

ARTICLE

# The impact of automated feedback on L2 learners' written causal explanations

Aysel Saricaoglu

TED University, Turkey (aysel.saricaoglu@tedu.edu.tr)

## Abstract

Even though current technologies allow for automated feedback, evaluating content and generating discourse-specific feedback is still a challenge for automated systems, which explains the gap in research investigating the effect of such feedback. This study explores the impact of automated formative feedback on the improvement of English as a second language (ESL) learners' written causal explanations within two cause-and-effect essays and across pre- and post-tests. Pre- and post-test drafts, feedback reports for first and revised drafts from the automated writing evaluation system, and screen-capturing videos collected from 31 students enrolled in two sections of an advanced-low-level academic writing class were analyzed through descriptive statistics and the Wilcoxon signed-rank test. Findings revealed statistically significant changes in learners' causal explanations within one cause-and-effect essay while no significant improvement was observed across pre- and post-tests. The findings of this study offer not only insights into how to further improve automated discourse-specific feedback but also pedagogical implications for better learning outcomes.

**Keywords:** automated writing evaluation; causal explanations; cause-and-effect essays; discourse-specific feedback

## 1. Introduction

Causal explanations not only dominate the written language of science (Wellington & Osborne, 2001), but they also go beyond science and “are part of academic literacy generally” (Mohan & Slater, 2004: 255–256). Despite their centrality in academic success, causal explanations have not received enough attention from writing researchers. The existing studies are mostly descriptive, examining causal language development over time. There is a huge gap in research on how students develop causal language with formative feedback in instructional settings.

Given the challenging amount of work that evaluating learners' causal explanations and generating discourse-specific feedback manually would result in, such provision of detailed feedback may be through the use of automated writing evaluation (AWE) programs. “AWE programs ... are designed to foster learner autonomy by performing error diagnosis of learner input, generating individualized feedback, and offering self-access resources such as dictionaries, thesauri, editing tools, and student portfolios” (Chen & Cheng, 2008: 97). Since the entrance of AWE programs into the field of second language (L2) writing, several studies have demonstrated positive findings regarding their effectiveness in improving learners' grammatical and mechanical correctness (e.g. Chodorow, Gamon & Tetreault, 2010; Grimes & Warschauer, 2010; Rock, 2007). However, there is a lack of studies on their effectiveness to improve learners' discourse due to the inability of AWE tools to evaluate meaning. Addressing this gap, this study explores the

---

Cite this article: Saricaoglu, A. (2019). The impact of automated feedback on L2 learners' written causal explanations. *ReCALL* 31(2): 189–203, doi: 10.1017/S095834401800006X

effectiveness of an automated causal discourse evaluation tool (ACDET), which was specifically developed for the purpose of improving learners' written causal explanations.

## 2. Automated feedback on written causal explanations

### 2.1 Causal explanations

As causal relations have a crucial role in general knowledge (Lakoff & Johnson, 1999), writers' ability to express causal relations in writing is also fundamental. This ability requires the knowledge of causal language features, which are the linguistic structures that express causal meaning and causal relationship between events (Chukharev-Hudilainen & Saricaoglu, 2016). In writing, causal explanations are made either implicitly or explicitly (Stefanowitsch, 2001). The focus of this study is explicit causal expressions: the linguistic forms that imply causal relationship/meaning. For example, the sentence "Last month the vet gave us the bad news: There was a tumor the size of a golf ball near her heart, which caused her to die within a month" (Stefanowitsch, 2001: 25) includes the causal verb *cause*, which is an explicit causal language form.

In this study, causal explanations are evaluated based on six categories that were identified through corpus analysis in an earlier study during the development of ACDET (Saricaoglu, 2015). According to this categorization, causal relations can be expressed using (a) causal conjunctions such as *for*, *if*, or *so that*; (b) causal adverbs such as *fatally* or *in response*; (c) causal prepositions (including prepositional phrases) such as *through* or *as a consequence of*; (d) causal verbs (including phrasal verbs) such as *freeze* or *result from*; (e) causal adjectives such as *beneficial*; and (f) causal nouns such as *influence*. Sophisticated causal writing displays more of nouns rather than conjunctions or adverbs as nouns represent higher development according to the developmental path of cause (Halliday & Martin, 1993; Mohan & Beckett, 2003). Therefore, learners would benefit from feedback that would help them move along the developmental path of cause; that is, from conjunctions to nouns.

### 2.2 Developmental path of causal language

To learn how to write about causal explanations in a more sophisticated way, English as a second language (ESL) learners need feedback on their writing: "If teachers are consistently and reflectively assessing student explanations, focusing on aspects that students are having trouble with, they can provide successful assessment-learning cycles for teaching the forms and meanings of causal explanations" (Slater & Mohan, 2010: 267).

According to Slater and Mohan (2010), the developmental path of cause can guide the formative assessment of causal explanations. Throughout their causal language development learners demonstrate a shift from conjunctions to verbs and nouns, which characterizes the causal developmental path (Halliday & Martin, 1993). At the early stages of causal language development, learners use conjunctions to express causal relations (e.g. *because*); later on, they also choose verbs (e.g. *cause*); and finally, they add nominalizations to their causal repertoires (e.g. *the cause*) (Mohan & Beckett, 2003). In their evaluations of learners' explanations from early childhood to late adolescence, Christie and Derewianka (2008) found out that the students whose causal explanations were more congruent were between the ages of seven and 12 while the students whose causal explanations were less congruent or incongruent were between the ages of 15 and 17. Halliday (1994) refers to this development as a shift from more congruent to less congruent (also referred to as grammatically metaphorical) expression of meaning. "Man clean car" exemplifies a child's congruent language (Halliday, 2003: 19). "Man" is the doer (subject) and he does the cleaning. The action of cleaning is expressed with a verb that follows the subject. This pattern is a congruent pattern, and congruent patterns are characteristic of children's early language (Halliday, 2003).

In causal explanations, conjunctions represent the congruent expressions of causal relations, prepositions represent less congruent expressions, and verbs and nouns incongruent expressions (Mohan & Beckett, 2003). For example, the sentence "My plane was late so I had to run across

the terminal” would be the most congruent way of explaining the situation. The development of causal language reflects the transition from congruent expressions to less congruent or incongruent (or more grammatically metaphorical) expressions: “The late plane was the cause of my running.” Slater and Mohan (2010) claim that the path from *so* to *the cause* can inform the formative assessment of causal explanations. However, performing this assessment manually by identifying causal language forms in student drafts and giving formative feedback based on the causal developmental path would be a very time-consuming task. Providing automated formative feedback instead may help writing instructors overcome issues of practicality.

### 2.3 Computer-assisted language learning and automated writing evaluation

AWE tools are built based on artificial intelligence, natural language processing, and statistical techniques that enable them to accomplish evaluations of written texts in a much shorter time than manual evaluations (Grimes & Warschauer, 2010). Since their entrance into computer-assisted language learning, AWE tools have been investigated for their accuracy in detecting language errors and scoring essays. Liu and Kunnan (2016) found out that WriteToLearn was not a reliable tool because it could not identify students’ errors in certain categories as articles, prepositions, word choice, and expression. According to Lavolette, Polio and Kahng (2015), Criterion also failed to identify many errors of students, and out of all the errors it identified, 75% was correctly identified. Despite such limitations, several other studies have found AWE feedback to be helpful for writing improvement, especially in grammar and mechanics (e.g. Chodorow *et al.*, 2010; Grimes & Warschauer, 2010; Lai, 2010; Rock, 2007). Li, Link and Hegelheimer (2015) showed that Criterion feedback significantly reduced the error rates from the first to the final drafts for three papers written by lower level students and for four papers written by higher level students. Liao (2016) also showed that using Criterion improved learners’ accuracy in fragments, subject–verb disagreement, run-on sentences, and ill-formed verbs when both revising texts and constructing new texts. Ma and Slater (2016), who connected Criterion scores to ESL learners’ use of causal language, revealed that Criterion scores reflected students’ developmental levels of causal language.

The feedback that current AWE tools generate for the macro-level aspects of language such as organization, content, and development is more generic than the feedback for micro-level aspects including punctuation, spelling, mechanics, grammar, and usage. For example, Criterion evaluates learner writing in terms of grammar, usage, mechanics, style, organization, and development (Burstein, Chodorow & Leacock, 2003). However, its generic feedback on discourse elements as “Is this part of the essay your thesis? The purpose of a thesis is to organize, predict, control, and define your essay. Look in the *Writer’s Handbook* for ways to improve your thesis. (Criterion feedback)” (Hegelheimer & Lee, 2013: 293) does not address the content of the essay discourse elements.

Using AWE tools only for micro-level textual aspects (e.g. grammar, mechanics, usage, etc.) is “against the very social and interactive nature of writing” (Hegelheimer & Lee, 2013: 293). Writing “takes place within a context, that accomplishes a particular purpose, and that is appropriately shaped for its intended audience” (Hamp-Lyons & Kroll, 1997: 8). Therefore, the capabilities of AWE systems, as Cotos (2012: 88) emphasizes, need to be expanded to the “contextual richness and functional meanings of the discourse.” ACDET, which is a recently developed automated causal discourse evaluation tool, analyzes a wide range of causal language forms and provides formative feedback on causal explanations, and therefore has been chosen as the AWE tool for this study.

## 3. The current study

### 3.1 Theoretical framework

This study is informed by the interaction hypothesis (IH). IH hypothesizes that language learners need to interact with others and to receive feedback from them for learning to happen as a result (Long, 1983). Interaction enhances L2 learning by providing learners with access to linguistic

input, drawing their attention to linguistic form, giving feedback on their language, and creating opportunities for output and interactional modifications (Gass, 1997; Long, 1996; Pica, 1994). Taking into account that today people interact not only with other people but also with computers, Chapelle (2003) expanded the definition of interaction to “the activity between person and computer” (p. 56), which may also enhance language development (Chapelle, 1998). She suggests that learners’ output can be marked to draw learners’ attention to the errors in their output. Such feedback can draw learners’ attention to language form, give them a chance to notice the gap between their forms and the target forms to correct their errors, and ultimately enhance language learning in instructional settings (Robinson, Mackey, Gass & Schmidt, 2012). “[V]isually enhancing a particular structure in the input” (Robinson *et al.*, 2012: 249) exemplifies attention-drawing feedback.

This study aims to draw learners’ attention to causal explanations through automated feedback and to help them notice what needs to be improved during their interactions with ACDET. In this study, interaction is defined as the activity between learners and ACDET. In this activity, learners revised their cause-and-effect essays to improve their causal explanations through ACDET feedback.

### 3.2 Research questions

This study aims to answer the following research questions:

1. To what extent does automated formative feedback provided by ACDET lead to improvement of ESL learners’ written causal explanations within essays?
2. To what extent does automated formative feedback provided by ACDET lead to improvement of ESL learners’ written causal explanations across pre- and post-tests?

In this study, improvement in written causal explanations within essays is defined as a shift in learners’ causal explanations from congruent expressions of causal meaning (i.e. causal conjunctions) to less congruent expressions (i.e. causal verbs and nouns), the latter being grammatical metaphor. Improvement in students’ written causal explanations across pre- and post-tests is defined as a decrease in the number of more congruent expressions of causal meaning and an increase in the number of less congruent expressions.

## 4. Methodology

This study employs a pre-experimental pre-test/post-test design. In the pre-test, students wrote a cause-and-effect essay (henceforth pre-test drafts). In the treatment, they wrote two different cause-and-effect essays (henceforth Essay 1 and Essay 2 drafts), received automated feedback on their causal language, and revised their drafts based on ACDET feedback. Their revisions, as captured by screen-capturing recordings in Essay 1 and Essay 2 as well as the automatically generated feedback reports for first and revised drafts of both essays, were analyzed for causal language improvement within essays. In the post-test, students wrote a cause-and-effect essay (henceforth post-test drafts). Their causal language in the pre-test drafts was compared with their causal language in their post-test drafts for improvement across pre- and post-tests.

### 4.1 Context and participants

Participants of this study were 32 first-year undergraduate ESL learners (11 female, 21 male) from two sections of an advanced-low-level academic writing class at a Midwestern university in the USA. The two sections were taught by the same writing instructor following the same classroom procedures. The instructor was a fifth-year female PhD student in the applied linguistics and technology program at the same university. She had prior experience of teaching academic writing and using AWE tools in classroom settings.

**Table 1.** Summary of data collected

Implementation	Data sets	<i>n</i>
Pre-	Pre-test essay drafts	31
While-	ACDET feedback reports on the 1st essay drafts	25
	ACDET feedback reports on the 1st essay revised drafts	25
	Screen-capturing recordings of students' interactions with ACDET	25
	ACDET feedback reports on the 2nd essay drafts	27
	ACDET feedback reports on the 2nd essay revised drafts	27
Post-	Screen-capturing recordings of students' interactions with ACDET	22
	Post-test essay drafts	31

Note. The number in each data set is different due to absent students or the technical problems students had in recording their screens.

Participants were placed into the academic writing class based on their essay writing scores from the university's English Placement Test. They were native speakers of different languages: Chinese (66%,  $n = 21$ ), Malay (13%,  $n = 4$ ), Spanish (9%,  $n = 3$ ), Hindi (3%,  $n = 1$ ), Korean (3%,  $n = 1$ ), Portuguese (3%,  $n = 1$ ), and Thai (3%,  $n = 1$ ). Their ages ranged from 18 to 25, and they were from a variety of majors with the majority being from business, civil engineering, electrical engineering, food science, mechanical engineering, and nutritional science.

The academic writing class was an undergraduate-level English class for non-native speakers of English. In this class, students were required to write five essays: an expository essay, a classification essay, a comparison and contrast essay, and two cause-and-effect essays. For each essay, students first received textbook instructions. Then they followed a process approach in which they wrote their first drafts, received peer and/or teacher feedback, and revised their drafts based on the feedback received.

## 4.2 Materials and instruments

This study was conducted in three stages as pre-implementation, while-implementation, and post-implementation. Students' cause-and-effect essay drafts from both pre- and post-test, ACDET feedback reports on students' first and revised drafts of two cause-and-effect essays, and screen-capturing recordings of students' interactions with ACDET for revising their written explanations were collected (see Table 1 for a summary of data collected).

Given that written tasks are differentiated by the communicative goal such as comparison and contrast, narration, argumentation, or cause and effect (Ruiz-Funes, 2015), the same written task (cause and effect) was used in all data collection stages with different topics to ensure that the levels of complexity were not different across tasks (Yasuda, 2011).

### 4.2.1 Pre-tests

In pre-implementation stage, students were asked to write a cause-and-effect essay around 350–500 words, and their essays were collected as pre-test drafts to compare their written explanations in the pre-test with those in the post-test. The pre-test was administered in class, and students were given 40 minutes following the standard time limit for the writing exams in the program. The topic for the pre-test was as follows: "Write an essay about the causes and effects of poverty (not having enough money to pay for one's needs) for a family or a city or a country." The same topic was used in both sections, and 31 pre-test drafts were collected in total.

Prior to the pre-test, students had completed the expository, classification, and comparison and contrast essays. Given that this study assessed only causal language development as a shift

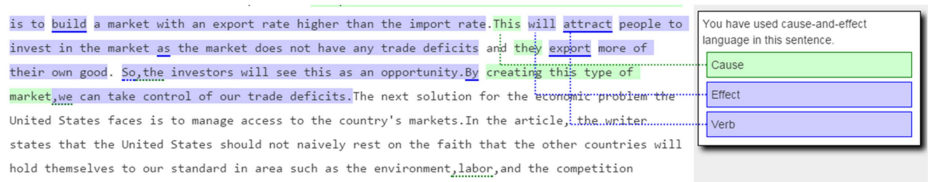


Figure 1. Sentence-level feedback by ACDET.

from more congruent to less congruent expressions of causal meaning, students' background in essay writing was assumed not to have any impact on the pre- and post-test findings. The cause-and-effect chapter that students completed after the pre-test and before the use of ACDET, as a requirement of the course, included exercises that addressed only a few cause-and-effect items (e.g. nouns: cause, effect, factor, reason, result; verbs: cause, lead to, result in), which helped learners gain the grammatical knowledge they needed to understand ACDET feedback (e.g. causal noun, causal verb, causal conjunction). All the cause-and-effect exercises done were based on isolated items, and no instruction was given on causal language development, thus creating no impact on the pre- and post-test findings.

#### 4.2.2 ACDET feedback reports

Students used ACDET in two cause-and-effect essays. They wrote their first drafts on ACDET outside class without receiving feedback, and they received automated feedback on their essays in class. They used ACDET for 50 minutes for each essay in class and revised their written explanations based on automated feedback. In both essays, students were asked to write an essay of about 700 words on one of the following three topics (different in each essay): the effects of globalization on a country, region, or city; the reasons why a country has a strong, weak, or a variable economy; the effects of a specific event (like an earthquake or flood) that brought about positive, negative, or mixed economic results in a country, region, or city (like modernization, or new industry, a political change, a treaty agreement, a war, or other action).

ACDET evaluates written causal explanations and provides sentence-level feedback highlighting causal meaning and form in the text, and formative text-level feedback aiming to help learners improve their causal explanations. ACDET was developed using a hybrid natural language processing combining a statistical approach (automatic tagging of sentences and words by the Stanford CoreNLP) with a rule-based approach (hand-coded linguistic rules written in Prolog based on part-of-speech tags and type dependencies). ACDET was developed based on expert writing and was tested and improved based on learner writing (for the details of ACDET development and its analysis of cause-and-effect sentences, see Chukharev-Hudilainen & Saricaoglu, 2016).

ACDET highlights the cause in green and effect in blue, and underlines the explicit causal language form (see Figure 1) to draw learners' attention to causal meaning and form. When students click on one highlighted sentence, they receive sentence-level feedback presented in a box in the left margin. The comment in the box includes elaboration of the color-coded feedback.

ACDET also provides learners with text-level feedback consisting of two parts (see Figure 2). In the first part, the causal language features are summarized in a table with two columns as "the casual language features that you have used" and "the casual language features that you have not used." The analysis presents the word counts of causal language features in each category.

The second part of the text-level feedback gives learners suggestions for improvement and provides examples for revision. The feedback is offered based on the frequency of repetition, and the goal is to help learners move from conjunctions to prepositions, from prepositions to verbs, or from adjectives to nouns in their causal language instead of repeatedly using the same causal forms from the same category within the causal developmental path.

Causal analysis of your text

The causal language features that you have used						The causal language features that you have not used
Causal Conjunction(s)	Causal Adverb(s)	Other Causal Construction(s)	Causal Prepositional Phrase(s)	Causal Adjective(s)	Causal Verb(s)	Causal Noun(s)
if = 5 times because = 2 times	therefore = 2 times thus	why	with = 3 times	bad	influence = 6 times create = 2 times transmit = 2 times make change produce	

Suggestions for improving your causal language

You have used the causal conjunction: "if" repeatedly. Please consider converting your repeated causal conjunctions to causal prepositional phrases. For example:  
Emerging economies may be hit harder by a spike **since** they use a lot of oil per unit of output.  
Emerging economies may be hit harder by a spike **because** of the amount of oil they use per unit of output.

You have used the causal prepositional phrase: "with" repeatedly. Please consider converting your repeated causal prepositions to causal verbs. For example:  
The recent recession happened **due** to the financial crisis.  
The financial crisis **led** to the recent recession.

You have used the causal verb: "influence" repeatedly. Please consider converting your repeated causal verbs to causal adjectives. For example:  
The sound of running water **distracted** them.  
The sound of running water was **distracting** for them.

Figure 2. ACDET’s text-level feedback.

The performance of ACDET to identify causes and effects in 585 sentences written by 17 undergraduate ESL students was measured based on precision (i.e. the ratio of correctly identified causal language features to the total number of identified features), recall (i.e. the ratio of causal language features identified to the total number of causal language features), accuracy (i.e. the percentage of correctly identified causal language features), and *F*-score (i.e. the harmonic mean of precision and recall). Its precision was found to be .93 (extracted 93% of the causal language features correctly) and recall 71% (captured 71% of the features that human annotators manually captured), which resulted in accuracy of .76, an *F*-score of .81 (for the details of ACDET’s performance, see Chukharev-Hudilainen & Saricaoglu, 2016).

To make a comparison between learners’ drafts before and after using ACDET in terms of causal explanations, the text-level feedback reports generated by ACDET for both drafts were collected in two cause-and-effect essays. The total number of feedback reports collected was 104: 50 reports of 25 students (25 for first drafts and 25 for revised drafts) in Essay 1 and 54 reports of 27 students (27 for first drafts and 27 for revised drafts) in Essay 2.

4.2.3 Screen-capturing recordings

Screen recordings of participants’ use of ACDET in class were also collected to analyze their modifications of causal explanations during their interactions with the tool. Learners’ modifications of causal explanations were recorded for convenience purposes as analyzing their revisions during their use of the tool was more cost-effective than comparing each sentence across drafts. Two screen-capturing programs were used: Quick Time Player and Camtasia. In total, 47 screen-capturing videos (25 from Essay 1 and 22 from Essay 2) were collected. The length of recordings ranged from five minutes to 48 minutes.

4.2.4 Post-tests

In post-implementation stage, students wrote another cause-and-effect essay of around 350–500 words, and their essays were collected as post-test drafts. The post-test was given in class as the final exam of the course, and students were given 40 minutes to complete the post-test. The topic for the post-test was: “What can cause close friends to become enemies and what are the consequences?” In total, 32 post-test drafts were collected, but the draft of the student who was absent in the pre-test was excluded from the data analysis.

### 4.3 Procedure

This study was exempted by the institutional review board of the university. At the beginning of the study, students were informed about the study, and even though formal documentation was not required, their consent was obtained through consent forms. Data collection started with the pre-test in the first week of the first cause-and-effect essay. In the second and third weeks, the cause-and-effect chapter of the textbook was completed, and students wrote their first drafts out of class without any time limits. In the fourth week, ACDET was introduced and explained to the students. They were shown a demo in which the teacher typed in some sentences with causal language forms on ACDET, explained the sentence-level feedback and the text-level feedback in detail. Students were asked to type in their essays into ACDET in the first class of the week. ACDET's feedback was not activated in this class as the class time would not allow for revisions and the goal was to record their screens. Students were informed that they would receive causal-discourse feedback in the second class and would revise their essays. Students received the automated feedback two days later in class and made revisions, for which their screens were recorded. They made their revisions in class (i.e. 50 minutes) during which the teacher had the role of a monitor making sure students did not have any technical issues and answering students' questions on how to use ACDET. The same process was followed in Essay 2 except for the textbook activities and the ACDET demo. Text-level feedback reports for both first and revised drafts in both essays were also collected. In week seven, the second essay was completed and in week eight, the post-test was administered.

### 4.4 Data analysis

For causal language improvement within essays, first data from ACDET's text-level feedback reports were analysed for the first and revised drafts in both Essay 1 and 2 ( $n=104$ ). Raw frequency counts of causal features in each category were obtained to identify the changes in the explanations. For accurate comparison between drafts of different length, raw frequencies were normalized per 1,000 words. Descriptive statistics were calculated for group findings. Normalized frequencies were further tested for normality using Shapiro–Wilk test. Because data were not normally distributed, the mean scores of each causal feature category on first and revised drafts were compared using the Wilcoxon signed-rank test. A statistically significant decrease in the mean scores of more congruent expressions and a statistically significant increase in the mean scores of less congruent expressions would indicate improvement in causal explanations within essays.

Second, improvement in causal explanations within essays was evaluated analyzing the data from the screen-capturing recordings ( $n=47$ ). Recordings were coded and analyzed in NVivo. The causal language modifications learners made were coded in terms of congruence. When a learner revised the causal form in a sentence by changing it to another causal form, this modification was coded in one of the three codes as less congruence, same congruence, or more congruence. If a congruent causal expression (e.g. So, economy is something magical; the global economy will always find a way out to keep it in balance) was changed to a less congruent expression (e.g. Thanks to magical economy; the global economy will always find a way out to keep it in balance), this modification was coded as less congruence. If a causal expression (e.g. These could be hard tasks and challenges) was changed to a more congruent expression (e.g. These could be hard tasks and will challenge them), this modification was coded as more congruence. If the congruence was the same before modification (e.g. which may and will cause conflicts to the bonds of families) and after modification (e.g. which may and will generate conflicts to the bonds of families), this modification was coded as same congruence. Descriptive statistics were calculated for all types of modifications. Only modifications with less congruence would indicate improvement in causal explanations.



**Table 2.** Frequencies of causal language features in first and revised drafts

Causal language features	Essay 1 (N = 25)				Essay 2 (N = 27)			
	FD reports		RD reports		FD reports		RD reports	
	Total	%	Total	%	Total	%	Total	%
Conjunctions	181	17.3	142	13.6	162	14.7	149	14.1
Adverbs	35	3.3	48	4.6	45	4.1	46	4.3
Prepositions	62	5.9	70	6.7	60	5.5	60	5.7
Verbs	592	56.6	598	57.3	670	61.0	641	60.5
Adjectives	30	2.9	42	4.0	28	2.5	35	3.3
Nouns	146	14.0	144	13.8	134	12.2	128	12.1
Total	1046	100	1044	100	1099	100	1059	100

Note. FD reports = first draft AWE feedback reports; RD reports = revised draft AWE feedback reports.

For improvement in causal explanations across pre- and post-tests, pre-test drafts ( $n = 31$ ) and post-test drafts ( $n = 31$ ) were analyzed. Causal explanations in the drafts were coded for causal language features and frequency counts of features in each category were conducted and normalized per 1,000 words. For determining the reliability of the coding, over 20% of the total 62 drafts (nine from the pre-test and nine from the post-test) were coded by a second trained coder. Inter-coder reliability, calculated by Cohen's kappa, was good ( $k = 0.73$  for the pre-test and  $k = 0.71$  for the post-test). Descriptive statistics were calculated for the whole group and the mean scores of each causal feature category were compared through the Wilcoxon signed-rank test for statistical significance. A statistically significant decrease in the mean scores of more congruent expressions and a statistically significant increase in the mean scores of less congruent expressions would indicate improvement in causal explanations across pre- and post-tests.

## 5. Results and discussion

### 5.1 Improvement in causal explanations within essays

The effect of ACDET feedback on improving learners' causal explanations within essays was first investigated by analyzing causal language features in learners' first draft AWE feedback reports and revised draft AWE feedback reports in both Essay 1 and Essay 2. Table 2 summarizes frequency counts of causal language features in six categories in total numbers and percentages.

For both essays, frequency counts of causal language features display changes from first drafts to revised drafts after participants used ACDET. The changes are also observed in the descriptive statistics results presented in Table 3.

The Wilcoxon signed-rank test results of Essay 1 revealed a statistically significant decrease in the means of conjunctions ( $Z = -2.58$ ,  $p = .01$ ) from the first to the revised drafts and an increase in the means of adverbs ( $Z = -3.11$ ,  $p = .00$ ) and adjectives ( $Z = -2.43$ ,  $p = .02$ ). In Essay 2, the mean of causal verbs significantly decreased from the first draft to the revised draft ( $Z = -1.96$ ,  $p = .05$ ).

The present study found that automated feedback from ACDET led to a statistically significant reduction in the number of causal conjunctions and an increase in the number of adverbs and adjectives in learners' first cause-and-effect essays. Given that conjunctions and adverbs are more congruent causal expressions and adjectives are less congruent than causal

**Table 3.** Causal language features in first and revised drafts

Causal language features	Essay 1 (N=25)						Essay 2 (N=27)					
	FD reports		RD reports		Wilcoxon signed-rank test		FD reports		RD reports		Wilcoxon signed-rank test	
	M	SD	M	SD	z	p	M	SD	M	SD	z	p
Conjunctions	7.24	4.93	5.70	3.37	-2.58	.01**	6.01	4.37	5.51	4.02	-.38	.70
Adverbs	1.39	2.15	2.81	2.58	-3.11	.00**	1.68	1.96	1.69	1.74	-.78	.44
Prepositions	2.49	2.72	3.08	3.34	-1.32	.19	2.23	1.93	2.23	1.78	-.28	.78
Adjectives	1.21	1.56	1.67	2.00	-2.43	.02*	1.04	1.61	1.29	1.61	-1.84	.07
Verbs	23.67	10.80	23.91	8.92	-.26	.80	24.96	10.36	23.74	9.85	-1.96	.05*
Nouns	5.83	5.41	5.74	4.27	-.04	.97	4.96	3.67	4.72	3.66	-.32	.75

Note. FD reports = first draft AWE feedback reports; RD reports = revised draft AWE feedback reports.

conjunctions and adverbs, the decrease in the number of causal conjunctions and adverbs and the increase in the number of adjectives are positive. ACDET's feedback was able to create changes in learners' causal explanations within essays in line with the causal developmental path. Immediate changes in particular structures in learners' output as a result of feedback being associated with learning by interaction researchers (Gass & Mackey, 2015), learners' modifications in causal explanations in this study could indicate the potential of ACDET to improve learners' written causal explanations. The modifications might be the consequence of learners' noticing causal form, although this claim cannot be substantiated due to the lack of qualitative data explaining students' decisions of causal modifications.

In Essay 2, there were no statistically significant changes in causal language features from the first draft to the revised draft that would indicate improvement even though students made several causal modifications. These findings might point to a possible novelty effect: the innovative look of the AWE tool may have excited learners (Phakiti, 2014).

Improvement in learners' causal explanations within essays was also evaluated for congruence by analyzing data from the screen-capturing recordings of students' causal modifications. Table 4 presents the descriptive statistics of learners' modifications with less, same, and more congruence.

In both essays, students made more causal modifications that did not change the congruence than the modifications with less congruence or more congruence. These findings demonstrate the capacity of ACDET for interactions with learners. However, in the majority of these modifications, students mostly substituted certain causal words or phrases with others from the same category. For example, they changed *effect* to *consequence* or *because* to *since*. This suggests that learners might have focused more on the numerical feedback than the revision suggestions offered by ACDET. Such tendency was also observed by Cotos (2012) in her study on the impact of automated feedback on the rhetorical quality of learners' research article drafts. Cotos (2012: 103) concluded that the impact was "likely to be negative when learners relied only on numerical feedback" because such focus decreases cognitive involvement. As emphasized in IH, the interaction between the learner and the computer program is supposed to facilitate the individual cognitive processes of the learner for learning to occur, which is less likely when the learner's focus is on numerical feedback.

**Table 4.** Descriptive statistics for causal modifications in Essay 1 and 2

Causal modifications	Essay 1 (N = 25)		Essay 2 (N = 22)	
	M	SD	M	SD
Less congruence	0.78	1.17	0.50	0.80
Same congruence	3.91	3.03	3.86	2.62
More congruence	0.17	0.39	0.14	0.35

**Table 5.** Frequencies of causal language features in pre- and post-tests (N = 31)

Causal language features	Pre-test		Post-test	
	Total	%	Total	%
Conjunctions	372	20	330	19
Adverbs	57	3	63	4
Prepositions	110	6	99	6
Verbs	824	44	613	36
Adjectives	151	8	212	12
Nouns	366	19	392	23
Total	1880	100	1709	100

### 5.2 Improvement in causal explanations across pre- and post-tests

Learners' causal language improvement pre- and post-tests was investigated by analyzing their pre- and post-test drafts. Frequencies of causal language features in pre- and post-test drafts were counted for each student. Means and standard deviations of each causal feature category were calculated for group findings and compared across pre- and post-test drafts. Table 5 presents frequencies of causal language features in pre- and post-tests.

Table 6 provides descriptive statistics of causal language features in pre- and post-tests. Even though the findings demonstrate a slight decrease in causal conjunctions and verbs from pre-test to post-test and a slight increase in adverbs, adjectives, and nouns, the Wilcoxon signed-rank test results presented in Table 6 yielded statistical significance for only causal verbs ( $Z = -2.70$ ,  $p = .01$ ).

Although the results illustrate ACDET's potential for helping learners revise their causal explanations, no development was observed across pre- and post-tests. The number of causal verbs decreased significantly from the pre-test to the post-test, but there was no significant increase in the number of causal nouns. The lack of transfer of immediate outcomes to long-term learning benefits is in line with other studies on automated feedback, which show no or very limited long-term learning gains compared to immediate learning outcomes (e.g. Li, Feng & Saricaoglu, 2017). Referring back to the important role of cognitive processes involved in language learning (Long, 1996), the fact that the effect of ACDET feedback did not extend from immediate improvement in learners' written products to learning in the long run is not surprising. However, evidence from qualitative data is needed for further insights into why there are no long-term effects.

**Table 6.** Causal language features in pre- and post-tests ( $N = 31$ )

Causal language features	Pre-test		Post-test		Wilcoxon signed-rank test	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>z</i>	<i>p</i>
Conjunctions	12.01	7.80	10.66	5.12	-.73	.47
Adverbs	1.84	2.33	2.02	2.88	-.36	.72
Prepositions	3.55	3.65	3.20	3.08	-.56	.58
Adjectives	4.88	4.87	6.85	4.94	-1.53	.14
Verbs	26.58	9.82	19.78	8.51	-2.70	.01*
Nouns	11.82	7.10	12.64	7.14	-.27	.79

## 6. Conclusion

This study was an attempt to address the need for automated formative assessment of learners' written causal explanations. As Slater and Mohan (2010) point out, learners need feedback on their causal writing, for which AWE tools can be helpful. There is, however, not enough evidence in the AWE literature showing that automated feedback leads to improved discourse as opposed to the evidence showing its positive effect on improving accuracy. Addressing the gaps in both areas, this study investigated to what extent automated formative feedback provided by ACDET led to improvement of ESL learners' written causal explanations (a) within essays and (b) across pre- and post-tests. Opposed to the common practice of using AWE for error correction in the early stages of a writing process, "freeing the teacher up to concentrate on higher-level meaning-oriented, genre-oriented and audience-oriented aspects of writing" (Stevenson, 2016: 12), this study attempted to use AWE in a later stage in writing to address genre-oriented aspects of writing.

The results of this study revealed limited improvement within essays and no improvement across pre- and post-tests. Although revision at the desired level was not observed, the fact that students revised is alone pleasing given the findings of previous AWE studies that showed lack of student revision. The important question is how ACDET feedback can stimulate students to revise for less congruence, whose answer might lie in the way ACDET is pedagogically used. As Stevenson (2016: 13) argues, "it is not sufficient to ... consider the effects of AWE systems on students' writing without considering their contexts of use." Given that how AWE is integrated into classroom instruction affects students' perceptions and reaction to AWE feedback, it is important to pay equal attention to classroom instruction as to the capabilities of AWE alone. In this study, AWE feedback on explanations was used alone without further teacher feedback, which is not students' preference as was found by Chen and Cheng (2008). Chapelle and Voss (2016: 121) add that "feedback from the system along with human guidance and feedback based on a sound pedagogical foundation shows the most promise to support the assessment of and for learning." In their study, *Li et al.* (2015) found that students' positive perceptions of the AWE tool were closely related to the manner the instructors used the tool. The lack of data from student views of their experiences with ACDET is a major limitation in this study, thus it is difficult to explain the lack of desired revision in relation to the context of use. Further research could look into the ways different teachers integrate ACDET in classroom instruction and how different integrations affect students' use and perceptions of ACDET and reactions to ACDET feedback.

The results of this study support the fact that it is difficult for learners to modify their causal explanations using grammatical metaphor. "[The teacher] suggests moving to a less congruent causal statement, but it is too difficult for [the student]" wrote Mohan and Beckett (2003: 428) based on their observations during teacher–student interactions of grammatical scaffolding of learners' causal explanations. Considering the amount of time it takes children to move from

congruence to incongruence, the difficulty that ESL students have in educational settings is understandable. When compared to learners who complete the causal developmental path in around seven years in natural language acquisition settings (Christie & Derewianka, 2008), it is very normal for ESL students to have difficulty learning grammatical metaphor of causal explanations in classroom settings. Students need more time, more feedback, and repeated practice; the more consistent integration of AWE in classroom instruction leads to more revisions by the students (Li *et al.*, 2015). Students' AWE access and use in this study was restricted to two different revisions in class, each limited to 50 minutes. As expressed by Steinhart (2001), students should have unlimited access to AWE after being trained on how to use it.

Although it is clear that ACDET feedback stimulated causal revision at the same level of congruence, it is unclear why it did not lead to revision towards less congruence. Did students not revise because they did not understand the text-level feedback or because they did not know how to revise, as it was found in a study by Steinhart (2001)? Follow-up data from stimulated recalls would be reflective of learners' responses to the feedback. This could also provide insights into the weaknesses and strengths of the AWE tool and would be informative for the further refinement of the tool.

The results of this study have implications for automated discourse-specific formative feedback. As the feedback that students receive has to be correct, the accuracy of automated writing evaluation systems has been among the primary concerns of researchers: How well does the automated system identify the target errors or features (e.g. Lavolette *et al.*, 2014)? Although performance of ACDET (precision .93, recall .71, and *F*-score .81) is considered to be good for automated systems, it should be noted that .71 recall means there were causal language features in students' essays that ACDET was not able to identify and generate feedback on. If students received more feedback, this would, as a result, lead to more interactions with the tool and more revisions to their writing. Increasing ACDET's recall up to .90 or higher would be ideal for obtaining better learning outcomes with more written explanations identified.

ACDET's text-level feedback was indirect in that it provided learners with numerical feedback on their causal language features and general suggestions for improvement. The fact that students responded to the feedback in the form of causal modifications that actually led to limited improvement implies that numerical feedback draws more attention than improvement feedback. On the contrary, automated feedback needs to be presented to learners in a way that will stimulate more cognitive involvement and focus on the meaning and form at the same time.

Another explanation for the limited improvement might be related to the proficiency level of the participants. Because learners in this study were in the same level of academic writing class and were considered to have the same level of language proficiency, differences in causal language improvement according to proficiency levels were not investigated. Further research is needed to find out how ACDET feedback works with students from different proficiency levels. Learners, in particular with lower level language proficiency, might have had difficulty in responding to the feedback in the way it was desired (Ferris & Hedgcock, 2005). Making discourse-level revisions focusing on the causal meaning requires more cognitive involvement and is probably more difficult for learners than making grammatical or mechanical revisions. Referring to the studies that have demonstrated the effectiveness of direct feedback on writing, especially in the long term (e.g. Bitchener & Knoch, 2010), ESL learners might also benefit more from direct feedback on causal explanations than indirect feedback as of ACDET's.

Although this study has yielded important information regarding automated formative assessment of causal explanations, more studies on the effectiveness of ACDET are needed. Whether direct text-level feedback rather than indirect text-level feedback would increase ACDET's potential in helping learners improve their written causal explanations could be studied in future research. The study of students' use of ACDET in two essays, in a total period of eight weeks during which they used ACDET twice, was helpful in gaining some understanding of ACDET's potential for improving causal explanations. However, longer studies in which students

have more exposure to ACDET might yield more findings on whether or not ACDET feedback has a lasting effect.

**Acknowledgments.** I would like to thank Dr Evgeny Chukharev-Hudilainen, Dr Ruslan Suvorov, and the anonymous reviewers for their constructive feedback on the earlier drafts of this manuscript.

**Ethical statement.** I declare that I have no conflict of interest. This work is entirely original and it has cited others' work appropriately. Institutional requirements were followed in data collection procedures, and informed consent was obtained from the participants.


## References

- Bitchener, J. & Knoch, U. (2010) Raising the linguistic accuracy level of advanced L2 writers with written corrective feedback. *Journal of Second Language Writing* 19, 207–217. <https://doi.org/10.1016/j.jslw.2010.10.002>
- Burstein, J., Chodorow, M. & Leacock, C. (2003) Criterion<sup>SM</sup>: Online essay evaluation: An application for automated evaluation of student essays. *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*. Acapulco, Mexico, 3–10.
- Chapelle, C. A. (1998) Multimedia CALL: Lessons to be learned from research on instructed SLA. *Language Learning & Technology*, 2(1): 21–39.
- (2003) *English language learning and technology: Lectures on applied linguistics in the age of information and communication technology*. Philadelphia: John Benjamins. <https://doi.org/10.1075/lllt.7>
- Chapelle, C. A. & Voss, E. (2016) 20 years of technology and language assessment in language learning & technology. *Language Learning & Technology*, 20(2): 116–128. <http://www.lltjournal.org/item/2950>
- Chen, C-F E & Cheng, W-YE (2008) Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology*, 12(2): 94–112. <http://www.lltjournal.org/item/2631>
- Chodorow, M., Gamon, M. & Tetreault, J. (2010) The utility of article and preposition error correction systems for English language learners: Feedback and assessment. *Language Testing*, 27(3): 419–436. <https://doi.org/10.1177/0265532210364391>
- Christie, F. & Derewianka, B. (2008) *School discourse: Learning to write across the years of schooling*. New York: Continuum.
- Chukharev-Hudilainen, E. & Saricaoglu, A. (2016) Causal discourse analyzer: Improving automated feedback on academic ESL writing. *Computer Assisted Language Learning*, 29(3): 494–516. <https://doi.org/10.1080/09588221.2014.991795>
- Cotos, E. (2012) Towards effective integration and positive impact of automated writing evaluation in L2 writing. In: Kessler, G., Oskoz, A. & Elola, I. (eds.), *Technology across writing contexts and tasks (CALICO Monograph Series Vol. 10)*. San Marcos: CALICO, 81–112.
- Ferris, D. R. & Hedgcock, J. S. (2005) *Teaching ESL composition: Purpose, process, and practice* (2nd ed.). Mahwah: Lawrence Erlbaum.
- Gass, S. M. (1997) *Input, interaction, and the second language learner* Mahwah, NJ: Lawrence Erlbaum.
- Gass, S. M. & Mackey, A. (2015) Input, interaction, and output in second language acquisition. In: VanPatten, B. & Williams, J. (eds.), *Theories in second language acquisition: An introduction* (2nd ed.). New York: Routledge, 180–206.
- Grimes, D. & Warschauer, M. (2010) Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Language, and Assessment*, 8(6): 1–43.
- Halliday, M. A. K. (1994) *An introduction to functional grammar* (2nd ed.). London: Hodder Arnold.
- (2003) *On language and linguistics*. New York: Continuum.
- Halliday, M. A. K. & Martin, J. R. (1993) *Writing science: Literacy and discursive power*. London: Falmer Press.
- Hamp-Lyons, L. & Kroll, B. (1997) *TOEFL 2000 - writing: Composition, community, and assessment* (ETS Research Report No. RM-96-05). Princeton: Educational Testing Service.
- Hegelheimer, V. & Lee, J. (2013) The role of technology in teaching and researching writing. In: Thomas, M., Reinders, H. & Warschauer, M. (eds.), *Contemporary computer-assisted language learning*. New York: Bloomsbury, 287–302.
- Lai, Y-H (2010) Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology*, 41(3): 432–454. <https://doi.org/10.1111/j.1467-8535.2009.00959.x>
- Lakoff, G. & Johnson, M. (1999) *Philosophy in the flesh: The embodied mind and its challenge to western thought*. New York: Basic Books.
- Lavolette, E., Polio, C. & Kahng, J. (2015) The accuracy of computer-assisted feedback and students' responses to it. *Language Learning & Technology*, 19(2): 50–68. <http://www.lltjournal.org/item/2903>
- Li, J., Link, S. & Hegelheimer, V. (2015) Rethinking the role of automated writing evaluation (AWE) feedback in ESL writing instruction. *Journal of Second Language Writing*, 27, 1–18. <https://doi.org/10.1016/j.jslw.2014.10.004>
- Li, Z., Feng, H-H & Saricaoglu, A. (2017) The short-term and long-term effects of AWE feedback on ESL students' development of grammatical accuracy. *CALICO Journal*, 34(3): 355–375. <https://journals.equinoxpub.com/index.php/CALICO/article/viewArticle/26382>

- Liao, H-C (2016) Using automated writing evaluation to reduce grammar errors in writing. *ELT Journal*, 70(3): 308–319. <https://doi.org/10.1093/elt/ccv058>
- Liu, S. & Kunnan, A. J. (2016) Investigating the application of automated writing evaluation to Chinese undergraduate English majors: A case study of WriteToLearn. *CALICO Journal*, 33(1): 71–91. <https://doi.org/10.1558/cj.v33i1.26380>
- Long, M. H. (1983) Native speaker/non-native speaker conversation and the negotiation of comprehensible input. *Applied Linguistics*, 4, 126–141.
- (1996) The role of the linguistic environment in second language acquisition. In: Ritchie, W. & Bhatia, T. (eds.), *Handbook of second language acquisition*. San Diego: Academic Press, 413–468. <https://doi.org/10.1016/B978-012589042-7/50015-3>
- Ma, H. & Slater, T. (2016) Connecting Criterion scores and classroom grading contexts: A systemic functional linguistic model for teaching and assessing causal language. *CALICO Journal*, 33(1): 1–18. <https://doi.org/10.1558/cj.v33i1.26562>
- Mohan, B. & Beckett, G. H. (2003) A functional approach to research on content-based language learning: Recasts in causal explanations. *The Modern Language Journal*, 87(3): 421–432. <https://doi.org/10.1111/1540-4781.00199>
- Mohan, B. & Slater, T. (2004) The evaluation of causal discourse and language as a resource for meaning. In: Foley, J. A. (ed.), *Language, education and discourse: Functional approaches*. New York: Continuum, 255–269.
- Phakiti, A. (2014) *Experimental research methods in language learning*. New York: Bloomsbury Academic.
- Pica, T. (1994) Research on negotiation: What does it reveal about second language learning conditions, processes, and outcomes? *Language Learning*, 44(3): 493–527.
- Robinson, P., Mackey, A., Gass, S. M. & Schmidt, R. (2012) Attention and awareness in second language acquisition. In: Gass, S. M. & Mackey, A. (eds.), *The Routledge handbook of second language acquisition*. New York: Routledge, 247–267.
- Rock, J. (2007) *The impact of short-term use of Criterion on writing skills in 9th grade* (Research Report RR-07-07). Princeton: Educational Testing Service.
- Ruiz-Funes, M. (2015) Exploring the potential of second/foreign language writing for language learning: The effects of task factors and learner variables. *Journal of Second Language Writing*, 28, 1–19. <https://doi.org/10.1016/j.jslw.2015.02.001>
- Saricaoglu, A. (2015) *A systemic functional perspective on automated writing evaluation: Formative feedback on causal discourse*. Iowa State University, unpublished PhD.
- Slater, T. & Mohan, B. (2010) Towards systematic and sustained formative assessment of causal explanations in oral interactions. In: Paran, A. & Sercu, L. (eds.), *Testing the untestable in language education*. Bristol: Multilingual Matters, 256–269. <https://doi.org/10.21832/9781847692672-015>
- Stefanowitsch, A. (2001) *Constructing causation: A construction grammar approach to analytic causatives*. Rice University, unpublished PhD.
- Steinhart, D. (2001) *An intelligent tutoring system for improving student writing through the use of latent semantic analysis*. University of Colorado, unpublished PhD.
- Stevenson, M. (2016) A critical interpretative synthesis: The integration of automated writing evaluation into classroom writing instruction. *Computers and Composition*, 42, 1–16. <https://doi.org/10.1016/j.compcom.2016.05.001>
- Wellington, J. & Osborne, J. (2001) *Language and literacy in science education*. Buckingham: Open University Press.
- Yasuda, S. (2011) Genre-based tasks in foreign language writing: Developing writers' genre awareness, linguistic knowledge, and writing competence. *Journal of Second Language Writing*, 20(2): 111–133. <https://doi.org/10.1016/j.jslw.2011.03.001>

### About the author

Aysel Saricaoglu (PhD, Applied Linguistics and Technology, Iowa State University) is an assistant professor in English Language Education, TED University, Ankara, Turkey. She investigates academic writing, automated formative assessment, corpus linguistics, telecollaborative learning, and project-based learning. Her work has appeared in journals such as *CALL* and *CALICO*.

Author ORCID.  Aysel Saricaoglu, <http://orcid.org/0000-0002-5315-018X>