










# Semiautomatic Grading of Short Texts for Open Answers in Higher Education

Luis de-la-Fuente-Valentín<sup>1</sup>  , Elena Verdú<sup>1</sup> , Natalia Padilla-Zea<sup>1</sup> ,  
Claudia Villalonga<sup>2</sup> , Xiomara Patricia Blanco Valencia<sup>1</sup> ,  
and Silvia Margarita Baldiris Navarro<sup>1</sup> 

<sup>1</sup> Universidad Internacional de La Rioja (UNIR), Logroño, Spain  
{luis.delafuente, elena.verdu, natalia.padilla,  
xiomarapatriacia.blanco, silvia.baldiris}@unir.net

<sup>2</sup> Universidad de Granada, Granada, Spain  
cvillalonga@ugr.es

**Abstract.** Grading student activities in online courses is a time-expensive task, especially with a high number of students in the course. To avoid a bottleneck in the continuous evaluation process, quizzes with multiple choice questions are frequently used. However, a quiz fails on the provision of formative feedback to the student. This work presents PLeNTaS, a system for the automatic grading of short answers from open domains, that reduces the time required for the grading task and offers formative feedback to the students. It is based on the analysis of the text from the point of view of three different levels: orthography, syntax, and semantics. The validation of the system will consider the correlation of the assigned grade with the human grade, the utility of the automatically generated feedback and the pedagogical impact caused by the system usage in the course.

**Keywords:** Automated grading · Semantic similarity · Feedback · Readability · Short-answers

## 1 Introduction

Nowadays, the lifelong learning paradigm is consolidated, allowing to adapt students' knowledge and skills to the evolving labor market. This is being especially relevant during the Covid-19 pandemic, which has meant an abrupt change in the way of working. In this context, online learning is a key alternative for many companies and people looking for self-sustainability.

The success in online learning requires an early and well-justified feedback, which implies a work overload for teachers. Moreover, open-answer questions allow better evaluating the knowledge [1, 2] but evaluating this type of questions is a complex task performed with considerable variability among teachers [3, 4]. This problem of variability can be mitigated by using rubrics, which facilitate a more consistent and transparent evaluation process [5]. On the other hand, techniques based on artificial intelligence may support teachers in the extensive evaluation processes involved in online learning.

Specifically, open-answer questions imply writing a text by the student, which can be automatically processed by techniques of Natural Language Processing (NLP), machine learning or semantic based systems, techniques widely used in many different fields [6–8].

The reviewed literature shows that there are several tools that allow the automatic evaluation of short-answer questions. An interesting solution that allows the evaluation of different aspects of a text considering different metrics is ReaderBench [9]. It has integrated the indexes proposed by the systems E-rater [10], iSTART [11] and CohMetrix [12]. This open-source tool allows to evaluate the complexity of a text, summaries and explanations, as well as measuring the social collaboration within a group. It uses text mining techniques, NLP and social network analysis tools. ReaderBench has been used in different experiments. For example, Westera et al. [1] used NLP methods of ReaderBench to generate up to 200 indexes, with the aim of evaluating texts in a game environment in the context of online learning. The result of this evaluation is failing or passing. Panaite et al. [13] used ReaderBench to generate a grade for short answers based on four categories (poor, pass, good, excellent). In the 97% of the cases the grade was close to the grade given by a human expert. Recently, deep learning based solutions have been explored for automatic text complexity evaluation, obtaining limited performance due to the lack of a reliable corpus [14]. The limitations of all the above mentioned systems are the lack of explanations in the evaluations so that the students understand the obtained grades. Giving explainable feedback in artificial intelligence is very important and there is much literature about it in recommender systems and decision support systems, for example [15]. In the educational field, a few examples are found focused on decision makers, as an explainable tool that facilitates decision-making processes incorporating textual and graphical explanations when predicting students' outcomes [16], or a dashboard based approach for explainable student agency analytics [17]. However, there are almost no works related to explainable feedback for automatic evaluation systems in the educational field. From the literature review, we can conclude that the evaluation of learning results via analytics methods is a non-mature field [18].

The PLeNTaS project intends to provide students and teachers with valid explanations as formative feedback. A software platform is being developed to provide a semi-automatic evaluation of activities based on short open-answer questions. For each evaluated characteristic of a given answer, a textual explanation of the grade will be provided starting from the information previously provided by teachers.

The rest of the chapter is organized as follows: Sect. 2 presents a brief overview of the PLeNTaS project, including the different levels of analysis and the validation design. In Sect. 3, the main conclusions are outlined.

## 2 The PLeNTaS Project

With the focus on the provision of an automatic grade of open questions with a quick and accurate formative feedback, PLeNTaS (“Proyectos I+D+i 2019”, PID2019-111430RB-I00) is a three-year project that proposes an evaluation process for short-answer questions for Spanish language. For this purpose, the project requires research from the pedagogical and technological points of view, both described as follows.

## 2.1 Design Decisions

The automatic evaluation requires the definition of a solid pedagogical approach, which should be the basis of any further development. The first required decision is the type of question to be evaluated. The project is focused on open-answer questions but, within such frame, some relevant decisions have still to be made: Will PLeNTaS support specific domain or open domain questions? Will the questions follow a given template? To answer these questions, during the pedagogical work phase of the project the metrics to be automatically evaluated are being defined. From a catalogue of metrics, teachers will compose the rubrics to evaluate the specific answers, so the technological research goal is the automatic fulfill of those rubrics. Then, to validate the system, automatic and human evaluations based on rubrics will be compared. Some relevant decisions are:

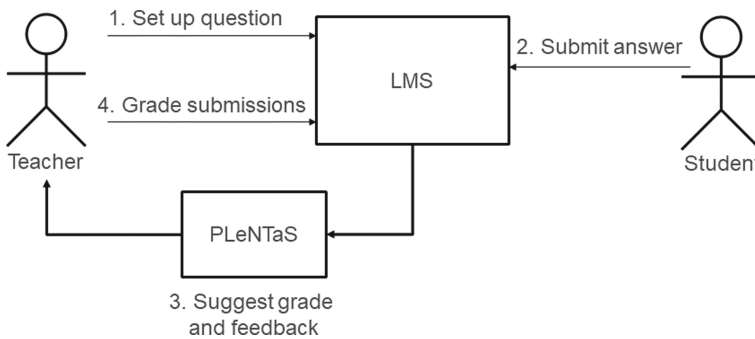
- The grading system will consider different levels of analysis, namely orthography, syntax, and semantics. We consider that none of the three levels is enough by itself to grade an activity. However, through the combination of these three levels, the grade is likely to be more accurate. For example, one of the possible techniques for the semantic level is checking if some key terms are present in the text. Therefore, it would be possible to game the system by only writing a list of key terms without any further explanation. A combination of this analysis with a readability analysis from the syntactic level deactivates the chance to game the system this way and forces a good writing style. Moreover, when the text is manually graded by the teacher, a student that offers a readable answer is more likely to obtain better grades because the teacher better understands its meaning.
- PLeNTaS is oriented towards the evaluation of short answers, with a limit of 200 words. Despite Burrows et al. [19] set the length of short answers in a range from one phrase to one paragraph, it is hard to evaluate the syntax when the text contains only one phrase [20]. In short answers, syntax cannot be used as the only grading criterion. Furthermore, in higher education, syntax is not usually a frequent element of the evaluation model. However, given that a student who writes a more readable text is more likely to obtain a higher grade, it seems reasonable to analyze the readability and clarity of the short answers. Therefore, the PLeNTaS system will promote one or two paragraph answers, so that the readability can be graded by the system as a complement for the semantic analysis.
- The system must provide useful feedback so that the grade can be understood by both teachers and students. This is important for the students who need to learn from the received comments, and for the teacher who should accept or refuse the automatic grade, or even explain it to the students.
- Open Domain questions are candidates to be evaluated. In a domain specific approach, a knowledge model is difficult to build. A system that requires a high effort for the creation of a set of questions is less likely to be widely adopted for instructors and institutions. Furthermore, recent works [21, 22] reveal that deep learning techniques are capable of determining similarity between two texts, and therefore it is possible to determine if the student answer is similar to a known solution.
- The three levels of analysis (orthography, syntax and semantics) will consist in a list of lower-level metrics. For example, the semantic level will be determined by

methods such as checking the occurrence of a set of given keywords and measuring the similarity of the student response with a known valid solution. None of these methods is expected to provide a perfect solution, while their combination might produce a fair evaluation. All the provided metrics will set up the rubric used for the evaluation, and the teacher will be able to configure them, for example by setting their weight on the final grade, the number of allowed spelling mistakes, etc. In other words, the rubric is the instrument that allows the teacher to decide which analysis will be considered for the calculation of the final grade, and what is the weight of each of the analysis. One important fact is that the low-level metrics allow for a more detailed feedback, with at least one sentence for each of the elements being graded.

With all these above-described elements, the PLeNTaS system will be able to answer questions as the following one:

“In less than 200 words, explain what a cell is. In your response, you have to give answer to the following questions: what parts does a cell have? What types of cells are there? How many cells are required to form a living being?”

This sample question contains two relevant ideas. First, it starts with the limit of words so that the students know that the text should not be long, and that they should put some effort on summarizing their knowledge. Second, it splits the question in several smaller questions. With this, the teacher guides the specific content expected in the answer and allows for a more precise application of the semantic analysis. In other words, text similarity and text analysis with deep learning approaches work better with short sentences. A text of 200 words is hard to analyze, while the division of the question in smaller ones allows for a topic-by-topic analysis.



**Fig. 1.** Pedagogical model for the PLeNTaS system with LMS integration.

Finally, it is important to note that the PLeNTaS system will be used to support teacher’s tasks. That is, for each of the answers to grade, the teacher will receive a grade proposal and the feedback sentences. With this, the teacher may accept or refuse the grade, and accept or refuse the feedback comments. This described workflow is depicted in Fig. 1.

## 2.2 Levels of the Analysis

The automated grade proposed by PLeNTaS is obtained by analyzing the text from three different perspectives, as shown in Fig. 2. The figure summarizes the methodology of the analysis, which is described in this section and can be described in general terms as follows:

- The teacher proposes a question, which divides into sub questions. (S)he writes a set of keywords that should be in the student's response, and provides valid responses for each sub question separately.
- Each student response is analyzed from the orthographical point of view with a spellchecker, which counts the number of mistakes made. This number is one of the values considered in the rubric, which applies the corresponding weight for the final grade calculation.
- The readability of the student response is analyzed with two different instruments ( $\mu$  index [23] and Fernandez Huerta [24]). The result is averaged and used as an input for the rubric, where the corresponding weight is applied. The calculated ratios are used to build the sentences that serve as feedback.
- The student response is split into sentences, each of which is analyzed separately. First, by comparing a set of keywords with the response; second, with the use of NLP techniques for the calculation of semantic similarity with a given valid response. Each of these analyses are correspondingly weighted, with weights previously defined by the teacher in the configuration options, for the calculation of the final grade.
- The final grade, calculated by applying the rubric specification with the values obtained in each of the levels, and the feedback sentences are then offered to the teacher, who uses this input as a support to assign the grade.

### Orthography

Despite orthography is not usually graded by teachers in higher education, a recent study in the Spanish National University of Distance Learning reveals that there is considerable room for improvement in the orthography of university students' asynchronous digital writing, where a total of 71.3% of errors were conditioned by ignorance of the orthographic rules or incorrect use of the language [25].

Kukich [26] divided the types of possible spelling errors depending on their inclusion in a dictionary of correct words. That is:

- non-word errors, where the incorrect word form is not in the dictionary of correct words.
- real-word errors, where the word is spelled incorrectly but its form is in the dictionary of correct words.

While the first type is easy to detect, the second type is more difficult and can be detected with the contextual analysis of the word, that is, with syntactic and semantic information of the surrounding words. Furthermore, correctly spelled words can be syntactically or semantically incorrect in a sentence. This means that a result of zero spelling errors found

in the automated analysis does not imply a text with zero errors and thus, as previously stated, the score obtained in the orthographical level cannot represent the grade by itself.

Since many years ago, automatic spellcheckers are very common in word processors and are also integrated in web browsers and almost any software that allows to introduce text as an input. Despite challenges are still present such as improving recommendations that take context into account, or producing spellcheckers for low resource languages [27], automatic spelling correction is a well established field.

The output of current spellcheckers is the list of words identified as misspelled. This output allows for very specific feedback, also identifying the words and the evaluation criteria. For example: “you misspelled two words: ‘telephone’ and ‘dadabase’”. With these mistakes you have lost 2 points in the final grade”.

In summary, spelling errors should not be allowed in higher educational level. Despite automated spellcheckers do not guarantee a perfect detection, the maturity of the field and the simplicity of its integration in the PLeNTaS system are good arguments for its inclusion as a level of analysis. However, orthographic analysis is considered in PLeNTaS a complementary, optional module.

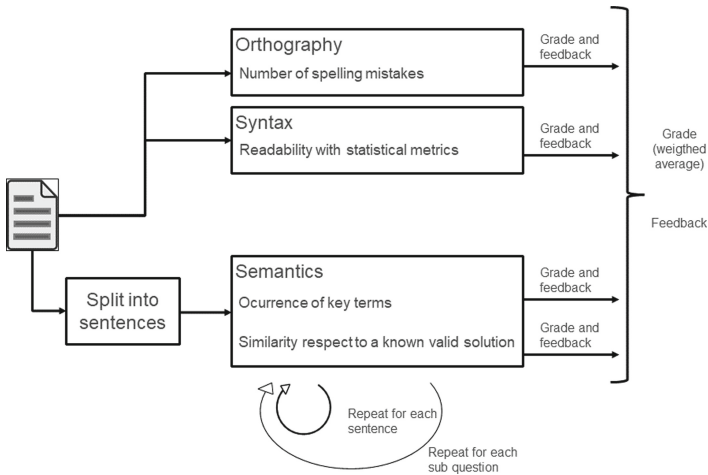


Fig. 2. General workflow of the analysis.

**Syntax**

George R. Klare [28] defined readability as the ease with which a reader can understand a document due to the style of writing. Of course, understanding a document is a subjective matter that depends on the reader. Therefore, readability should be understood from a statistical viewpoint. Readability is related to the complexity of the sentences and the words used in them. For example, long sentences with several dependent clauses demand more attention and they are therefore harder to understand.

Classical readability formulas such as the Fry Graph [29] or the Raygor estimate graph [30] are based on the sentences length, total number of sentences and number

of characters as the most relevant indicators. These statistical methods are language dependent, while they can be easily adapted to other languages by simply adapting the ratios to the proper values.

Other formulas such as Dale Chall [31] rely on the idea that a text is complex to read if the used words are complex to understand. This family of formulas count the average length of the words in the text, the number of longer words and the number of words that appear in a list of unfamiliar words. Depending on the formula, the list of unfamiliar words can be built with different criteria and depends on the educational level of the readers. Given that the list of unfamiliar words is highly language dependent, it is difficult to adapt such formulas to different languages.

Readability formulas has been used for years in many domains such as journalism, insurance, health, laws and, of course, education [32, 33]. According to [34], readability formulas are used with two main goals: first, to predict the readability of a given text; second, to assist in the production of readable writing.

Well known formulas set thresholds for different educational levels, and usually focus on secondary education. Students that enter higher education are supposed to have acquired enough writing skills (which is not necessarily true [35]) and therefore readability is not directly graded at the university. Furthermore, university teachers may have good writing skills, but they are not supposed to have the competence to grade them.

However, readability is still important for students while producing their short-answer texts, because the more readable is a text, the easier the teacher will understand its concepts. In other words, if the teacher does not understand the text, it is unlikely that the student gets a high grade. Furthermore, it is always a good practice to promote the production of readable text, no matter the knowledge field.

In PLeNTaS, readability is measured by applying two classic models for Spanish language, the  $\mu$  index [23] and Fernandez Huerta [24]. Both models are applied to the whole text provided by the student and will classify it in a scale that ranges from “very easy” to “very difficult”. Two indexes will be measured in order to obtain redundancy that ensures the validity of the diagnostic.

When measured from a statistical point of view, readability allows for the composition of formative feedback that explains the application of the grading criteria. It is possible, for example, to identify too long sentences as examples of bad practice, and to mention ratios that are poorly fulfilled. For example: *“the average length of your sentences is 34 words. Good readability requires an average of less than 20 words per sentence. Please consider splitting your long sentences”*.

Thus, being readability formulas a widely used method in many domains, PLeNTaS includes such analysis as part of the grading system. Therefore, a valid solution must contain the relevant concepts required in the question and must also have a readable writing style.

### **Semantics**

Given that a correctly written text should be readable and without orthographic errors, the actual grade should come from the semantic analysis. At this level, the following questions can be analyzed: 1) is the answer in the domain of knowledge of the posed question? 2) Is the text answering the question? And 3) is the answer correct? For

example, in the question about cells given as example in Sect. 2.1, we could have the following answers:

- “cells and living beings are both in the Earth”. This is in the domain of knowledge of the question, but it is not answering the question.
- “living beings require at least twelve cells to be considered as such”. This is in the domain of knowledge, provides an answer, but the answer is not correct.

The proposed method for the semantic analysis combines simple techniques such as keyword matching with state-of-the-art techniques for text similarity identification based on Deep Learning NLP.

In PLeNTaS, the teachers create the questions and do the set up (Fig. 1), which means that they provide a set of key terms that are supposed to be included in a valid answer, and a reference response considered correct. There is not a minimum number of key terms required, as this depends on the specific domain and question, although the general recommendation is to include between 4 and 6 keywords. This additional data included by the teachers is latter used to assess the validity of the student answers. Furthermore, as the question should be divided into sub-questions, the reference answer should be also divided into sub-responses, and therefore the analysis is guided by the sub-questions.

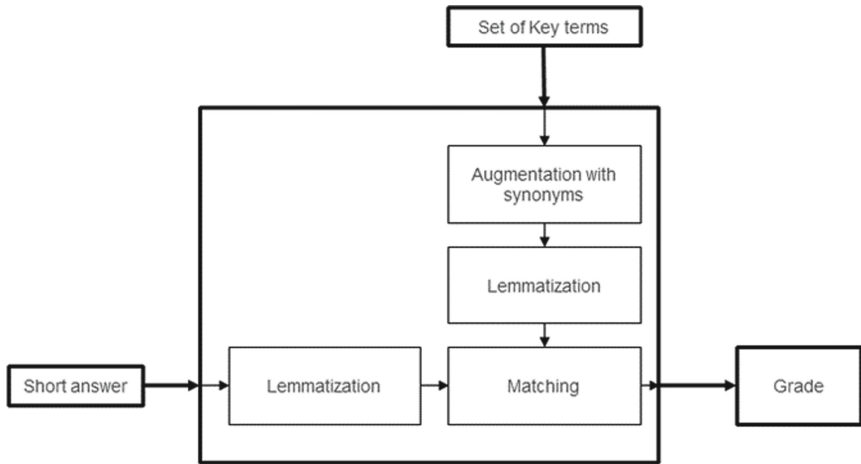
In PLeNTaS, we propose a sub-questions based analysis that first analyzes single sentences and then groups of consecutive sentences. For example, in a four-sentence paragraph, the analysis will be executed in the following order: {1}, {2}, {3}, {4}, {1, 2}, {2, 3}, {3, 4}, {1, 2, 3}, {2, 3, 4}, {1, 2, 3, 4}. This division alleviates the complexity of NLP (which is harder in longer texts) at the time that measures the accuracy of the specific sentences.

One existing approach in automated grading is key term matching. That is, given that the teacher provides a set of key terms that are expected to appear in the student’s answer, key term matching can be used. Such type of analysis only provides the answer to the first of the sematic questions: “is the answer in the domain of knowledge of the question?”.

As stated in [36], key terms matching has usually been considered a poor method, given that it is difficult to tackle problems such as synonymy or polysemy in the student answers. Despite this known problem, it is a simple approach that provides good enough results. For example, Omran and Ab Aziz [37] use Common Words (COW) as a metric for the distance between the student answer and the reference answer, which counts the number of words that are present in both texts and divide the result by the total of words in the sum of the two answers. Another example of use is the existing module for the Moodle LMS [38], that allows the teacher to configure a set of target phrases that should appear in the answer, and executes the matching with regular expressions.

As stated above, key terms matching has the problem of synonymy in the given terms. The PLeNTaS approach is the augmentation of the catalogue of key terms by checking them in a dictionary and including in the catalogue the direct synonyms. As depicted in Fig. 3, this produces an augmented catalogue that is lemmatized and compared with the lemmatized version of the student answer.





**Fig. 3.** Key term matching procedure in PLeNTaS.

However, in very specific domains of knowledge, key terms tend to be field specific, and they are unlikely to appear in general dictionaries. The potential result will be an augmented set of key terms equal to the original one. Thus, in the set up of the question the teachers should make their effort on creating a catalogue of key terms as complete as possible, including all the existing correct ways to mention a concept.

Another method to check the semantic correctness of the answer is the calculation of the semantic similarity with respect to the reference solution. There are many techniques reported in the literature for semantic similarity calculation. The existing methods can be divided in Knowledge-based methods, Corpus-based methods and Deep Learning approaches [39]. Knowledge-based methods consider the actual meaning of the words in the text by using knowledge models such as ontologies, thesauri, dictionaries, etc. This approach requires a high effort on building the knowledge model, and the result cannot be adopted in other knowledge domains. The approach of Corpus-based methods is based on the idea that ‘similar words occur together, frequently’ [40]. With this idea in mind, a large amount of text is processed so that the system finds statistical relations among words, without considering their actual meaning. Those relationships, called word embeddings, can be used to determine if two texts are similar, given that they will be similar if their words have shown to be similar in other texts, which is statistically measured. For example, the log-bilinear regression model GloVe provides word representations that are effective for word analogy, word similarity and named entity recognition tasks [41]. Learning high dimensional word vectors from large datasets, which have a huge number of words in the vocabulary, implies high computational costs. To alleviate these costs, simple model architectures requiring low computational complexity have been proposed, proving to be effective in word similarity tasks [42]. When word embeddings are analyzed with neural network techniques, usually Long Short-Term Memory (LSTM) or Bidirectional LSTM, this is considered a Deep Learning approach. A recent work combines Relational Graph Convolutional Networks (R-GCNs) with Bidirectional Encoder Representations from Transformers (BERT) to learn the contextual similarity

between sentences [43]. The latter has shown better performance, but Deep Learning approaches show the problem of their lack of interpretability. That is, given a result, it is difficult to explain how it was calculated.

The semantic similarity approach in PLeNTaS for automated grading offers an approximation for the second and third questions to be solved:

- Is the text answering the question? The approach can tell if the text is similar or not to a given correct answer. Considering that the analysis is performed sentence by sentence and that the whole question is divided into several sub-questions, it can be considered that a sentence with high similarity to the correct sub-answer to a given sub-question is answering the sub-question.
- Is the answer correct? With the current state of the technology in Natural Language Processing, semantic similarity will hardly provide a Yes/No response to this question. However, it can be reworded as “how far is this text from a valid answer?”, which is exactly the meaning of semantic similarity.

Feedback provision is a challenging task in the semantic analysis. For example, when applying Deep Learning approaches for Question Answering, the system is able to identify the specific words that are likely the response to a specific question. However, in the case the system finds nothing, it is not able to justify why. This output leads, in most cases, to vague and unspecific feedback. The PLeNTaS project will split the total answer in smaller sentences and analyze them separately. This fact allows identifying the similarity of specific sentences with the real answer. Therefore, the PLeNTaS system will be able to create feedback such as “the question ‘What types of cells are there?’ has not been answered. Sentence 3 is near to the actual response, but not enough accurate. The expected response is ‘(...)’.” Although far from the capabilities of a human teacher, this type of feedback allows for the identification of strong and weak aspects in the response.

The PLeNTaS system proposes a hybrid approach that will combine the better results of Deep Learning NLP approaches with the explainability of knowledge-based analysis. As the analysis is repeated for each of the sub-questions and executed at the level of feedback sentence, the grading process is expected to determine the meaningful sentences and therefore provide the useful feedback for the sentences poorly graded. Given that deep learning solutions require large datasets, transfer learning will be adopted. Therefore, a pre-trained model like BERT will be fine-tuned with the set of real answers collected in the project [44].

### 2.3 Validation

Most of the automated grading systems in the literature are validated from a statistical point of view. That is, they measure the correlation of the grade calculated by the system with the grade assigned by the human teachers. While this is a fair and important measure, it is also important to measure the pedagogical effectiveness of the tool. In other words, the PLeNTaS validation stage will measure the achievement of the project from three different perspectives: 1) the system precision on the automatic evaluation task when compared to human evaluations, 2) the appropriateness of the elaborated feedback, and

3) the overall impact on the teaching-learning process. Although the project is still being developed, the validation plan includes the following stages:

In the first stage, in a laboratory setting, we compiled a corpus of 660 short-answers collected from previous real exams. These exams belong to four different master's degrees in an online university in the first semester of 2021. The students completed their answers in a fully online mode, where a proctoring system supervised the proper conduct of the students during the examination. The collected questions were written according to the template given for the PLeNTaS question type, and the answers were valid for the final grade. The teachers graded such answers without any consideration coming from the researchers. This corpus is the base to do preliminary tests of the different aspects of the PLeNTaS system. The *system precision* seeks for the correlation between machine and human grading. To do this evaluation, every question in the corpus includes the grade assigned by the teacher, which will be compared to the one returned by the system. The results obtained will be contrasted to other tools in the state of the art.

For the continuous evaluation pedagogical approach, where formative grading is important, a fair system precision is not enough for a tool to be adopted. It is also important to measure the *validity of the feedback* offered by the system. The validation of the feedback will combine quantitative and qualitative methods. First, the teachers will receive the feedback and the grade for the automatically graded answers and will be asked if (i) the feedback is in agreement with the calculated grade and (ii) the feedback allows for the identification of the aspects for improvement. A qualitative analysis of those answers will give the opportunity to adjust the system.

In a second stage, further tests of validity of the feedback will be developed, together with the *educational impact*. To widen the validity of the feedback, the same questions will be presented to a set of volunteer students. The evaluation of the *effects on educational quality and student experience*, which is highly relevant for automated grading systems [45], will be performed in a pilot course that incorporates automated grading and feedback as a part of the course. During a 15-weeks course, the students will answer 3–4 PLeNTaS-templated questions and the teacher will grade them with the support of the developed system. During the course, the researchers will annotate any observation from the teacher and, after the course, they will interview the teacher to understand the impact of the system in the course. Additionally, a TAM (Technology Acceptance Model) [46] test will be delivered both to teachers and students in order to evaluate their perceived ease of use and usefulness.

### 3 Conclusions

PLeNTaS is an on-going project that finalizes in June 2023 and whose main objective is the development and validation of an automated grading system. This project is especially relevant for online, higher education, where frequent sub-missions would create better engagement in the course experience.

Automated grade is calculated with a hybrid approach that considers three levels of analysis: orthography, syntax and semantics. While the first two levels are well established in the state of the art, their use in conjunction to the semantic level will offer more complete feedback to the students.

The analysis of the semantic level is divided in two ways: firstly, the question posed by the teachers is divided into sub-questions, according to the template proposed by the project PLeNTaS. Each of these sub-questions are analyzed separately, which allows for a more accurate analysis and feedback. Secondly, the 200-words answer provided by the student is divided into separate sentences, which are analyzed one by one. This strategy seeks a more accurate detection of the semantics in the text, and therefore more accurate feedback. The semantic analysis proposes a hybrid approach that combines knowledge-based and deep-learning approaches for the seek of semantic similarity.

The project is still in an initial stage, where next steps are the implementation of the three-levels model and the use of the already collected student answers to validate the proposed model.

**Acknowledgements.** This Work is partially funded by the PLeNTaS project, “Proyectos I+D+i 2019”, PID2019-111430RB-I00, by the PL-NETO project, Proyecto PROPIO UNIR, projectId B0036, and by Universidad Internacional de la Rioja (UNIR), through the Research Institute for Innovation & Technology in Education (UNIR iTED, <http://ited.unir.net>).

## References

1. Westera, W., Dascalu, M., Kurvers, H., et al.: Automated essay scoring in applied games: reducing the teacher bandwidth problem in online training. *Comput. Educ.* **123**, 212–224 (2018). <https://doi.org/10.1016/J.COMPEDU.2018.05.010>
2. McNamara, D.S., Crossley, S.A., Roscoe, R.D., et al.: A hierarchical classification approach to automated essay scoring. *Assess. Writ.* **23**, 35–59 (2015). <https://doi.org/10.1016/J.ASW.2014.09.002>
3. Campbell, J.R.: Cognitive processes elicited by multiple-choice and constructed-response questions on an assessment of reading comprehension. Temple University (UMI No. 9938651) (1999)
4. Rodrigues, F., Oliveira, P.: A system for formative assessment and monitoring of students’ progress. *Comput. Educ.* **76**, 30–41 (2014). <https://doi.org/10.1016/J.COMPEDU.2014.03.001>
5. Brame C.J.: Rubrics: tools to make grading more fair and efficient. In: *Science Teaching Essentials*, pp. 175–184. Academic Press (2019)
6. Prasad Mudigonda, K.S., Sharma, P.: multi-sense embeddings using synonym sets and hypernym information from wordnet. *Int. J. Interact. Multimed. Artif. Intell.* **6**, 68 (2020). <https://doi.org/10.9781/ijimai.2020.07.001>
7. Zhou, S., Chen, B., Zhang, Y., et al.: A feature extraction method based on feature fusion and its application in the text-driven failure diagnosis field. *Int. J. Interact. Multimed. Artif. Intell.* **6**, 121 (2020). <https://doi.org/10.9781/ijimai.2020.11.006>
8. Rao, S.B.P., Agnihotri, M., Babu Jayagopi, D.: Improving asynchronous interview interaction with follow-up question generation. *Int. J. Interact. Multimed. Artif. Intell.* **6**, 79 (2021). <https://doi.org/10.9781/ijimai.2021.02.010>
9. Dascalu, M.: readerbench (1) - cohesion-based discourse analysis and dialogism, pp. 137–160 (2014)
10. Ramineni, C.: Automated essay scoring: psychometric guidelines and practices. *Assess. Writ.* **18**, 25–39 (2013). <https://doi.org/10.1016/J.ASW.2012.10.004>

11. McNamara, D.S., Levinstein, I.B., Boonthum, C.: iSTART: interactive strategy training for active reading and thinking. *Behav. Res. Methods Instr. Comput.* **36**, 222–233 (2004). <https://doi.org/10.3758/BF03195567>
12. Graesser, A.C., McNamara, D.S., Kulikowich, J.M.: Coh-metrix. *Educ. Res.* **40**, 223–234 (2011). <https://doi.org/10.3102/0013189X11413260>
13. Panaite, M., Dascalu, M., Johnson, A., et al.: Bring it on! Challenges encountered while building a comprehensive tutoring system using ReaderBench. In: Penstein, R.C., et al. (eds.) AIED 2018. LNCS (LNAI and LNB), vol. 10947, pp. 409–419. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-93843-1\\_30](https://doi.org/10.1007/978-3-319-93843-1_30)
14. Cuzzocrea, A., Bosco, G.L., Pilato, G., Schicchi, D.: Multi-class text complexity evaluation via deep neural networks. In: Yin, H., Camacho, D., Tino, P., Tallón-Ballesteros, A.J., Menezes, R., Allmendinger, R. (eds.) IDEAL 2019. LNCS, vol. 11872, pp. 313–322. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-33617-2\\_32](https://doi.org/10.1007/978-3-030-33617-2_32)
15. Zhang, Y., Chen, X.: Explainable Recommendation: A Survey and New Perspectives (2018)
16. Alonso, J.M., Casalino, G.: Explainable artificial intelligence for human-centric data analysis in virtual learning environments. In: Burgos, D., et al. (eds.) HELMeTO 2019. CCIS, vol. 1091, pp. 125–138. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-31284-8\\_10](https://doi.org/10.1007/978-3-030-31284-8_10)
17. Saarela, M., Heilala, V., Jaaskela, P., et al.: Explainable student agency analytics. *IEEE Access* **9**, 137444–137459 (2021). <https://doi.org/10.1109/ACCESS.2021.3116664>
18. Kent, C., Laslo, E., Rafaeli, S.: Interactivity in online discussions and learning outcomes. *Comput. Educ.* **97**, 116–128 (2016). <https://doi.org/10.1016/J.COMPEDU.2016.03.002>
19. Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answer grading. *Int. J. Artif. Intell. Educ.* **25**, 60–117 (2015)
20. Pérez-Marín, D., Pascual-Nieto, I., Rodríguez, P.: Computer-assisted assessment of free-text answers. *Knowl. Eng. Rev.* **24**, 353–374 (2009). <https://doi.org/10.1017/S026988890999018X>
21. Mohler, M., Mihalcea, R.: Text-to-text semantic similarity for automatic short answer grading (2009). (3AD)
22. Gautam, D., Rus, V.: Using neural tensor networks for open ended short answer assessment. In: Bittencourt, I.L., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12163, pp. 191–203. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-52237-7\\_16](https://doi.org/10.1007/978-3-030-52237-7_16)
23. Muñoz Baquedano, M.: Legibilidad y variabilidad de los textos. *Boletín Investig. Educ.* **21**, 13–25 (2006)
24. Fernandez Huerta, J.: Medidas sencillas de lecturabilidad. *Consiga* **214**, 29–32 (1959)
25. Vázquez-Cano, E., González, A.I.H., Sáez-López, J.M.: An analysis of the orthographic errors found in university students' asynchronous digital writing. *J. Comput. High. Educ.* **31**(1), 1–20 (2018). <https://doi.org/10.1007/s12528-018-9189-x>
26. Kukich, K.: Techniques for automatically correcting words in text. *ACM Comput. Surv.* **24**, 377–439 (1992). <https://doi.org/10.1145/146370.146380>
27. Hládek, D., Staš, J., Pleva, M.: Survey of automatic spelling correction. *Electronics* **9**, 1–29 (2020)
28. Klare, G.R.: *The Measure of Readability*. University of Iowa Press, Ames (1963)
29. Fry, E.: A readability formula that saves time. *J. Read.* 513–516, 575–578 (1968). (8 pages)
30. Raygor, A.L.: The Raygor readability estimate: a quick and easy way to determine difficulty. *Read. Theory Res. Pract.* **1977**, 259–263 (1977)
31. Dale, E., Chall, J.S.: A formula for predicting readability. *Educ. Res. Bull.* **27**(1), 11–28 (1948). <http://www.jstor.org/stable/1473169>
32. Crossley, S.A., Skalicky, S., Dascalu, M.: Moving beyond classic readability formulas: new methods and new models. *J. Res. Read.* **42**, 541–561 (2019). <https://doi.org/10.1111/1467-9817.12283>

33. Morato, J., Iglesias, A., Campillo, A., Sanchez-Cuadrado, S.: Automated readability assessment for spanish e-government information. *J. Inf. Syst. Eng. Manag.* **6**, em0137 (2021). <https://doi.org/10.29333/jisem/9620>
34. Klare, G.R.: A second look at the validity of readability formulas. *J. Read. Behav.* **8**, 129–152 (1976). <https://doi.org/10.1080/10862967609547171>
35. Taylor, Z.W.: College admissions for L2 students: comparing L1 and L2 readability of admissions materials for U.S. higher education. *J. Coll. Access.* **5**(1) (2020). <https://scholarworks.wmich.edu/jca/vol5/iss1/6>. Article 6
36. Selvi, P., Bnerjee, D.A.K.: Automatic short-answer grading system (ASAGS) (2010)
37. Ben, O.A.M., Ab Aziz, M.J.: Automatic essay grading system for short answers in English language. *J. Comput. Sci.* **9**, 1369–1382 (2013). <https://doi.org/10.3844/jcssp.2013.1369.1382>
38. Essay (auto-grade) question type - MoodleDocs
39. Chandrasekaran, D., Mago, V.: Evolution of semantic similarity – a survey. *ACM Comput. Surv.* **54** (2020). <https://doi.org/10.1145/3440755>
40. Gorman, J., Curran, J.R.: Scaling distributional similarity to large corpora. In: *COLING/ACL 2006 - 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pp. 361–368. Association for Computational Linguistics (ACL), Morristown (2006)
41. Pennington, J., Socher, R., Manning, C.D.: GloVe: global vectors for word representation. In: *EMNLP 2014 – Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543 (2014). <https://doi.org/10.3115/V1/D14-1162>
42. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013)
43. Xu, S., Shen, X., Fukumoto, F., et al.: Paraphrase identification with lexical, syntactic and sentential encodings. *Appl. Sci.* **10**, 4144 (2020). <https://doi.org/10.3390/APP10124144>
44. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: Pre-trained models for natural language processing: a survey. *Sci. China Technol. Sci.* **63**(10), 1872–1897 (2020). <https://doi.org/10.1007/s11431-020-1647-3>
45. Hahn, M.G., Navarro, S.M.B., De La Fuente, V.L., Burgos, D.: A systematic review of the effects of automatic scoring and automatic feedback in educational settings. *IEEE Access* **9**, 108190–108198 (2021). <https://doi.org/10.1109/ACCESS.2021.3100890>
46. Davis, F.D.: Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* **13**, 319–339 (1989). <https://doi.org/10.2307/249008>