G-Rubric: Writing 92%, Content **6.56**
  - Tutor's Grade: 9.5

As shown in Table 5, all the evaluations scored higher in Writing than in Content. This difference reflects an unusual discrepancy that has not been supported or justified by the four external experts. The authors claim that, despite claims on the part of the developers regarding the multilinguality of the system, the formal parametres (i.e., the linguistic corpus and the concordancer, which operates statistically to extract information in the form of collocations, n-grams, etc., following a lexical-grammatical approach) were largely Spanish-based and, hence, not transportable to the English language. A similar argument could be applied to the Natural Language Processing mechanism of the system which, as described by the developers (Jorge-Botana, Olmos, & Luzón, 2019), employed statistical methods to train the system in a language-specific way for the identification (and penalization) of formal inaccuracies and agrammaticalities. The improvement of this aspect of the system's performance has been reported to the system developers by the authors as part of the list of improvements to be considered in the following version of the system.

Table 5. *Subset of Cases Showing the Greatest Numerical Grading Discrepancies*

Student ID	Tutor's Evaluation	G-Rubric's Evaluation
21089	5.00	7.20 (W: 8.20 / C: 5.56)
21172	6.75	7.62 (W: 8.50 / C: 6.30)
21198	6.25	6.73 (W: 9.30 / C: 2.89)
21124	5.75	6.65 (W: 8.90 / C: 3.28)

21386	7.00	7.03 (W: 8.80 / C: 4.39)
21427	5.00	6.48 (W: 7.80 / C: 4.51)

The reason for showing only one decimal in Writing is because marks were given in percentages of whole numbers. In contrast, the Content marks contained two decimals, and those in tutor's grades were rather regular (mostly 25, 50, and 75), which suggests a qualitative type of evaluation. Regarding content scoring, G-Rubric proved to be far more sensitive than tutors. There were no 10/10 scores and they were consistently lower than the corresponding ones for Writing throughout the sample. The underlying reason is related to the system's reasoning and weighting of lexical-semantic relations between its knowledge repository and the composition in terms of coincidences/discrepancies.

Regarding the tutor feedback, partly due to the size limitations of the evaluation tool and partly due to the lack of consistent supervision on the part of the teaching team, it showed certain deficiencies, which include the following:

- Evidence of quick and careless revision revealed in very brief comments (much shorter than the 300 characters available), telegraphic style text with inaccurate punctuation, misspellings and indications to the student to read an attached file with a lengthy revision that was not available! Examples include the following: "Well done!! Check the attached file for a few minor corrections. Congrats!"; "Be careful: Write all in English conclusion. Probably it was an automatic change in typing [*sic*]." Furthermore, some scores were not accompanied by any form of feedback.
- Irrelevant or inappropriate comments, such as "REMEMBER: NEXT PEC FINAL ORAL EXAM: 18th May Classes are only in Microsoft Teams on Mondays and Wednesdays at 20:15. Try to attend if possible." This diminished the focus on the formative purpose of the activity.
- Lack of deep diagnosis analysis. Regarding error interpretation, a holistic examination of the text would reveal that at least part of the inaccuracies was due to writing typos (mainly because they were sporadic and not consistent with other features). Evidence of knowledge of the norm was largely ignored.
- Little specification of students' strengths, only vague remarks such as: "Congratulations," "Excellent job," "Good work," and so on. The following are more explicit: "Good work, easily explained and argued. Careful with some grammar mistakes"; "Sara, it is a very well written piece of text, making use of the standards of professional English and an adequate vocabulary, I would recommend you to avoid one-line paragraphs and revise the punctuation of longer paragraphs. Congratulations!"
- Failure to select the types of aspects that impaired the quality of the text, either because of their recurrency or because of how they affected its intelligibility and caused confusion and misunderstanding to the reader. Feedback focused on illustrative rather than representative mistakes and did not follow any apparent consistent prioritization of criteria.
- Occasional use of the rubric provided by the teaching team. Only four compositions out of the 26 were given feedback with specific reference to the rubric and the assessment criteria. The following is one of the feedback texts which refers to the rubric: "CRITERIA FOR WRITING ASSIGNMENTS LEVEL OF PERFORMANCE D (0-0.5) C (0.5-1) B (1-1.5) A (1.5-2) • Successful accomplishment of the contents of the units 0.5 (you have not completed the task properly) • Adequate use of the vocabulary and expressions studied in the units (1 many mistakes, words wrongly written or even words in Spanish!) • Adequate use of the grammar rules learned in the units 0.5 (very serious grammar mistakes, sentences with two subjects or without any at all) • Fluency 0.2 (the text is difficult to be understood) • Discourse cohesion (use of connectors and other cohesive devices) 0 (lack of connectors) TOTAL SCORE 2.20".
- Strong focus on formal aspects, which were mostly grammatical ("Grammatically only 2 mistakes: is required an adequate investment [an adequate investment is required] and "in very short time" [in a...]; "Very good job. Careful with some grammar mistakes"; "Very good job, Amparo. Well explained and with a wide use of vocabulary. But be careful with the use of prepositions and you should use better the verb pattern form of verb+object+to+verb."), followed by far by remarks on the lexis, the use of English, and formal nuances ("very well written, but you must write 'nowadays' as a single word"; "The use of capital letters and it is quite short"). Pragmatic, discourse and content considerations were practically absent from tutors' comments, with exceptions like the following: "I miss some more concretion of the bullet points demanded, especially on the last one focusing on your company"; "The greatest problem is that you have written a very generic text, and it should be more focused on a specific company."

Regarding G-Rubric's feedback, unlike tutor feedback, it was more useful for feedback provision than summative evaluation. As mentioned above, the mean grade between Writing and Comment rendered the result closer to the evaluation of the external experts who participated in the project.

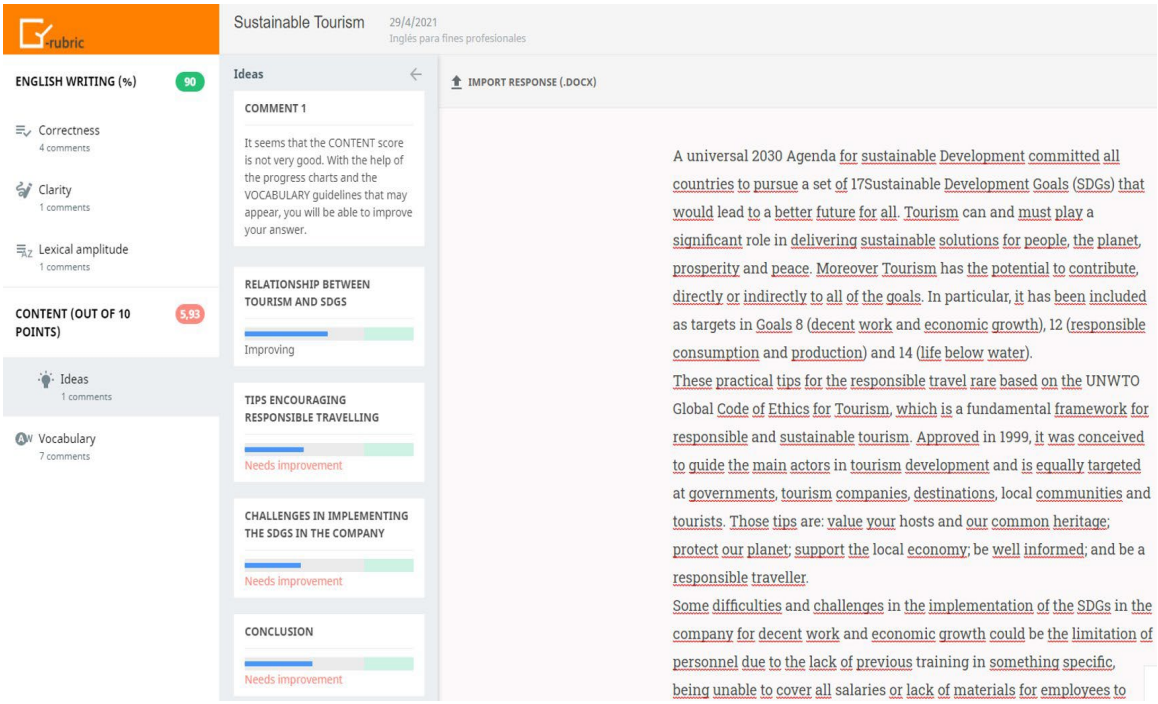
In the current version of the system, there was no student log and intermediate feedback texts were not stored either, which has been reported by the authors as part of the scope of improvement for future research. The following is an example of the system's feedback on Clarity in the final version of a student's composition. It shows aspects present in interpersonal communication, such as the use of the 1st and 2nd person singular ("I can confirm", "you have taken", "your writing"); expressions of encouragement ("congratulations and keep that up"), of motivational value; reference is made to the process, not only the result, through the use of comparatives ("more care") and lexis ("achieved"; "acceptable"); and a general informal tone (achieved through the choice of words ["OK"] and imperatives ["keep that up"]). Furthermore, reference is made to the high score (90/100), which establishes a link between scoring and feedback, reflecting good teacher practice:

> "OK, I can confirm that you have taken more care of your writing and achieved an acceptable STYLE score. Congratulations, and keep that up."

In order to provide an example of other formal criteria, in Lexical amplitude, a student received the following feedback, with a suggestive tone, that included an explicit diagnosis: word repetition, and offered four pieces of advice: word deletion, word paraphrase, using a paraphrase, and using a stylistic resource:

> "You may be repeating words in your answer, so you can use new stylistic resources, delete or vary some words, or simply use a new phrase."

Regarding system feedback on Content, as can be seen in Figure 6 on the Ideas subcriterion, it was a synthesis of the results obtained in the independent analysis of the four sections of the composition (provided in the form of progress charts): the relationship between tourism and SDGs, tips to encourage responsible travelling, challenges in the implementation of SGDs in a company, and the conclusion. In this case, it was the first version of the composition and the score given by the system was 5,93/10. Feedback consisted of diagnosis followed by tips on vocabulary and an indication of the parts that needed more attention from the student:



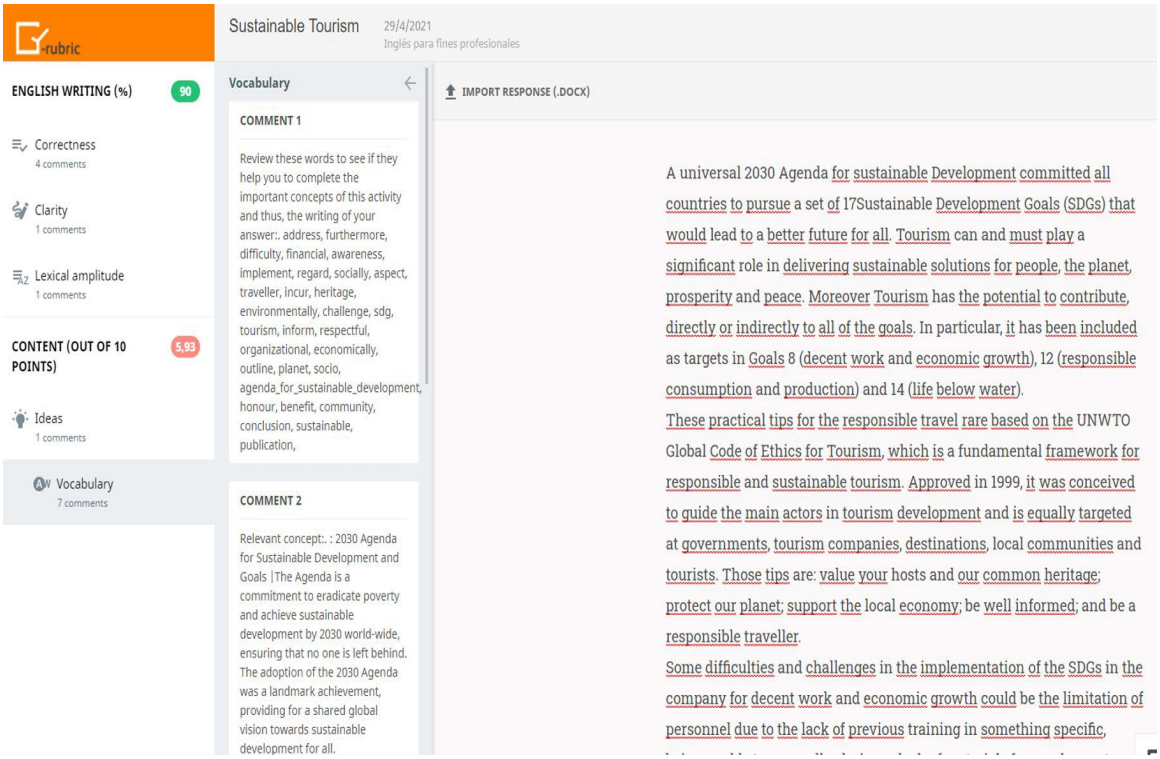


Figure 6. *Example of Content Feedback, Subcriteria Ideas and Vocabulary, Provided by G-Rubric* (Martín-Monje & Castrillo, 2022)

Feedback could sometimes consist of several independent comments. As shown in Figure 6, in the Vocabulary part of Content, G-Rubric provided two pieces of independent remarks to improve the content of the composition: one of them was provided through key words and terms, which might trigger ideas to the student (common nouns, adjectives, verbs, modal adverbs, discourse markers, and locutions), and the second one provided an elaborate idea for them to complete their composition.

Preliminary results obtained in the quasi-experiment threw light on an aspect of utmost importance in distance learning education: the detection of plagiarism or fraudulent strategies. In Table, 6 a tutor justified a drastic reduction of the grade due to plagiarism of almost one fourth of the composition:

Table 6. *Example of Detection of Plagiarism on the Part of a Tutor*

21089
General comments: Dear NAME, You did a good job; however, your grade is 5 as we have found that 23% of your work has been copied from the following websites: https://www.shaalaa.com/question-bank-solutions/short-notes-tourism-and-gdp-importance-of-transport-in-trade_164670 https://www.athensjournals.gr/tourism/2017-4-1-1-Jones.pdf https://www.huffpost.com/entry/what-is-responsible-trave_b_9730434 Please, remember that originality is important. Good luck with your exams.
GRADE: 5

G-Rubric does not detect bad praxis on the part of the students, such as copying from other sources without quoting the reference or using online translators (e.g., in compositions no. 21427 and 21105, the tutor identifies a literal translation). Tutors, on the other hand, identify plagiarism easily and capture idiolect nuances, such as an abrupt change in register, interpreting the case in hand expertly. However, further guidance and/or control must be exerted on the part of the teaching team, so that academic fraud detection follows institutional regulation and receives a consistent penalization. See, for example, the tutor feedback text in Table 6 where the student passed despite plagiarism detection, despite institutional recommendations to mark plagiarised assignments with 0.

In sum, the effectiveness of G-Rubric in providing automatic focused feedback has been confirmed, as well as its efficiency in grading, since it reduces significantly the time needed for grading assignments. This analysis has shown tutors' grades to be higher (Tables 1 & 2) and it has also revealed some inconsistencies between the feedback provided by tutors and the grade awarded to the assignment. G-Rubric, as a computer tool, is not inconsistent, which is an improvement to tutor grading, but it does not detect plagiarism, and the tutors do. So, it could be said that G-Rubric offers more standardized and consistent grading and automatic focused feedback, but tutors offer more nuanced assessment.

5. Conclusion

This article aimed to gain insights on the customization of state-of-the-art text evaluation by machine, and particularly, automatic feedback, as a form of DDL. The application of this research revolved around second language teachers and students in mass online learning institutions, where the unbalanced student-teacher ratio together with the external pressure of the latter to cover the whole syllabus, etc. cause formative writing correction to be a highly challenging task. A case scenario was considered, namely that of English for Tourism in the Spanish national distance university, where teachers can rely on the support of a larger group of tutors in regional centers for formative evaluation. However, this solution was far from ideal, and other problems were introduced, such as heterogeneity of evaluation criteria and the lack of rigor in the application of the proposed rubric. Also, despite the improved student-tutor ratio, the challenging scenario did not allow for draft revisions, which greatly impaired the formative value of the assignment. Therefore, given the existence of a popular in-house standalone system that had been successfully developed for composition evaluation in specialized domains in Spanish with an emphasis on iterative draft revision, G-Rubric, a quasi-experiment was set up to analyze contrastively the performance of tutors vs. a new adaptation of the tool for English for Tourism in a mass online tertiary education environment, with a view to the automatization of the process.

The adaptation process to the new case of application in the domain consisted of pattern-matching and machine reasoning against a specialized corpus and associated knowledge previously inserted in the tool. Regarding system evaluation, the analysis of the results revealed affordances that coincided with limitations in human evaluation and vice versa. The tool integrated well in the given online educational context and was robust and reliable (did not provide spurious advice). It aimed at helping students with their writing process by undertaking progressive revisions of composition drafts, which covered diagnosis, a reference to textual strengths, and recommendations for future improved performance at various linguistic and communicative levels. It extended students' exposure time to the target language and intensified their dedication to the task in hand, by enabling them to work informatively on their compositions once and again before final delivery. However, the system lacked linguistic and real-world knowledge at times for the identification of certain key mistakes. This could be improved in the future by incorporating wider corpora in the system. Therefore, important shortcomings were identified in its performance, such as its inability to detect bad praxis on the part of students (e.g., plagiarism or the use of Spanish-English machine translations).

Tutors' focused feedback lacked consistency and precision and failed to cover the spectrum of criteria included in the rubric at their disposal. It also failed to perform expert diagnosis analyses of the students' strengths and competence gaps; failed to prioritize the mistakes that mostly impaired the quality of the text (i.e., affected its intelligibility or created misunderstanding on the part of the reader); showed frequent evidence of quick and careless revision; and made irrelevant or inappropriate comments. More importantly, the feedback showed a strong bias towards grammatical issues, followed by far by vocabulary and form, at the expense of other levels of analysis, such as pragmatics and discourse, or the content or subject matter, all of which are of great significance in languages for specific purposes. Regarding scoring, tutors showed a tendency to over-grade, confirmed by external expert evaluators. Preliminary interpretations were undertaken in the context of the institutional dynamics and regulations and compromised the reliability of tutors' task.

However, although external evaluators' scoring was closer to that of G-Rubric, a closer analysis showed severe operational limitations with the tool, particularly in Writing (form), becuase the system appeared to be insufficiently adapted to the English linguistic domain under study. Moreover, the research work undertaken demonstrated that the system performed better for feedback provision than for scoring. Therefore, the authors claim that these functionalities require further adaptation to ESP-specific knowledge to reflect its writing norms and practices. Another research line that has become apparent so that the system gains in usability by all parties involved (i.e., students, tutors, teachers, and researchers) includes keeping a student log with the previous versions of the compositions and their corresponding evaluations.

As can be seen, this preliminary experience with G-Rubric has led to number of findings and opened research lines for a future improved version of the system. As for the answer to the research question raised at the beginning of the article regarding whether G-Rubric will substitute or complement current human formative evaluation, the combination of complementary abilities and shortcomings of the two types of evaluations explored hints towards a hybrid scenario of joint evaluation of written compositions by tutor and machine. To this end, it is fundamental for tutors to both be trained on best practices on mass composition evaluation involving the explicit use of the rubric, and gain a better understanding of the system's functionality and capabilities for better integrated performance. Similarly, usability will arguably improve if students are provided with the necessary guidance to interpret the current disparities between both types of evaluation. In that sense, it would be advisable to use G-Rubric in both yearly written assignments so that students can have a more prolonged opportunity to master its use and learn to produce compositions in a structured, sequential, and self-regulated fashion. Furthermore, given the fluent communication between students of consecutive years, a more informed use of the system can be expected in the future. These substantial technological and pedagogical improvements will arguably take the proposed hybrid evaluation scenario forward to a more advantageous position in order to deal with further research questions, such as the perceptions of the parties involved in the learning process and the potential correlations between the levels of engagement and strategies used and the improvement of composition quality and grades.

Notes

[1]The G-Rubric solution and its linguistic-mathematical system versions and developments are owned by Semantia Lab, a technology-based spin-off specialized in Natural Language Processing (NLP) and Cognitive Science, and were developed with the academic and institutional support of UNED (https://www.grubric.com/nuestro-equipo.html)

Acknowledgments

This research paper stems from the project "Data-Driven Learning (DDL) for the improvement of written competence in English," funded by Universidad Nacional de Educación a Distancia (UNED). This project was awarded as the "Best Teaching Innovation Project" in 2021 by UNED.

Information on Informed Consent or any Data Privacy Statements

The instruments used in the data collection have followed the UNED's General Data Protection Regulation: https://www.uned.es/universidad/inicio/institucional/proteccion-datos.html

Conflict of Interest

The authors declare no conflict of interests.

Funding

The project "Data-Driven Learning (DDL) for the improvement of written competence in English," was funded by the Vice-rectorate of Educational Innovation at Universidad Nacional de Educación a Distancia.

References

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 316-329). The Guilford Press.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* W.H. Freeman/Times Books/Henry Holt & Co.

Bárcena Madera, E., Martín-Monje, E., & Jordano de la Torre, M. (2016). Methodological and technological innovation in distance teaching of English for tourism. *Iberica*, *2016*(31), 39-61.

Barrantes Montero, L. (2009). A brief view of the ESP approach. *LETRAS, 46*, 125-143. https://doi.org/10.15359/rl.2-46.7

Bird, B., Downs, D., McCracken, I. M., & Rieman, J. (Eds.). (2019). *Next steps: New directions for/in writing about writing.* University Press of Colorado.

Bitchener, J., & Knoch, U. (2008). The value of written corrective feedback for migrant and international students. Language *Teaching Research, 12*(3), 409-431. https://doi.org/10.1177/1362168808089924

Bruton, A. (2009). The vocabulary knowledge scale: A critical analysis. *Language Assessment Quarterly, 6*, 288-297. http://dx.doi.org/10.1080/15434300902801909

Boulton, A. (2017). Research timeline: Corpora in language teaching and learning. *Language Teaching, 50*(4), 483-506. http://dx.doi.org/10.1017/S0261444817000167

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning, 67*(2), 348–393. http://dx.doi.org/10.1111/lang.12224

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

Crossley, S. A. (2020). Linguistic Features in writing quality and development: An overview. *Journal of Writing Research, 11*, 415-443. Https://doi.org/10.17239/jowr-2020.11.03.01

Curado, a. (2017). Form-focused data-driven learning for grammar development in ESP contexts. *Revista de Lenguas para Fines Específicos, 23*(1), 13-30. https://doi.org/10.20420/rlfe.2016.327

Dağ Akbaş, R. (2021). A systematic review of the ESP (English for specific purposes)-based post-graduate research in Turkey. *Sosyal Bilimler Dergisi*, 22, 12-31. https://www.researchgate.net/publication/357735869_A_systematic_review_of_the _ESP_English_for_specific_purposes-based_post-graduate_research_in_Turkey

Díez-Arcón, P., & Martín-Monje, E. (2021). G-Rubric: The use of open technologies to provide personalized feedback in languages for specific purposes, in *Edulearn21 Proceedings*, 2635-2643. http://dx.doi.org/10.21125/edulearn.2021.0574

Dudley-Evans, T., & St John, M. J. (1998). *Developments in English for specific purposes*. Cambridge University Press.

Ellis, R. (2009). Task-based language teaching: Sorting out the misunderstandings. *International Journal of Applied Linguistics, 19*(3), 221-246. http://dx.doi.org/10.1111/j.1473-4192.2009.00231.x

Fleckenstein, J., Leucht, M., & Köller, O. (2018). Teachers' judgement accuracy concerning CEFR levels of prospective university students. *Language Assessment Quarterly, 15*(1), 90-101.

Fletcher-Wood, H. (2021). *Habits of success: Getting every student learning*. Routledge. https://doi.org/10.4324/9781003010074

Flowerdew, J. (2016). English for specific academic purposes (ESAP) writing: Making the case. *Writing & Pedagogy, 8*(1), 5-32.

Fox, J. & Arteneva, N. (2017). From diagnosis toward academic support: Developing a disciplinary, ESP-based writing task and rubric to identify the needs of entering undergraduate engineering students. *ESP Today, 57,* 148-171.

Ghanbar, H., & Rezvani, R. (2024). Four decades of publications in English for specific purposes: Mapping the trajectory of empirical research. *Journal of Research in Applied Linguistics, 15*(1), 32-49. http://dx.doi.org/10.22055/rals.2023.44222.3102

Harris, Z. (1968). *Mathematical structures of language.* New York: John Wiley & Sons.

Hernández Benítez, M. H., & Santamaría Lancho, M. (2016). G-Rubric: Una aplicación para corrección automática de preguntas abiertas. Primer balance de su utilización. In *Nuevas perspectivas en la investigación docente de la historia económica* (pp. 473-494). Editorial de la Universidad de Cantabria.

Hutchinson, T. & Waters, A. (1987). *English for specific purposes*. Cambridge University Press.

Hyland, K. (2022). English for specific purposes: What is it and where is it taking us? *ESP Today, 10*(2), 202-220. https://doi.org/10.18485/esptoday.2022.10.2.1

Hyland, K., & Hyland, F. (2006). *Feedback in second language writing: Contexts and issues*. Cambridge University Press. https://doi:10.1017/CBO9781139524742

Jorge-Botana, G., Olmos, R. & Luzón, J.M. (2019). Bridging the theoretical gap between semantic representation models without the pressure of a ranking: some lessons learnt from LSA. *Cognitive Processing, 21,* 1-21. https://doi.org/10.1007/s10339-019-00934-x

Kao, C.-W., & Reynolds, B. L. (2016). Analyses of EFL teacher feedback on word choice errors from an on-line writing platform. *The Academic Conference of Foreign Language Teaching and Intercultural Studies, May*, 35-51.

Karpenko-Seccombe, T. (2020). *Academic writing with corpora: A resource book for data-driven learning*. London: Routledge. https://doi.org/10.4324/9780429059926

Keuning, H., Jeuring, J., & Heeren, B. (2018). A systematic literature review of automated feedback generation for programming exercises. *ACM Transactions on Computing Education (TOCE), 19*(1), 1-43.

Kittredge, R. (1982). Variation and homogeneity of sublanguages, In R. Kittredge & J. Lehrberger (Eds.), *Sublanguage: Studies of Language in Restricted Semantic Domains,* Berlin: Walter de Gruyter, 107-137.

Knight, S., Martinez-Maldonado, R., Gibson, A., Buckingham Shum, S. (2017). Towards mining sequences and dispersion of rhetorical moves in student written texts. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK '17).* Association for Computing Machinery, New York, NY, USA, 228-232. https://doi.org/10.1145/3027385.3027433

Knight, S., Shibani, A., Abel, S., Gibson, A., Ryan, P., Sutton, N., Wight, R., Lucas, C., Sándor, Á., Kitto, K., Liu, M., Vijay Mogarkar, R., & Buckingham Shum, S. (2020). Acawriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research, 12*(1), 141-186. https://doi.org/10.17239/jowr-2020.12.01.06Kurdi

Limpo, T., Nunes, A., & Coelho, A. (2020). Introduction to the special issue on technology-based writing instruction: A collection of effective tools. Journal of *Writing Research, 12*(1), 1-7. https://doi.org/10.17239/jowr-2020.12.01.01

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276. https://doi.org/10.1191/0265532202lt230oa

Martín-Monje, E. & Castrillo, M.D. (2021, May 12). Data driven learning en entornos de aprendizaje formales e informales. *XI Jornadas de Investigación en Innovación Docente de la UNED.* (online conference). https://www.youtube.com/watch?v=mAajXIK6vrA

Martín-Monje, E. & Castrillo, M.D. (2022). Data-driven learning to improve writing skills in foreign languages. In K.D. Rossade, J. Janssen, C. Wood, & G. Ubachs, *Designing Online Assessment — Solutions that are Rigorous, Trusted, Flexible and Scalable*. Maastricht: EADTU.

Miao, Y., Badger, R., & Zhen, Y. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing Class. *Journal of Second Language Writing, 15*, 179-200. https://doi.org/10.1016/j.jslw.2006.09.004

Mitchell, R., Myles, F., & Marsden, E. (2019). *Second language learning theories.* (4th ed.). London: Routledge. https://doi.org/10.4324/9781315617046

Roscoe, R.D., Allen, L.K., Weston, J.L., Crossley, S.A. & McNamara, D.S. (2014). The writing pal intelligent tutoring system: Usability testing and development. *Computers and Composition, 34*(1), 39-59.

Sadler, R. (2010) Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education, 35*(5), 535-550. http://dx.doi.org/10.1080/02602930903541015

Smelstor, M. (1978). *A guide to teaching the importance of audience and subject*. https://eric.ed.gov/?id=ED176274

Stevenson, M. & Phakiti, A. (2019). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51-65.

Swales, J. M. (2000). Languages for specific purposes. *Annual Review of Applied Linguistics, 20*(1), 59-76. https://doi.org/10.1017/S0267190500200044

Trace, J., Hudson, T., & Brown, J. D. (Eds.). (2015). *Developing courses in languages for specific purposes*. National Foreign Language Resource Center.

Van Beuningen, C. (2021). Focused versus unfocused corrective feedback: from part IV-feedback provider, feedback intensity, and feedback timing. In *The Cambridge handbook of corrective feedback in second language learning and teaching* (pp. 300-321). Cambridge University Press.

Wang Z, Han F. (2022). The effects of teacher feedback and automated feedback on cognitive and psychological aspects of foreign language writing: A mixed-methods research. *Frontiers in Psychology*, 13, 909802. https://doi.org/10.3389/fpsyg.2022.909802

Warren, M. J. (2015). Introduction to data-driven learning. In F. Farr & L. Murray (Eds.), *The Routledge handbook of language learning and technology*. Routledge.

Wilson, J., Ahrendt, C., Fudge, E.A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. *Computers & Education, 168*, 104208.

Appendixes

Appendix A: Instructions provided

Written Assignment 2 / Units 4-6

Your company has become involved in the 2030 Agenda for Sustainable Development and the Sustainable Development Goals (SDG, http://tourism4sdgs.org/tourism-for-sdgs/what-are-the-sdgs/). Your boss has asked you to write a report including the following information:

- Relationship between Tourism and SDGs (http://tourism4sdgs.org/tourism-for-sdgs/)
- Tips encouraging responsible travelling (https://www.responsibletravel.com/copy/tips-for-responsible-travel)
- Difficulties and challenges in implementing the SDGs to the company. Mention financial, organizational and socio-cultural aspects.
- Your conclusion: Why is it important to incorporate the 2030 Agenda for Sustainable Development and the SDGs to your company?

The report should be 290-330 words long.

You can also check these websites to find out about sustainable tourism development:

- https://www.unwto.org/sustainable-development
- https://www.weforum.org/agenda/2019/09/global-tourism-sustainable/

If you need guidance when writing your report, these websites may be useful:

- http://learnenglishteens.britishcouncil.org/skills/writing/upper-intermediate-b2-writing/report
- https://www.skillsyouneed.com/write/report-writing.html

Appendix B: Grades Obtained by Students Who Decided to Do the Composition Assignment and Use G-Rubric

ID	Assessment by Tutor	G-Rubric Assessment	G-Rubric Weighted Score (60% Writing/40% Content)
11959	GRADE: 10	English Writing 82% Content: 6.48	4.92 2.59 Total: 7.51
21047	GRADE: 9	English Writing 94% Content: 4.12	5.64 1.64 Total: 7.28
21073	GRADE: 9.5	English Writing 86% Content: 3.42	5.16 1.37 Total: 6.53
21085	GRADE: 9	English Writing 83% Content: 3.5	4.92 1.4 Total: 6.32
21089	GRADE: 5	English Writing 82% Content: 5.56	4.98 2.22 Total: 7.20
21105	GRADE: 7.5	English Writing 91% Content: 2.37	5.46 0.94 Total: 6.40
21123	GRADE: 9	English Writing 93% Content: 1.57	5.58 0.62 Total: 6.20
21172	GRADE: 6.75	English Writing 85% Content: 6.3	5.1 2.52 Total: 7.62
21188	GRADE: 9.5	English Writing 92% Content: 6.56	5.52 2.62 Total: 8.14
21198	GRADE: 6.25	English Writing 93% Content: 2.89	5.58 1.15 Total: 6.73
21224	GRADE: 5.75	English Writing 89% Content: 3.28	5.34 1.31 Total: 6.65
21228	GRADE: 9.25	English Writing 87% Content: 6.42	5.22 2.56 Total: 7.78
21244	GRADE: 9.5	English Writing 87% Content: 3.88	5.22 1.55 Total: 6.77
21256	GRADE: 8.8	English Writing 88% Content: 1.45	5.28 0.58 Total: 5.86
21297	GRADE: 9.4	English Writing 89% Content: 5.94	5.34 2.37 Total: 7.71
21318	GRADE: 9.4	English Writing 82% Content: 4.31	4.92 1.72 Total: 6.64
21319	GRADE: 7	English Writing 88% Content: 4.16	5.28 1.66 Total: 6.94
21326	GRADE: 8	English Writing 85% Content: 2.94	5.1 1.17 Total: 6.27
21341	GRADE: 7.5	English Writing 90% Content: 2.97	5.4 1.18 Total: 6.58
21359	GRADE: 10	English Writing 90% Content: 4.92	5.4 1.96 Total: 7.36
21386	GRADE: 7	English Writing 88% Content: 4.39	5.28 1.75

			Total: 7.03
21396	GRADE: 9.7	English Writing 90% Content: 4.48	5.4 1.79 Total: 7.19
21419	GRADE: 9.5	English Writing 90% Content: 2.65	5.4 1.06 Total: 6.46
21420	GRADE: 9	English Writing 93% Content: 4.79	5.58 1.91 Total: 7.49
21427	GRADE: 5	English Writing 78% Content: 4.51	4.68 1.8 Total: 6.48
21428	GRADE: 8.5	English Writing 92% Content: 1.16	5.52 0.46 Total: 5.98