Grade Prediction Is Not Grading

On the Limits of the e-rater

Jan Georg Schneider & Katharina A. Zweig

Abstract: *We use the example of the so-called e-rater to show how automated essay grading systems work and where we see their limitations, but also their potentials. From the perspective of a speech act theory that follows Austin and the late Wittgenstein, we show that the prediction of an essay evaluation as provided by the e-rater is subject to completely different felicity conditions from those of the evaluation itself. We find that the e-rater is not suited to capture cultural meaning. It analyzes cohesion without considering coherence, and for this reason it cannot be used to evaluate essays as a 'grader'. Nevertheless, we explore the question of whether it could be integrated into the evaluation process as a corrective under certain circumstances. Thus, using a specific example as an illustration, we show in detail the conditions under which machine predictions can have their role in social processes, even if the prediction of the outcome of a speech act is not equivalent to the performance of that speech act itself. On the basis of these findings, we reflect on the social nature of machine learning systems and their embeddedness in society and culture.*

1. Introduction

In this contribution we are concerned with the question of whether artificial intelligence can replace human decision-makers by predicting their decision on a new case. We address this question for a specific task: the grading of an essay. For this purpose, we analyze the so-called e-rater, a machine learning system for grading essays, which was developed in the United States and is already widely used in the educational system there. We take it as an example to show how automated essay grading systems work and where we see their limitations, but also their potentials. In the following three sections, we describe

how the e-rater functions and what issues it raises. In the last section, we further develop our central speech act argument by showing how predicting an essay grade is fundamentally different from the grading itself. Finally, we address the question of whether, despite its limitations, the e-rater can be used in some way to support essay grading.

2. Training an essay scoring system

The patent of the e-rater, entitled "System and Method for Computer-based Automatic Essay Scoring",[1] was approved in 2002. The description essentially also applies to the current version of the e-rater, in which nothing has changed fundamentally (cf. Perelman 2020). The system serves the purpose of replacing gradings by human reviewers (patent: 1). Primarily, the e-rater is used in so-called language proficiency courses, such as the TOEFL test. Here is an example of a typical task that could appear in such tests:

> "Everywhere, it seems, there are clear and positive signs that people are becoming more respectful of one another's differences." In your opinion, how accurate is the view expressed above? Use reasons and/or examples from your own experience, observations, or reading to develop your position. (patent: 10)

The task is then to write an essay of about 400–500 words, which is very often part of the final exam of language proficiency courses. They aim to find out whether the student's English skills are good enough to study at university, for instance.

The underlying scoring system by humans is very elaborate and uses a matrix of so-called rubrics, which assign grades according to quality criteria as described in the patent:

> For example, the scoring guide for a scoring range from 0 to 6 specifically states that a "6" essay "develops ideas cogently, organizes them logically, and connects them with clear transitions". A human grader simply tries to evaluate the essay based on descriptions in the scoring rubric. This technique, however, is subjective and can lead to inconsistent results. (patent: 1)

1 Cf. Burstein et al. 2002; hereafter cited as "patent" with the indication of the column number.

Thus, according to its inventors, the attractiveness of the e-rater lies on the one hand in its greater accuracy and objectivity, and on the other hand in the cost-efficient replacement of human reviewers. However, previous human evaluations form the basis for the calculations of the e-rater. The general approach of the system is to check which features are common in essays that have been positively evaluated by humans before. So, how does the e-rater do this?

Technically speaking, the e-rater is a combination of curated rules, so-called *expert systems*, and a learnt component to predict essay grades. In the patent, the approach of learning how to predict a grade is described for two kinds of essays. Here, we focus on so-called *argument essays*, in which the student is provided with an argument and asked to analyze it. The grading process for a given test question to be answered in the argument essay is based on an electronic version of the essay. This electronic version is read by a standard language parser assigning word categories to each word and also identifying larger syntactic structures such as *infinitive clauses* or *relative clauses*. Another expert system tries to identify the beginning and end of individual arguments by searching for a list of keywords such as *otherwise*, *conversely* or *notwithstanding* (patent: column 11). This heuristic thus annotates the text and splits it into individual arguments based on particular words and phrases. The partition of the text and the text as a whole then provide the basis for further calculations, which are all of a very simple nature, e.g., counting the total number of infinitive clauses.

Eventually, each essay results in four sets of numbers.

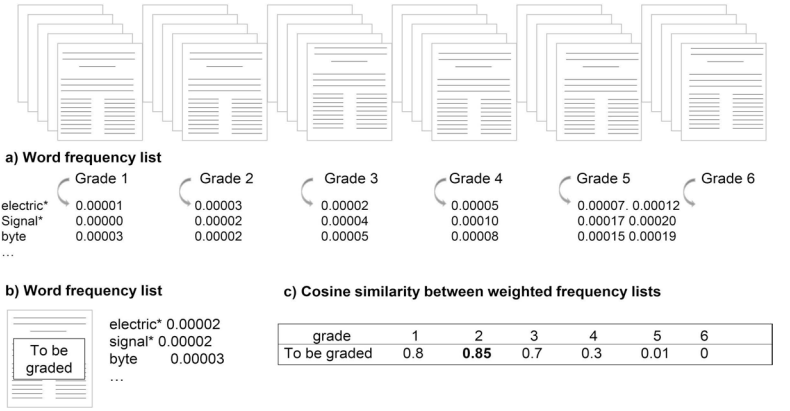
- The first set of results contains only two values: The total number of *modal auxiliary verbs* (such as *can*, *must*, *will*) and the relative portion of *complement clauses* per sentence.
- The second set of results is supposed to score the *rhetorical structure* of the essay; this is based on the output of the argument-identification component described above. It contains four values obtained from a count of specific markers, e.g., the total number of occurrences of subjunctive forms of modal verbs (*would*, *could*, *should*) in the final paragraph of the essay.
- The third set of results contains a weighted value for words depending on their salience (see Fig. 1). The salience of a word is calculated by its relative frequency in any essay and by the inverse frequency in the set of all essays. For example, any word directly related to the test question will be very frequent in an individual essay, but at the same time its salience is very low because it occurs in all of the essays. Technically, the third set of results

is based on the product of the token frequency of the particular word in one document and its inverse token frequency across all documents.[2] This gives a long list of numbers for each individual essay. The same is done for the concatenation of all essays graded with a 1, for the concatenation of all essays graded with a 2, and so on. In the last step, the weighted word list is compared to the weighted word list for essays from each of the six different grades by a measure called *cosine similarity*. The result of this third analysis is the grade of the essay collection that is most similar to the essay under scrutiny with respect to this similarity measure. That is, if in class 6, for example, some rare words are used by many essays and the new essay does so as well, while other, more frequent, words are not used so much in essays of this class and in the new essay, the essay is more likely to be assigned to class 6 than to some other class. Figure 1 illustrates the issue just described, i.e., the e-rater's grading approach based on word frequency lists, which leads to the third result mentioned. a) The essays graded by humans are sorted by grade. Over all essays with the same grade, the frequencies of the stemmed words are counted. Thus words like *electrical* and *electricity* are counted as *electric**. For each stem of a word, all occurrences in the essays of the same grade are counted as a frequency. b) The same is done for the essay to be graded by the system. All words on all word lists are then weighted by the inverse frequency of the (stemmed) words in all texts. c) Finally, the resulting sequences of numbers from all essays with a grade 1, those of all essays with a grade 2, and so on are compared to the sequence of numbers of the essay to be graded in terms of cosine similarity. The highest value determines the outcome of the third score in the grading approach of the e-rater.

- For the fourth output, the procedure described for the third is similarly performed for the words in each 'argument' as identified by the expert system with its keywords created by humans: That is, a grade is now assigned to each 'argument' by comparing word usage with arguments from essays in the different categories. All these scores for the individual 'arguments' are then averaged in a way not described in detail in the patent to form the final score.

2 Stop-words are removed and the words are stemmed.

Figure 1: e-rater's grading approach based on word-frequency lists (3^{rd} number computed in the approach)



The entire procedure of creating these four sets of scores is applied to 250–300 essays already graded by humans (see also Burstein et al. 2013a: 61) – thus the eight resulting numbers and the grade given to each of these essays are known. These numbers are then the input to a *linear regression*, a simple machine learning method. The method finds the weights for each of the inputs so that the predicted grade from a linear equation based on the inputs is not too far from the actual grade, averaged over all 250–300 graded essays. The resulting formula is then used for all essays to be re-scored. Thus in summary, the e-rater does the following:

- Based on a set of 250–300 essays on one test question which were graded by humans, it learns which of the very simple syntactical features are most associated with a good or bad grade. These features can be easily identified, e.g., the number of tokens of modal auxiliary verbs.
- It also learns which words are most and least often used in essays on the given question, both in the overall text and in the individual arguments identified by an expert system, based on a list of keywords created by human experts.

It can thus be stated that the e-rater is not designed and hence not able to identify coherence, logical arguments and the reasonability of their interconnec-

tion. It can only count individual words[3] and very simple syntactical structures, and it has "learned" which number of tokens of these structures is often associated with a high or low grade.

3. Criticism of the e-rater

The patent states that the e-rater-system "automatically rates essays using features that reflect the 6-point holistic rubrics used by human raters" (patent: 3). For example, an essay would be worth the highest grade 6 according to such a rubric matrix if it "develops ideas cogently, organizes them logically, and connects them with clear transitions", as already quoted above (patent: 1). The patent-holders base their quantification of these features on the identification of specific words that may be used in a discourse to structure arguments. However, these word lists are not complete, nor is the identification of mere words a substitute for the semantic and pragmatic analysis of the argumentative structure. Thus the heuristics used may or may not identify all subtexts containing an argument. In any case, the logic of their organization or the use of clear transitions cannot be grasped by machines. It is not measured whether the ideas are coherently developed – even the inventors of the e-rater concede this (Burstein et al. 2013b). Nor is the organization of the arguments assessed in terms of content.

The second vector, which is said to represent "rhetorical structure", is just a count of simple syntactic properties of the text, e.g., the "total occurrences of argument development using belief words", based on a heuristic for identifying keywords that statistically indicate arguments, i.e., that are taken as "indices" or "symptoms" of arguments (cf. Keller 2018: 155–168, with reference to Peirce; see also section 5 below). Again, it is not at all the students' rhetorical skills that are determined since the accomplished quantification is not based on a semantic analysis of the text. Perelman (2020) puts it like this: "Testing companies freely use the term artificial intelligence, but most of the systems

3 In the *Handbook of Automated Essay Evaluation*, Burstein et al. (2013a: 60) state that "topic-specific vocabulary" would also be identified in advance – by humans – as characteristic of better-rated essays; thus it is not a purely quantitative approach. This does indeed bring into play a human categorization ("topic-specific vocabulary") which as such cannot be quantitative but rather constitutes a semantic property; however, this does not change the fact that the meaningfulness and correctness of the use of these expressions cannot be checked by the e-rater.

appear to produce a holistic score largely through summing weighted proxies." To prove this, Perelman built a text generator called BABEL which generates grammatically correct but otherwise non-sensical texts. Here is an example of a text generated by such a system (quoted after Perelman 2020):

> Theatre on proclamations will always be an experience of human life. Humankind will always encompass money; some for probes and others with the taunt. Money which seethes lies in the realm of philosophy along with the study of semiotics. Instead of yielding, theatre constitutes both a generous atelier and a scrofulous contradiction. As I have learned in my reality class, theatre is the most fundamental analysis of human life. Gravity catalyzes brains to transmit pendulums to remuneration. Although the same gamma ray may receive two different pendulums at the study of semiotics, a plasma processes interference. Simulation is not the only thing an orbital implodes; it also inverts on theater. […]

According to Perelman, this text and similar ones earned the highest grade even with the newest e-rater system. Perelman's results are often used to show that students might be able to "learn to the test", e.g., by adding rare words to their essay whether they fit or not, or by learning lists of special "cue words" by heart; they can also be taught to use specific syntactic structures independent of their semantic quality and fit. It is obvious that memorizing and using proxies does not mean knowing how to write meaningful essays, while it elicits the highest grades from the Automated Easy Scoring (AES) since the AES cannot capture whether a text makes sense or not. Thus a self-fulfilling prophecy can set in: It is possible that the predictions become more and more precise, while the texts become more and more meaningless, since a meaningful coherence is no longer an evaluation criterion. If the e-rater were actually used as a *substitute* for essay evaluations by humans, as the authors of the patent envision, then it would basically be honest and transparent to test, very explicitly, only the knowledge of the expected proxies and thus also make explicit the underlying 'teaching to the test'.[4]

However, since the required competence of each student consists or should consist precisely in writing texts that can be understood by other humans in terms of content, and since the e-rater is not suitable for handling situations in which human feedback is needed, the use of such a system *alone* is already ruled

4 However, it should not be forgotten here that successful candidates do also consider what *human* raters might 'want to hear' (in ideological terms, for example).

out for didactic reasons. For the following discussion, we will thus assume that students need to expect at least control samples by human graders so that they cannot afford to write non-sensical texts.

4. Quality of the prediction

Can the e-rater and similar systems be used as the main grading system if human graders control some sample of the essays? To answer this question, the accuracy of the grade prediction must be considered first. A study in Germany and Switzerland showed that for two different types of tasks the e-rater, when specifically trained, achieves between 13% and 42% absolute agreement, i.e., it hits exactly the same grade as human graders in this percentage range on a 0–5 point scale (Rupp et al. 2019). Furthermore, if a deviation of the grade by no more than one point on the 0–5 point scale is considered, the accuracy of the system increases to up to 99% (between 73.8% and 99.4%; Rupp et al. 2019: Table 5). In the light of these results the authors estimated that the e-rater's predictions were within an acceptable range – although the agreement between human raters was significantly higher than that between the system and human raters.

Other studies on the e-rater show similar accuracy values or even slightly better ones (cf. Meyer et al. 2020: 4), so the acceptance level for its use has increased in recent years. In the following, we will argue that the results provided by a system such as the e-rater, even if a 100% agreement can be achieved, are not *qualitatively* the same as an essay assessment, but something categorically and qualitatively (cf. Becker 2021: 9–30) different. It is not even an essay grade, but exclusively a prediction of such a grade. This is our core argument, which we develop in the next section using the perspective of speech act theory. What are the characteristics of the act of evaluating an essay? Here, we are not interested in denying the power of the e-rater as a predictive tool; rather, we want to explore the limits of its applicability. Under what conditions can such a system be used in a supportive manner, and what requirements of the social process can it *not* meet? In order to arrive at an assessment that is as robust and fair as possible, we assume a best-case scenario for the use of an e-rater system as follows:

- The system is trained for each test question separately, based on 250–300 essays rated by human raters.

- The students know or at least assume that their essay may be graded by a human rater – thus they have to write an intelligible essay (no 'learning to the test').

5. Grade prediction is not grading

Summarizing what we have presented in the last two sections, we can say that the e-rater is not able to apply quality criteria but only to count 'symptoms'[5] such as the frequency of words and particular syntactic structures. Thus it cannot *evaluate* essays, which are culturally anchored semiotic phenomena. As such, the comprehension of essays and other texts is culturally embedded and dependent on cultural knowledge the e-rater does not have – for fundamental reasons. Cultural knowledge, which significantly includes the understanding of cultural artifacts, is "a complex constellation of acquired abilities" (Goodman/Elgin 1988: 114) that cannot be acquired by machines trained in the way outlined above.

In order to make this idea clearer, let us undertake a small detour via the automated *translation* program DeepL. Since the culture-dependency of essays and other texts seems almost self-evident in some respects, the recent progress of DeepL is astonishing. How is it able to produce translations that can – to some extent and within limits – impress even experienced human translators? The software employs Convolutional Neural Networks (CNN), a method commonly used in image recognition. The advantage is that – unlike so-called recurrent neural networks – all words are translated (cf. Merkert 2017). In the case of DeepL, the CNNs were trained with the associated Linguee system. With this database, DeepL was able to collect extremely extensive, very high-quality training data (Schmalz 2018: 200). The company bases its success in part on the fact that it has access to "billions of high-quality translations" from its corporate history (Schmalz 2018: 203).

The fundamental qualitative difference between predicting a grade of an essay and translating a text with an automatic translator is twofold. First, DeepL produces something that is categorically the same as a human translation: a linguistic product that can be evaluated according to the same criteria

5 In semiotics, a symptom is an outward indicator (*Anzeichen*) of something. For example, red spots on the face can be a symptom of measles (cf. Keller 2018: 161). In informatics, the analogous term 'proxy variables' is used.

as a human translation. Second, the cultural embedding or the cultural circumstances of the words in the text are potentially represented, although there may of course be errors in these representations of cultural meaning offered by DeepL, so the ultimate decision and responsibility must rest with the human translator. However, the large number of human translations that serve as reference texts ensure inclusion of the relevant criteria which effectively orient people when translating. This way, for instance, human intentionality and taste as well as the 'zeitgeist' together with its fluid, group-specific differentiating conventions can, and often do, find expression in the translations produced by DeepL.

The e-rater, by contrast, makes a rating prediction, more precisely a grade prediction,[6] based on indices or symptoms alone. And from the perspective of speech act theory, the prediction of a grade is, as we will show below, something categorically different from an evaluation itself.

5.1 The basic idea of speech act theory

Speech act theory was developed by John L. Austin in his 1955 Harvard lectures and published posthumously under the title *How to Do Things with Words* (Austin 1975 [1962]). In everyday situations, we often think of speaking and acting as opposites. But for Austin, speaking in many cases means doing something, namely, performing speech acts. When I say, "I christen this ship the Queen Elizabeth", I am performing the speech act of ship christening, provided the circumstances fit and I am authorized to do so. When children re-enact such a christening in play, it will thus not have the same effect. Only if a set of "felicity conditions" is met can the speech act be successful. Austin calls utterances of this kind *performatives* or *illocutionary acts*.[7] A judge's verdict is also such a

6 Of course, one could object that DeepL also technically generates a translation *prediction*. However, the de facto difference is that here a product is created which can be evaluated in the same way as a human translation and into which the cultural context has implicitly entered, as already mentioned. The e-rater, on the other hand, does not base its grade prediction on criteria but only on symptoms, on the basis of which the quality of the evaluation and grading cannot be assessed, but only the precision of the prediction in comparison with a human evaluation.

7 On the surface, Austin's argument could be understood as abandoning the distinction between performatives and constatives, replacing it with the distinction between locutionary, illocutionary and perlocutionary acts, in which the notion of illocutionary act captures the performative aspect of speech acts. However, if one includes the

performative utterance. If the court is legitimate and the procedure is carried out correctly and completely, then the verdict applies with all its consequences. Furthermore, the question arises of whether the verdict was appropriate and fair. All these aspects are considered by Austin.

Each speech act has its specific set of felicity conditions, and the individual conditions often differ from those of other speech acts. Such differences may include, for instance, whether a speech act requires a justification and why a justification is necessary. With this in mind, let us now present and discuss the felicity conditions of evaluating and grading.

5.2 Felicity conditions of evaluation and grading of essays

Our hypothesis is that an essay evaluation must include a justification, preferably even an explicit one. But how can we justify why it must contain a justification? Why is a prediction not sufficient here? Why do we also need a qualitative evaluation?

As philosophers of language have made very clear, cultural meaning is constantly renegotiated by the collective, that is, by a community of sign users (cf. Wittgenstein[8] 1984; Goodman/Elgin 1988). But how can we involve this collective in the grading process? Essay evaluation can only be legitimized by knowing how sign usages are entrenched and established within the collective. The existence of this knowledge, then, is one of the felicity conditions in essay evaluation, for only on this basis can a valid justification be given for an evaluation. And only a justification provides the possibility of checking whether, e.g., an evaluation is intersubjectively legitimated or – at least – legitimizable and not arbitrary.

With regard to felicity conditions, the questions that arise are a) exactly which procedures are to be chosen here, and b) which persons fit these procedures and therefore should be authorized for an evaluation. In scientific and educational contexts, we employ *experts* for this purpose: i.e., people we authorize as reviewers because we believe they have been part of the collective long

---

fact that Austin kept revisiting the distinction between performatives and constatives, even though he had long since 'deconstructed' the dichotomy between the two, then there are good reasons to suppose that the notion of the performative remained important to him. Even if all speech acts are ultimately performative, there are those in which the performative character is more prominent than in others.

8 On Wittgenstein's pragmatic conception with regard to linguistics cf. Schneider 2008.

enough to be able to speak for it, or more precisely, to be able to provide good candidates for justifications that are then in turn intersubjectively verifiable (e.g., by other experts).

In the following, we go through Austin's (1975 [1962]: 14–18 and 25–46) six felicity conditions in detail and apply them systematically to essay grading:

- *A 1*: A speech act can only be accomplished at all if there is a corresponding *conventional procedure* which involves certain persons uttering certain words under certain circumstances. In the case of essay grading, this is a procedure that requires a close reading of the submitted essay and an assignment of the essay to a score level according to specific criteria within the framework of the underlying grading scheme, e.g., those given by a rubric. The central linguistic act, usually a writing act in the case of essay evaluation, has certain similarities with a judge's verdict and can be put into an explicitly performative form (cf. Austin 1975 [1962]: 69) of the following kind: 'I hereby evaluate the present essay with the grade x.' As mentioned, essay evaluation also includes justification of the grade by the reviewer or at least the assumption that it can be justified by the reviewer upon request.
- *A 2*: The respective persons, objects and circumstances must fit the speech act to be performed. In our case, the persons authorized and qualified to perform the evaluation are the experts employed by the collective, e.g., teachers or professors.
- *B 1 and B 2*: All persons involved must carry out the procedure correctly and completely. In our case, this means that on the basis of the respective essay and with the help of transparent criteria (cf. rubric descriptions), an assignment to one of the intended grading categories must be made unambiguously. The correctness and completeness of the procedure also includes, for example, checking as well as possible whether the submitted essay is valid, i.e., that it is, for instance, not plagiarized.
- If one or more of the conditions A 1 to B 2 are not met, then the speech act of essay grading does not proceed. It can also happen that such an evaluation is null and void in retrospect, e.g., because the essay only later turns out to be plagiarized.
- But even if the evaluation has come about and is thus valid, it can still fail in two other ways. Since these conditions are of a categorically different kind, Austin does not continue his list with the third letter of the Latin alphabet, but with a Greek gamma.

- *Γ 1*: The speech act must not be untrustworthy or insincere. With respect to grading in general, this requirement can be deduced from the perspective that a grade must also be a "signal" (see Spence 1973) to the author (i.e., the student) and also to potential future employers. The grade signals how the student's performance was assessed by an expert with regard to the qualification aimed at and possibly also with regard to the student's career opportunities. If the expertise is not carried out in good faith, then this condition is not met. We see that it is precisely here that the ethical and moral dimension of grading is located.
- *Γ 2*: Afterwards, all participants must behave in a way that fits the respective completed speech act. For example, if one has given a very good grade, it is not appropriate to reprimand the student afterwards. It would be equally inappropriate for an expert to cast doubt on the student's evaluation afterwards. Here we can see, by the way, how closely Γ 1 and Γ 2 can be related.

If all six felicity conditions with respect to the grading are fulfilled, then the probability is very high that the evaluation was successful as a speech act.[9] Beyond that, however, the question can of course still arise as to whether it was fair, appropriate, etc. If a grade is being challenged on substantive rather than procedural grounds, it may be appropriate to bring in additional reviewers. How many reviewers are required depends, generally speaking, very much on the type of evaluation procedure. In the case of a post-doctoral habilitation in the German university system, for example, three to five reviewers are involved, while a simple exam in school is usually graded by just one teacher.

In contrast to the procedure of human essay evaluation and grading described above, the symptoms that the e-rater identifies and then uses for its predictions can never be used as *reasons* for gradings. However, the justification is the most important factor in the procedure of essay evaluation, which consists of both grading and justification. The justification serves to stabilize the procedure for the future, and only in this way can the felicity conditions be maintained here.

Thus the procedure must remain anchored in the corresponding cultural practice. It is essential that the criteria which make up a good essay are explicitly taken into account in the evaluation procedure and that their aggregation

9 However, this is not absolutely certain, because owing to the fundamental unpredictability of human communication Austin gives only necessary, but not sufficient felicity conditions.

leads to the specification of a grade, because only in this way can a comprehensible and adequate justification be given. These criteria include, in particular, coherence, argumentative plausibility, truthfulness, originality and aesthetic value. As shown by the software BABEL, the generator of non-sensical texts, the e-rater cannot capture any of these aspects but can only consider surface phenomena of vocabulary and cohesion. It analyzes cohesion without coherence, symptoms but not criteria. Machine learning cannot distinguish between rational and senseless inferences (Goodman/Elgin 1988: 108f.; Anson/Perelman 2017: 279); AES systems cannot 'read between the lines', they do not capture allusions and irony, and they are not able to assess complex novel metaphors and humor (Balfour 2013: 42).

5.3 Can predictions nevertheless play a meaningful role in the evaluation process?

As we have pointed out, grading is a completely different speech act from grade predicting, and grading can only be done by educated, selected members of the respective language collectives – which restricts the performance of such an act to humans for the time being. Strictly speaking, the e-rater cannot perform the speech act of prediction either – because "machines are not actors" (cf. Becker 2021: 19, our translation) – but at least it can substitute such a prediction under certain circumstances (cf. Janich 2015: 314; Becker 2021: 182). This raises the question of whether such automated grade predictions can nevertheless be helpful in the process of grading and evaluating an essay. A prediction is successful if it is as statistically accurate as possible. This statistical accuracy must be established in many instances in order to say that the prediction could be viable at all. Obviously, this is the case with the e-rater, so its predictions, if used correctly, may have some value for the process after all.

With all due caution, the e-rater could perhaps help to filter out the 'outliers' heuristically in a large set of essays to be evaluated. Those essays where the human evaluation deviates significantly from that of the e-rater might then deserve special attention. They could – and we do not think this is unlikely – actually be particularly good, although they do not exhibit the statistical symptoms. Conversely, however, the presence of the statistical symptoms could also indicate that a human rater may have rated an essay too negatively or too positively. An automated comparison of the essays with regard to the symptoms mentioned cannot determine whether the respective 'outlier' was caused by the specificity of the essay or by the specificity of the expert opinion. This, in turn,

can only be verified by the judgement of a human employed by the collective, because the e-rater's predictions are based on the assumption of a 'normality' of the essays to be graded.

So even if the e-rater could be used carefully as a convenient tool, this would by no means – and this is the crucial point – authorize it to perform the speech act of grading with all its consequences. As we have argued above, only humans who are part of the respective collective can actually evaluate a text and determine whether it is written in accordance with the culture of that collective. But not *all* humans are allowed to speak for the entirety of this collective: only evaluators chosen on the basis of their qualifications or other prior achievements are capable of doing so. Thus a grade as the result of an evaluation can only be given by a human. Here again, the comparison with a translation by a system like DeepL is quite interesting. It reveals differences as well as similarities: In the case of the translation, it is immaterial whether the rough version came from a human being or a machine, since the cultural context tends to be translated in the process. In the end, however, it is again crucial that for the released version responsibility is taken by a competent human translator who can check whether the nuances have been correctly captured, etc. Again, this has to do with the assumption of the duty to justification and is part of the language game, for example, in book translations by publishers.

Certainly, with the e-rater, a comparison can always be made between the real evaluation by a human and the automated prediction. This prediction always remains dependent on such comparisons, since it is not, after all, based on the quality criteria that humans use to evaluate an essay meaningfully. Without human raters, there is as yet no way to provide a justification for the particular grade, and as we have argued, an essay grade must always be intersubjectively legitimated.

In the course of this analysis, we have identified some methodological preconditions that might allow for a useful application of AES. The first two were already noted at the end of section 4, and we add a third point (c) here:

- Individual training based on tests graded by humans: The system would need to be trained separately for each test question, based on 250–300 essays scored by human reviewers, in line with Austin's felicity conditions.
- Safeguards against learning to the test: It has to be made explicit to the students that they can assume their essay is to be graded by humans, i.e., they need to know that writing a meaningful essay is what matters.

- Legitimacy of the speech act: The prediction by the e-rater is used only as a measure of deviation, a guide to attention, and a possible corrective to supplement human evaluations; it is intended to support human evaluators, but in no way can or should it completely replace them without harming the legitimacy of grading as a speech act.

Generally, it must always be considered and appropriately taken into account that this type of software has massive economic incentives, in terms of saving time and human resources. It is therefore necessary to reflect ethically and politically on the implications this may have (cf. Zweig 2018; Zweig et al. 2021). The more the evaluators are under pressure of time and economic considerations, the more the 'sincerity condition' (Γ 1) is challenged and possibly compromised: honest 'signals' can only be given in the long run if both sides have to prove and can verify honesty. To this end, it must be ensured on the one hand that the reviewer provides grade justification when asked to do so, and, on the other hand, to the greatest possible extent the achievement of good grades with plagiarized or non-sensical texts must be prevented.

6. Conclusion

For the future, the question remains whether the e-rater could become as convincing a tool as, say, DeepL if it learned the grading process itself by being supplied *en masse* with human essay evaluations in text form. This would require, however, linking these evaluations to the respective essays during the e-rater's machine learning process. Then, perhaps, the e-rater could generate an assessment in text form that would be linked and matched to a specific new essay, and that might then even substitute, as in the case of DeepL, the human action in a near-equivalent way (cf. Becker 2021: 182, following Janich 2015[10]). But is such a process of essay evaluation technically even possible, given that

10 In the German original, the relevant term is "leistungsgleiche Substitution", which is hard to translate into English. It means that the substitute is not the same but something with the same output as the substituted action or process: computers, for example, do not calculate, because calculating is an intentional action. But they can have the same output as human calculating. In this sense they substitute human action in a near-equivalent way (cf. Becker 2021: 19).

the content relationship between essay and evaluation text is much more complex and 'loose' than that between a text and its translation? And even if it were possible, we could not do without human reviewers, because only they could further explain the justification if this is required, and take responsibility for evaluating the essays. However, as argued above, even with automated translation, it is necessary that justification by a human can be provided upon request.

The application of speech act theory in our analysis has shown why the prediction of a speech act result does not match the correct performance of the speech act in all its facets. The prediction cannot substitute the speech act of grading in a near-equivalent way. On the one hand, our comparison of grading and grade prediction shows that we are dealing with two categorically different processes; on the other hand, however, the findings also leave room for the assumption that even AI systems that are actually inappropriate can sometimes represent an interesting corrective in a social process. In the intelligent engagement with so-called artificial intelligence, we have the opportunity to reflect on our cultural practices and examine them in terms of their viability: In which cases does a decision need an explicit substantive justification? What are historically evolved practices worth to us in particular? Which ones do we want to retain for ethical or cultural reasons and which can be modified with the help of AI? What happens in communicative processes when machines are involved? Machines rarely perform in the same way as humans; in some cases they perform something functionally equivalent (e.g., a calculator), in many other cases something categorically different (e.g., the e-rater), and in still other cases something somewhere in between (e.g., DeepL). It becomes particularly interesting when a machine learning system uses large amounts of data to uncover quantitative aspects that humans do not see because they interpret qualitatively and in this sense proceed *intelligently*. When humans 'interact' with such technologies, i.e., deal with them intelligently and use them as a corrective, they can optimize their results and decisions, whether in translation, essay evaluation or other cultural practices. In this respect, as Elena Esposito (2017) points out, it is indeed more appropriate to speak not of artificial intelligence but rather of "artificial communication". When people learn with the help of machines by 'interacting' with them, machine learning can potentially help to improve social decisions.

We have proposed an approach that could also be suitable for the assessment of similar AI systems which are intended to complement human evaluations or decisions in social processes. We therefore see potential for generalization in this approach, which we will explore in future research.

Bibliography

Anson, Chris M. and Les Perelman. 2017. Machines Can Evaluate Writing Well. In *Bad Ideas About Writing*, Eds. Cherryl E. Ball and Drew M. Loewe, 278–286. Morgantown, W. Va.: West Virginia University Libraries.

Austin, John L. 1975 [1962]. *How To Do Things with Words*. 2nd edition. Oxford: Oxford University Press.

Balfour, Stephen P. 2013. Assessing writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review™. *Research & Practice in Assessment* 8:40–48.

Becker, Ralf. 2021. *Qualitätsunterschiede. Kulturphänomenologie als kritische Theorie*. Hamburg: Meiner.

Burstein, Jill C., Joel Tetreault and Nitrin Madnani. 2013a. The e-rater Automated Essay Scoring System. In *Handbook of Automated Essay Evaluation. Current Applications and New Directions*, Eds. Mark D. Shermis and Jill Burstein, 55–67. London: Routledge.

Burstein, Jill C., Lisa Braden-Harder, Martin S. Chodorow, Bruce A. Kaplan, Karen Kukich, Chi Lu, Donald A. Rock and Susanne Wolff. 2002. *US 6,366,759 B1* [United States Patent, April 2, 2002: System and method for computer-based automatic essay scoring].

Burstein, Jill, Joel Tetreault, Martin Chodorow, Daniel Blanchard and Slava Andreyev. 2013b. Automated Evaluation of Discourse Coherence Quality in Essay Writing. In *Handbook of Automated Essay Evaluation. Current Applications and New Directions*, Eds. Mark D. Shermis and Jill Burstein, 267–280. London: Routledge.

Esposito, Elena. 2017. Artificial communication? The production of contingency by algorithms. *Zeitschrift für Soziologie* 46(4):249–265.

Goodman, Nelson and Catherine Z. Elgin. 1988. Confronting Novelty. In *Reconceptions in Philosophy & Other Arts & Sciences*, Eds. Nelson Goodman and Catherine Z. Elgin, 101–120. Indianapolis, Ind./Cambridge, Mass.: Hackett Publishing Company.

https://www.heise.de/newsticker/meldung/Maschinelle-Uebersetzer-DeepL-macht-Google-Translate-Konkurrenz-3813882.html. Last access: 3 March 2022.

Janich, Peter. 2015. *Handwerk und Mundwerk. Über das Herstellen von Wissen*. Munich: Beck.

Keller, Rudi. 2018. *Zeichentheorie. Eine pragmatische Theorie semiotischen Wissens*. 2nd edition. Tübingen: Narr.

Merkert, Pina. 2017. *Maschinelle Übersetzer: DeepL macht Google Translate Konkurrenz.*

Meyer, Jennifer, Thorben Jansen, Johanna Fleckenstein, Stefan Keller and Olaf Köller. 2020. Machine Learning im Bildungskontext: Evidenz für die Genauigkeit der automatisierten Beurteilung von Essays im Fach Englisch. *Zeitschrift für Pädagogische Psychologie* 2020:1–12, https://doi.org/10.1024/1010-0652/a000296.

Perelman, Les. 2020. The BABEL Generator and e-rater: 21st century writing constructs and Automated Essay Scoring (AES). *Journal of Writing Assessment* 13(1). http://journalofwritingassessment.org/article.php?article=145. Last access: 3 March 2022.

Rupp, André A., Jodi M. Casabianca, Maleika Krüger, Stefan Keller and Olaf Köller. 2019. Automated Essay Scoring at Scale: A Case Study in Switzerland and Germany. *TOEFL Research Report Series and ETS Research Report Series 1/2019:1–23.*

Schmalz, Antonia. 2018. Maschinelle Übersetzung. In Volker Wittpahl (Ed.). *Künstliche Intelligenz. Technologie, Anwendung, Gesellschaft*, 194–208. Berlin and Heidelberg: Springer.

Schneider, Jan Georg. 2008. *Spielräume der Medialität. Linguistische Gegenstandskonstitution aus medientheoretischer und pragmatischer Perspektive*. Berlin and New York: de Gruyter.

Spence, Andrew Michael. 1973. Job market signaling. *Quarterly Journal of Economics* 87(3):355–374.

Wittgenstein, Ludwig. 1984. Philosophische Untersuchungen. In Ludwig Wittgenstein. *Werkausgabe in 8 Bänden. Vol. 1: Tractatus logico-philosophicus*, 225–580. Frankfurt a. M.: Suhrkamp.

Zweig, Katharina A., Tobias D. Krafft, Anita Klingel and Enno Park. 2021. *Sozioinformatik: Ein neuer Blick auf Informatik und Gesellschaft*. Munich: Hanser.

Zweig, Katharina. 2019. *Ein Algorithmus hat kein Taktgefühl. Wo künstliche Intelligenz sich irrt, warum uns das betrifft und was wir dagegen tun können*. 9th edition. Munich: Heyne.