JELCS
Journal of English Literature & Cultural Studies
Quarterly
Eurasian Applied Linguistics Society
ISSN: 2667-6214
Editor - in - Chief: Farid Parvaneh

Journal of English Literature and Cultural Studies

jelcsjournal 2022, 3 (2), 46–63

Original Research

doi: 10.26655/JELCS.2022.2.6

Human Vs. Machine Assessment of Essay Writings of B.A. Students of English Language Translation

Parisa Ahmadi Ardakani[1]
Mohammad Reza Falahati Qadimi Fumani[2],

[1]Graduate student of English Language Translation, M.A.
[2]Associate professor, Faculty Member of the Department of Computational Linguistics, Regional Information Center for Science and Technology (RICeST), Shiraz, Iran

Corresponding Author: Mohammad Reza Falahati Qadimi Fumani

E-mail: mrfalahat@yahoo.com

Article Citation: Ahmadi Ardakani, P. & Falahati Qadimi Fumani, M. R. (2022). Shobhaa De's Idea of Feminism: Human Vs. Machine Assessment of Essay Writings of B.A. Students of English Language Translation, *Journal of English Literature and Cultural Studies, 3* (2): 46–63.

Received Date: August 15, 2022
Accepted Date: September 3, 2022
Online Date: Septembre 18, 2022

Publisher: Kare Publishing

E-ISSN: 2667-6214

ABSTRACT

Writing assessment, esp. automated assessment, is a difficult job. Yet, due to its benefits, it has already found its way into educational settings. Accordingly, the present study intended to compare human vs. machine assessment of essay writings of B.A. students of English Language Translation in an EFL setting. To undertake the study, the final exam essay writings of 30 female B.A. level students of English Language Translation were collected based on availability sampling. These papers were corrected once by a human assessor, class instructor, and once by a software (PaperRater) in terms of spelling, grammar, word choice, style, and overall grade. The scoring system of PaperRater was also used to compare the scores given by human and PaperRater. Wilcoxon signed ranks test as well as tests of correlation were used to analyze the significance of difference and relationship between the scorings of machine and human assessors. Based on the results, for spelling errors, no significant difference was observed between the human rater and the machine. For style and overall grade, the scores assigned by the human rater were significantly lower than those assigned by PaperRater. For grammar errors, the human rater found significantly more errors and hence assigned a lower score and finally for word choice the human assessor assigned a significantly higher score to the papers compared to the machine. For the second question (significance of correlation) the findings revealed that for spelling, there was a very strong positive correlation between the errors found by human and the machine. For grammar errors, a rather strong positive correlation was observed. For overall grade, the correlation was weak but positive. Unlike these three sub-parts, for word choice and style no significant correlation was observed between the human and the machine. The finding of this research are in line with the findings of [33] and [34], who found agreement between machine and human scorings of essays. What is apparent is that machine assessing needs to be researched further so as to increase its validity and performance, but even in its current state, it can be used as a great help by teachers and class instructors at least in combination with their manual scorings.

Keywords: Essay Writing, Automated Writing Assessment Tools, Human Assessment of Writing, PaperRater, Spell Cheking, Grammar Cheking, Automated Scoring Systems.

1. Introduction

Assessment has found its way into almost all aspects of our lives including education. This term is defined by authors in [1], as "an essential component of classroom work … which can raise standards of achievement" (p. 12).

In the realm of language teaching, for instance, as we have different methods of teaching, we also have different methods of language assessment.

Authors in [2] state that a good assessment must help students improve their own work, develop a particular piece of writing, gain control of a personal writing process, and set new goals for future effort. Language assessment targets both active (speaking and writing) and passive (listening and reading) language skills. From among the four, writing is very difficult for students [3]. Generally speaking, students' writings are assessed by a class teacher/instructor, a software, or by a combination of the two. Each method of assessment – human or machine – has its own cons and pros. For instance, human assessors are often criticized for subjectivity and being biased [4] while machine assessors or better automated scoring systems are often claimed to lack the human-like precision, commit fake assessment [5, 6], be a threat to human raters [7], and need a huge corpus of sample texts [8]. Each method has its own advantages as well. For example, machine assessment has the advantage of being fast and consistent [9]. In fact, a major advantage of automated writing assessment is that it saves a lot of time on the part of teachers and hence teachers can allocate more time to teaching the content [7]. Human assessment also has the advantage of using an array of linguistic and paralinguistic elements, i.e. cognition, world knowledge etc.

An overview of the works undertaken by different researchers reveals that there are two broad views on the agreement between human and machine assessment of essay writings. Some researchers like [10], by referring to issues like fake assessment, etc., do not see much agreement between the two while others like [11, 12] report high rates of agreement between human and machine assessments of essays.

What is apparent is that although writing assessment tools are not perfect and have their own problems, they have already found their way into educational settings [13, 14] and hence cannot be ignored. Accordingly, the present researchers decided to implement a study and compare human vs. machine assessment of essay writings of B.A. students of English Language Translation. More specifically, the following research questions were introduced in this study:

Q1. Is there any significant difference between machine and human assessment of essay writings of B.A. students of English Language Translation in each error type (spelling, grammar, word choice, style and overall grade)?

Q2. Are the (spelling, grammar, word choice, style and overall grade) scores submitted by the software and the human assessor positively correlated?

To analyze the above research questions, two null-hypotheses were formulated as follows:

H01: There is not any significant difference between machine and human assessment of essay writings of B.A. students of English Language Translation in each error type (spelling, grammar, word choice, style and overall grade).

H02: The (spelling, grammar, word choice, style and overall grade) scores submitted by the software and the human assessor are not positively correlated.

2. Literature Review

2.1 Background

The origin of writing assessment in academic settings dates back to at least fifty years ago [15]. From that period onward writing assessment has undergone different shifts. Yancey [16] explained that there were totally three important shifts in methods used to assess writing: The first wave was observed from 1950 to 1970 and tried to use objective tests to assess the writing indirectly [16]. The second wave (between 1970 and 1986) included tests which focused on students' actual writing skill [17]. Finally, the third wave which began in 1986 focused on assessing a collection of students' works and programmatic assessment [18].

The first attempt to automatize writing assessment was carried out by [19] who believed that computers could be used to score writing tasks of students and in 1968 he designed a program called Project Essay Grade™ (PEG™). However, the system was not considered to be cost-effective [20]. By 1990, personal computers became powerful and hence Automated Essay Scoring (AES) was a possibility [14]. In 1990s, Page collaborated with different companies, updated his program, and performed several successful trials in the field of computerized writing assessment [20].

Later, Foltz and Landauer used the scoring engine called the Intelligent Essay Assessor™ (IEA) and developed a scoring system, which was first launched in 1997 and used to score essays of undergraduate students [21]. In 1998, Vantage Learning launched its own AES engine, called IntelliMetric® (Elliot, 2003). As stated by [22], in another attempt, Mitzel and Lottridge developed a constructed response automated scoring engine, called CRASE®. This technology has been used in large-scale formative and summative assessment environments since 2007 [22] (for a detailed account of the automated essay scoring systems c.f. [23].

2.2 Human Assessment of Texts

Text and rating quality are important aspects of writing assessment. In an attempt to examine these issues, [24] investigated the relationship between textual characteristics and rating quality in rater-mediated writing assessments. The aim of their study was to suggest a method of exploring the influence of textual characteristics of essays on rating quality in the context of rater-mediated writing performance assessments. The data of the study consisted of copies of handwritten essays and quantitative ratings that were collected during rater-mediated writing assessments. The data were collected from a population of L1 students. The authors wanted to achieve a comprehensive understanding of rating quality. To perform the study, they employed rating quality as their framework and defined it as "adherence to the principles of invariant measurement" (ibid, p. 3). In addition, they used the rater reliability coefficients, rater agreement statistics and generalizability theory as the framework of their study. The results showed that raters used empirical evidence of relationships among textual characteristics and the acquired data to identify errors.

Writing task efficacy can be improved by using self and peer assessment. Authors in [25] focused on students of a first year geography course to show that writing performance could improve through self and peer assessment. To do so, the researchers investigated 50 first year students of geography about their experiences of essay writing and essay assessment. The aim of their study was to make a comparison between self-assessment and peer assessment. Based on the results, 60 percent found self-assessment difficult, and 67 percent explained that the peer assessment was a hard job. Also, 57 percent agreed that self-assessing their essay made them put more thought into how they were writing it. Further, 51 percent said that the self-assessment experience helped their understanding of assessment, while 45 percent said the same of the peer assessment. Similarly, 18 percent stated that peer and self-assessments were less appropriate than tutors' assessment. Finally, 64 percent mentioned that the assessment procedure helped them to write better essays.

Writing assessment, comparative judgement and students' evaluative expertise is another aspect of human assessment. In this context, [26] focused on a sample of 3000 students from 24 schools in Norway. The sample was divided into two groups of control and experimental. The aim of this study was to enhance the writing skill of students. Therefore, the researcher used Wheel of Writing, which is a research-based writing tool that offers norms of expectations for writing proficiency. The experiment group was taught through this method. Pre-test and post-test data were collected from students' responses to standardized writing tasks. To collect the data, the researcher used the Norwegian Sample-Based Writing Test (NSBWT). This test is the standard writing task that is used in Norway. After the test was taken by the students, a team of professional teachers were employed to score and assess these writing tasks. The results of the study proved that students in experiment schools had improved their writing skills more than students in control schools. In addition, results showed substantial changes in writing quality of schools, classes and individual students.

Regarding the influence of vocabulary and spelling on assessment of writing tasks, [27] explored the influence of vocabulary and spelling on teachers' perceptions of essay writings of ESL students. In this study, the researchers asked 69 teachers from Switzerland and Germany to assess 4 upper-intermediate ESL essays. In addition, the teachers were asked to mention their comments on the weaknesses and strong points of these writing tasks. The results of their study illustrated that when students used less sophisticated vocabulary in their writings, the assessors provided negative comments about the grammar. Besides, when the spelling was not good enough the assessors made negative comments about vocabulary and grammar. Therefore, the researchers concluded that the perception of teachers on positive or negative attitudes of the tasks was influenced by their holistic and analytic assessment of the texts.

Content of writing plays also an important role in writing assessment. In this regard, [28] focused on the content of written stories by children. The data of this study included stories written by children in English which were scored by human assessors. The criteria of the researchers for assessment included coherence, originality, grammar, text length, and vocabulary diversity. The participants of the study were 175 students from grade 2. The results of the study showed that content could be defined as sum of five elements i.e. coherence, originality, grammar, text length, and vocabulary diversity.

2.3 Machine Assessment of Texts

In the area of machine assessment of writing many works have been undertaken a few of which are mentioned below:

Authors in [29] investigated a new approach to formative writing assessment. They evaluated the writing tasks done by Grade 6 and Grade 8 students based on word, sentence, and discourse. They used automated measures of word choice, syntax, and cohesion. The required data for their study were derived from a database that compiled students' writing samples, demographics, and achievement data. To analyze the data, the researchers used Mplus V.7.4 software program. The results showed that new levels of language detection algorithms were identified that could be used within automated writing evaluation software programs to expand automated teacher assessment and feedback approaches. Further, [15] investigated problems and issues of transition from human (traditional) to machine (computerized) writing assessment. He made a comparison between computer and human assessment and mentioned that some issues such as the way to use computers can be a problem for writing assessment. The researcher mentioned that validity of machine writing assessment was under debate and that computer effects and performance was another problem meaning that it was not obvious how the anxiety of using computer could influence students' writing ability. In [30] the author investigated whether automated writing assessment could help students improve their writing skills. He focused on 735 essays written by 53 Taiwanese college students and analyzed them by a machine assessor. The researcher used descriptive statistics, paired-samples t-tests, Pearson correlation, effect size, and regression to analyze the data. The results revealed that the writings improved significantly in terms of the length and that the scores of the students mentioned by the machine were better than those released by human raters. Also, [31] investigated the influence of word processing on the writing of ESL students and writing assessment. Li focused on 21 advanced Chinese students of English. Each student wrote two comparable writings and the thoughts of the students were recorded through think-aloud protocols. The results showed that higher order thinking activities were of more importance to the students. This proved that teachers needed to pay more attention to the impact of computers on writing assessment.

2.4 Comparison of Machine and Human Assessment of Texts

Research on replacing human assessors with machine assessors has attracted the attention of many scholars. For instance, [32] performed a study to find if machine scoring of writing test responses agreed with human readers as much as humans agreed with other humans. The results showed that the reliability of machine scoring was lower than the human assessor and that the human outperformed the machine. In contrast, [33] used three groups of raters including naïve, (untrained raters) potential (but untrained raters) and trained (experienced raters). The material of their study consisted of 60 essays written by different GRE test takers. These essays were both corrected by the human assessors and the OSN system. The results showed that there was a machine-human agreement and that this agreement was dependent upon experience and expertise of the raters. Authors in [34] reported a similar finding when they compared human and machine assessors according to gender, ethnicity, and country. The data of this study consisted of all the essays written for GRE and TOEFL exams from January 2008 through October 2008. The researchers used Ellis Page's Project Essay Grade for the machine scoring and grouped the essay writers according to their gender, ethnicity, and country. The results showed that human and machine scores were very similar between most subgroups. The researchers concluded that there would be the possibility that in the future the scoring of human and machine assessors would have little impact on the final score.

In Iran, little attention has been paid to comparison of human and machine assessors in writing. For example, [35] explored the effect of online summative and formative assessments on 130 Iranian EFL junior university students. The data of the study were gathered from students' writings in both online summative and portfolio formative assessments as well as collaborative writing in online collaborative formative assessment in e-writing forum. The results showed that technology could improve the writing efficacy of students. Authors in [36] investigated diagnostic and developmental potential of dynamic assessment for writing skill. The researchers focused on three students of English literature and asked them to write a composition individually. The essays were corrected separately by the authors and the results of the study showed that the use of dynamic assessment for writing skill could significantly improve this skill in language learners. Writing assessment is an important aspect of language teaching. In this regard, [37] examined EFL writing tasks in IELTS, TOEFL, FCE, and CAE of Iranian applicants. To perform this study, the authors asked 114 learners to rate EFL writing tasks according to a checklist previously designed and validated. The independent-sample t-test was employed to compare the mean scores of ratings for all items of the checklist. The participants of the study included 11 classes of ESOL exam preparation courses – 3 IELTS, 2 TOEFL, 3 FCE and 3 CAE classes. The results showed significant differences in quality of writing procedures in the students. Some researchers studied the variable nationality. For example, [38] investigated whether Iranian raters differed from NES raters in their severity when they rated students' essays. In addition, existence of any bias in both groups toward a certain feature of writing was investigated. Multi-faceted Rasch measurement results showed that Iranian raters were significantly more severe than NES raters in rating Iranian students' writings. In addition, no significant bias was found in Iranian or NES raters toward a certain feature of writing.

This brief review revealed that comparing human and machine assessors is not yet a well-researched area, especially in Iranian context. To summarize, the research conducted in this area could be classified into three classes. The first class covered traditional writing assessment by human assessors. The second class dealt with modern computerized approaches and software programs designed for machine writing assessment.

Finally, the third class focused on comparing the first two groups. This brief literature review revealed that little attention had been paid to the third area, esp. in Iran, and hence more studies must be conducted in this area. Accordingly, the aim of the current study was to compare human vs. machine assessment of essay writings of B.A. students of English Language Translation in Iranian context.

3. Methodology

3.1 Data

This research is a descriptive-comparative study that focuses on comparing the performance of human and machine assessors in evaluating the essay writings of B.A. students of English Language Translation in an Iranian context. The original data included the final term essay writings of 30 female senior B.A. students of English Language Translation collected from Zand University, Shiraz, Fars Province, Iran. The ultimate data comprised of corrections of these papers as made by PaperRater and a human assessor (the class instructor) each at a separate time. The original data was in handwritten form which was later typed by the present researchers to enable the analysis of the data by PaperRater. During this process, of course, all the spelling and grammatical problems of the original papers were retained in the typed version to avoid manipulation of the data.

3.2 Participants

Participants of the study included 30 Iranian female senior B.A. students of English Language Translation. The reason for choosing only female students was that the researchers did not have access to enough number of male participants. The participants were not homogeneous in terms of proficiency, which was not a problem since the same group was assessed by two different assessment methods – manual assessment by a human expert and automatic assessment by PaperRater. To select the sample participants, first, the total number of senior B.A. level students of English Language Translation, who had attended the final exam of the essay writing course, in Zand University, was determined. There were 40 students all from the same class of which 4 were male who were discarded. From 36 female students, 30 had selected the same topic – on the final exam two topics had been introduced by class instructor – and hence these 30 students comprised the ultimate participants of the study. Also, the human assessor, in this study, had more than 10 years of teaching experience in essay writing.

3.3 Instrument

The instruments used in this study were as follows:

1) **The writing task.** This was the final exam produced by class instructor and with two topics of which one had to be selected by the students. The two topics were:

a) *Some people believe that robots will play an important role in future societies, while others argue that robots might have negative effects on society,* and

b) *Some people believe that the salaries paid to professional sportspeople are too high, while others argue that sports salaries are fair*.

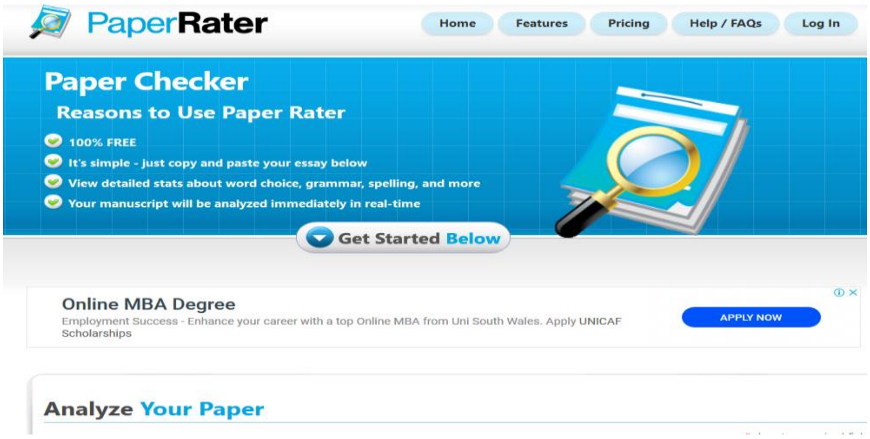
No word limit had been set by class instructor on the volume of the essay and hence the students were free to write as much as they wished.

Table 1. *Length of students' essays in words.*

St .No	Word Count	St. No	Word Count	St. No	Word Count	St. No	Word Count
1	280	9	353	17	316	25	229
2	293	10	319	18	295	26	311
3	293	11	319	19	333	27	291
4	398	12	446	20	258	28	320
5	240	13	370	21	317	29	274
6	312	14	293	22	378	30	345
7	364	15	372	23	262	31	375
8	278	16	346	24	341		

As indicated in Table 1, the length of the essays written by the students ranged between 229 and 446 words.

2) PaperRater. This software was used for machine scoring of the essay writings of the students. To access this software, users need to log onto www.paperrater.com. By so doing, a window will appear as shown in Figure 1.



Paper Checker

Reasons to Use Paper Rater

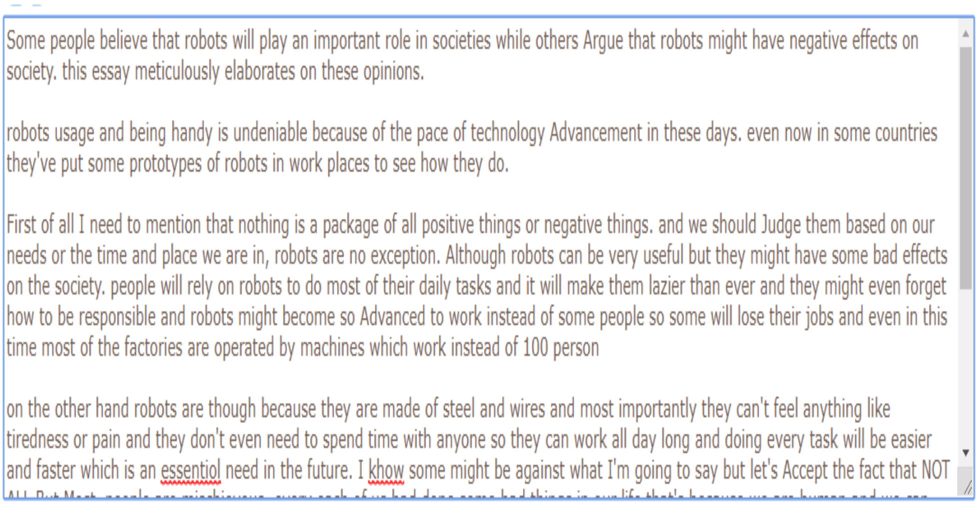
- 100% FREE
- It's simple - just copy and paste your essay below
- View detailed stats about word choice, grammar, spelling, and more
- Your manuscript will be analyzed immediately in real-time

Online MBA Degree
Employment Success - Enhance your career with a top Online MBA from Uni South Wales. Apply UNICAF Scholarships

Analyze Your Paper

Figure 1. *The main window of PaperRater software.*

This software evaluates writings of students and submits a report on the volume of errors as well as error types. PaperRater uses artificial intelligence to reveal the errors available in texts. To do so, the text must be copied in the text box of the software as shown in Figure 2.



Some people believe that robots will play an important role in societies while others Argue that robots might have negative effects on society. this essay meticulously elaborates on these opinions.

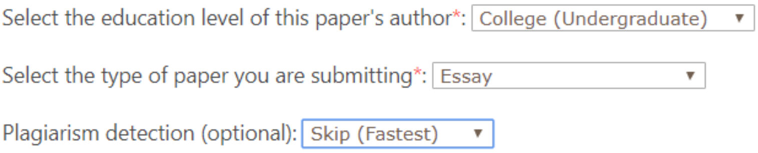
robots usage and being handy is undeniable because of the pace of technology Advancement in these days. even now in some countries they've put some prototypes of robots in work places to see how they do.

First of all I need to mention that nothing is a package of all positive things or negative things. and we should Judge them based on our needs or the time and place we are in, robots are no exception. Although robots can be very useful but they might have some bad effects on the society. people will rely on robots to do most of their daily tasks and it will make them lazier than ever and they might even forget how to be responsible and robots might become so Advanced to work instead of some people so some will lose their jobs and even in this time most of the factories are operated by machines which work instead of 100 person

on the other hand robots are though because they are made of steel and wires and most importantly they can't feel anything like tiredness or pain and they don't even need to spend time with anyone so they can work all day long and doing every task will be easier and faster which is an essentiol need in the future. I khow some might be against what I'm going to say but let's Accept the fact that NOT

***Figure 2.** Text box of PaperRater.*

Having uploaded or copied the text into the text box, the user must take three more steps (Figure 3). First, the user must select the level of education of the writer of the essay, i.e. 1st grade, 6th grade, 12th grade, B.A., M.A., etc. Then, the type of the text written should be determined, i.e. 'essay', 'article', etc. Finally, the user may use the fastest version or skip (fastest). If the user presses the skip button, s/he will be exposed to the free version and if s/he does not then s/he will be able to use the premium version which needs prior subscription and payment. This premium version also analyzes the text for possible plagiarism.



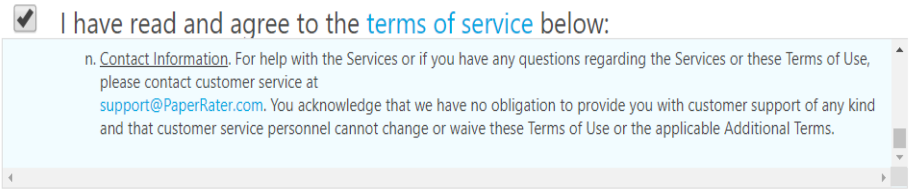
Select the education level of this paper's author*: College (Undergraduate) ▼

Select the type of paper you are submitting*: Essay ▼

Plagiarism detection (optional): Skip (Fastest) ▼

Figure 3. *Other settings in PaperRater.*

Next, the user should put a checkmark on the box before the sentence *"I have read and agree to the terms of service below"* as indicated in Figure 4.

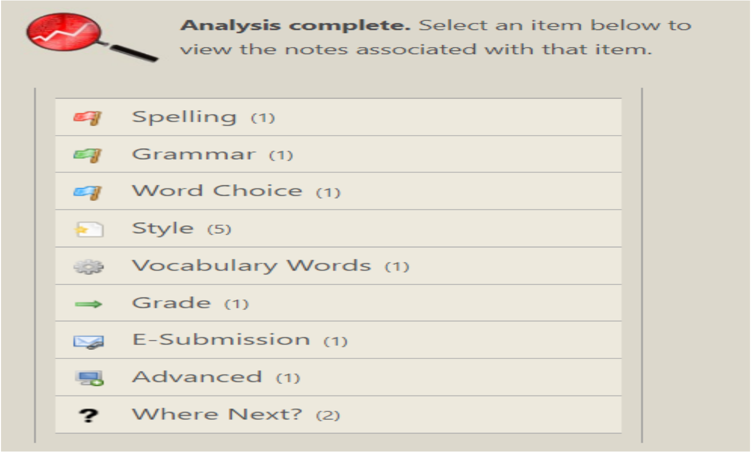


☑ I have read and agree to the terms of service below:

n. Contact Information. For help with the Services or if you have any questions regarding the Services or these Terms of Use, please contact customer service at support@PaperRater.com. You acknowledge that we have no obligation to provide you with customer support of any kind and that customer service personnel cannot change or waive these Terms of Use or the applicable Additional Terms.

Figure 4. *Agreement to terms of service.*

Having completed these steps, the user should click on *"Get Report"* in the free version and *"Advanced Check"* on the paid version. This will activate the software and within seconds the software will submit the report indicating the volume and types of errors as indicated in Figure 5.

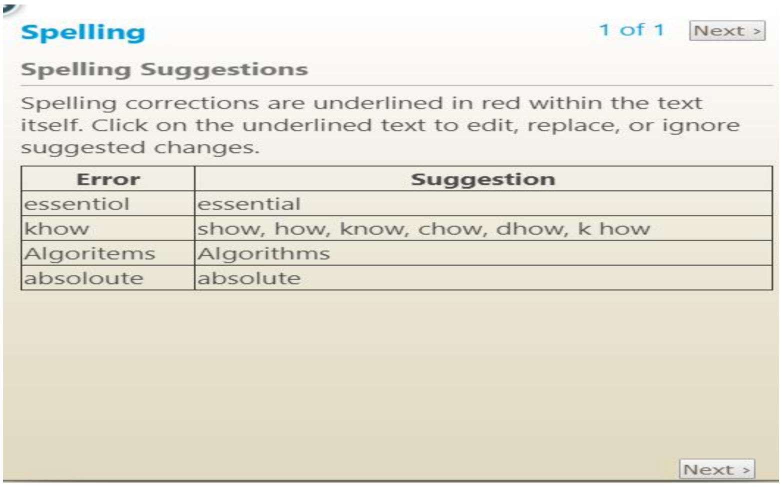


Analysis complete. Select an item below to view the notes associated with that item.

Spelling (1)
Grammar (1)
Word Choice (1)
Style (5)
Vocabulary Words (1)
Grade (1)
E-Submission (1)
Advanced (1)
Where Next? (2)

Figure 5. *PaperRater's report submission sample.*

As indicated in Figure 5, PaperRater analyzes the text and extracts errors and produces a table embodying error types and quantity. By clicking on each item, i.e. 'spelling', 'grammar', etc. a new window pops out in which the complete report of that error type is provided. For example, by clicking on 'spelling' a window will appear as in Figure 6.



Spelling

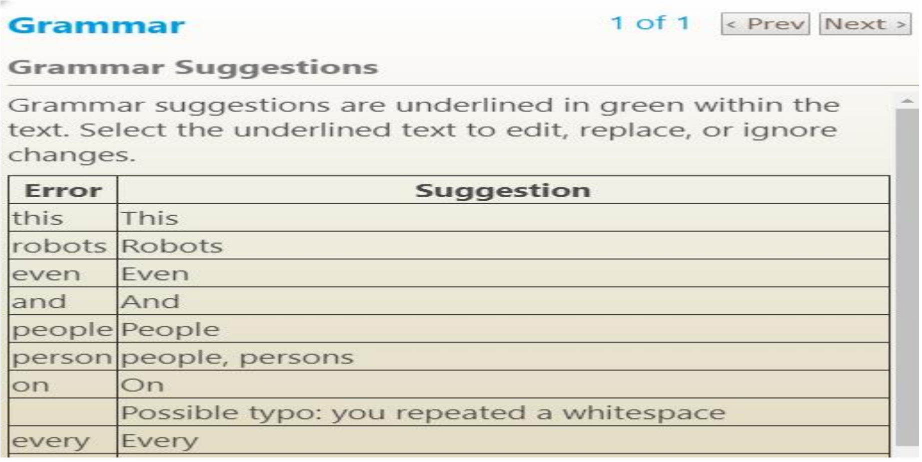
Spelling Suggestions

Spelling corrections are underlined in red within the text itself. Click on the underlined text to edit, replace, or ignore suggested changes.

Error	Suggestion
essentiol	essential
khow	show, how, know, chow, dhow, k how
Algoritems	Algorithms
absoloute	absolute

Figure 6. *Report submitted by PaperRater on 'spelling'.*

Here in Figure 6, possible errors along with suggestions made by the software are provided. The user needs to review these words and the equivalents suggested and decide if s/he wishes to accept the suggestion or skip it. For instance, the software has identified the word 'essentiol' as a possible error and has suggested 'essential'. Or, the word 'khow' has been found and a number of suggestions have been introduced by the software including 'show', 'how', 'know', 'chow', 'dhow' and 'k how'. The user may review these suggestions and select one as the right answer – the word 'know' in this case. The report formats of PaperRater on grammar, word choice, style, and overall grade have been presented in Figures 7 to 10.



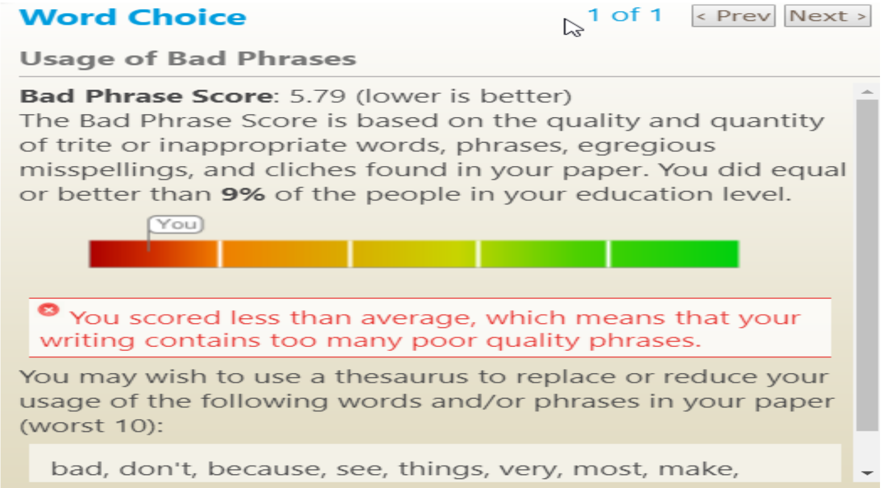
Grammar

Grammar Suggestions

Grammar suggestions are underlined in green within the text. Select the underlined text to edit, replace, or ignore changes.

Error	Suggestion
this	This
robots	Robots
even	Even
and	And
people	People
person	people, persons
on	On
	Possible typo: you repeated a whitespace
every	Every

Figure 7. *Report submitted by PaperRater on 'grammar'.*



Word Choice

Usage of Bad Phrases

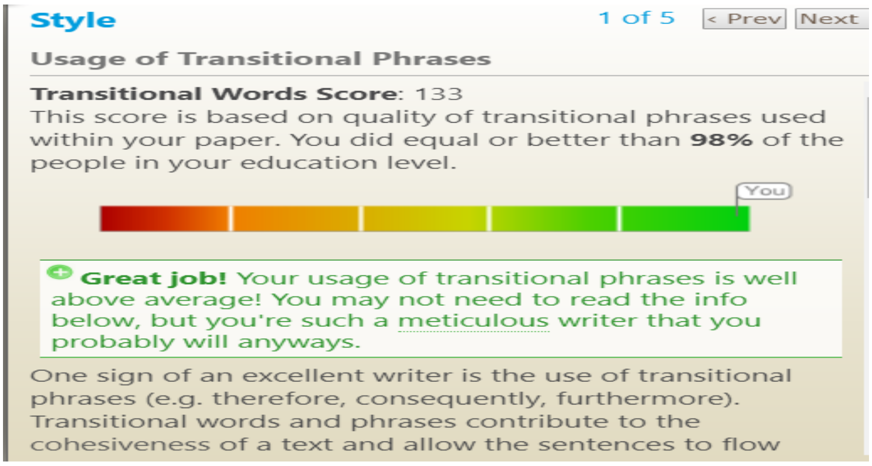
Bad Phrase Score: 5.79 (lower is better)
The Bad Phrase Score is based on the quality and quantity of trite or inappropriate words, phrases, egregious misspellings, and cliches found in your paper. You did equal or better than **9%** of the people in your education level.

You scored less than average, which means that your writing contains too many poor quality phrases.

You may wish to use a thesaurus to replace or reduce your usage of the following words and/or phrases in your paper (worst 10):

bad, don't, because, see, things, very, most, make,

Figure 8. *Report submitted by PaperRater on 'word choice'.*



Style

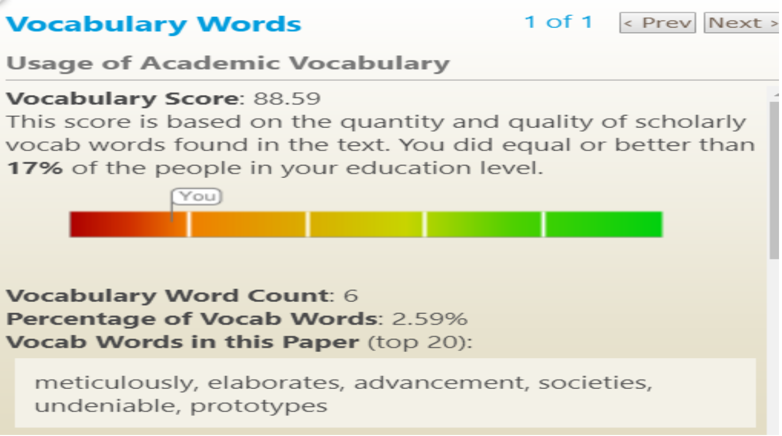
Usage of Transitional Phrases

Transitional Words Score: 133
This score is based on quality of transitional phrases used within your paper. You did equal or better than **98%** of the people in your education level.

Great job! Your usage of transitional phrases is well above average! You may not need to read the info below, but you're such a meticulous writer that you probably will anyways.

One sign of an excellent writer is the use of transitional phrases (e.g. therefore, consequently, furthermore). Transitional words and phrases contribute to the cohesiveness of a text and allow the sentences to flow

Figure 9a. *Report submitted by PaperRater on 'style1'.*



Vocabulary Words

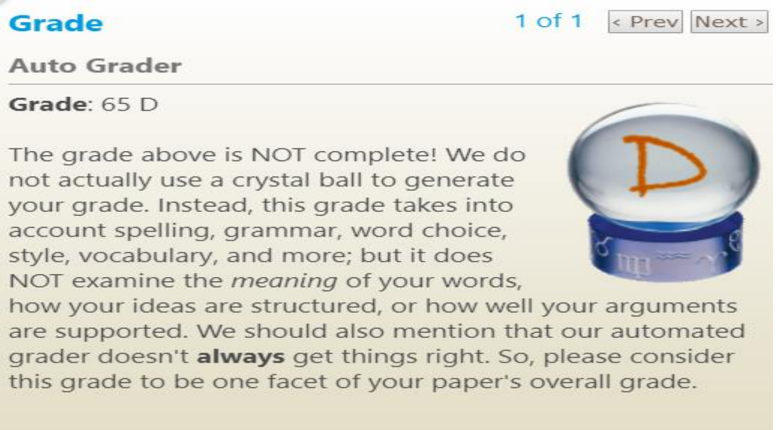
Usage of Academic Vocabulary

Vocabulary Score: 88.59
This score is based on the quantity and quality of scholarly vocab words found in the text. You did equal or better than **17%** of the people in your education level.

Vocabulary Word Count: 6
Percentage of Vocab Words: 2.59%
Vocab Words in this Paper (top 20):

meticulously, elaborates, advancement, societies, undeniable, prototypes

Figure 9b. *Report submitted by PaperRater on ‘style2’.*



Grade

Auto Grader

Grade: 65 D

The grade above is NOT complete! We do not actually use a crystal ball to generate your grade. Instead, this grade takes into account spelling, grammar, word choice, style, vocabulary, and more; but it does NOT examine the *meaning* of your words, how your ideas are structured, or how well your arguments are supported. We should also mention that our automated grader doesn't **always** get things right. So, please consider this grade to be one facet of your paper's overall grade.

Figure 10. *Report submitted by PaperRater on ‘overall grade’.*

In this research, the performance of the software on ‘spelling, ‘grammar’, ‘word choice’, ‘style’ and ‘overall grade’ was considered to assess the essays written by the participants. For style, the average score of style1 and style2 was used.

3.4 Scoring

The scores given by the human assessor (class instructor) were normalized by 100 points because on the final exam, the instructor had used a score range of 0 to 10. So, a score of 5 was changed into 50. The score range used in PaperRater was 0 to 100. The scoring system and the sub-parts of PaperRater have been presented in Table 2.

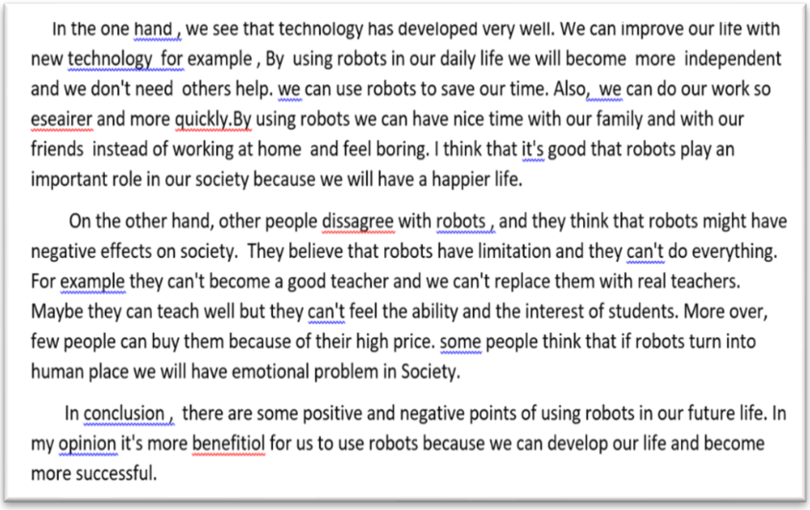
Table 2. *Demonstration of PaperRater’s subparts and their specifications.*

Error Type	Explanation	Example
Spelling	This part deals with the spelling errors found in the text. The software highlights all the spelling errors. The higher the spelling errors the lower the score of the student.	Hapy rather than Happy
Grammar	This part deals with the grammatical errors found in the text. The software highlights all the grammatical errors. The higher the grammatical errors the lower the score of the student.	Goed
Style 1	Here the score is between 0 and 100 and is given based on text cohesion. In other words, attention is paid to transition words such as ‘hence, thus etc. Correct application of these words results in higher scores. The text which meets these points receives a 100.	Any deviation from cohesion

Style 2	In Style2, the score is between 0 and 100 and is given based on the skill of the student regarding opening sentences with different words and expressions. Starting sentences with repetitive or similar words lowers the score.	
Word Choice	The score range is again between 0 and 100. Here, the words chosen by a given student are compared with those used by other students of the class. A higher score indicates a higher quality of word choice.	Spoiled eggs instead of Addled eggs
Grade	The ultimate score of the students is provided here and it ranges again between 0 and 100.	

*.For style, the average scores of style1 and style2 were used.

In Figure 11 a sample text written by a student is shown. Here, the spelling and grammatical problems have been highlighted by red and blue colors respectively.



In the one hand , we see that technology has developed very well. We can improve our life with new technology for example , By using robots in our daily life we will become more independent and we don't need others help. we can use robots to save our time. Also, we can do our work so eseairer and more quickly.By using robots we can have nice time with our family and with our friends instead of working at home and feel boring. I think that it's good that robots play an important role in our society because we will have a happier life.

On the other hand, other people dissagree with robots , and they think that robots might have negative effects on society. They believe that robots have limitation and they can't do everything. For example they can't become a good teacher and we can't replace them with real teachers. Maybe they can teach well but they can't feel the ability and the interest of students. More over, few people can buy them because of their high price. some people think that if robots turn into human place we will have emotional problem in Society.

In conclusion , there are some positive and negative points of using robots in our future life. In my opinion it's more benefitiol for us to use robots because we can develop our life and become more successful.

Figure 11. *Sample paper written by a student.*

3.5 Procedure

In this study, first, 30 Iranian female senior B.A. level students of English Language Translation were selected using availability sampling as the participants. Next, their final term exam papers on essay writing were collected from the university. Then, the papers were typed to enable inputting them into PaperRater. After that the papers were input and scored by PaperRater in terms of spelling, grammar, word choice, style and overall grade. The human assessor was also asked to score the papers using the same sub-parts (spelling, grammar, word choice, style and overall grade). The error quantity and types extracted by the software and the human assessor formed the basis of the analysis. The reports by the human assessor and PaperRater on error types and quantity were used to answer the research questions.

4. Results

Two broad research questions were introduced in this study as follows:

Q1. Is there any significant difference between machine and human assessment of essay writings of B.A. students of English Language Translation in each error type (spelling, grammar, word choice, style and overall grade)?

Q2. Are the (spelling, grammar, word choice, style and overall grade) scores submitted by the software and the human assessor positively correlated?

To address the questions of the study, both descriptive and inferential statistics were used. To start the analysis, first descriptive statistics regarding the data of the study have been presented. The sample of the study consisted of 30 students.

Table 3. *Statistics for research variables assessed by machine and human.*

	Variable	N	Mean	Std. Deviation	Min	Max
Machine	Spelling (count)	30	6.867	5.5939	0	21.0
	Grammar (count)	30	4.533	3.9456	0	14.0
	Word Choice	30	15.000	11.5311	0	47.0
	Style	30	90.967	17.4168	23.0	100.0
	Overall Grade	30	67.933	3.4535	62.0	76.0
Human	Spelling (count)	30	7.467	7.1907	0	33.0
	Grammar (count)	30	15.833	6.7573	2.0	31.0
	Word Choice	30	59.167	15.0907	25.0	85.0
	Style	30	66.267	16.2033	30.0	94.0
	Overall Grade	30	59.567	15.4979	25.0	90.0

Table 3 presents statistics for research variables including spelling (count), grammar (count), word choice, style and overall grade as assessed by PaperRater and the human assessor. As indicated in this table, with regard to spelling, the mean score obtained by PaperRater was 6.86 with a standard deviation of 5.59. For PaperRater, the mean scores for grammar, word choice and style were 4.53, 15 and 90.96 with a standard deviation of 3.94, 11.53 and 17.41 respectively. Finally, the overall grade mean for PaperRater was 67.93 with a standard deviation of 3.45. For the human assessor, the mean scores for spelling, grammar, word choice and style were 7.46, 15.83, 59.16 and 66.26 with a standard deviation of 7.19, 6.75, 15.1 and 16.2 respectively. The overall grade mean score for the human assessor was 59.56 with a standard deviation of 15.49.

In what follows, the results of data analysis are presented. The two research variables, spelling and grammar, are countable variables for which non-parametric tests were used, while word choice, style and overall grade are scale variables. In this case, parametric tests can be used provided that the variables have normal distribution. Thus, first, normality of the variables' distribution was investigated by the Kolmogorov-Smirnov Test.

Table 4. *Kolmogorov-Smirnov Test of normality.*

	Machine		Human	
Variables	Z	Sig.	Z	Sig.
Word Choice	0.169	**0.029**	0.138	0.147
Style	0.346	**0.000**	0.168	**0.030**
Overall Grade	0.192	**0.006**	0.084	0.200

Table 4 shows that from among the six cases (word choice, style and overall grade for PaperRater and the human assessor), only in two cases (word choice and overall grade for the human assessor) the distribution is normal and hence parametric tests could be used. In the other four cases, the test is significant (p<0.05), that is, the distribution is not normal and hence non-parametric tests should be used.

The first research question of the study was as follows:

Q1. Is there any significant difference between machine and human assessment of essay writings of B.A. students of English Language Translation in each error type (spelling, grammar, word choice, style and overall grade)?

To answer this question, each sub-part was dealt with separately. Regarding the spelling, the errors found by the human (spelling_h) and the machine (spelling_m) were compared using the

Wilcoxon signed ranks test. The Wilcoxon test examines which group, human or machine, found more spelling errors.

Table 5. *The Wilcoxon test for comparison of spelling errors count between human and machine.*

	N	Mean Rank	Sum of Ranks	Z	Sig.
Spelling_h < Spelling_m	13	14.42	187.50	-0.306	**0.759**
Spelling_h > Spelling_m	13	12.58	163.50		
Spelling_h = Spelling_m	4				

Results of the Wilcoxon test in Table 5 indicated that in 13 essays the machine found more spelling errors than the human. In 13 other essays, the situation was the opposite and the number of spelling errors found by the human was higher than the machine. In the remaining 4 essays, the human and the machine found an equal number of errors. In the whole, the test was not significant (p=0.759>0.05) which means that the number of spelling errors found was not significantly different between the human rater and PaperRater.

Regarding the sub-part grammatical errors, the grammar error counts by human (grammar_h) and the machine (grammar_m) were compared again using the Wilcoxon signed ranks test.

Table 6. *The Wilcoxon test for comparison of grammar errors count between human and machine.*

	N	Mean Rank	Sum of Ranks	Z	Sig.
grammar_h < grammar_m	1	1.00	1.00	-4.764	**0.000**
grammar_h > grammar_m	29	16.00	464.00		
grammar_h = grammar_m	0				

As indicated in Table 6, in 29 essays the number of grammar errors found by the machine was lower than that by the human. Only in one essay, the machine found more errors. The test was significant (p=0.000<0.05) which means that the number of grammar errors found by human was significantly greater than those found by the machine.

Regarding word choice, the distribution for the human assessor was normal which was not so for the machine. For this reason, the non-parametric (Wilcoxon) test was used again to answer this question.

Table 7. *The Wilcoxon test for comparison of word choice assessed by human and machine.*

	N	Mean Rank	Sum of Ranks	Z	Sig.
Word Choice_h < Word Choice_m	0	0.00	0.00	-4.783	**0.000**
Word Choice_h > Word Choice_m	30	15.50	465.00		
Word Choice_h = Word Choice_m	0				

Results of the Wilcoxon test in Table 7 indicated that in all cases, word choice scores given by the human assessor were greater than the scores given by the machine. The test was significant (p=0.000<0.05) which means that the scores given by the human were significantly greater than those given by the machine.

Regarding the Style, the distributions of the scores as assessed by human and the machine were not normal. Accordingly, the non-parametric (Wilcoxon test) test was used.

Table 8. *The Wilcoxon test for comparison of style assessed by human and machine.*

	N	Mean Rank	Sum of Ranks	Z	Sig.
style_h < style_m	27	16.26	439.00	-4.248	**0.000**
Style_h > style_m	3	8.67	26.00		
style_h = style_m	0				

Results of the Wilcoxon test in Table 8 indicated that in 27 essays, the style scores given by human were lower than the scores given by the machine. The test was significant (p=0.000<0.05) which means that the scores given by human were significantly lower than those reported by the machine.

Finally, regarding the overall grade, the distribution of the scores for the human assessor was normal, which was not, of course, normal for the machine. For this reason, the (non-parametric) Wilcoxon test was performed.

Table 9. *The Wilcoxon test for comparison of overall grade assessed by human and machine.*

	N	Mean Rank	Sum of Ranks	Z	Sig.
overall _h < overall_m	21	15.48	325.00	-2.779	**0.005**
overall_h > overall_m	7	11.57	81.00		
overall_h = overall_m	2				

Results of the Wilcoxon test in Table 9 indicated that in 21 essays, overall grades given by human were lower than the grades given by the machine. The test was significant (p=0.005<0.05), that is, the overall grades given by human were significantly lower than those given by the machine.

In the second research question of the study, the correlation between the scores given by the human assessor and the machine was dealt with. The question was as follows:

Q2. Are the (spelling, grammar, word choice, style and overall grade) scores submitted by the software and the human assessor positively correlated?

To answer this question, the Spearman correlation test was used which is a non-parametric test.

Table 10. *Spearman correlation between spelling errors count by human and the machine.*

Variables	Spearman Correlation	Sig.	N
Spelling_m & Spelling_h	0.716	**0.000**	30

As presented in Table 10, the Spearman correlation computed between the spelling error counts of the human and the machine was significant (p<0.05). The Spearman correlation was 0.716 which shows a very strong positive correlation.

Regarding the grammar errors, again the Spearman correlation test was employed.

Table 11. *Spearman correlation between grammar error counts by human and the machine.*

Variables	Spearman Correlation	Sig.	N
grammar_m & grammar_h	0.47	**0.009**	30

As presented in Table 11, the Spearman correlation between grammar errors found by human and the machine was significant (p<0.05). The result of the correlation was 0.47 which is a rather strong positive correlation.

Regarding word choice, Spearman and Pearson correlation tests were used.

Table 12. *Spearman and Pearson correlations between word choice scores by human and the machine.*

Variables	Test	Correlation	Sig.	N
Word Choice_m & Word Choice_h	Spearman	0.080	**0.673**	30
	Pearson	-0.004	**0.983**	30

As presented in Table 12, the correlation between word choice scores as assessed by human and the machine was not significant in either case (Spearman and Pearson) (p>0.05). Thus, there was no significant correlation between human and the machine in the assessment of word choice scores.

Regarding style, the Spearman correlation test results have been summarized in Table 13.

Table 13. *Spearman and Pearson correlations between style scores by human and the machine.*

Variables	Test	Correlation	Sig.	N
style_m & style_h	Spearman	-0.163	**0.390**	30
	Pearson	0.151	**0.427**	30

As presented in Table 13, the correlation between style scores as assessed by human and the machine was significant neither in Spearman nor in Pearson tests (p>0.05). Thus, there was no significant correlation between human and the machine in the assessment of style scores.

Finally, regarding the overall grade, again Spearman and Pearson correlation tests were used.

Table 14. *Spearman and Pearson correlations between overall grades by human and the machine.*

Variables	Test	Correlation	Sig.	N
overall_m & overall_h	Spearman	0.382	0.037	30
	Pearson	0.414	0.023	30

As illustrated in Table 14, the Spearman and Pearson correlations between overall grades as assessed by human and the machine were both significant (p<0.05). The correlation was a moderate positive correlation.

5. Discussion

The main intention of this study was to compare the scorings of a human assessor and PaperRater on essay writings of 30 Senior B.A. level students of English Language Translation to see if there were any significant differences between the scores assigned by the human assessor and the machine and if the scores were correlated.

For the first question (significance of difference) five sub-parts were considered namely spelling errors, grammar errors, word choice, style and overall grade. Regarding spelling errors, no significant difference was observed between the human rater and the machine. For style and overall grade, the scores assigned by the human rater were significantly lower than those assigned by PaperRater. For grammar errors, the human rater found significantly more errors and hence assigned a lower score and finally for word choice the human assessor assigned a significantly higher score to the papers compared to the machine.

For the second question (significance of correlation) the findings for the five sub-parts were as follows: For spelling errors there was a very strong positive correlation between the errors found by the machine and the human. For grammar errors, too, a rather strong positive correlation was observed. For overall grade, the correlation was weak but positive. Unlike these three sub-parts, for word choice and style no significant correlation was observed between the human and the machine.

Looking at the above findings, the following discussions could be made: First, if we look at the overall grades assigned by the human assessor and the machine we can say that although the scores assigned by the human were significantly lower than those submitted by the machine, there was a positive though weak correlation between the two. This shows the going togetherness of the scores assigned by human and the machine for the overall grade. That is, the overall grades computed by PaperRater could be valid and trustable since they agreed with the scores submitted by the human assessor. Maybe the only change they may wish to make is to put the scores on the curve chart and accordingly increase the scores since the machine scores were significantly lower than the human assessor’s scores. So, one way to use PaperRater will be to use the overall grade that it submits.

Second, if we go to the sub-parts, a number of points will be revealed. For example, if we consider the spelling errors, then we will see that the human and machine performed very similarly. No significant difference was observed between them in terms of the number of spelling errors they found. Further, the scores they introduced were positively correlated. This shows that if students and instructors want to use PaperRater for spell check, the software will prove useful since an agreement was observed between human and machine ratings.

Third, if only the grammar component of the software is considered, again the software is relatively trustable since a rather strong positive correlation was found between the scores submitted by the human rater and PaperRater, although the number of grammar errors found by PaperRater was significantly lower than that found by the human. This shows that the grammar component of the software should, of course, be revised and promoted in order to enable it to reveal a human-like performance in detecting the grammatical errors.

Fourth, for word choice and style no correlation was found between the scores assigned by machine and the human assessor. For style, the human assessor's scores were significantly lower than the machine, but for word choice the human assessor's scores were significantly higher. One possible reason for this finding could be lack of full familiarity of human assessors with the criteria of assessing the style. These criteria in the software are clear and straightforward, and so the scoring is easy but human assessors look at this sub-part in a general term and as a vague concept and hence the scores they assign are somewhat low. For word choice another scenario could be introduced. Here, no correlation was observed between the human and the machine and the scores assigned by the human were lower. One possible reason for this could be that vocabulary knowledge is something that varies from one human assessor to another. Further, each human assessor might assess the word choices made by the students in a different way and in reference to his/her vocabulary repository. For example, since the instructors are not native and teach in an EFL environment, and differ in their levels of vocabulary knowledge, they may assess the same paper differently and in a subjective way. So, because of the good stock of vocabulary items in PaperRater, this software seems more trustable in judging the real word choice power of the participants in essay writing.

In conclusion it could be said that neither human assessment nor machine assessment is complete and each has its own advantages and disadvantages. For example, [32] reported that the reliability of machine scoring is lower than scoring by humans. In contrast, [33] and [34] reported an agreement between machine and human scoring of essays. What is apparent is that machine assessing needs more research so as to increase its validity and performance, but even in its current state, it can be used as a great help by teachers and class instructors.

The findings in this research have a number of implications for educational systems as well as for instructors, students, syllabus designers and policy makers. For example, instructors can use a number of text scoring software and then select one in their scoring of the students' papers. Of course, teachers and instructors might wish to use a hybrid model in which the ultimate assessment is based on a combination of feedback delivered by the instructor and the software. Anyway, application of a software is advantageous since as mentioned by [39] assessment by humans is costly and time consuming, and hence machine assessment can be used to reduce time and the cost. Students will also benefit from the findings of the present study. For example, they can use software for self or pre-assessment purposes. Such software could also raise awareness regarding the students' most common writing mistakes [40]. Further, they will be able to revise their writings so as to increase their scores in formal exams. In fact, as asserted by [41] automated scoring is significantly important in essay writing. Therefore, the results of this study could be helpful in writing courses as it could be a great help to students to improve their writing skills. Also, the results of this research could be beneficial for language learning and language teaching curricula in that the error analysis shows the major areas of concern about the writing skills of ESL students [7]. Finally policy makers and syllabus designers can also embed a demonstration of text scoring software in the syllabi so that both teachers and students will be familiarized with the advantages and of also disadvantages of text scoring software.

While doing the present research, the researchers encountered a number of limitations. For example, due to limited availability of final exam papers and also the outbreak of Covid-19, which made access to the students a difficult job, the researchers used a limited number of final exam papers that were available. Further, only papers written by females were used since there were few male students in the class. In Iran, the majority of students of English Language Translation and also TEFL are female. Finally, only a single text scoring software, PaperRater, was used in this study.

There is no end to research and the present study dealt only with a couple of issues. So, there are a lot to be done by other researchers. As an example, other researchers can use the findings of this study and undertake a number of more comprehensive studies. For example, here the present researchers used only one software, PaperRater, to compare it to a human assessor. Other researchers can use other software or they may use more than one software which will result in more interesting findings. Here in this study, the final exam papers of students in essay writing were used as data. Other researchers can use other types of writings of students, e.g. letter writing, etc. Further, in this study B.A. students of English Language Translation were used. Other researchers may use students from other majors or from other levels like M.A. or even Ph.D. Or some researchers might wish to use multiple human raters and multiple software at the same time and make in-group and out-group comparisons between the scorings of human assessors and the text scoring software.

References

[1] Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. Assessment in Education. Assessment in Education, 5, 7-74. http://dx.doi.org/10.1080/0969595980050102.

[2] Townsend, J. S., Fu, D., & Lamme, L. L. (1997). Writing assessment: Multiple perspectives, multiple purposes. *Preventing School Failure: Alternative Education for Children and Youth*, *41*(2), 71-76.

[3] Javed, M. S., Juan, W. X., & Nazli, S. (2013). A Study of Students' Assessment in Writing Skills of the English Language. *International Journal of Instruction*, *6*, 129-144.

[4] McNamara, T. F. (1996). Measuring second language performance. London, England: Longman.

[5] Jones, E. (2006). Accuplacer's essay scoring technology. *Machine scoring of student essays*. Retrieved September 21, 2019 from https://www.jstor.org/stable/j.ctt4cgq0p.9.

[6] Rudner, L. & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. Retrieved on Feb. 19, 2021, from https://www.learntechlib.org/p/95815/

[7] Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. Pedagogies, 3(1), 52-67.

[8] Chung, K. W. K. & O'Neil, H. F. (1997). Methodological approaches to online scoring of essays. Retrieved on 22, Nov. 2021, from https://cresst.org/wp-content/uploads/TECH461.pdf

[9] Vojak, C., Kline, S., Cope, B., McCarthey, S., Kalantzis, M. (2011). New spaces and old places: An analysis of writing assessment software. Computers and Composition, 28(2), 97–111.

[10] Kukich, K. (2000). Beyond automated essay scoring. In M. A. Hearst (ed.), *The debate on automated essay grading*. Retrieved on Oct. 16, 2021, from http://que.info-science.uiowa.edu/~light/research/mypapers/autoGradingIEEE.pdf

[11] Attali, Y. (2004). Exploring the feedback and revision features of criterion. Retrieved Nov. 26, 2021, from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.214.9047&rep=rep1&type=pdf.

[12] Nicholes, P. (2005). Evidence for the interpretation and use of scores from an automated essay scorer. Retrieved Sep. 4, 2021, from https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.183.3141&rep=rep1&type=pdf.

[13] Munday, J. (2016) Introducing Translation Studies Theories and Applications. Routledge, London. https://doi.org/10.4324/9781315691862

[14] Shermis, M. D., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Lawrence Erlbaum Associates Publishers.

[15] Hunsu, N. J. (2015). Issues in transitioning from the traditional blue-book to computer-based writing assessment. *Computers and Composition*, *35(2)*, 41-51.

[16] Yancey, K. B. (2013). Composing, networks, and electronic portfolios: Notes toward a theory of assessing ePortfolios. *Digital Writing Assessment and Evaluation*, *12(3),* 121-132.

[17] Bell, B. (2001). The characteristics of formative assessment in science education. *Science Education*, *85*(5), 536-553.

[18] O'Neill, P., Moore, Cindy, Huot, Brian (2009). *Guide to colledge writing assessment*. United States: Utah State University Press.

[19] Page, E. B. (1966). The imminence of... grading essays by computer. The Phi Delta Kappan, 47(5), 238-243. https://doi.org/10.2307/20371545

[20] Perelman, L. (2014). When "the state of the art" is counting words. *Assessing Writing*, *21(2)*, 104-111.

[21] Rudner, L. (2013). An evaluation of Intellimetric™ essay scoring system using responses to GMAT AWA prompts. *McLean, VA: GMAC.* Retrieved October 16, 2019, From https://www.gmac.com/~/media/Files/gmac/Research/research-report-series/RR0508_IntelliMetricAWA.pdf.

[22] Keith, T. Z. (2003). Validity and automated essay scoring systems. In M. D. Shermis & J. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (pp. 147-168). Mahwah, NJ: Lawrence Erlbaum Associates, Inc

[23] Dikli, S. (2006). An overview of automated scoring of essays. Retrieved on Dec. 10, 2021, from https://files.eric.ed.gov/fulltext/EJ843855.pdf

[24] Wind, S. A., Stager, C., & Patil, Y. J. (2017). Exploring the relationship between textual characteristics and rating quality in rater-mediated writing assessments: An illustration with L1 and L2 writing assessments. *Assessing Writing*, *34(2)*, 1-15.

[25] Mowl, G., & Pain, R. (1995). Using self and peer assessment to improve students' essay writing: A case study from geography. *Innovations in Education and Training International*, *32*(4), 324-335.

[26] Hopfenbeck, T. N. (2019) Writing assessment, comparative judgement and students' evaluative expertise. *Assessment in Education: Principles, Policy & Practice*, *26*(1), 1-5.

[27] Vögelin, C., Jansen, T., Keller, S. D. & Möller, J. (2021) The impact of vocabulary and spelling on judgments of ESL essays: an analysis of teacher comments, The Language Learning Journal, *49*(6), 631-647.

[28] Bae, J., Bentler, P. M., & Lee, Y. S. (2016). On the role of content in writing assessment. *Language Assessment Quarterly*, *13*(4), 302-328.

[29] Wilson, J., Roscoe, R., & Ahmed, Y. (2017). Automated formative writing assessment using a levels of language framework. *Assessing Writing, 34(4)*, 16-36.

[30] Wang, E. (2013). Assessing students' skills at writing analytically in response to texts. *The Elementary School Journal*, *114*(2), 142-177.

[31] Li, J. (2006). The mediation of technology in ESL writing and its implications for writing assessment. *Assessing Writing*, *11*(1), 5-21.

[32] McCurry, D. (2010). Can machine scoring deal with broad and open writing tests as well as human readers? *Assessing Writing*, *15*(2), 118-129.

[33] Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay scoring: A (modest) refinement of the "gold standard". *Applied Measurement in Education*, *28*(2), 130-142.

[34] Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27-40.

[35] Mohammadi, Z. (2018). Comparative effect of online summative and formative assessment on EFL student writing ability. *Studies in Educational Evaluation*, *59(3)*, 29-40.

[36] Kooshki. A., Rahimi, B., & Mehrpour, N. (2013). D*iagnostic and developmental potentials of dynamic assessment for writing skill* (Unpublished M.A. thesis). Shiraz University, Shiraz, 2013).

[37] Khodadadi, A., Hashemi. S., & Yazdanmehr. H. (2009). *Writing assessment of non-English students: Teachers and students' perspective* (Unpublished M.A. thesis, Ferdowsi University, Mashhad, 2009).

[38] Heidari. G., Maerefat. Y., & Keyvan Panah. T, (2013). *An investigation into Iranian teachers consistency and bias in evaluation of students writing.* (Unpublished M.A. thesis, Allameh Tabatabaei University, Tehran, 2013).

[39] Paul, M., Finch, A., & Sumita, E. (2007). Reducing human assessment of machine translation quality to Binary classifiers. Retrieved Dec. 14, 2021 from https://aclanthology.org/www.mt-archive.info/TMI-2007-Paul-ppt.pdf.

[40] Schraudner, M. (2014). The online teacher's assistant: Using automated correction programs to supplement learning and lesson planning. Retrieved Nov. 23, 2021, from https://www.semanticscholar.org/paper/The-Online-Teacher%27s-Assistant-%3A-Using-Automated-to-Michael/fe21cca6250d6776fb60a4cfbd771918a463116a.

[41] Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. Retrieved Feb. 2022 from https://ejournals.bc.edu/index.php/jtla/article/view/1650/1492.

Appendix A: A Student's Sample Essay.

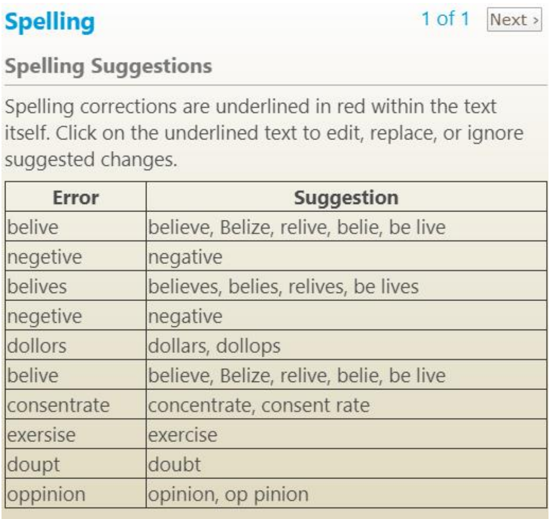
Some people believe that robots will play an important role in societies while others Argue that robots might have negative effects on society. this essay meticulously elaborates on these opinions.

robots usage and being handy is undeniable because of the pace of technology Advancement in these days. even now in some countries they've put some prototypes of robots in work places to see how they do.

First of all I need to mention that nothing is a package of all positive things or negative things. and we should Judge them based on our needs or the time and place we are in, robots are no exception. Although robots can be very useful but they might have some bad effects on the society. people will rely on robots to do most of their daily tasks and it will make them lazier than ever and they might even forget how to be responsible and robots might become so Advanced to work instead of some people so some will lose their jobs and even in this time most of the factories are operated by machines which work instead of 100 person on the other hand robots are though because they are made of steel and wires and most importantly they can't feel anything like tiredness or pain and they don't even need to spend time with anyone so they can work all day long and doing every task will be easier and faster which is an essentiol need in the future. I khow some might be against what I'm going to say but let's Accept the fact that NOT ALL But Most people are mischievous. every each of us had done some bad things in our life that's because we are human and we can think but robots are not like that, they are operated upon some numbers and Algoritems so they will do exactly what they should and they will do whatever they supposed to do neatly , completely and most importantly without cheating in addition to that robots can work in dangerous places which has high risk of accident and severe injury and this will save the life of some person whom might worked them and putting them in food factories in even better because you won't see a hair in your food but you might see a wire which is less disgusting than a hair

to sum it up as I've mentioned robots will be an absoloute and undeniable helper in the near future which is based on fast pace and maximum organization in every task but after all people are in charge of robots so they will decide whether they should be useful or not.

Appendix B: A Sample Writing Task Assessed by PaperRater.

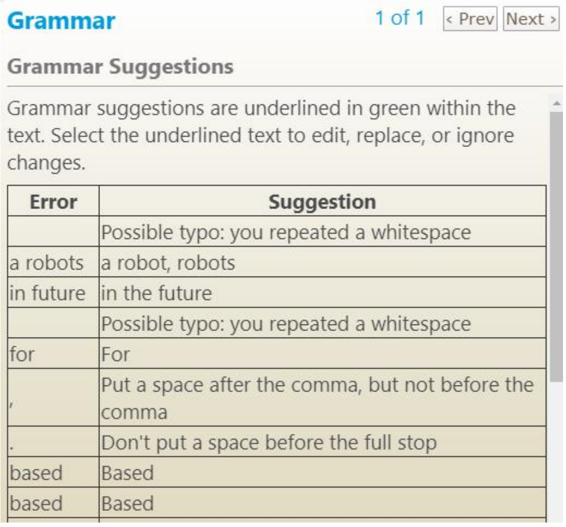


Spelling

Spelling Suggestions

Spelling corrections are underlined in red within the text itself. Click on the underlined text to edit, replace, or ignore suggested changes.

Error	Suggestion
belive	believe, Belize, relive, belie, be live
negetive	negative
belives	believes, belies, relives, be lives
negetive	negative
dollors	dollars, dollops
belive	believe, Belize, relive, belie, be live
consentrate	concentrate, consent rate
exersise	exercise
doupt	doubt
oppinion	opinion, op pinion

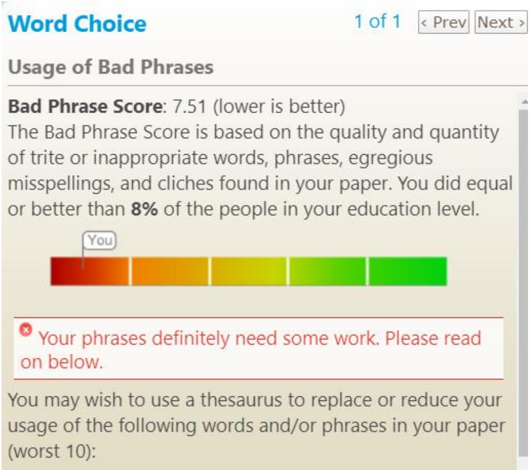


Grammar

Grammar Suggestions

Grammar suggestions are underlined in green within the text. Select the underlined text to edit, replace, or ignore changes.

Error	Suggestion
	Possible typo: you repeated a whitespace
a robots	a robot, robots
in future	in the future
	Possible typo: you repeated a whitespace
for	For
,	Put a space after the comma, but not before the comma
.	Don't put a space before the full stop
based	Based
based	Based



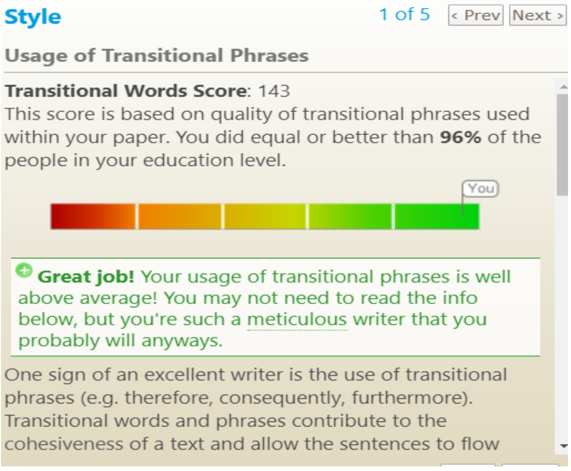
Word Choice

Usage of Bad Phrases

Bad Phrase Score: 7.51 (lower is better)
The Bad Phrase Score is based on the quality and quantity of trite or inappropriate words, phrases, egregious misspellings, and cliches found in your paper. You did equal or better than **8%** of the people in your education level.

Your phrases definitely need some work. Please read on below.

You may wish to use a thesaurus to replace or reduce your usage of the following words and/or phrases in your paper (worst 10):



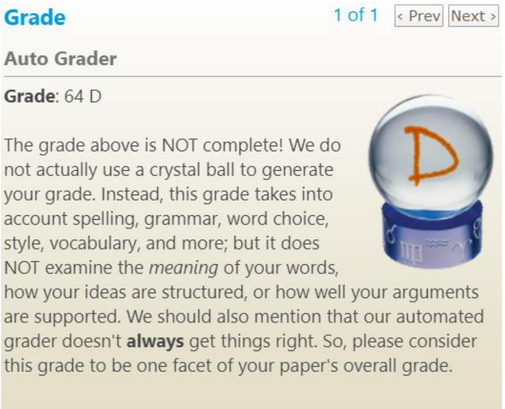
Style

Usage of Transitional Phrases

Transitional Words Score: 143
This score is based on quality of transitional phrases used within your paper. You did equal or better than **96%** of the people in your education level.

Great job! Your usage of transitional phrases is well above average! You may not need to read the info below, but you're such a meticulous writer that you probably will anyways.

One sign of an excellent writer is the use of transitional phrases (e.g. therefore, consequently, furthermore). Transitional words and phrases contribute to the cohesiveness of a text and allow the sentences to flow



Grade

Auto Grader

Grade: 64 D

The grade above is NOT complete! We do not actually use a crystal ball to generate your grade. Instead, this grade takes into account spelling, grammar, word choice, style, vocabulary, and more; but it does NOT examine the *meaning* of your words, how your ideas are structured, or how well your arguments are supported. We should also mention that our automated grader doesn't **always** get things right. So, please consider this grade to be one facet of your paper's overall grade.