Writing Instruction Tips For Automated Essay Graders: How To Design an Essay for a Non-human Reader

Writing Instruction Tips For Automated Essay Graders: How To Design an Essay for a Non-human Reader

Robo Grader - When Artificial Intelligence (AI) Becomes the Evaluator of Writing

ALISE LAMOREAUX

Contents

Introduction

ALISE LAMOREAUX

Photo by Alise Lamoreaux

I've been involved with writing as a teacher and an author much of my life. I've written for academia and personal interest. I've had my writing evaluated for entrance into programs; resumes I've written documents that have been scrutinized for the workplace; I've had stories evaluated by editors for publication, but I have never had my writing evaluated by artificial intelligence (AI). In my training as a teacher, I wasn't taught to teach writing for an audience that wasn't human. I wasn't taught to look at writing as it would be seen by artificial intelligence. Yet, that is what the students in my classroom will face at the end of their coursework. It will not be up to me to evaluate their writing. They will face a standardized test and an automated essay grader. The purpose of this book is not to debate the use of Automated Essay Graders (AEG) or the pros and cons of AEG. I am creating this book in 2020 because AEG is a fact of life for the students I teach.

Automated essay graders, or Robo-graders as they are sometimes called, are cheaper and faster than human readers, and testing is a rapidly growing industry. Automated essay graders are programmed to "read" for certain types of words that signal the content and structure of the essay. AEG is looking for a specific type of organization and is limited in the types of essays

it can effectively score. Using automated essay graders puts an emphasis on argumentative and informational essays, styles that are evidence-based. Building from that concept, I began researching the type and organizational structure that would best suit an automated essay grader. The difficulty in trying to discover the "best practices" for helping students prepare to face Robo-grading is that much of the information regarding how the systems are designed is proprietary. It is the intellectual property of the testing industry and not something to be shared. The testing industry is privatizing the educational preparation for the tests they administer.

As I began digging deeper into the topic of AEG, the topic took on a new meaning to me. At first it was a kind of sadness about teaching what seems like a joyless writing format. From my perspective, writing has always been a rhetorical art, the transfer of information, feelings, and opinions from one mind to another. Writing has been about communication and interaction and teaching writing has always been fun for me, but with automated essay graders as the final evaluators of my students' skill level, that does not seem to be the priority, because it is not something artificial intelligence can assess. My students' futures may depend on the score they receive on their standardized test. The score may impact college placement or workplace job offers. I began to see the issue of AEG as one of social justice and something I need to better understand, so I can help students understand the nature of the "audience" they will be facing when their final writing topic will not be assessed by a human reader.

In the process of writing this book, I have felt like a detective. The information students need to be successful is not easy to find, unless you pay the company providing the test fees for their private information. And even then, the information provided is not transparent. As a result, I have gone behind the scenes looking at research from the artificial intelligence programming side of the house as well as literature regarding linguistics as it relates to artificial intelligence. AEG is based on comparing artificial intelligence scoring essays to humans scoring the same essays. It has been a challenge to try to find out where the sample essays came from and the diversity of the essay writers is in question. In addition, the background of the human graders is proprietary and not disclosed.

Based on the research I have been able to find, my goal for this book is to create an understanding of what AEG can assess and provide tips for the best practices and skills to develop when facing AEG systems. There are many arguments regarding teaching to a test, and that Robo-grading is

harming writing instruction, but regardless of those opinions, students are being evaluated on the basis of artificial intelligence and their transition to college or the workplace is being impacted. The testing industry is the clear winner in the standardized testing movement. Rather than making software recognize "good" writing, they will redefine "good" writing according to what the software can recognize. Considering the resources being put into perfecting Robo-grading, it's likely that we will see rapid expansion in the use of artificial intelligence as an evaluation tool. It's important to give students a chance to learn to "think" like a Robo-grader.

Learning Objectives

ALISE LAMOREAUX

Upon completion of reading this book, the reader will be able to:

1. Recognize the historical development of automated essay grading industry.
2. Describe important technological changes that happened and what results occurred over time.
3. Understand the components of "reading" when the reader is an automated essay grader.
4. Describe the common assumptions about "good" writing upon which the programming for an automated essay reader can be based.
5. Identify 3 models of argumentation and the audience expectation of each model.
6. Identify the 7 components of the Toulmin Method of organization for an argument.
7. Reiterate the argumentation model best suited for a Robo-Grader and why.
8. Evaluate an argument based on the components of the Toulmin Method of Argumentation.
9. Understand the role of word choice and the impact it has on "reading" for automated essay graders, including the notion of "academic vocabulary".
10. Appraise how an essay unfolds naturally using signposts, discourse markers, transitional phrases, and other components of sentence structure, to manage a sequence of events.

1. Robo-Grader: Artificial Intelligence As An Automated Essay Grading System, The Backstory

ALISE LAMOREAUX

The idea of Automated Essay Graders (AEG or Robo-graders) has been around since the early 1960's. A former English teacher, Ellis B. Page, began working on the idea of helping students improve their writing by getting quick feedback on their essays with the help of computers. In December of 1964, at the University of Connecticut, Project Essay Grade (PEG®) was born (Page, 1967). At that time, 272 trial essays were written by students grades 8-12 in an "American High School" and each was judged by at least 4 independent teachers. A hypothesis was generated surrounding the variables, also referred to as features, that might influence the teachers' judgement. The essays were manually entered into an IBM 7040 computer by clerical staff using keypunch cards. The process was time consuming and labor intensive due to the limitations of computers at that time, but the results were impressive.

Page believed that writing could be broken down into what he called a "trin" and a "prox". The Trin was a variable that measured the intrinsic interest to the human judge, for example, word choice. The Trin was not directly measurable by the computer strategies of the 1960's. The Prox was an approximation or correlation to the Trin, for example, the proportion of "uncommon words" used by a student (Page, 1967). Thirty variables were identified as criterion for Project Essay Grade (PEG®). Page found that "the overall accuracy of this beginning strategy was startling. The proxes achieved a multiple-correlation of .71 for the first set of essays analyzed, and by chance, achieved the identical coefficient for the second set." (Page, 1967) While the results were impressive, the technology of the time was too cumbersome for practical applications, and computers were not readily accessible to most people. Page's ideas may have seemed outlandish at the

time, but it could be argued that they were prophetic. His work with AEG came years before students would have computers to write essays with.

Page continued to work on PEG for the next 30 years and his research consistently showed high correlations between Automated Essay Graders (AEG) and human graders. One study, (Page, 1994) analyzed 2 sets of essays: one group of 495 essays in 1989, and another group of 599 in 1990. The students involved in the analysis were high school seniors participating in the National Assessment of Educational Progress who were responding to a question about recreational opportunities and whether a city should spend money fixing up old railroad tracks or convert an old warehouse to a new use. Using 20 variables, PEG reached 87% accuracy compared with targeted human judges.

In May of 2005, Ellis B. Page passed away at the age of 81. Two years earlier, he sold Project Essay Grade (PEG®) to a company called Measurement Incorporated. PEG® is currently being used by the State of Utah as the sole essay grader on the state summative writing assessment. According to Measurement Incorporated's website (www.measurementinc.com) 3 more States are considering adapting the program. PEG® is currently being used in 1,000 schools and 3,000 public libraries as a formative assessment tool. Ellis B. Page could be considered the forefather of Automated Essay Graders.

What changed since Ellis B. Page began his Project Essay Grade in 1964? Personal computers and the Internet! The onset of personal computers in the 1990's changed the face of possibility for Automated Essay Graders. With electronic keyboards in the hands of students and the Internet to provide a universal platform to submit text for evaluation, (Shermis, Mzumara, Olson, & Harrington, 2001), a new industry, testing, was born.

In 1997, *Intelligent Essay Assessor® (IEA®)* was introduced as another type of automated essay grading system developed by Thomas Landauer and Peter Foltz. In 1989, the system was originally patented for indexing documents for information retrieval. The indexing programming was subsequently applied to automated essay grading. Intellectual property rights became a factor in the marketplace of automated essay grading. The *Intelligent Essay Assessor®* program was designed to use what's known as Latent Semantic Analysis (LAS), which determines similarity of words and passages by analyzing bodies of text. Developers using LAS create code that estimates how close the vocabulary of the essay writer is to the targeted vocabulary set (Landauer, Foltz, & Laham, 1998). Like most automated essay

grading systems, documents are indexed for information retrieval regarding features, such as proportion of errors in grammar, proportion of word usage errors, proportion of style components, number of discourse elements, average length of sentences, similarity in vocabulary to top scoring essays, average word length, and total number of words. Typically, these features are clustered into sets. The sets may include content, word variety, grammar, text complexity, and sentence variety. In addition to measuring observable components in writing, the IEA® system uses an approach that involves specification of vocabulary. Word variety refers to word complexity and word uniqueness. Text complexity is similar to determining the reading level of the text. As with Project Essay Grader®, IEA® has reported high correlations with human scored essays (Landauer, Foltz, & Laham 1998). IEA® has become the automated grading system used by Pearson VUE. In 2011, Pearson VUE and the American Council on Education (ACE) partnered and launched GED® Testing Services (GEDTS) which provides students with a high school equivalency (HSE) program.

Around the same time period as IEA® was being developed, Educational Testing Services (ETS®), was developing the Electronic Essay Rater knows as *e-rater*®. This system uses a "Hybrid Feature Identification Technique" (Burstein et al, 1998) that includes syntactic structure analysis, rhetorical structure analysis, and topical analysis to score essay responses via automated essay reading. The *e-rater*® system is used to score the GRE® General Test for admission to Graduate, Business, and Law school programs. ETS also provides testing for HiSET®, and TOEFL®. The e-rater® measurement system counts discount words (words that help text flow by showing time, cause and effect, contrast, qualifications etc.), the number of complement, subordinate, infinite, and relative clauses, as well as the occurrence of modal verbs (would, could, etc.) to calculate ratios of syntactic features per sentence and per essay. The structural analysis uses 60 different variables/features similar to the proxes used in Project Essay Grader® to create the essay score (Ruder & Gagne, 2001).

The e-rater® was the initial AEG used by the GMAT® (Graduate Management Assessment Test) when the test added an essay component to the testing format in 1999. In January 2006, ACT, Inc. became responsible for development and scoring of the written portion of the GMAT®test. At that point, ACT, Inc. partnered with Vantage Learning and a new automatic essay grading system was introduced, IntelliMetric™, for use with the Analytic Writing Assessment. Vantage Learning's corporate policy treats

IntelliMetric™ as an intellectual property asset. Many of the details regarding this automated essay grader remain trade secrets (Rudner, Garcia, & Welch, 2005). However, the general concepts behind the AEG system used in IntelliMetric™ have been described by Shermis and Burstein in their book, *Handbook of Automated Essay Evaluation* (2013). According to their research, the IntelliMetric™ model selects from 500 component features (proxes) and clusters them into 5 sets: content, word variety, grammar, text complexity, and sentence variety.

One thing is true across all the major automated essay grading systems: due to the proprietary nature of the artificial intelligence surrounding the exact algorithms used to create these automated essay grading systems, the exact weighting of the system's features, or exactly how the clusters and what features are in them are created, cannot be known. It's important for test examinees to find out which automated essay grading system is being used by the company administering the test to be taken because that is the "audience" for the essay that is to be graded. Essays have traditionally been thought of as school-related assignments, something to use for college admission or a scholarship application, but the nature of the workplace is changing and automated essay graders are also used to determine the writing skills of future employees. Automated essay graders are impacting more than just academics.

It's important to remember that AEGs can't read for understanding when evaluating text. That is beyond the capabilities of artificial intelligence currently. For example, an automated essay reader could not "understand" the following joke:

> Did you hear about the Mathematician who is afraid of negative numbers?
>
> He'll stop at nothing to avoid them.

Or the following play on words:

> No matter how much you push the envelope, it will still be stationary.

Artificial intelligence (AI) cannot make inferences or judge cleverness of word choice. Artificial intelligence would not understand that I feel like I have been chasing squirrels, herding cats, and falling down rabbit holes in the process tracking down the information used in this book.

Artificial intelligence cannot understand polysemy and so does not

understand whether the word *mine* is being used as a pronoun, or an explosive device, if it is referring to a large hole in the ground from which ore is produced, or part of the name of the 2009 Kentucky Derby winner, Mine That Bird. It can count how many times the word shows up in a text. Understanding what automated essay graders can "read", and how they "read" is important for helping test examinees learn to think like their audience and write for that audience. But if the details behind the "thought process" of automated essay graders is proprietary, what can be found out about how an AEG thinks? Research can be found that provides general details about the major AEG systems currently in use, and like a puzzle, things become clearer as more pieces are added to the picture.

In early 2012, The William and Flora Hewlett Foundation sponsored a competition geared towards data scientists and machine learning specialists called the Automated Student Assessment Prize (ASAP). The goal of this competition was to "...help solve an important social issue. We need fast, effective and affordable solutions for automated grading of student written essays" (www.kaggle.com). The competition had 2,500 entries and 250 participants who made up 150 teams. The competitors were provided with essays that had been scored by human readers and that varied in length and skill level of the writers. The competition sought to find a winner who could come closest to the results of the human scorers. "Software scoring programs do not independently assess the merits of an essay; instead they predict, very accurately, how a person would have scored the essay" (www.gettingsmart.com). In May of 2012, a winning team was announced, but no information was provided as to the algorithms behind the winning software. That was proprietary information. However, by the Autumn of 2012, students involved in studying artificial intelligence at universities in the US began producing "final projects" for their classes that tried to duplicate the results of the ASAP competition. The students used the same sample sets of essays used in the competition. Their studies provided many more details into the process of developing automated essay graders.

2. Thinking Like A Robo-Grader: What The Research Tells Us… Words Matter!

ALISE LAMOREAUX

Photo by Alise Lamoreaux

While it's not possible to know the actual coding behind the proprietary rights of the major testing companies, it is possible to find research from the people involved in creating the technology driving the industry. Specific algorithms were not discussed in the professional research projects, but general methodology was, especially in the early research studies before proprietary rights were involved. It is also possible to find student projects in artificial intelligence attempting to recreate the Hewlett Foundation's ASAP competition results. The information provided in this chapter is based

on the knowledge available in 2019. Based on the history of the automated testing movement, it is likely that there will be improvements to the systems, but unless there is a major technological advancement, like personal computers and the Internet were to the 1990's, the basics will likely remain the same.

The framework of the features to be evaluated by the automated essay grader is based on assumptions about what indicates good writing. However, essay grading can be plagued with inconsistencies in determining what good writing really involves. Ellis Page began with the premise that there was an intrinsic aspect to good writing that couldn't be measured by a computer. He called the intrinsic components the "Trins". He believed that approximations could be developed to represent those intrinsic features and call them the "Proxes". The concepts he developed are foundational components to understanding the coding behind automated essay graders (AEG). The basis for determining good writing has bias built into it. Assumptions or beliefs on the part of the program developers are fundamental to the baseline from which the AEG is created.

Some Common Assumptions Underlying Good Writing:

- People who read and write more frequently have broader knowledge and larger vocabularies which correlates to higher essay scores
- The more people read and write the greater their exposure to a larger vocabulary and more thorough understanding of how to properly use that information in their own writing.
- The number of sentences in an essay equates to the quality of the essay
- Complex sentences are of more value than simple sentences
- Longer essays are more likely to have more unique words which shows a bigger vocabulary
- Short essays have a low word count and fewer words means lesser writing ability
- Correct spelling indicates a command over language and facility of use
- Good punctuation is an indicator of a well-structured essay
- Good essays will use similar vocabulary to high scoring essays in the data set used to judge the essay against
- Intrinsic features such as style and fluency cannot be measured, but can be approximated with measurable qualities like sentence length, word length, and essay length

What happens to an essay submitted to an AEG?

The process of "reading" for AEG is about analyzing components within the essay submitted. The essay will be "tokenized" or broken into individual tokens/features to be assessed. In reviewing the research on the development of AEG, different terminology was used to explain the process associated with the feature selection, but the basics are similar because of the limitations of artificial intelligence at this time. How the tokens are valued or weighted will vary and this methodology will not be shared based on intellectual property rights and ownership of the data set essays. There are several different coding platforms for determining the grammatical correctness of sentences and how much variance is allowed from the standard set. This is another area that is not shared information. The testing companies don't reveal the source of their essay data sets or who the human essay scorers were or even whom the human readers may have worked for in the past. The standardization information is proprietary.

A glimpse into the process can been seen in the research projects of the students studying artificial intelligence who were trying to recreate the results from the ASAP competition. In their projects they explain their assumptions and features selected to measure as well as how their program compared to the results they are trying to match.

In 2012 at Stanford University, a group of students (Mahanna, Johns, & Apte), reported a final project for their CS229 Machine Learning course involving automated essay grading. The purpose of their project was to develop algorithms to assess and grade essay responses. They used the essays provided for the Hewlett Foundation ASAP competition. In the explanation of the data they used, they stated the essays were written by students from grades 7-10. Each essay was approximately 150-550 words in length. The essays were divided into 8 sets and had different types of essays associated with each set. They used a linear regression model to assess the essays.

The assumptions/hypothesis behind their study was that a good essay would involve features such as language fluency and dexterity, diction and vocabulary, structure and organization, orthography and content. They stated they were unable to test for content. They use the *Natural Language Toolkit* (NLTK) and *Textmining* to process the language. The process to prepare the essays for assessment involved removing all placeholder for proper nouns and stripping all the punctuation from the essays.

Machine learning algorithms cannot work with raw text directly; the text

must be converted into numbers. The students used a model for extracting features regarding the words in the essays called a "Bag-of-Words" (BOW). A bag-of-words is a representation of text that describes the occurrence of words within a document. It is called a "*bag*" of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where they occur in the document. BOW involves two things: a vocabulary of known words and a measure of the presence of known words. A set of top words was created for the BOW and the" Stop Words" were discarded. Stop words are commonly used words: the, a, of, is, at, and so on. Search engines are commonly programmed to ignore these words as they are deemed irrelevant for searching purposes because they occur frequently in language. To save space and time, stop words are dropped at indexing time. Once the BOW is established, the top words are assigned a numerical value or a "weight".

Numerical features were also assigned to total word count per essay, average word length per essay, number of sentences in the essay (which was used to indicate fluency and dexterity) and character count per essay. Various parts of speech (nouns, adjectives, adverbs and verbs) were determined to be good proxes for vocabulary and diction. The essays were tokenized or split into sentences before the tagging process. Correct spelling is an indicator of command of language, so the ration of misspelled words is another feature assessed.

The Stanford study also revealed some information about the 8 sets of essays that were provided for use in the study. Sets 1-2 were persuasive/ informative essays and were relatively free from contextual text. Sets 3-6 expected a critical response after reading a text provided (story, essay, excerpt from a book) and therefore were expected to have more specific content. Sets 7-8 were narrative essays based on personal experiences or imagination. The students' results showed that their model of analysis performed relatively well on sets 1-2, the persuasive essays where the content was easier to control for. The model suffered on sets where content could vary more and they stated, "Our model does not work well on narrative essays."

In the fall of 2016, at Harvard University, a group of students, (Gupta, Hwang, Lisker, & Loughlin) reported a final project to their CS109A course. They also studied machine learning as it related to automatic essay grading. Like the Stanford study, these students used the essays from the Hewlett

Foundation ASAP competition. Their assumption/hypothesis was that word count would positively correlate to a good essay and that longer essays were reflective of deeper thinking and stronger content. They assumed a skilled writer would use a greater variety of words.

The study at Harvard used similar methodology to the Stanford students' study. The Natural Language Toolkit was used and stop words were removed from indexing as well. They based their analysis on number of sentences per essay, percent of misspellings, percentages of each part of speech, the likelihood of a word appearing that matched the test data essays, total word count, and unique word count. Their results also showed better scoring results on persuasive essays.

A project from Rice University (Lukic & Acuna,2012) also used the Hewlett Foundation ASAP essays to develop and evaluate an AEG system. Their study assumption/hypothesis was that numerical data could be a good predictor of an essay score. They measured: word count, character count, average word length, misspelled word count, adjective count, transition/analysis word count, and total occurrence of words in the prompt in the essay. As with the projects from Stanford and Harvard, the Natural Language Tool Kit was used to tokenize essays and strip the essays of punctuation and stop words and for parts of speech tagging.

The Rice study featured noun-verb pairs within sentences and gave weight to those. In addition, weight was given to the number of words found in both the prompt and the essay. Nouns were weighted and it was believed that they would demonstrate a focus surrounding topics and also demonstrate if an essay became "off-topic". The 200 most frequently occurring words within a data set were selected. Word pairing was evaluated and believed to indicate topical association between words. Nouns that were "personally identifying information" were censored and censored nouns were stripped from the essay. In the results, the authors felt that censoring for nouns with personal identifiers may have affected the noun-verb pairings and thus effected results within the project.

Automated essay grading has turned into a reality now. In helping students prepare to take an examination that will be scored by an AEG, it's important to know what the "rules" of the grader might be. After reviewing several studies, it seems that inferences can be made as to how the AEG will be "reading" the essay. Summing up the results of the projects at Stanford, Harvard, and Rice Universities the following inferences can be made about

the basis of the algorithms used the automated essay graders and what their measurement capabilities are as of the writing of this book:

- Lexical complexity rewards bigger vocabulary words and usage of unique words
- Text complexity is similar to assessing the reading level of the text
- Proportion of errors in grammar, usage, and mechanics can be rated
- Essay length matters
- Tokenizing components and ignoring stop words are part of indexing to "read"
- Matching examinee's essay vocabulary to data test sets vocabulary can matter

In February of 2012, Douglas D. Hesse, Executive Director of Writing at the University of Denver, published a paper entitled, "Can Computers Grade Writing? Should They?" In this paper, Hesse states that for automated essay graders, "Content analysis is based on vocabulary measures and organizational development scores are fairly related." AEG depends on chains of words, and words those words are associated with. Sophistication of vocabulary may be determined by the collection of terms. For example, the word "dog" could be replaced with the word "canine" and the sentence it was used in would rate higher. In addition, the word canine could be viewed as a more unique word because dog is more commonly used.

An ambiguous aspect of the vocabulary usage is the component of grade-level appropriateness. A teacher/human grader is responding to student work and can judge appropriate vocabulary for the essay. An AEG has a Bag of Words in some form it is "reading" for, and the words commanding the highest value may or may not be grade-level appropriate. What is the appropriate grade level for college transition or high school equivalency (HSE) or a workplace? Who determines vocabulary level appropriateness? It appears that it's the assumptions of the people writing the computer code behind the AEG and the test set of essay samples that are being compared to as the standard.

Hesse also provides examples of how longer sentence length is rewarded. Longer sentences may be valued as having more "style" than short sentence. According to Hesse, a short sentence combination like, "Dogs are interesting animals. Dogs are friendly to their owners. Dogs show affection by wagging their tails." would score lower with an AEG than the following sentence,

"Friendly to their owners, wagging tails to show affection, dogs are interesting animals." Chaining words matters. The Bag of Words created for the essay evaluation will contain words associated with other words. For example, owner and wag are words associated with dogs.

Hesse's paper seems to support the student projects from Stanford and Harvard that AEG is better suited to grade certain types of writing than others. He says, "Computer scores tend to be more valid and reliable – in relation to scores from expert human readers – when the tasks are very carefully designed and limited in length. The SAT® writing sample, for example, gives students 25 mins to write on a limited task…"

On the SAT® website, (https://collegereadiness.collegeboard.org/sample-questions/essay), the examples of topic prompts for preparing to take their test, the prompts ask students to develop an argument, thus examine persuasion, as their response. Once again, this seems to support the Stanford and Harvard student project findings in which their results were more accurate for test sets 1-2 from the Hewlett Foundation ASAP essay sets, which were the persuasive essays.

It's interesting to think about tokenization of an essay. As human readers we are not used to looking at essays or extended responses without paragraphs and punctuation. Thinking like a Robo-grader creates a new word awareness. What words demonstrate sentence complexity? What are the "Sign Posts" or "Cue terms" that indicate organization? It's not enough to be a "transition word" when being evaluated by an automated essay grader. Word choice matters, but not "stop words" yet for human readers those "commonly used" words are essential parts of communication not to be ignored.

3. Organizational Style & Structure of Response for a Robo-Grader

ALISE LAMOREAUX

Automated essay graders (AEG) are programmed to "read" for certain types of words that signal content and structure of the essay. AEG is limited in the type of essay it can effectively score. AEG does not independently assess the merits of an essay. AEG is designed to mirror or predict the score an expert human reader would assign to the essay or extended response.

The software programming behind the AEG system has been trained to look for similarities between the test set data and the response being currently evaluated. The basis for the trait analysis is that good writing should look like good writing. Automated essay graders are good at evaluating writing with specific parameters and defined vocabulary selection. Essays with a narrative format are difficult for AEG to evaluate due to the wide-open possibilities of language use. Using automated essay graders puts an emphasis on rhetorical essays of the argumentative/ persuasive or informational styles. Unfortunately for teachers and students, the argumentative response can be one of the most difficult essay styles to teach. William Jolliff (1998) in his faculty publication, for George Fox University, states that, "Most of [his] students have apparently seldom witnessed *how* real argument works…." It could be argued that in today's (2020) environment of conflict, it's hard to find a model for understanding the basis of evidence and persuasion via argument.

Allison Rose Greenwald, in her 2007 Thesis research at Iowa State University, cites four major difficulties in teaching the argumentative style of writing to students:

1. Students' limitation in comprehending logic
2. Argument is a difficult form of discourse to teach
3. Lack of guidance provided by standard textbooks
4. Poor teacher training as most composition teachers lack the background and skills in rhetoric and logic to teach argumentation

effectively

Three major types/models of argumentation formats are academically hailed. Each one has a different methodology and audience expectation.

1. **Classical**: argues an issue using evidence and refutation expecting an "opponent" with an open mind to change.
2. **Rogerian**: argues an issue emphasizing similarities with the "opponent's" beliefs attempting to establish a "win-win" outcome with no losers
3. **Toulmin**: argues an issue emphasizing the strength of evidence to a close-minded audience. Practical arguments are comprised of probability. Best used when the audience is logical and rational.

Different types of arguments lend themselves to each of the 3 models of argumentation styles. For example, the topic of Universal Health Care is a topic with "gray" areas of debate. The Rogerian model, looking for a "win-win" outcome, might be better suited for this topic. When facing a Robo-Grader, the Toulmin model may be the best choice. Toulmin arguments are most effective where there is a clear split between ideas on both sides of the issue. The Toulmin model focuses on removing the credibility of the opposition and showing the strength of the position supported. The topic of environmental damage being done by humans would fit the Toulmin model of argumentation.

The Toulmin method is good to use when the goal is to put facts at the forefront of the argument. It is also a good format for addressing the scientific community. Toulmin's ideas about logical arguments are relatively easy to explain to students and lend themselves to the tokenization that AEG will apply to the sentences.

One of the benefits to the Toulmin method is that it offers a non-complicated system for presenting an argument. It offers a structural model for building and analyzing rhetorical arguments. A complaint about the Toulmin method may be that it seems a bit like a formula for organizing an essay; however, that could also be its strength. The Toulmin method is a way to help writers think about connections and how to link the evidence to the claim. It's important to remember that evidence allows for judgement. It is not the same as proof, which means something is absolute, and not able to be contested.

In June of 2015, a teacher training session provided by GEDTS® designed for preparing students to take the GED® Reasoning Through Language Arts test, which includes an Extended Response that is scored by AEG, the presenters specifically suggested using the Toulmin method for structuring the response (https://www.youtube.com/watch?v=DAwXSOan3KQ). A more recent training by the same organization (Aug. 2019), still suggests using the Toulmin method, but they no longer refer to the format by name.

Components of the Toulmin Method for creating a structured argument:

1. Make a **claim.**
   The claim answers the question of "So, what's the point?"
2. What are the **grounds**/data?
   The grounds/data to answer the question of "How come?" or "Why?"
3. State the **warrant**/bridge that connects the claim to the grounds.
   "Why do these things go together?"
4. Provide **backing** to the warrant.
   Provide additional logic or supporting evidence for the warrant.
5. Include **qualifiers** to show the strength of the argument.
   Examples: so, some, many, in general, usually, typically, 75%
6. Create a **counterclaim** to the claim.
   Anticipate the opposing perspective and state it. Responding to counterclaims make you seem unbiased.
7. State a **rebuttal** which provided evidence to disagree with the counterclaim.

Example #1 of the foundation of the Toulmin Model (Simplified)

Claim: There is a forest fire nearby.
Grounds: Smoke is in the air.
Warrant: Fires produce smoke…
Qualifier: …so chances are, where there is smoke there is fire.
Backing: It is summer and that's fire season.
Counterclaim: It rained all last week and the ground is wet.

Rebuttal: A helicopter with a water bucket just flew overhead heading in the direction of the smoke.

Example #2 of the foundation of the Toulmin Model:

Grounds: My thoroughbred horse was born in the state of New York
Claim: ,so my horse is eligible for extra winnings in races specifically for New York bred horses
Warrant: ,since an equine born in New York will be considered a New York bred horse.

How the sentence would read to AEG:

My thoroughbred horse was born in the state of New York, so my horse is eligible for extra winnings in races specifically for New York bred horses, since an equine born in New York will be considered a New York bred horse.

Analyzing the above sentence regarding the horse, as an AEG might "read" it, the sentence has the following features:

- 42 words and 181 characters
- Reading level of 16.1
- The word "so" signals additional information and acts as a "qualifier"
- The word "since" signals a relationship
- The word "equine" is less commonly used/unique word and a synonym for horse
- Longer sentences are correlated to higher skills in language usage
- There are no misspelled words

As a writing instructor, I may not like the sentence construction, and feel it uses too many words, but looking at it from the perspective of its "features", I might think differently. The sentence will likely score high in the "eyes" of a Robo-grader.

Organization, Style, & Word Value

We know from the research that automated essay graders can't make judgements about evidence or content within the writing being scored. We know from the original research around Project Essay Grader (PEG®) that the intrinsic value associated with good writing can't be measured, so instead features that approximate the intrinsic qualities are defined and quantified. Key words signal the complexity of the writing. One such word is "because" and it is also linked to style (Shermis, Burstein, Higgins, & Zechner 2013).

One method of measurement used by AEG is to tie two or more features together to assess the complexity of the writing. It appears that AEG likes clauses, and especially dependent clauses, because they show relationships and can be qualifiers. Dependent clauses also make sentences longer. Dependent clauses signal that more information is coming. They also signal reasoning. Subordinating conjunctions are almost always associated with dependent clauses and can be interpreted as "cue" words. A word like "before" can cue the AEG into potential sequencing and organization. "Rather than" can cue a turn in reasoning or topic. The word "because" implies that a reason for the action or behavior will follow. To a Robo-grader, "power words" like "because" not only show a relationship, but also increase sentence length, which increases reading level, which increases sentence complexity, and subsequently, equals a higher score.

Robo-graders also like discourse markers, the words that help the text flow by showing time, cause and effect, contrast, comparison, qualification, and so on. Examples of discourse markers can be words like however, likewise, until, consequently, and therefore. Discourse markers are words that help connect sentences and ideas. They are basically transition words or conjunctions. These words match features that the AEG is looking for.

The AEG can look for words with similar meanings. The coding behind AEG will have clustered these similar meaning words together. The words that are longer, thus containing more letters, will have a higher value. The synonyms that are used less often and therefore considered more unique will also have a higher value.

AEG cannot detect polysemy, the coexistence of many meanings to the same word. Strictly counting the appearance of the word by AEG could be misleading. For example, the word, mine. It could be a personal pronoun,

a hole in the ground, an explosive device, or part of the name of the 2009 Kentucky Derby winner, Mine That Bird.

Another aspect of automated essay grading is looking for word matches to the test/sample essays that set the basis for the scoring. AEG has been trained to look for words that look like the highest scoring essays and then award value to the essay being scored based on its similarity to the sample set.

Length of the essay another vital component to pay attention to. If a suggested word length is given, for example, 300-500 words, it is an important piece of information. Failing to meet the suggested minimum number of words may trigger the AEG to not be able to find a paper in the sample set it is trying to match. Essays that are short will lack many of the features the AEG is looking for. Assumptions behind the programming from the research available suggest that shorter essays will equate to lower quality writing. At the opposite end of the spectrum, going beyond the suggested length may not gain the writer additional value as the essay has already been assigned the value of length and the assumption would be that more words aren't necessary. Demonstrating what 300 words looks like in print can be a helpful tool for students to increase their awareness and understanding of the expectation for the AEG.

Organizational style and structure can take on a different meaning when an automated essay grader is the final evaluator of an essay. An argumentative essay may not be the style of essay the student is familiar with, and therefore may need additional guidance in "thinking like a Robo-grader". The style of essay students will be asked to write will involve Evidence Based Writing (EBW), which is not the typical essay students learn to write. Traditionally, students are taught to engage with writing on a personal level. Robo-graders cannot handle the nuances of expression and may penalize the writer for vocabulary choices. Longer writing will rate higher, while fragments will decrease the score, even if they are stylistically appropriate. Word choice can take on a different meaning and significance when preparing to write for an automated essay grader.

4. Read Like A Robo-Grader: Developing Audience Awareness

ALISE LAMOREAUX

Before beginning to think about what words would influence a Robo-Grader, think about how people are persuaded. For example, how do kids get their parents to do what they want? Or, how do families decide what products to buy? Consider the following situations and what type of evidence might be important for each of the cases below:

Situation #1

Your school is looking at some of its policies to make changes. You have been asked to provide evidence about whether homework is harmful or helpful.

1. How would you present this to your school? Consider what type of information you would want to provide. What sort of information would make good evidence in this case?
2. Whose point of view should the presentation use? Students or teachers?
3. How would you present the same information to students as you would to teachers? Would you change anything you said based on your audience?
4. Is there a difference in effectiveness of evidence based on the audience? Why?

Situation #2

The City Council for where you live is considering a ban on "vaping". You have been asked to provide evidence on this topic.

1. How would you present this information to the City Council?
2. What type of information and evidence do you think would be important to include for this type of a decision? What would make good evidence for this decision?
3. How would you present this information to your friends to let them know what was happening with the City Council?
4. Would there be a difference in your presentation based on the audience you were trying to influence? Why?

Situation #3

Your school is working on revising its *Student Code of Conduct*. Plagiarism has been a problem that students don't seem to understand. The committee you are serving on is dealing with the question, "Is copying someone else's work ever acceptable?"

1. What type of information and evidence do you think would be important to include for this type of decision? What would make good evidence for this decision?
2. How would you present this information to the students of the school?
3. How would you present this information to the local community?
4. Would there be a difference in your presentation based on the audience you were trying to influence? Why?

Evidence can be categorized in many ways. Think about the above situations and the type of evidence best suited to each of those situations. Here are examples of 6 major types/categories.

- **Pathos:** Pathos involves emotional appeals. Language that shows emotions or feeling conveys the pathos. The goal of the evidence is to sway the emotions of the decision maker. An example of pathos might be, "Don't be the last person on the block to have their lawn treated – you don't want to be the laughing-stock of your community!" Or, "You'll make the right decision because you have something that not many people do: you have heart."
- **Ethos:** Ethos tries to show that the person providing the evidence is

believable. Expert witnesses in a trial are an example of ethos-the insinuation is that a psychiatrist's opinion about a person's state of mind should carry more weight with a jury, or that a forensic scientist should be able to interpret evidence better than the jury.

- **Logos**: Logos involves studies, data, charts, and logic to back up the statements being put forth. An example of Logos would be, "More than one hundred peer-reviewed studies have been conducted over the past decade, and none of them suggests that this is an effective treatment for hair loss."
- **Kairos:** Kairos involves an argument that creates a state of urgency. An example is this quote from Sir Thomas Moore, "*This is the right time, and this is the right thing.*"
- **Big Names:** Big names involves using names of experts or well-known people who support your position. Think of celebrity product endorsements or causes they support.
- **Testimony:** Testimony can be a personal story as support for why something should happen. It can be referred to as anecdotal evidence. Personal proof as a way of supporting the claim.

How will the automated essay grader recognize evidence?

The AEG will be looking for key words that signal the presence of evidence, such as words that link ideas together and show progression of thoughts, or words that show relationships. That is why the word "because" has now been called a "power" word. It links 2 ideas together *and* implies a relationship of dependency.

Examples of words that link argumentative or persuasive stances together could be:

Not only that, but …

Not only are they …, they are also …

They are not ..., nor are they ...
There are various/several/many reasons for this.
First, ... / Firstly, ...
Second, ... / Secondly, ...
Moreover, ... / Furthermore, ... / In addition, ...
Another significant point is that ...
Finally, ...
On the one hand, ... On the other hand, ...
In contrast to this is ...
Because of ...
That is why ...
After all, ...
The reason is that ...
In that respect ...
The result of this is that ...
Another aspect/point is that ...
It is because ...
Although it is true that ... it would be wrong to claim that ...
That may sometimes be true, but ...
One could argue that ..., but ...

Examples of words that show additional information is being added or a conclusion being drawn would be:

Most probably ...
It appears to be ...
It is important to mention that...
As indicated ...
In other words, ...
So, all in all it is believed that...
(In) summing up it can be said that ...
In conclusion
Therefore,
In short,
To conclude,
The evidence highlights, or has shown

The strength of … is that

These examples are not meant to be complete lists of words demonstrating connection of ideas, but rather ideas to build from. AEG will be able to recognize synonyms for these words as well. It's important to understand the patterns of words the AEG will be looking for.

Signposting Sentences

Signposting sentences can be thought of as explanations to the logical organization of the argument. Signposts help guide a reader, human or Robo, through the response. Signposts are helpful elements of each paragraph. They can be thought of as linking words or short phrases. Signposts are a good way to quantify what the response will do. Signposts are like symbols on a road map that make it easier for a reader to know at what stage the response is currently, and where it is going next. Examples of signpost include, but are not limited to, the following lists of words:

Highlighting or emphasizing a point

Importantly, … Indeed, … In fact, … More importantly, … Furthermore, … Moreover, … It is also important to highlight …

Changing direction or creating a comparison

However, … Rather, … In contrast, … Conversely, … On one hand, … On the other hand, … In comparison, … Compared to … Another point to consider is …

Adding a similar point

Similarly, … Likewise, … Again, … Also, …

Summarizing

Finally, … Lastly, … In conclusion, … To summarize, … In summary, … Overall, … The three main points are

Being more specific

In particular, … In relation to … More specifically, … With respect to … In terms of …

Giving an example

For instance, … For example, … this can be illustrated by … …, namely, … …, such as …

Acknowledging something and moving to a different point

Although … Even though … Despite … Notwithstanding …

Following a line of reasoning

Therefore, … Subsequently, … Hence … Consequently, … Accordingly, … As a result, … As a consequence, To this end,

Stop Words

Another important list of words to be aware of when an automated essay grader is involved are stop words. These are commonly used words that search engines are programmed to ignore when indexing word entries. Stop words are deemed irrelevant for searching purposes because they occur frequently in the language. To save both time and space, stop words are ignored. On the website GitHub Gist, a site that is commonly used to house open source projects (https://gist.github.com/sebleier/554280), a list of the stop words for Natural Language Tool Kit (NLTK) can be found. The initial list provided there includes 127 words, many of which are pronouns and conjunctions. Automated essay graders can be trained to read for word pairings, which may be important to realize since a word like "because" appears on the NLTK word list, indicating this "power" word may not actually be read due to the frequency of its occurrence in the English language. A more unique synonym could be a better word choice when the Robo-grader is the audience. Once again, due to proprietary information, there is no way to be sure of the stop words that are being filtered for by the automated essay reader; however, becoming aware of the existence of stop words can help students think carefully about the vocabulary they select. For instance, the words "I" and "think" are both on the stop word list for NLKT, which indicates that those words would not even be indexed and thus not "read".

One tip to help people increase awareness of their word selection and usage is to use a word processing program, like Microsoft Word, to assess the reading level associated with the piece of writing the person has created. In addition, students can assess the number of sentences used in the writing, the average number of words used per sentence, the average number of characters per word used, for example, to critically examine the details of their writing.

Another aspect of sentence organization to consider is where the key

terminology is presented in the writing. AEG will be looking for matches, like the game of *Concentration*, and the sooner it finds matches to the data it is looking for, the sooner value can be associated with the writing being evaluated. Sentence frames may assist people in configuring their writing in a manner that will be positively seen by the automated essay grader in a fast and efficient manner.

Sentence Frames

Sentence frames and starter sentences are a common suggestion for use in crafting the format of argumentative essays. The frameworks are designed to help students move through the organizational language to create fluency for the reader. It's important to remember that AEG is looking for patterns it sees in the test essays from which it has been trained. Without really knowing anything about the test set of essays, definitive statements are difficult to make. Intellectual property rights keep the test set of essays elusive; however, an inference could be made that it is likely that some of the test set essays include sentence frames. If that is the case, the AEG reader will rate those sentences as a match to the test set of essays and points will be generated based on the match that occurs.

One question that comes to mind is, what level of test set essays used sentence frames? Top-scoring essays? Essays that receive minimum passing scores? AEG likes unique words, so one strategy could be to examine the sentence frame suggestions and develop alternative ways to state the same framework. The following sentences are examples of using sentence frames. This list is not intended to be a complete list of sentence frames, just a place to start from.

Sentence Frames to Introduce the topic:

- The general argument made by _____ in his/her work _____ is that _____ because _____.
- Although _____ (believes, demonstrates,

argues) that _____, _____ supports/provides the clearest evidence _____.

- A key factor in both _____ can be attributed to _____.
- When comparing the two positions in this article, _____ provides the clearest evidence that _____.
- Looking at the arguments regarding _____, it is clear that _____.

Sentence frames to introduce issues:

- The issue of _____is a complex one. What it is about is_____.
- The question everyone's asking is ,_____? Here's the controversy: _____.
- We all need to consider_____. The debate is about _____.
- The issue to grapple with is_____. The problem is _____.
- We need to determine if _____because_____.
- It will be important to decide _____because_____.

Sentence frames to demonstrate the counter argument (Opposing Side's Position):

- People who disagree may claim that ... "state the opposing side's position"...
- Critics may claim that ..." state the opposing side's position" ... .
- Some people may argue that _____.
- A possible concern they may raise is that _____.

Demonstrate Why The Opposing Argument is Strong

- This opinion could be possible due to _____.
- They may have a strong argument as a result of _____.

Rebuttal (Explain Why Their Argument is Weak)

- This argument is wrong since _____.
- The evidence, however, overwhelmingly supports the argument that _____.
- On the contrary _____

It becomes clear that word selection, vocabulary depth, and essay organization are key components of success when the audience is automated essay graders. Understanding who the target audience of the writing will be is always crucial to receiving positive judgement from the reader; however, while AEG can mirror the results of human graders, it is not proven that they arrive at the similar conclusions for the same reasons. It's important to remember that some of the fundamental principles of "good" writing will be ignored by automated essay graders. Even though the "stop words" can be ignored by AEG, they are essential to human communication in standard English. Additionally, they may impact the total word count of the piece of writing, which is another aspect of the overall feature rating system.

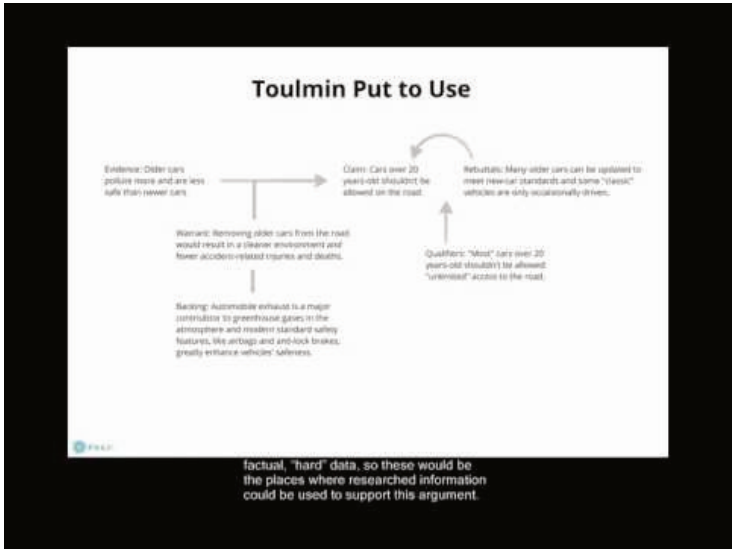
5. Writing For A Robo-Grader: Understanding the Toulmin Method

ALISE LAMOREAUX

Toulmin Model of Argument

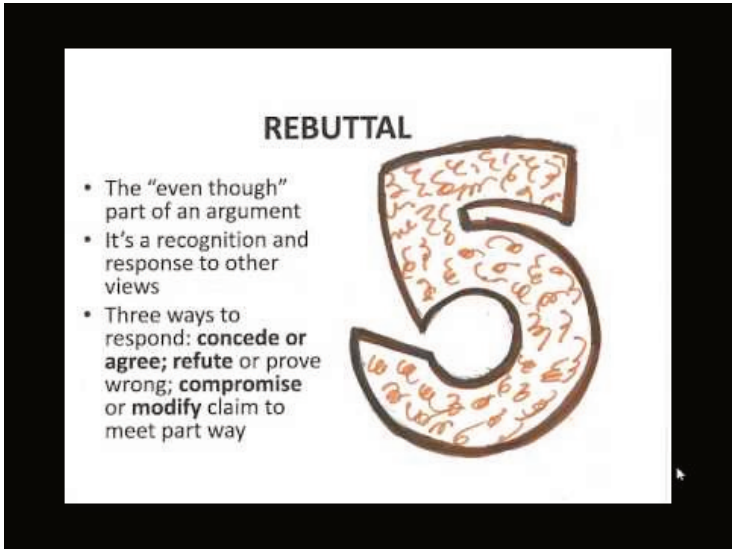
Watch the following 2 videos that explain the Toulmin Model of Argument. Compare the presentation styles of both videos. The videos are designed for students in different classes at different schools, who are all learning about this style of argument. Which style do you like better? Why? What don't you like about the other video? What makes one video's presentation better than the other? Why might someone like the video you don't prefer?

Parts of an Argument: Simple Example



A YouTube element has been excluded from this version of the text. You

can view it online here: https://openoregon.pressbooks.pub/robograders/?p=40



A YouTube element has been excluded from this version of the text. You can view it online here: https://openoregon.pressbooks.pub/robograders/?p=40

1. The essence of an argument is the **claim**. Without a claim, there is not an argument. The claim is the "umbrella" that all the other parts of the argument fit under. For example:
    - Classes at Lane Community College are fun to take.

2. **Warrants** are underlying assumptions that are the foundation for the

claim being made. They are usually unstated.

- People want to go to college and have fun while doing it.

3. **Qualifiers** are the words that put limits around the claim. Without qualifiers the word "all" is implied. Qualifiers make an argument easier to defend.

    - Many classes at Lane Community College are fun to take.
    - Some classes at Lane Community College are fun to take.
    - In general, classes at Lane Community College are fun to take.
    - Usually, classes at Lane Community College are fun to take.
    - Typically, classes at Lane Community College are fun to take.
    - Probably, classes at Lane Community College are fun to take.

4. **Identified Exceptions** from an argument give the author the ability to make exceptions to the claim. An exception is different from a qualifier in that it comes in the form of an example rather than a single word.

    - After struggling through an Anatomy and Physiology class at Lane Community College, I would not rank that as a fun experience.

5. In the Toulmin style of argument, the **reasons/grounds** why the author gives for making the claim answer 2 main questions: Is the reason relevant to the claim it supports? And Is the reason effective? If the reason gives the reader a sense of value, that it is believable, something to be agreed with, it is a "good" reason. Reasons cause the reader to make value judgements. When crafting an argument, it can be a good idea to restate the value invoked as clearly as possible.

    - Time management skills are important to learn to be successful in college.

***The reason*:** A student's daily schedule will require awareness of time.

- Getting started in college can seem like an uphill battle.

The ***reason***: College has many rules and policies to learn.

Note: If you are tasked with the job of identifying reasons, being able to restate them in your own words is important.

6. For a claim to be believable and convincing, it must supply evidence to satisfy 3 conditions in the Toulmin Method of argumentation. **Evidence** must be *sufficient*, *credible*, *and accurate.* Evidence can come in many forms: facts, examples, statistics, expert testimony, big name endorsements, emotional triggers, and more. The following statements are examples of evidence. Think about which ones have the best evidence in the statement and why?

   - If you examine the course syllabus of many college classes, you will find some element of time management involved in most of them.
   - Examination of 200 college syllabi revealed that 95% of them included some form of time management.
   - Student comments on a recent survey indicated that time management was part of their experience in their college courses.

7. An argument expects opposition to the claim made. The Toulmin Model of argumentation expects the author to anticipate the opposition and be able to state what the opposition might be thinking. The Toulmin Method expects the **counter argument** to the claim to be identified.

   Anticipating the opposition gives the author a chance to refute the opposition, also known as a **rebuttal**.

   - Many students believe that time management skills are important to college success; however, a recent study appearing in the *College Journal of Student Success* indicates that time management is the number one reason cited by faculty for student success in college.

8. Concluding statements finish the argument presented. Similar to a court case, a closing statement concludes the argumentation process.
    - In conclusion, the evidence for time management being an important aspect of college success is supported by students, faculty, and current research.

Activity:

Create simple examples of the Toulmin Method of Argumentation using the following claims:

1. Dogs make better pets than cats.
2. Aliens probably exist.
3. Boredom leads to trouble.
4. Robo-calling should be outlawed.
5. Education should be free for everyone.
6. Energetic drinks should be banned and made illegal.
7. Technology is limiting creativity
8. Online friends are the same as imaginary friends
9. Graffiti should be legal artwork.
10. Drinking soda has negative effects on health.

6. Practice Activities For Reading Like A Robo-Grader: Become A Reading Detective

ALISE LAMOREAUX

Practicing reading like a Robo-grader will involve reading and analyzing 5 separate essays on different topics. The essays to be read all come from the openly licensed collection 88 Open Essays.
. The essays used will be:

- #16 Misinformation and Biases Infect Social Media, Both Intentionally and Accidentally
- #6 Tools and Tasks
- # 39 The Dirt on Soil Loss from the Midwest Floods
- # 10 How to Save The Middle Class
- #57 How to Increase Your Chances of Sticking with Your Resolutions

There are many types of essays included in the "88 Open Essays". For the purposes of this activity, the essays selected will be of the argumentative or informative type, since those are the types of essays automated essay readers are most successful at reading. They are also the types most likely to be used by the testing industry.

The activities associated with reading the essays will involve 2 parts. The first part will focus on examining the essays from the level of word/sentence selection and usage, and looking at the sentences from the perspective of their individuality. Examination of the texts for signpost, words' cuing evidence, uniqueness of language, sentence structure, use of clauses, and looking for sentence frames will be part of this activity. The second part of the activity will involve using the Toulmin Model of argument to analyze the components of the essay from that standpoint. Finding the claim and assumption behind the claim, then determining the type of evidence being used to support the claim, finding qualifying words to show the degree of support for the claim presented, analyzing what a counter argument to the

claim could be, and then analyzing for a potential rebuttal for the argument or information presented.

Misinformation and Biases Infect Social Media, Both Intentionally and Accidentally

By Giovanni Luas Ciampaglia and Filippo Menzer

Activity Part 1:

1. Read the Misinformation and Biases Infect Social Media, Both Intentionally and Accidentally essay once through as a human reader would evaluate the essay. Make notes about any observation that you notice during your first read of the essay Misinformation and Bias Infect Social Media
2. After reading the essay, what words would you use to "tag" the essay? For example, #clickbait or #socialmedia. Tags are like thinking about the "key words" or main points of the essay.
3. Find the section of the essay titled, "Bias in the brain". Estimate how many words that section of the essay involves.
4. Next, using that same section, "Bias in the brain" re-write the section as individual sentences. Remove the punctuation. Now read the section as a Robo-grader might see it. Examine the word choices used by the author. Look for stop words. Are there any words that signpost the organization of the section? Were any sentence frames used by the authors?
5. Does this section read differently to you when you look at in terms of individual sentences? Explain your response. Do any words change their meaning without punctuation? Would a Robo-Grader notice the word meaning change? Is punctuation necessary for a human reader to understand the sentence?
6. Next, examine the entire essay for signposts, sentence structure, unique words, organization, and sentence frames, etc.
7. What is your overall impression of this essay at the word/sentence

level? Remember, AEG cannot evaluate content.

Activity Part 2: Examining the essay from the Toulmin Model

1. What is the claim of the essay?
2. What is an assumption (warrant) made by the authors of the essay?
3. Examine the essay for the type of evidence used (pathos, ethos, logos, kairos, big names, testimony). List the evidence and supporting details you uncover.
4. Were any qualifiers used to demonstrate a level of support for the claim?
5. What could a counter argument be to this essay's claim?

Bonus Activity:

Try to create a chart showing how the claim is connected to the assumption/warrant and the evidence. There may be more than one warrant based on the evidence you find.

Tools and Tasks

By Anonymous

Activity Part 1:

Read the Tools and Tasks essay once through as a human reader would evaluate the essay. Make notes about any observation that you notice during your first read of the essay Tools and Tasks.

1. After reading the essay, what words would you use to "tag" the essay? For example, #technology or #automotives. Tags are like thinking about the "key words" or main points of the essay.
2. Read paragraphs 1-5 of this essay and estimate how many words that section of the essay involves.
3. Next, using the same paragraphs 1-5, re-write the section as individual sentences. Remove the punctuation and paragraphs. Now read the section as a Robo-grader might see it. Examine the word choices used by the author. Look for stop words. Are there any words that signpost the organization of the section? Were any sentence frames used by the authors?
4. Does this section read differently to you when you look at in terms of individual sentences? Explain your response.
5. Do any words change their meaning without punctuation? Would a Robo-Grader notice the word meaning change? Is punctuation necessary for a human reader to understand the sentence?
6. Next, examine the entire essay for signposts, sentence structure, unique words, organization, and sentence frames, etc.
7. What is your overall impression of this essay at the word/sentence level? Remember, AEG cannot evaluate content.

Activity Part 2: Examining the essay from the Toulmin Model

1. What is the claim of the essay?
2. What is an assumption (warrant) made by the authors of the essay?
3. Examine the essay for the type of evidence used (pathos, ethos, logos, kairos, big names, testimony). List the evidence and supporting details you uncover.
4. Were any qualifiers used to demonstrate a level of support for the claim?
5. What could a counter argument be to this essay's claim?

Bonus Activity:

Try to create a chart showing how the claim is connected to the assumption/warrant and the evidence. There may be more than one warrant based on the evidence you find.

The Dirt on Soil Loss from the Midwest Floods

By Jim Ippolito and Mahdi Al-Kaisi

Activity Part 1:

Read the Dirt on Soil Loss from Mid-west Floods essay once through as a human reader would evaluate the essay. Make notes about any observation that you notice during your first read of the essay The Dirt On Soil Loss From Midwest Floods.

1. After reading the essay, what words would you use to "tag" the essay? For example, #floods or #soil. Tags are like thinking about the "key words" or main points of the essay.
2. Read paragraphs 1-5 of this essay and estimate how many words that section of the essay involves.
3. Next, using the same paragraphs 1-5, re-write the section as individual sentences. Remove the punctuation and paragraphs. Now read the section as a Robo-grader might see it. Examine the word choices used by the author. Look for stop words. Are there any words that signpost the organization of the section? Were any sentence frames used by the authors?
4. Does this section read differently to you when you look at in terms of individual sentences? Explain your response.
5. Do any words change their meaning without punctuation? Would a Robo-Grader notice the word meaning change? Is punctuation necessary for a human reader to understand the sentence?

6. Next, examine the entire essay for signposts, sentence structure, unique words, organization, and sentence frames, etc.
7. What is your overall impression of this essay at the word/sentence level? Remember, AEG cannot evaluate content.

Activity Part 2: Examining the essay from the Toulmin Model

1. What is the claim of the essay?
2. What is an assumption (warrant) made by the authors of the essay?
3. Examine the essay for the type of evidence used (pathos, ethos, logos, kairos, big names, testimony). List the evidence and supporting details you uncover.
4. Were any qualifiers use to demonstrate a level of support for the claim?
5. What could a counter argument be to this essay's claim?

Bonus Activity:

Try to create a chart showing how the claim is connected to the assumption/warrant and the evidence. There may be more than one warrant based on the evidence you find.

How to Save the Middle Class When Jobs Don't Pay

By Peter Barnes

Activity Part 1:

Read the How to Save the Middle Class When Jobs Don't Pay essay once through as a human reader would evaluate the essay. Make notes about any

observation that you notice during your first read of the essay Essay How To Save The Middle Class When Jobs Don't Pay

1. After reading the essay, what words would you use to "tag" the essay? For example, #middleclass or #personalfinance. Tags are like thinking about the "key words" or main points of the essay.
2. Read paragraphs 1-5 of this essay and estimate how many words that section of the essay involves.
3. Next, using the same paragraphs 1-5, re-write the section as individual sentences. Remove the punctuation and paragraphs. Now read the section as a Robo-grader might see it. Examine the word choices used by the author. Look for stop words. Are there any words that signpost the organization of the section? Were any sentence frames used by the authors?
4. Does this section read differently to you when you look at in terms of individual sentences? Explain your response.
5. Do any words change their meaning without punctuation? Would a Robo-Grader notice the word meaning change? Is punctuation necessary for a human reader to understand the sentence?
6. Next, examine the entire essay for signposts, sentence structure, unique words, organization, and sentence frames, etc.
7. What is your overall impression of this essay at the word/sentence level? Remember, AEG cannot evaluate content.

Activity Part 2: Examining the essay from the Toulmin Model

1. What is the claim of the essay?
2. What is an assumption (warrant) made by the authors of the essay?
3. Examine the essay for the type of evidence used (pathos, ethos, logos, kairos, big names, testimony). List the evidence and supporting details you uncover.
4. Were any qualifiers used to demonstrate a level of support for the claim?
5. What could a counter argument be to this essay's claim?

Bonus Activity:

Try to create a chart showing how the claim is connected to the assumption/warrant and the evidence. There may be more than one warrant based on the evidence you find.

How to Increase Your Chances of Sticking with Your Resolutions

By Camilla Nonteerah

Activity Part 1:

Read the How to Increase Your Chances of Sticking with Your Resolutions essay once through as a human reader would evaluate the essay. Make notes about any observation that you notice during your first read of the essay Increase Your Chances of Sticking to Your Resolutions

1. After reading the essay, what words would you use to "tag" the essay? For example, #resolutions or #advice. Tags are like thinking about the "key words" or main points of the essay.
2. Read paragraphs 1-4 of this essay and estimate how many words that section of the essay involves.
3. Next, using the same paragraphs 1-4, re-write the section as individual sentences. Remove the punctuation and paragraphs. Now read the section as a Robo-grader might see it. Examine the word choices used by the author. Look for stop words. Are there any words that signpost the organization of the section? Were any sentence frames used by the authors?
4. Does this section read differently to you when you look at in terms of individual sentences? Explain your response.
5. Do any words change their meaning without punctuation? Would a Robo-Grader notice the word meaning change? Is punctuation

necessary for a human reader to understand the sentence?
6. Next, examine the entire essay for signposts, sentence structure, unique words, organization, and sentence frames, etc.
7. What is your overall impression of this essay at the word/sentence level? Remember, AEG cannot evaluate content.

Activity Part 2: Examining the essay from the Toulmin Model

1. What is the claim of the essay?
2. What is an assumption (warrant) made by the authors of the essay?
3. Examine the essay for the type of evidence used (pathos, ethos, logos, kairos, big names, testimony). List the evidence and supporting details you uncover.
4. Were any qualifiers use to demonstrate a level of support for the claim?
5. What could a counter argument be to this essay's claim?

Bonus Activity:

Try to create a chart showing how the claim is connected to the assumption/warrant and the evidence. There may be more than one warrant based on the evidence you find.

7. Postscript: Closing Thoughts

Before publishing this book, I taught the information presented here for 9 months with students and gathered feedback (6 of those 9 months were entirely online). When I first got the idea for this book, I was opposed to the idea of artificial intelligence evaluating writing. I believed that automated essay graders (AEG) compromised the social nature of writing. As I dug into the process and began to unlock the mystery of how the writing was actually evaluated, I began to see a way to use this technology to help students. COVID19 sent everyone into some form of remote or online learning format. Teachers, students, and parents became overwhelmed by learning and working at home. Parents suddenly became teachers and writing took on a new emphasis in a digital environment. It can be argued that digital writing environments present a more complex communication environment than print. Teaching and learning took on a different level of time commitment. As school systems grapple with how they will "open back up" and teachers and students grapple with the decision to return to the classrooms, artificial intelligence and its role in education may be taking on a new meaning. Consequently, my thoughts on the topic of automated essay graders (AEG) have come full circle.

I presented the information in this book at an Oregon GED Summit 2019 conference, and saw hope in the eyes of teachers who had previously been frustrated, because they didn't understand the "reading" mechanism behind automated essay graders (AEG). I have seen hope in the eyes of students because AEG "reads" the same way every time, and the consistency is like a video game. Students feel like they can "level up" by learning what matters to the AEG, the rules of the game so to speak. For some students, knowing how the AEG will "read" takes the fear out of writing. Students can submit a paper to AEG, receive a score, re-write the paper, and see if they can improve the score. Human graders can be inconsistent and not always evaluate writing in the same manner. Human graders can vary their interpretation of what is "good" writing. The automated essay grader has been programmed to compare an essay written by a student to sample essays which are the foundation of the programming. It's true that the sample essays are proprietary, and there is likely a bias built into the sample essays, but understanding the nature of how an automated essay grader will "read,"

and what is valuable to the reader, is just another way of "knowing your audience".

During the time I started working with the concepts of the book, COVID-19 hit the country and schools were forced to go online or remote as a delivery model. Teachers became overwhelmed with trying to deliver a classroom to students who were now learning at home. An automated essay grader suddenly had a different role. I found a free AEG source (paperrater.com) that students could use to evaluate their own writing and give themselves feedback in much the way Ellis B. Page had once envisioned. Students began to understand why word count and word choice mattered in a way they hadn't previously understood. Students could understand the importance of how a sentence is started and how the use of transitional words would impact the grade their paper would get. The automated essay grader also added a level of consistency to my own evaluation process and gave me specific directions to help my students improve their writing in a definitive manner.

Overall, the creation of this book on automated essay graders has been an evolutionary journey for me. I now see automated essay graders in much the same way I see using my phone for help or directions. Automated essay graders are not useful for all types of writing. Creativity will never be measured through programming, but not all writing is meant to be creative. Writing this book has helped me think about strategies to help my students be more successful in their written communications. It has also helped me think about what words really "matter" and the importance of helping students understand the relationship between cause and effect. It seems especially important to develop those skills in students during this historical moment of COVID-19. As schools, as well as the workplace, become more automated, and remote or distance learning/working becomes the "new normal" understanding and leveraging artificial intelligence will become a critical skill.

References

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris, M. (1998). Automated Scoring Using A Hybrid Feature Identification Technique. *Annual Meeting of the Association of Computational Linguistics, August* 1998, 1-7. Retrieved 8/1/2019 from https://www.ets.org/Media/Research/pdf/erater_acl98.pdf

Greenwald, A. R. (2007). *Learning how to argue: experiences teaching the Toulmin model to composition students* [Retrospective Theses and Dissertations 14521, Iowa State University]. https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=15520&context=rtd

Gupta, A., Hwang, A., Lisker, P. & Loughlin, K. (2016). *Automated Essay Grading* [CS109a, Harvard University]. https://openoregon.pressbooks.pub/robograders/chapter/thinking-like-a-robo-grader-what-the-research-tells-us-words-matter/

Hesse, D. (2012). *Can Computers Grade Writing? Should They?* [Writing, The University of Denver]. https://www.du.edu/writing/media/documents/hesse-can-computers-grade-writing.pdf

Jollif, W. (1998). *Text As Topos: Using the Toulmin Model of Argumentation in Introduction to Literature*. [Faculty Publications – Department of English, George Fox University]. https://digitalcommons.georgefox.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1022&context=eng_fac

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, *25*, 259-284. Retrieved 8/1/2019 from http://lsa.colorado.edu/papers/dp1.LSAintro.pdf

Lukic, A. and Acuna, V. (2012). *Automated Essay Scoring* [Comp540, Rice University]. https://www.thefourthcomic.com/comp540/aes_report.pdf

Mahana, M., Johns, M., & Apte, A. (2012). *Automated Essay Grading Using Machine Learning* [CS229, Stanford University]. http://cs229.stanford.edu/proj2012/MahanaJohnsApte-AutomatedEssayGradingUsingMachineLearning.pdf

Page, E. B. (1967). Statistical and Linguistic Strategies in the Computer Grading of Essays. *COLING '67: Proceedings of the 1967 conference on Computational linguistics*, 1-13. Retrieved 8/1/2019 from https://www.aclweb.org/anthology/C67-1032.pdf

Page, E. B. (1994). Computer Grading of Student Prose, Using Modern Concepts and Software. *The Journal of Experimental Education*, *Vol. 62, No. 2*, 127-142. Retrieved 8/1/2019 from https://www.jstor.org/stable/20152405?seq=1

Rudner, L. M. and Gagne, P. (2000). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research, and Evaluation*, *Vol. 7, Article 26*, 1-5. Retrieved 8/1/2019 from https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1109&context=pare

Rudner, L. M., Garcia, V., & Welch, C. (2005). An Evaluation of IntelliMetric™ Essay Scoring System Using Responses to GMAT® AWA Prompts. *GMAC ® Research Reports*, *RR-05-08*, 1-13. Retrieved 8/1/2019 from https://www.gmac.com/~/media/Files/gmac/Research/research-report-series/RR0508_IntelliMetricAWA.pdf

Shermis, M. and Burstein J. (2013). Handbook of Automated Essay Evaluation, *Routledge*. 1-355.

Shermis, M., Burstein J., Higgins, D., & Zechner, K. (2013). *Automated Essay Scoring: Writing Assessment and Instruction* [Educational Testing Service, The University of Akron]. https://pdfs.semanticscholar.org/eed5/67622453f29b7b2d72955d07d8aad5e86daf.pdf

Shermis, M., Mzumara, H., Olson, J., & Harrington, S. (2001). On-line Grading of Student Essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education*, *Vol. 26, No. 3*, 247-258. Retrieved 8/1/2019 from https://www.academia.edu/6251154/On-line_Grading_of_Student_Essays_PEG_goes_on_the_World_Wide_Web