EssayCritic: Writing to Learn with a Knowledge-Based Design Critiquing System

Anders I. Mørch[1*], Irina Engeness[1], Victor C. Cheng[2], William K. Cheung[2] and Kelvin C. Wong[2]

[1]Department of Education, University of Oslo, Oslo, Norway // [2]Department of Computer Science, Hong Kong Baptist University, Hong Kong // anders.morch@iped.uio.no // irina.engeness@iped.uio.no // victor@comp.hkbu.edu.hk // william@comp.hkbu.edu.hk // kckwong@comp.hkbu.edu.hk

[*]Corresponding author

ABSTRACT

This article presents a study of EssayCritic, a computer-based writing aid for English as a foreign language (EFL) that provides feedback on the content of English essays. We compared two feedback conditions: automated feedback from EssayCritic (target class) and feedback from collaborating peers (comparison class). We used a mixed methods approach to collect and analyze the data, combining interaction analysis of classroom conversations during the writing process and statistical analysis of students' grades. The grades of students in both classes improved from pre-test to post-test but in different ways. The students in the target class included more ideas (content) in their essays, whereas the students in the comparison class put more emphasis on the organization of their ideas. We discuss our findings to identify strengths and weaknesses of our approach, and we end the paper by suggesting some directions for further research.

Keywords

Automated feedback, Collaboration, Decision tree learning, Design critiquing framework, English as a foreign language (EFL), Essay writing, Learning analytics (LA), Machine learning, Methods for using LA in EFL, Peer feedback

Introduction

Three components in successful foreign language learning according to the call of this special issue are: learners, goals, and contexts. We address them in this paper by formative assessment (process feedback). We focus on the communicative aspects of foreign language learning, writing and speaking, in terms of content (meaning or ideas) and organization of shorter texts, and not on grammar and spelling (acquiring linguistic competences). This is an under-researched theme in language learning and contemporary research explores semantic analyses tools, which we give an example of in this article. It addresses a challenge in many countries today that young people are overexposed to learning spoken English through new media without being activated by reading and writing. The overreliance on speaking for developing language skills may have a negative effect on vocabulary development (Weizmann & Snow, 2001). We address the discrepancy by a computer-based writing aid together with collaboration in small groups, which is informed by a theoretical framework for integrating action (writing) and reflection (thinking aloud or talking), the design-critiquing framework.

Learning analytics (LA) is a relatively new concept in the learning sciences, but it has existed under other names for more than 30 years in smaller scale. LA refers to the analysis of the learning process, in particular the traces of learning that can be captured by tools and adaptive teaching methods. LA data are dependent on observable data collected from any learning episode, including but not limited to educational technology and learning management systems. LA researchers also investigate impacts of learning traces on administrative policy and curriculum reform. Through analytics, institutions (e.g., universities, schools, online education providers) can collect large data sets and apply statistical techniques to predict success or failure, and give advice and suggestions. This may be through informing instructors how specific students are struggling so that they can contact those learners with advice (Baker & Siemens, 2014) or by informing the students directly by presenting automated feedback in the user interface of the educational technology (Fischer et al., 1991). LA departs from purely technical approaches that use data as their sole resource for analysis (educational data mining) and often involve theory to guide the selection of research methods, design of the educational technology, and the interpretation of usage data (Baker & Siemens, 2014).

Literature review

Feedback during the writing process (formative assessment) is an essential component of the teacher's role in English writing classes and is used to improve students' writing skills (Hattie & Timperley, 2007; Black & Wiliam, 2009). Previously, Black and Wiliam (2009) identified the following types of formative assessment in

English: (1) clarifying and sharing learning intentions and criteria for success, (2) engineering effective classroom discussions and learning tasks that elicit evidence of student understanding, (3) providing feedback that moves learners forward, and (4) activating students as instructional resources for one another and as owners of their own learning. Moreover, the effective teacher when giving feedback should address three major issues in a student's learning process (Hattie & Timperley, 2007): (1) Where am I going? (What are the goals?) (2) How am I doing? (What progress is being made toward meeting the goals?) (3) Where to next? (What activities need to be undertaken to make better progress?) Thus, effective feedback is able to bridge the gap between students' prior knowledge and the new knowledge encapsulated in a learning assignment or a learning goal.

Peer assessment is an alternative type of formative assessment often used in educational practice to stimulate student collaboration on texts in progress (Birenbaum, 2003). Previous studies on EFL learning have compared teacher feedback with peer feedback, and researchers have found that teacher feedback is more likely to be incorporated in redrafts than peer feedback (e.g., Yang, Badger, & Yu, 2006). However, to the best of our knowledge, very few studies have compared peer feedback and automated (computer based) feedback.

Research on computer-based feedback includes automated essay scoring (AES). AES systems assign scores to essays written for educational purposes. The score is dynamically computed by machine learning and statistical techniques on large data sets, often based on supervised learning algorithms (Hastie, Tibshirani, & Friedman, 2001). There are disagreements in the literature about the strengths and weaknesses of automated assessment (Foltz, 2014; Kukich, 2000; Lee et al., 2009; Mørch et al., 2005; Sireci & Rizavi, 2000). Proponents have argued that these digital aids successfully compare with the accuracy and reliability of human evaluation (Sireci & Rizavi, 2000). However, critics have pointed out that such systems do not encourage students to pursue novelty in their writing and instead lead to conformity; furthermore, these systems can be fooled by intentional gibberish and thus can give students inaccurate, higher scores (Kukich, 2000).

We aimed to fill a niche in the previous research and designed an experiment to address the following research questions: *How do the two types of feedback (automated vs. peer) assist EFL students during the writing process, and what impact does the feedback have on their final essay grades?*

Design critiquing framework

The design-critiquing framework (DCF) is a theoretical framework that integrates action (doing) and reflection (conversing and thinking) within a computational environment (Fischer et al., 1991; Robbins & Redmiles, 1998; Mørch et al., 2005). It was inspired by Donald Schön's theory of the reflective practitioner (Schön, 1983), which suggests that skilled performers toggle between action and reflection, as in thinking about alternatives and what to do next when they create. Novice designers need scaffolding to integrate action and reflection and this is provided by "back talk" from the environment (e.g., problematic situations, contingencies, nudges, hints, prompts) to trigger the shift from action to reflection (Schön, 1983). Within a computational environment the shift can be accomplished by automated feedback generated from domain-specific rules (Fischer et al., 1991). A goal of "back talk" is to bringing a degree of objectivity to the novices' highly subjective creative process of designing. Full computational support of DCF requires automated analysis of artifacts and deliberations. The current version of EssayCritic (version 3) supports artifact analysis.

The computer's role in DCF in this study is to analyze textual artifacts (essays) and make suggestions to students for improving the text, according to a writing assignment. Hattie and Timperley (2007) argue effective feedback connects students' prior knowledge with the desired knowledge. With DCF, this means using automated feedback to intervene in the action-reflection loop and create a shift from action to reflection, triggering talk (reflection) based on the current version of an essay ($action_n$), which may lead to subsequent revision ($action_{n+1}$). The writing process ends when the assignment has been completed to a sufficient degree or when the allocated time is up. We explored how EFL students used feedback produced by EssayCritic in two ways: (1) in small group discussions triggered by the feedback (surfacing students' prior knowledge) and (2) feedback incorporated as new sentences in their essays (approximating desired knowledge).

The EssayCritic system

The EssayCritic system is a web application for semantic analysis of short texts (< 500 words) (Lee et al., 2009; Mørch et al., 2005). The key feature of EssayCritic is its identification of the presence or absence of subthemes (specific topics/ideas written about) in individual essays. Its novelty is that it decomposes the subthemes into

simpler concepts and this makes concept identification, and subsequently topic/idea identification, easier and more accurate.

In preparing the knowledge base for a new essay topic, the first step is to create a concept tree representing the topic; this is done in collaboration between the teachers and researcher. The teachers suggest a set of desired subthemes for the topic, obtained from the analysis of the content of a textbook chapter, and provide a set of sample essays collected from previous students. The researcher decomposes each subtheme into simpler concepts, which can be precisely represented by a few phrases or keywords using synonyms from dictionaries, WordNet (Fellbaum, 1998), textbooks, and texts from the sample essays acquired from students, newspapers, and web pages.

During the system-training phase, the second step, the researcher manually labels the sentences of the sample essays to identify which concepts are contained in them. These instances act as the training data for the decision tree algorithms (Quinlan, 1986), which output a set of rules, i.e., logic combinations of the derived simple concepts, indicating the relationship between these concepts and the subthemes (Appendix A). In this study, eleven subthemes were identified for the essay assignment (Appendix B). The overall time required to prepare the knowledge base and train the system was approximately four weeks. Most of the time is spent on preparing the knowledge base. Training the system took about two to three days, including fine-tuning of the rules. In terms of supervised machine learning, the system-training phase is very short compared to many other applications.

The students invoke EssayCritic by uploading essays to a server that are examined sentence by sentence. If a sentence is found to satisfy a rule, it is presumed to contain the associated subtheme, and information can be presented as feedback in the user interface (Figure 1). The students can choose between two types of feedback: (1) covered subthemes (Figure 1: left) and (2) suggested subthemes (Figure 1: right).

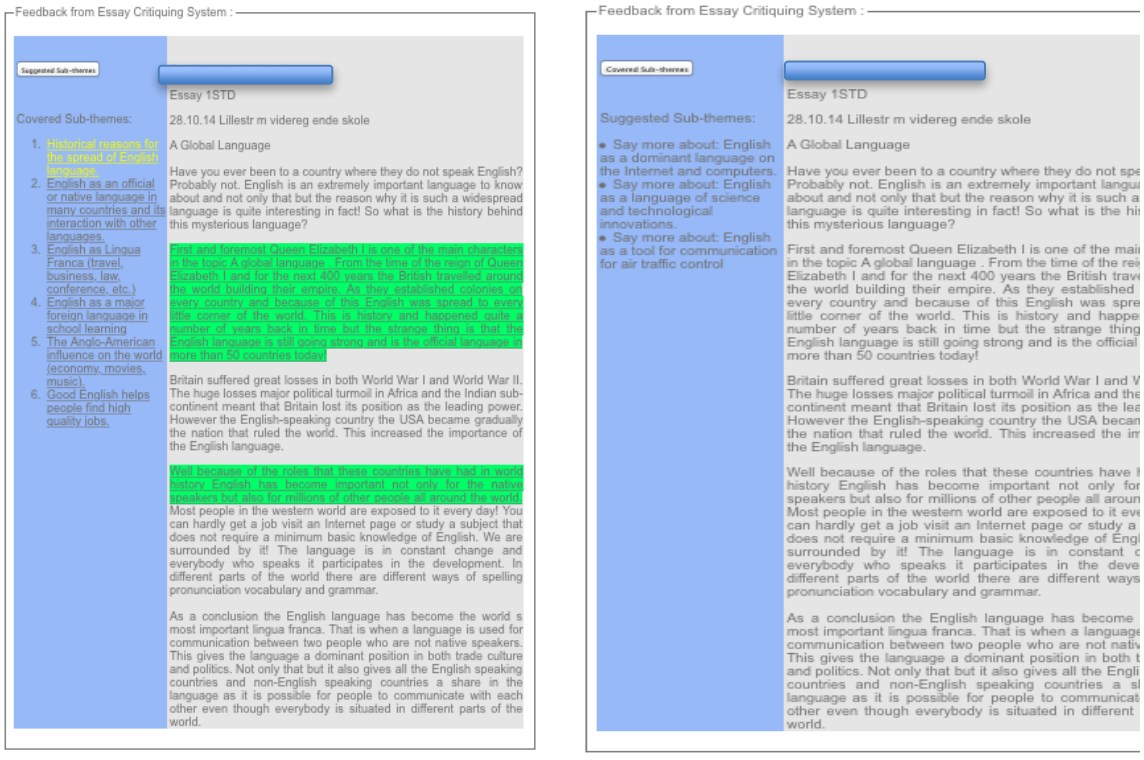

Figure 1. The user interface of EssayCritic: Covered subthemes (left) and absent (and suggested) subthemes (right)

Research design and mixed methods

Five teachers of a Norwegian upper secondary school created the following assignment given to 125 students aged 16-17 from five classes: “Write an essay on the topic of English as a global language. Explore how English was spread around the globe, and present the most important reasons for this development. About 300-400 words.” All students wrote the first draft using a word processor, and two teachers graded the essays by using an

assessment rubric aimed at evaluating the content of students' essays. Eleven best essays from three classes were selected to train the system. The remaining two classes of 24+24 students served as target and comparison classes. Students in the *target class* received feedback from EssayCritic, and the students in the *comparison class* provided feedback to each other. The students in both classes had Norwegian as a native language; they had been studying English for 10 years and followed similar teaching plans prior to the intervention. To make the comparison fair in terms of equal access to learning resources, all students had access to the same list of the eleven EssayCritic subthemes (Appendix B) and the assessment rubric. The students in both classes were asked to focus on content and use the resources they had available during the writing process, but it was only the target class students who received the feedback integrated with their essays in the user interface of EssayCritic (see Figure 1). The students in both classes revised their essays twice and handed version 3 for grading by an independent teacher.

The students of both classes worked in small groups of four. The purpose of these groups for the target class was to collaboratively discuss the meaning and implications of the EssayCritic feedback, and allowed researchers to observe students' deliberations. Figure 2 illustrates the setting in the target class. In the comparison class, the students first read each other's essays and then gave oral feedback. Using this experimental set-up the research team aimed to make a fair comparison of two forms of giving feedback on written work, and to increase our understanding of the role of feedback in bridging students' prior and desired (teacher specified) knowledge.

Figure 2. Students in the target class were grouped in four so that they could verbalize their ideas about what is meant by EssayCritic feedback. They interacted 2&2 and showed each other the feedback they received from the system by turning their laptop around and read it out

To analyze the data, we applied mixed methods, combining quantitative and qualitative data sets (Creswell, 2012). A paired *t*-test was used to analyze the improvement in grades from pretest to posttest and the improvement in the number of subthemes from the pretest to the posttest. Cohen's *d* (Cohen, 1992; Field, 2013) was calculated to evaluate the effect size of these improvements. We used an independent *t*-test (Cohen, Manion, & Morrison, 2011; Field, 2013) to calculate the statistical difference between results of the pre- and post-tests and between the number of subthemes in the pre- and post-tests in both classes.

We also counted the number of times the teacher intervened in the two classes. In the target class the teacher intervened six times and in the comparison class 13 times. Interventions were mainly in response to students' requests for information about the task, the ideas they talked about in groups, and if they had understood the feedback procedures.

We used three cameras to record the activity of three (out of six) groups from each class to collect qualitative data. A fourth camera followed the teacher. In total we have nine hours of video data. Field notes taken during class observations were used to contextualize the video data (Derry et al., 2010). All students communicated in English. We analyzed the conversations that occurred in two groups, one from each class, using a qualitative approach, inspired by the interaction analysis method (Jordan & Henderson, 1995). The groups were selected because the students were verbally active. We selected only two data extracts for inclusion in the analysis below due to space limitations. The transcript notation used to reproduce the verbal protocols is shown in Appendix C.

The combination of quantitative and qualitative methods allowed us to compare achievements at different levels of detail: final essays (grades and number of subthemes) and writing process (what the students were talking about).

Results and analysis

Quantitative data

We compared the results of versions 1 and 3 of the essays from both classes. We considered version 1 as a pre-test and version 3 as a post-test. In total, 96 essays marked by the classroom and independent teacher on a scale of 1-6 constituted our quantitative data. The inter-rater reliability (Cohen's Kappa) between the grades by the classroom teachers and the independent teacher has been computed (Table 1).

Table 1. Reliability for two coders (classroom and independent teacher) of pre-test and post-test for the target and comparison classes (calculator used: http://dfreelon.org/utils/recalfront/recal2/)

Class	Cohen's Kappa, pre-test	Cohen's Kappa, post-test
Target class	0.226	0.229
Comparison class	0.318	0.296

Measured on the scale 0.2-1.0 (Field, 2013) the obtained results of Cohen's Kappa indicate fair inter-rater reliability between the assessments of the classroom teacher and the independent teacher.

The results of the pre- and post-tests (average grades) are presented in Table 2.

Table 2. Average pre-test and post-test grades in the target and comparison classes and the observed groups

Population	Avg. grade pre-test (essay v.1)	Avg. grade post-test (essay v.3)	Difference	Paired *t*-test	Cohen's *d*
Target class ($N = 24$)	2.8 $M = 2.79$; $SD = 0.88$	4.5 $M = 4.49$; $SD = 1.41$	1.7	$t(24) = -7.62$ $p < .0005$	$d = 2.26$
Observed group	2.7	4.5	1.7		
Comparison class ($N = 24$)	2.9 $M = 2.86$; $SD = 0.74$	4.8 $M = 4.83$; $SD = .87$	1.9	t (24) = -8.86 p < .0005	$d = 2.64$
Observed group	3.5	4.5	1		
Independent *t*-test	$p = .725$	$p = .903$			

The paired *t*-test (Cohen et al., 2011) shows a significant difference between the average grades of the pre- and post-tests in both classes, and the Cohen's *d* (Cohen, 1992; Field, 2013) indicates a *large effect* in both classes, as *d*-values larger than 1.0 is considered large according to Plonsky and Oswald (2014). The independent *t*-test (Field, 2013) shows no significant statistical difference between the pre-test grades and no significant statistical difference between the post-test grades ($p = .725$ and $p = .903$).

We used another measure to identify differences in outcome between the two feedback conditions (Table 3).

Table 3. Average number of subthemes in the essays of the target and comparison classes and the observed groups

Population	Avg. # subthemes in pretest (essay v.1)	Avg. # subthemes in posttest (essay v.3)	Difference	Paired *t*-test	Cohen's *d*
Target class ($N = 24$)	3.25 $M = 3.25$; $SD = 0.79$	7 $M = 7$; $SD = 1.89$	3.75	$t(24) = -9.70$ $p < .0005$	$d = 4.72$
Observed group	4	8.25	4.25		
Comparison class ($N = 24$)	4.04 $M = 4.04$; $SD = 0.91$	5.92 $M = 5.92$; $SD = 1.14$	1.88	$t(24) = -6.91$ $p < .0005$	$d = 2.07$
Observed group	4.5	6	1.5		
Independent t-test			$p < .0005$		

Table 3 presents the average number of subthemes in the essays of both conditions. The paired *t*-test indicates a significant difference between the numbers of subthemes from pretest to posttest in both classes. Cohen's *d* reflects a large effect in both classes. In addition, the independent *t*-test shows a significant statistical difference between the increase of subthemes across the two classes ($p < .0005$). In other words, the students in the target class included significantly more subthemes in their final essays than the students in the comparison class.

In summary: The quantitative data showed no significant difference between the final grades in the two classes, but there was a significant difference in the number of subthemes. To analyze the results in more detail we performed a qualitative analysis of the writing process.

Qualitative data

Extract 1: Analysis of the interactions between students in the target class (automated feedback)

The extract below (Extract 1) was taken from a discussion between four students. They have uploaded the first version of their essays to EssayCritic. It is Jane's turn to share the feedback she received with the rest of the group, which included: "Say more about: Good English helps people to find high quality jobs."

1. Jane: OK and then I have to say about how English helps people to find high quality jobs. Well, I think high quality jobs are international, in business, and in companies. Many companies deal outside the country. My dad, he works with furniture, he has to travel to Asia and Europe and he has travelled so many times. And he knows what furniture we are going to sell in Norway, and he has to know English very well because he talks to Chinese people and Japanese people.
2. Carol: Well, I went to chiropractor, and only the secretary was Norwegian. All the doctors there spoke English; they were Australians and British and Americans.
3. Jane: So, if you are going to have a job, you have to have some knowledge…
4. Carol: Yes, basically you don't get a good job if you don't speak English.
5. Jane: Yes, you have to know some basic English; you have to know English words for the things that you work with.
6. Carol: Yes, at least.
7. Jane: Yes, at least.

((Later in the writing process))

8. Teacher: Have you got any questions on the feedback on your draft?
9. Carol: I have covered seven subthemes, and I need two more.
10. Jane: Same here. I covered three when I submitted my first draft, and now I have covered seven.
11. Teacher: Oh, that's very well. *((Jane invokes the covered subtheme function of EssayCritic))*
12. Jane: But I think the structure of my text is bad.
13. Teacher: Take a look at the structure as well.

The extract shows that the brief feedback was meaningful to the learners and triggered memories (lines 1-2). The girls conclude that knowledge of English is very important for working life (lines 3-7). Later in the conversation the teacher offers assistance (line 8), and the two girls report the number of subthemes they have included (lines 9-10). The teacher acknowledges with praise that they have written about most of the subthemes, but Jane is unsure about her essay's organization ("I think the structure of my text is bad," line 12). Her concern is raised after invoking EssayCritic's covered subthemes function (Figure 1: left), which shows that the text for one of the subthemes is scattered throughout her essay. The teacher suggests that Jane works more on the structure but without providing specific guidance.

Jane included the following text in version 2 of her essay, addressing the subtheme "English helps people to find high quality jobs": *"We use English everywhere, and we are surrounded with it. We use English in science and technological innovations and to get a good job."* Carol wrote, *"You can hardly get a high quality job, if you don't speak English well. Sometimes you can get a high quality job, but you have to speak at least a little English, just to know what and whom you are working with."*

In summary: The feedback from EssayCritic to the target group prompted the students to write more idea-rich essays, but the focus on content creation took time away from organizing their essays for readability.

Extract 2: Analysis of the interactions between students in the comparison class (peer feedback)

In Extract 2 another group of four students have received grades on version 1 of their essays from the teacher but without any feedback. Josh and Mike have read each other's essays, and Josh is in the middle of giving feedback to Mike.

1. Mike: We wrote mostly entire essay about history. It's easy to fix it; we have to just make this part bigger and that part smaller *((pointing to text on computer screen))*. This is an introduction; and that is finish.
2. Josh: And there is also a problem: different reasons in one paragraph, because often you should have one reason per paragraph.

3. Mike: I hate this, I had to put two reasons and I hate it. In my next draft I will do something about that. How do you feel about language to write about the reason number three? *((Referring the list of subthemes on paper))*
4. Josh: You should use sentences binders.
5. Mike: Yes, I should, but I didn't use it. It's like every test I have. If we are allowed to use something, I always bring a pile of helping stuff and I never use it. I don't know why.
6. Josh: I really hate giving feedback because it's so difficult.
7. Mike: Yes, it is, even if you are a teacher.
8. Josh: And it's worse with us. She said to focus on content, the flow…
9. Mike: I have looked at yours, and it's very good. There is very little to correct.
10. Josh: I will try to cut it down.
11. Mike: Try to cut it down. It was too long, but this was not a problem. Maybe focus a bit more on other things than history.

((Later in the writing process))

12. Josh: If you want, you can write about the upper class situation with learning English.
13. Mike: It isn't coming naturally here; that's the point. And that's about history, and I don't want more history.
14. Josh: No, you need other things.
15. Mike: Because *((takes the assessment rubric))* I have already five or six sentences about history, I don't think I need more.
16. Josh: It's so difficult not to talk about history because the whole book is about history *((referring to the textbook))*. Have you talked about music?
17. Mike: Music? I said "media," perhaps I should mention "music." I said "media and entertainment."
18. Josh: You can specify, but that's not needed really.
19. Mike: Then there will be producing of media, music, and films and TV series ((*laughing*)) I feel like I want to write "entertainment." It feels like it's easier.
20. Josh: Yes, that's more descriptive.
21. Mike: OK. But I feel if I want a better grade, there is something more I have to change.
22. Josh: But you can see it?
23. Mike: And I don't know what it is. That's kind of annoying.

Josh indicates there are some problems with the organization of Mike's essay (line 2). The two boys are uncertain whether they should divide a paragraph in two (line 3-4). In lines 6-12, Josh expresses his frustration about the difficulty of giving feedback *("I really hate giving feedback because it's so difficult"*) and Mike ends up saying Josh's essay is good with little to correct (line 9). Mike consults the assessment rubric (line 15). He realizes that he has written 5-6 sentences *"about history"* (line 15). Josh mentions *"music"* (line 16), and Mike says that he has written about media and entertainment but is unsure if he also needs to mention music. By elaborating on each other's thoughts, the two peers contribute to the development of a common understanding of the assigned topic of English as a global language, albeit in a somewhat arbitrary way. They are not at all sure if they are moving in the right direction (lines 21-23).

Mike included the following in his final draft: *"… the US has taken the role of a mass producer of media and entertainment in the world. After the Second World War, Europe was in ruins. The European industry and prosperity was dramatically slowed down, while in the US, the economy grew. Making the US the first and the only international superpower. Now, new Hollywood movies are displayed on the big screen all over the world, every month."* Josh wrote in his essay, *"With all new smartphones and streaming possibilities of films and music, we are hearing English more than we used to. With streaming programs like Netflix and Spotify, you can watch Titanic or listen to The Beatles on the bus. The massive information we receive from international news pages also influences us."* Mike chose to combine two subthemes in his essay, media and entertainment in a historical perspective, whereas Josh wrote about the opportunities offered by streaming English language media.

General discussion

We were surprised at first to find that both classes received approximately the same grades on the post-tests. We did aim for a fair distribution of learning resources (all students had access to the 11 subthemes and the assessment rubric), but we anticipated that the students who received automated feedback would improve their grades on the basis of producing more content rich essays. The qualitative differences provide a clue to why the grades did not differ more.

Following Black and Wiliam (2009), we can state that the feedback from EssayCritic enabled the learners in the observed target group to move forward, as the feedback prompted them to "say more about" a missing subtheme and improve their essays by including more content related to the assignment. This effect was not observed to the same extent in the comparison group as we elaborate below.

Hattie and Timperley (2007) suggested that effective feedback should address the discrepancy between prior and desired knowledge. Our findings suggest that the students in the target group were more successful at building bridges. Carol and Jane were able to connect their understanding of English as a global language with personal experiences (e.g., *"My dad, he works with furniture, he has to travel to Asia and Europe…"* and *"I went to chiropractor, and only the secretary was Norwegian, all the doctors there spoke English,"* lines 1-2 in Extract 1). In their attempts to connect their ideas with the assignment, Josh and Mike were only able to think of what the teacher would say to them when giving feedback, i.e., referring to what was expected of them. It seems that the type of learning that occurred in the comparison group is less meaningful in terms of personalization than the type of learning achieved by the target group. However, at this stage, this should be considered a tentative hypothesis; further research is needed to test and compare the two conditions with more data and in more detail.

Two differences are apparent based on analyzing the writing processes of the two observed groups based on Extracts 1 and 2: (1) the students' (lack of) certainty about how to progress, and (2) how the ideas developed in the group discussions were incorporated in the essays.

First, the students in the target group appeared more confident than those in the comparison group when discussing what ideas to include in their essays and when using personal stories to anchor their knowledge. However, the variety of ideas surfacing in these students' discussions appeared not to be entirely beneficial; this is revealed, for example, when Jane discovered that her text was not well organized (*"But I think the structure of my text is bad,"* line 12 in Extract 1). However, the teacher did not address this issue. In a similar vein, the students in the comparison group was uncertain about their knowledge of the assigned topic and if they were progressing, as revealed by Mike in lines 21 and 23 of Extract 2: *"OK. But I feel if I want a better grade, there is something more I have to change." "And I don't know what it is. That's kind of annoying."* However, the two boys showed more confidence when it came to applying knowledge of essay organization. They suggested different strategies, including cutting down text, building up a complex argument by splitting a paragraph, shortening a long passage or expanding another one.

Second, the ideas prompted by EssayCritic in Extract 1 were worked with twice in the observed target group (Carol and Jane): first when discussed in their group (both tell personal stories) to trigger and elaborate common ideas and later when incorporated in their essays as individual writing efforts. In Jane's essay, the ideas discussed in her group were incorporated as one phrase whereas Carol wrote two sentences. The subthemes did not surface explicitly in the discussions of the comparison group (after reading Josh's essay, Mike says in line 9, *"I have looked at yours and it's very good, there is very little to correct"*). The two students struggled to generate feedback for each other in terms of subject matter content, even though they had access to the assessment rubric and the list of subthemes on a sheet of paper. They seemed to take on the role of the teacher when trying to create feedback for each other, possibly imagining what the teacher would have said in that situation. However, they found this process to be out of their league, as revealed by Josh in line 6 of Extract 2: *"I really hate giving feedback because it is so difficult."* The students' lack of competency in this area might have contributed to their uncertainty. Despite this, Josh and Mike received on average the same grade as Carol and Jane on the final essay (see Table 2), and some of the sentences they wrote can be traced back to their discussions.

The comparison group's understanding of how to structure an essay might have been one factor to explain their final grades. Other possible explanations are as follows: (1) the comparison group had a better starting point, i.e. their version 1 essays received higher grades than those of the target group (3.5 vs. 2.7), (2) the teacher intervened twice as often in the comparison class than in the target class (13 times vs. 6), (3) it is possible that the two boys were not able to verbalize their ideas but were able to express them in writing because the ideas did not originate through collaboration, and (4) the covered subthemes feedback from EssayCritic (Figure 1: left) did not provide explicit guidance for essay reorganization for the target group students..

Writing and discussing to learn through the use of EssayCritic is supported by the design-critiquing framework (Fischer et al., 1991; Robbins & Redmiles, 1998; Mørch et al., 2005), which suggests the use of a cyclic process of action (writing) and reflection (conversing and thinking aloud) to explain how feedback is used in practice. When the action-reflection loop is sufficiently tight, the learners may effectively use the feedback generated, as it is fresh in mind and relevant to their task. This might be the factor that allows feedback to serve as a bridge

between prior knowledge (personal experiences in our case) and teacher set goals (desired knowledge defined by the assignment).

A possible weakness of the approach to learning analytics for foreign language learning we have shown in this study can be traced back to the eleven subthemes that were programmed in the EssayCritic learning algorithm (Appendix B). The choice of subthemes might have been constrained by the content of the textbook used by the school used in this study and the pedagogical preferences of the teachers. Further research ought to investigate possibilities for distinguishing between sufficient and insufficient number of subthemes to extensively cover a particular topic.

Summary and conclusions

We have developed an English essay writing aid through several iterations, applying different algorithms for comparing a student essay with good examples represented by a conceptual model. The latest (third) version uses a decision tree supervised learning algorithm, a carefully constructed concept tree of the topic ("English as global language"), and a carefully prepared knowledge base (teachers and researcher). This novel approach has a distinct advantage; even a complex or abstract subtheme can be represented by some easily evaluated logic rules, which was not possible with the algorithms used in the previous versions of EssayCritic. Our design and analysis efforts were informed by a theoretical framework (DCF) that takes into account individual (writing) and collaborative (discussion) activities around shared artifacts (essays), iterative design (versioning), and feedback. We compared two conditions of feedback provision (computer-generated and peer assessment), involving five classes at an upper secondary school in Norway (three classes to provide a knowledge base and two classes for direct involvement in the experiment). We used a mixed methods approach to gather and analyze the data: to compare quantitatively the outcomes according to grades and subthemes and another for "zooming" into the qualitative differences in the students' writing process with examples. The quantitative data shows that there was no significant difference between the final grades in the two classes, but there was a significant difference in the number of subthemes. The qualitative data shows that the target group put more effort into writing essays rich in content, i.e., including many ideas, some of which were triggered by feedback from EssayCritic and the students' personal experiences. The comparison group found it difficult to provide feedback on content and compensated by focusing on essay organization of fewer ideas.

Further research that would build on our findings could include, but is not limited to, the following:

- Automating feedback that would assist students in organizing their essays
- Computational support for analyzing deliberations (conversing) in addition to artifacts (essays) for a full support of the DCF. This could be achieved via a discussion forum and/or by capturing conversations
- Exploring whether the students who used the feedback from EssayCritic are able to transfer the processes (rather than content) to other writing assignments (such as feedback giving skills)
- Comparing the effects of discussion groups with individual work when using EssayCritic to identify other processes and mechanisms students use when writing to learn in groups other than collaboration
- Reanalyze the data material for multiple triggers of deliberations (external and internal), including deliberations triggered by EssayCritic, peers and self-triggered deliberations
- Exploring the tentative hypothesis that new knowledge is connected to prior knowledge that can be externalized (personal stories; verbalized experience). What methods can help to reveal the "gap closing" activity of prior and desired knowledge, and what role should the computer play in offloading (e.g., capturing traces of past activity for triggering memory, tests to determine proficiency according to levels, and so forth)?
- Study how the ideas developed by students change over time in response to feedback and revision, and how the ideas travel back and forth between written work and small group conversations.

Acknowledgements

The research design and data collection were carried out in the Ark&App project funded by the Ministry of Education of Norway. We thank Øystein Gilje, Sten Ludvigsen, and Anne Edwards for their constructive comments on an earlier version of this article.

References

Baker, R. S. J. D., & Siemens, G. (2014). Educational data mining and learning analytics. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (2nd ed., pp. 253-272). New York, NY: Cambridge University Press.

Birenbaum, M. (2003). New insights into learning and teaching and their implications for assessment. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimising new modes of assessment: In search of qualities and standards* (pp. 13-36). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31. doi:10.1007/s11092-008-9068-5

Cohen, J. (1992). A Power primer. *Psychological bulletin*, *112*(1), 155-159. doi:10.1037/0033-2909.112.1.155

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education*. Milton Park, UK: Routledge.

Creswell, J. W. (2012). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks, CA: Sage publications.

Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., Hall, R., Koschmann, T., Lemke, J. L., Sherin, M. G., & Sherin, B. L. (2010). Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. *The Journal of the Learning Sciences*, *19*(1), 3-53. doi:10.1080/10508400903452884

Fellbaum, C. (1998). WordNet: Wiley online library. In *The Encyclopedia of Applied Linguistics.* doi:10.1002/9781405198431.wbeal1285

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Thousand Oaks, CA: Sage publications.

Fischer, G., Lemke, A. C., Mastaglio, T., & Morch, A. I. (1991). The Role of critiquing in cooperative problem solving. *Transactions on Information Systems, 9*(2), 123-151.

Foltz, P. W. (2014, May). *Improving student writing through automated formative assessment: Practices and results*. Paper presented at International Association for Educational Assessment (IAEA) Conference, Singapore.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of statistical learning: Data mining, inference and prediction*. New York: NY: Springer.

Hattie, J., & Timperley, H. (2007). The Power of feedback. *Review of Educational Research*, *77*(1), 81-112.

Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. *Journal of the Learning Sciences*, *4*(1), 39-103.

Kukich, K. (2000). Beyond automated essay scoring. *IEEE Intelligent Systems*, *15*(5), 22-27.

Lee, C., Wong, K. C. K., Cheung, W. K., & Lee, F. S. L. (2009). Web-based essay critiquing system and EFL students' writing: A Quantitative and qualitative investigation. *Computer Assisted Language Learning*, *22*(1), 57-72.

Mørch, A. I., Cheung, W., Wong, K., Liu, J., Lee, C., Lam, M., & Tang, J. (2005). Grounding collaborative knowledge building in semantics-based critiquing. In R. H. Lau, Q. Li, R. Cheung, & W. Liu (Eds.), *Proceedings 4th international conference on advances in web-based learning* (pp. 244-255). Berlin Heidelberg, Germany: Springer-Verlag.

Plonsky, L., & Oswald, F. L. (2014). How big is "big?" Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878-912.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, *1*(1), 81-106.

Robbins, J. E., & Redmiles, D. F. (1998). Software architecture critics in the Argo design environment. *Knowledge-Based Systems, 11*(1), 47-60.

Schön, D. A. (1983). *The Reflective practitioner: How professionals think in action*. New York, NY: Basic Books.

Sireci, S. G., & Rizavi, S. (2000). *Comparing computerized and human scoring of students' essays* (Technical report). Amherst, MA: Laboratory of Psychometric and Evaluative Research, University of Massachusetts Amherst.

Weizman, Z. O., & Snow, C. (2001). Lexical input as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology, 37*(2), 265-279.

Yang, M., Badger, R., & Yu, Z. (2006). A Comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179-200.

Appendix A

Illustration of a subtheme, the decomposed simpler concepts, and the rule learned

Example subtheme: **English as a dominant language on the Internet and computers**. Table 4 shows the simpler concepts and the corresponding words or phrases (created manually by teachers and researchers), and Table 5 is the rule learnt by computer for identifying the subtheme in student essays.

Table 4. The simpler concepts of a subtheme and the corresponding words or phrases

Simpler concepts	Words or phrases
English	English
Communication	communicate; communication; talk; talked; discuss; speak; spoken;
IT areas	computer; programming; web; internet; www; information; cyberspace; website; webpage; YouTube; Netflix
Pronoun*	it
Example*	example; instance; e.g.,

Table 5. The logic rule learned for the example subtheme

Subtheme	Logic rule composed of simpler concepts
English as a dominant language on the Internet and computers	(“English” AND “IT areas”) OR (“Pronoun” AND “Communication” AND “IT areas”) OR (“Example” AND “IT areas”)

Appendix B

List of subthemes of English as a global language

1. Historical reasons for the spread of English language
2. Impact of English on other languages
3. English as Lingua Franca (travelling, international policies and diplomatic negotiations)
4. English as a dominant language on the Internet and computers
5. English as a language of science and technological innovations
6. English grammar and borrowed words
7. Advantages of English in University studies abroad
8. The US influence on the world (economy, film industry)
9. Knowledge of English helps in finding high quality jobs
10. English as an International standard for communication for pilots and air traffic controllers
11. English vs. other official UN languages

Appendix C

Transcript notation

[ ]	Text in square brackets represents clarifying information
=	Indicates the break and subsequent continuation of a single utterance
?	Rising intonation
:	Indicates prolongation of a sound
(.)	Short pause in the speech
[…]	Utterances removed from the original dialog
-	Single dash in the middle of a word denotes that the speaker interrupts herself
--	Double dash at the end of an utterance indicates that the speaker’s utterance is incomplete
((Italics))	Annotation of non-verbal activity