GRASPE: An Adaptive Grading Support for Persuasive English Essays

Faten Ghali
Master Semester Project, Fall 2022

Abstract – **The purpose of this work is to design and implement a tool to help English teachers with grading persuasive essays. We present the user stories on which we based the design of our tool as well as the different techniques used for the implementation. Our tool uses three Machine Learning language models to detect spelling mistakes, grammar incoherence, and claim and premises in arguments. This report presents the first version of the tool as it is still under construction and needs further insights from English teachers in order to improve it and have it tested.**

I. Introduction

Among all ways of language evaluations, writing assignments remain one of the most effective and reliable ways to capture the student's skills or, for the matter, lack of them. In fact, writing demonstrates not only the student's level of mastering the language, but also their critical thinking and their abilities to express themselves and explain their opinions. To the present, the evaluation of writing skills is being done only manually. Even though the task of manually grading writing essays takes between 20 to 30 minutes per assignment in the best scenario leading to a total of 40 hours per class per academic year at minimum[1], teachers cannot give up this exercise. Fortunately, Natural Language Processing (NLP) advances come in handy to read and understand text content and can be leveraged to automate the scoring of essays. However, automatic scoring techniques are still not showing promising results and cannot totally replace the teacher because they can be biased and not scalable. In this work, we propose GRASPE, an IT tool to help teachers with grading argumentative essays by automating certain steps of this task. The tool is aimed to be tailored to the teachers' needs and to be adaptive accordingly offering them the space to give the feedback and the final grade.

II. Related work

Multiple papers investigate the different NLP-based approaches to improve automated corrections. Yuan et al. in [8] propose a context-aware grammatical error correction (GEC) system outperforming previous GEC systems. Wang et al. on the other hand suggest a novel representation for BERT that improves its performance and outperformes other deep learning models such LSTM [6]. These are two of many examples of projects working towards making automated scoring faster and more accurate.

As examples of applications that use these state of the art NLP models, AL [5] and ALEN [4] are both IT tools designed to help English learners to improve their writing skills by giving them feedback on the quality of their argument structures. By leveraging argumentation mining techniques, these tools detect the claims and premises in the text and the links between them as a primary feature to judge the quality of the arguments.

Sevcikova in [2] makes the comparison between human and automated scoring. She shows that, even though the computerized feedback is fast and efficient in large classes, it lacks the ability to evaluate abstract features like comprehensibility, content and aesthetic style. Also, automated scoring techniques cannot give a feedback focusing on a specific grade range that is associated with a specific set of tasks and a specific rubric as most human raters are trained to do. Sevcikova suggests that these technologies should be used carefully to contribute to the efficient delivery of essay scoring and ease the burden of the grading task but not to entirely replace the student-instructor interaction.

Thus far, we have available various technologies capable of detecting grammar, spelling, and argumentation errors with very high performances that have been used in applications to give automated feedbacks to students. However, for what we know, these techniques have never been used in applications to help the teachers with

[1] 20 students per class, 6 assignments per year.

the grading task. Additionally, along with Sevcikova's critique, the tool we propose in the aim of filling this gap, is intended to be an automated support for the teachers to make parts of grading faster and easier but still provide them enough space to give their personal feedback.

III. User stories

To design a tool as tailored to the user's needs as possible, we intended to interview 12-to-15 English teachers from schools in Switzerland. The first interviews were designed to extract the user stories based on which we choose the different features of the tool. Our plan was then to have another one or two interviews with each teacher to try the tool and give us feedbacks to improve it. The last step of our plan was to evaluate the tool through a real-life experiment where the teachers would use the eventual final version of the tool to grade the assignments of their students. Unfortunately things did not go as planned. We managed to interview only three teachers who then did not follow-up for a second interview. These are the few stories we could draw from the interviews:

- I would like to be able to see two assignments of two different students on the same view for comparison.
- I would like to be able to deactivate the options (i.e. automatic checks) when I do not need them because otherwise the interface becomes cumbersome.
- I would like to know the percentage of faulty words in the text.
- I would like to see which assignments were already graded to keep track of my work.
- I would like to have access to the student's past assignments and feedbacks to see if they overcame their past mistakes.
- I would like to have key words highlighted in the text to see how well did the student stick to the topic, and how rich is the lexicon they are using.

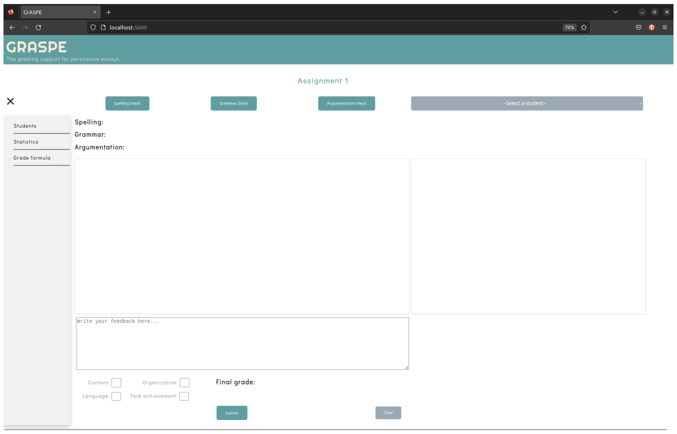
As you would notice from the description of the tool, we tried to rely as much as possible on these stories to design our tool, but also used literature. We would like to insist here on the importance of running the planned experiment even at a late stage of the development of the tool because from the three interviews we had, we could already see a number of differences between the teachers' needs and ways of doing when it comes to grading. For example, one interviewed teacher considers identifying the content and ideas in the text to be the most important part of grading, whereas another teacher thinks that, depending on the level of the student, grammar is sometimes more important than content. Also, one teacher would like to use past assignments of a student to check whether or not they are making the same mistakes, while another teacher would grade each assignment individually and would not penalize more a student for making the same mistake again. Such differences between the teachers can be linked to their experiences, to the levels of their students, or to their workload (number of students and writing assignments per year). This kind of hypothesis and other can only be validated or rejected when conducting the experiment with more teachers.

IV. Implementation

We decided for GRASPE to be a simple web interface for the ease of testing and also to save the teachers the struggle of a software installation. To design the interface, we relied on the few user stories that we collected from the interviews we had with the teachers, and mostly on literature and common sense. The interface consists of a main view used for the grading task, presented in Fig. 1, and two other views for statistic and final grade formula described later. All the features related to the task of grading are contented in the main ”grading view” on purpose to provide more flexibility and less going back and forth between pages for different small tasks.

The grading view contains three main correction tools: detection of spelling mistakes, detection of grammar incoherence, detection of claims and premises for argumentation. The three tasks rely on three different language models listed below. As the main purpose of the project is the design of the tool, we decided to leverage the variety of existing language models rather than build ones ourselves. The grammar and argumentation models have been pushed to the Hugging Face Hub [2] after training to avoid local storage of heavy files and speed up the loading time.

[2] https://huggingface.co/docs/hub/index



Spelling:
Grammar:
Argumentation:

Write your feedback here...

Content: □ Organization: □
Language: □ Task achievement: □

Final grade:

Fig. 1: Screenshot of the main grading view of GRASPE.

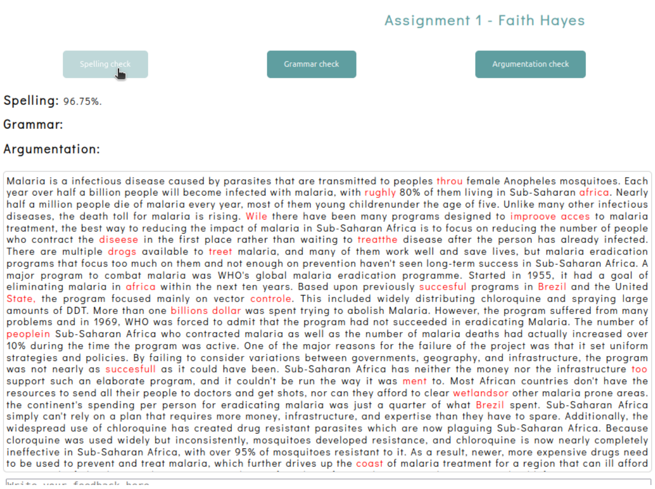
A. Language models

Spelling. For the detection of spelling mistakes, we used NeuSpell - A Neural Spelling Correction Toolkit[3]. It comprises several neural models such as BERT and ELMo that accurately capture context around the misspellings [1]. The final model was trained on 1.6M sentences with spelling mistakes injected synthetically and tested against 63K sentences extracted from real students' written essays. Details of the remarkable accuracy and correction rates of the toolkit are represented in the paper.

These strengths of NeuSpell are well confirmed when used on our test examples and compared to other models. It however doesn't seem to take into account punctuation mistakes. Even though it takes 2 to 3 seconds to generate the corrected text when run on an average long text, NeuSpell is by far faster than other pre-trained models that we tested.

When the spelling check is activated, the metrics area above the text shows the percentage of correct words of the total number of words in the text.

Grammar. For the grammar check, we used BertForSequenceClassification [4] from the Hugging Face Transformers package [5]. For the training of the model, we used 8,551 sentences from The Corpus of Linguistic Acceptability (CoLA) dataset [7] which is a set of sentences labeled as grammatically correct or incorrect.



Spelling: 96.75%.

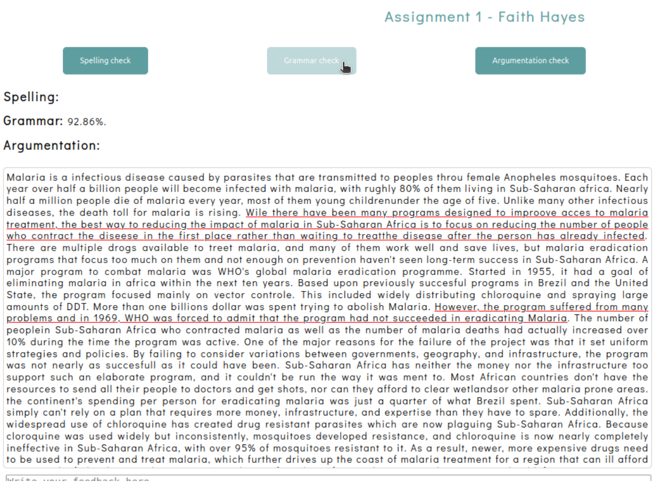
Grammar:

Argumentation:

Malaria is a infectious disease caused by parasites that are transmitted to peoples throu female Anopheles mosquitoes. Each year over half a billion people will become infected with malaria, with rughly 80% of them living in Sub-Saharan africa. Nearly half a million people die of malaria every year, most of them young childrenunder the age of five. Unlike many other infectious diseases, the death toll for malaria is rising. Wile there have been many programs designed to improove acces to malaria treatment, the best way to reducing the impact of malaria in Sub-Saharan Africa is to focus on reducing the number of people who contract the diseese in the first place rather than waiting to treatthe disease after the person has already infected. There are multiple drogs available to treet malaria, and many of them work well and save lives, but malaria eradication programs that focus too much on them and not enough on prevention haven't seen long-term success in Sub-Saharan Africa. A major program to combat malaria was WHO's global malaria eradication programme. Started in 1955, it had a goal of eliminating malaria in africa within the next ten years. Based upon previously succesful programs in Brezil and the United State, the program focused mainly on vector controle. This included widely distributing chloroquine and spraying large amounts of DDT. More than one billions dollar was spent trying to abolish Malaria. However, the program suffered from many problems and in 1969, WHO was forced to admit that the program had not succeeded in eradicating Malaria. The number of peoplein Sub-Saharan Africa who contracted malaria as well as the number of malaria deaths had actually increased over 10% during the time the program was active. One of the major reasons for the failure of the project was that it set uniform strategies and policies. By failing to consider variations between governments, geography, and infrastructure, the program was not nearly as succesfull as it could have been. Sub-Saharan Africa has neither the money nor the infrastructure too support such an elaborate program, and it couldn't be run the way it was ment to. Most African countries don't have the resources to send all their people to doctors and get shots, nor can they afford to clear wetlandsor other malaria prone areas. the continent's spending per person for eradicating malaria was just a quarter of what Brezil spent. Sub-Saharan Africa simply can't rely on a plan that requires more money, infrastructure, and expertise than they have to spare. Additionally, the widespread use of chloroquine has created drug resistant parasites which are now plaguing Sub-Saharan Africa. Because cloroquine was used widely but inconsistently, mosquitoes developed resistance, and chloroquine is now nearly completely ineffective in Sub-Saharan Africa, with over 95% of mosquitoes resistant to it. As a result, newer, more expensive drugs need to be used to prevent and treat malaria, which further drives up the coast of malaria treatment for a region that can ill afford

Write your feedback here

Fig. 2: Screenshot of GRASPE showing the spelling check feature.



Spelling:

Grammar: 92.86%.

Argumentation:

Malaria is a infectious disease caused by parasites that are transmitted to peoples throu female Anopheles mosquitoes. Each year over half a billion people will become infected with malaria, with rughly 80% of them living in Sub-Saharan africa. Nearly half a million people die of malaria every year, most of them young childrenunder the age of five. Unlike many other infectious diseases, the death toll for malaria is rising. Wile there have been many programs designed to improove acces to malaria treatment, the best way to reducing the impact of malaria in Sub-Saharan Africa is to focus on reducing the number of people who contract the diseese in the first place rather than waiting to treatthe disease after the person has already infected. There are multiple drogs available to treet malaria, and many of them work well and save lives, but malaria eradication programs that focus too much on them and not enough on prevention haven't seen long-term success in Sub-Saharan Africa. A major program to combat malaria was WHO's global malaria eradication programme. Started in 1955, it had a goal of eliminating malaria in africa within the next ten years. Based upon previously succesful programs in Brezil and the United State, the program focused mainly on vector controle. This included widely distributing chloroquine and spraying large amounts of DDT. More than one billions dollar was spent trying to abolish Malaria. However, the program suffered from many problems and in 1969, WHO was forced to admit that the program had not succeeded in eradicating Malaria. The number of peoplein Sub-Saharan Africa who contracted malaria as well as the number of malaria deaths had actually increased over 10% during the time the program was active. One of the major reasons for the failure of the project was that it set uniform strategies and policies. By failing to consider variations between governments, geography, and infrastructure, the program was not nearly as succesfull as it could have been. Sub-Saharan Africa has neither the money nor the infrastructure too support such an elaborate program, and it couldn't be run the way it was ment to. Most African countries don't have the resources to send all their people to doctors and get shots, nor can they afford to clear wetlandsor other malaria prone areas. the continent's spending per person for eradicating malaria was just a quarter of what Brezil spent. Sub-Saharan Africa simply can't rely on a plan that requires more money, infrastructure, and expertise than they have to spare. Additionally, the widespread use of chloroquine has created drug resistant parasites which are now plaguing Sub-Saharan Africa. Because cloroquine was used widely but inconsistently, mosquitoes developed resistance, and chloroquine is now nearly completely ineffective in Sub-Saharan Africa, with over 95% of mosquitoes resistant to it. As a result, newer, more expensive drugs need to be used to prevent and treat malaria, which further drives up the coast of malaria treatment for a region that can ill afford

Write your feedback here

Fig. 3: Screenshot of GRASPE showing the grammar check feature.

To evaluate the model on the test set, we used the "Matthews correlation coefficient" (MCC) metric because the classes are imbalanced. We got an MCC of 68.60% on 516 sentences. The model works well on the examples given in our interface. The output seems however to be impacted by the spelling mistakes; i.e. a grammarly correct sentence can be marked as incorrect when it contains a spelling mistake even if the overall structure is correct.

The percentage shown in the metrics area is output after the grammar check completes running and it represents the percentage of grammarly correct sentences from the total number of sentences in the text.

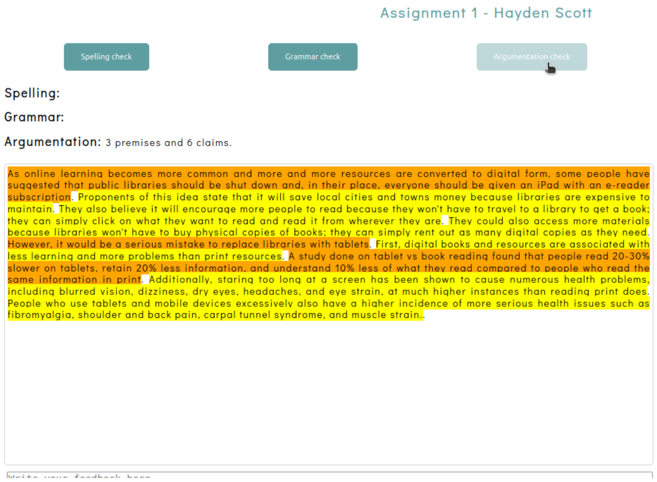
[3] https://github.com/neuspell/neuspell

[4] https://huggingface.co/transformers/v3.0.2/model_doc/bert.html#transformers.BertForSequenceClassification

[5] https://huggingface.co/docs/transformers/index

Argumentation. For the argumentation check, we used the AutoModelForSequenceClassification model class from the Transformers package to instantiate a sequence classification model class from a pre-trained configuration of the DistilBERT model. We choosed DistilBERT over BERT mainly for it being lighter and faster. We trained our text classification model on the Feedback Prize dataset from Kaggle. It is the biggest dataset of student writing ever released containing over 144K annotated sentences extracted from argumentative essays written by U.S. students in grades 6-12. The essays were annotated by expert raters for elements commonly found in argumentative writing. The labels originally stated in the dataset are: Lead, Position, Claim, Counterclaim, Rebuttal, Evidence, and Concluding Statement. To make things easier for a starter we narrowed the list down to 2 labels by considering both ”Claim” and ”Counterclaim” as Claim, and the remainder as Premise. This choice was made for the first, rather simple, version of the interface but we intend to implement the 7-label-classification if it is more appealing to the teachers.

When the ”Argumentation check” button is clicked, claims are highlighted in yellow and premises in orange, and the number of each is displayed in the metrics area above the text. We assume that this quantitative representation of the output can be helpful to the teacher to give them an idea on how dense or scarce is the argumentation, and the highlighting is more of a qualitative representation to show how balanced are the claims with respect to the premises and how well are they distributed through the text.



Assignment 1 - Hayden Scott

Spelling:

Grammar:

Argumentation: 3 premises and 6 claims.

As online learning becomes more common and more and more resources are converted to digital form, some people have suggested that public libraries should be shut down and, in their place, everyone should be given an iPad with an e-reader subscription. Proponents of this idea state that it will save local cities and towns money because libraries are expensive to maintain. They also believe it will encourage more people to read because they won't have to travel to a library to get a book; they can simply click on what they want to read and read it from wherever they are. They could also access more materials because libraries won't have to buy physical copies of books; they can simply rent out as many digital copies as they need. However, it would be a serious mistake to replace libraries with tablets. First, digital books and resources are associated with less learning and more problems than print resources. A study done on tablet vs book reading found that people read 20-30% slower on tablets, retain 20% less information, and understand 10% less of what they read compared to people who read the same information in print. Additionally, staring too long at a screen has been shown to cause numerous health problems, including blurred vision, dizziness, dry eyes, headaches, and eye strain, at much higher instances than reading print does. People who use tablets and mobile devices excessively also have a higher incidence of more serious health issues such as fibromyalgia, shoulder and back pain, carpal tunnel syndrome, and muscle strain..

Fig. 4: Screenshot of GRASPE showing the argumentation check feature.

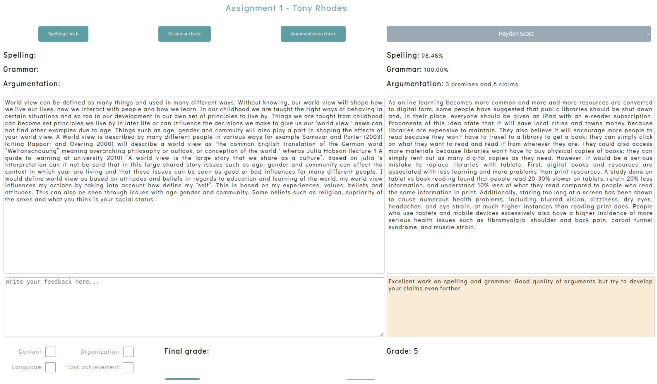
The model is accurate to 87%.

B. Other features

Feedback. As stated previously, this tool is not an automated scoring tool. It is a support that would help the teachers speed up the grading of essays while giving them enough space to accomplish their task in the ways they are used to. One of the most important parts of grading is the teacher's feedback. They precisely highlight the strengths and weaknesses of the student's writing, and can give them ways to improve. For these and other reasons, it only makes sens to provide a space on the interface for the teachers to type a personalized feedback.

Grade attribution. Under the span for the feedback, we put a section to give a grade. This section was mostly inspired from the analytic scoring method described in [3]. In this paper, Veloo et al. state that this type of scoring offers more detailed explanation on the writer's performance than one single score done in a holistic scoring. Even though Veloo et al. consider this method to be consisted of five different ratings (content, organization, vocabulary, language use and mechanics), we are here considering four categories only based on the needs of the teachers we interviewed: content, organization, language, task achievement. As a reminder, this is only the first version of the tool; and because of the lack of teachers' feedbacks, this version is prone to modifications. To make our tool as adaptive as possible, we can make this section to be adjustable to the user's will. I.e. we can implement a feature allowing the teacher to add, remove or edit the categories of sub-grades; or to remove the section completely if for example they are used to the holistic method of scoring which consists of giving a single and integrated score based on a general impression of writing [3]. In order to help the teacher save more time, and in case they are using the analytic method, we provide a separate view where they can define the formula to compute the final grade automatically based on the sub-grades of the different categories.

Statistics. Also inspired from the paper by Veloo et al., we thought of adding a second view to our tool where we would display statistics on the overall level of the class or on students individually. This feature goes along with the analytic method of grading because it would help the teachers to discriminate the students' weak and strong aspects of their writing performance from one test to another, and based on that they would adjust the



Assignment 1 - Tony Rhodes

Spelling:

Grammar:

Argumentation:

World view can be defined as many things and used in many different ways. Without knowing, our world view will shape how we live our lives, how we interact with people and how we learn. In our childhood we are taught the right ways of behaving in certain situations and so too in our development in our own set of principles to live by. Things we are taught from childhood can become set principles we live by in later life or can influence the decisions we make to give us our 'world view ' aswe can not find other examples due to age. Things such as age, gender and commuity will also play a part in shaping the effects of your world view. A World view is described by many different people in various ways for example Samovar and Porter (2003) (citing Rapport and Overing 2000) will describe a world view as 'the common English translation of the German word "Weltanschauung" meaning overarching philosophy or outlook, or conception of the world ' wheras Julia Hobson (lecture 1 A guide to learning at university 2010) "A world view is the large story that we share as a culture". Based on julia 's interpretation can it not be said that in this large shared story issues such as age, gender and community can effect the context in which your are living and that these issues can be seen as good or bad influences for many different people. I would define world view as based on attitudes and beliefs in regards to education and learning of the world, my world view influences my actions by taking into account how define my "self". This is based on my experiences, values, beliefs and attitudes. This can also be seen through issues with age gender and community. Some beliefs such as religion, supriority of the sexes and what you think is your social status.

Write your feedback here...

Content: □ Organization: □

Language: □ Task achievement: □

Final grade:

Hayden Scott

Spelling: 98.48%

Grammar: 100.00%

Argumentation: 3 premises and 6 claims.

As online learning becomes more common and more and more resources are converted to digital form, some people have suggested that public libraries should be shut down and, in their place, everyone should be given an iPad with an e-reader subscription. Proponents of this idea state that it will save local cities and towns money because libraries are expensive to maintain. They also believe it will encourage more people to read because they won't have to travel to a library to get a book; they can simply click on what they want to read and read it from wherever they are. They could also access more materials because libraries won't have to buy physical copies of books; they can simply rent out as many digital copies as they need. However, it would be a serious mistake to replace libraries with tablets. First, digital books and resources are associated with less learning and more problems than print resources. A study done on tablet vs book reading found that people read 20-30% slower on tablets, retain 20% less information, and understand 10% less of what they read compared to people who read the same information in print. Additionally, staring too long at a screen has been shown to cause numerous health problems, including blurred vision, dizziness, dry eyes, headaches, and eye strain, at much higher instances than reading print does. People who use tablets and mobile devices excessively also have a higher incidence of more serious health issues such as fibromyalgia, shoulder and back pain, carpal tunnel syndrome, and muscle strain.

Excellent work on spelling and grammar. Good quality of arguments but try to develop your claims even further.

Grade: 5

Fig. 5: Screenshot of GRASPE showing the comparison feature.

follow-up activities of consolidation after the tests [3]. The statistics view is still not implemented in this version of the interface mainly because we need more insights on what to include in it and how to measure certain writing skills and plot them on charts especially that the teachers we interviewed do not all agree on the idea as a whole.

Comparison. A major preoccupation of teachers when grading essays is to guarantee the fairness between the students when attributing the grades. 2 of the 3 teachers we interviewed use the comparison between different texts of students to decide whether the grades they have given are fair enough or need to be adjusted; and they expressed their will to have this feature implemented in the tool. One of the two teachers also uses the comparison between the student's actual work and their past assignments to check if they keep making the same mistakes in which case they might get more penalized.

To use this feature, the teacher simply needs to select a student's name from the dropdown list on the right-hand side of the interface to have their essay, feedback, grade and the different metrics of the spelling, grammar and argumentation checks displayed next to the essay they are currently grading.

V. Conclusion

Through this work, we have presented a first version of GRASPE, an adaptive grading support for persuasive English essays. We used three main language models to detect spelling mistakes, grammarly incorrect sentences, and the argumentation structure in students' writing essays. They all present high levels of accuracy. The tool also includes various other features based on user stories extracted from interviews with teachers. These features are meant to speed up the process of grading and make it easier on the teachers. The tool is currently still under construction and it would be useful to have further insights from teachers to continue the next steps of implementation and evaluation. As our aim is to make GRASPE as tailored as possible to the teachers' needs, existing literature in this case is not enough, and running the experiment with multiple interviews as planned would be the best way to finish the work. This process however might need more time than we think mainly due to the teachers' lack of free time during working periods of the year.

References

[1] Sai Muralidhar Jayanthi, Danish Pruthi, and Graham Neubig. "NeuSpell: A Neural Spelling Correction Toolkit". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, 2020, pp. 158–164. DOI: 10.18653/v1/2020.emnlp-demos.21. URL: https://aclanthology.org/2020.emnlp-demos.21.

[2] Beata Lewis Sevcikova. "Human versus Automated Essay Scoring: A Critical Review". In: *Arab World English Journal* 9 (2018), pp. 157–174.

[3] Arsaythamby Veloo, Noor Hashima Abd Aziz, and Aizan Yaacob. "The Most Suitable Scoring Method to Assess Essay Writing in ESL Classrooms". In: *Advances in Language and Literary Studies* (2018).

[4] Thiemo Wambsganß, Andrew Caines, and Paula Buttery. "ALEN App: Argumentative Writing Support To Foster English Language Learning". In: Jan. 2022, pp. 134–140. DOI: 10.18653/v1/2022.bea-1.18.

[5] Thiemo Wambsganss et al. "AL: An Adaptive Learning Support System for Argumentation Skills". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–14. ISBN: 9781450367080. DOI: 10.1145/3313831.3376732. URL: https://doi.org/10.1145/3313831.3376732.

[6] Yongjie Wang et al. "On the Use of Bert for Automated Essay Scoring: Joint Learning of Multi-Scale Essay Representation". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 3416–3425. DOI: 10.18653/v1/2022.naacl-

main.249. URL: https://aclanthology.org/2022.naacl-main.249.

[7] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. “Neural Network Acceptability Judgments”. In: *arXiv preprint arXiv:1805.12471* (2018).

[8] Zheng Yuan and Christopher Bryant. “Document-level grammatical error correction”. In: *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*. Online: Association for Computational Linguistics, Apr. 2021, pp. 75–84. URL: https://aclanthology.org/2021.bea-1.8.