Automated Essay Evaluation: The *Criterion* Online Writing Service

Jill Burstein, Martin Chodorow, and Claudia Leacock

■ In this article, we describe a deployed educational technology application: the *Criterion* Online Essay Evaluation Service, a web-based system that provides automated scoring and evaluation of student essays. *Criterion* has two complementary applications: (1) *Critique* Writing Analysis Tools, a suite of programs that detect errors in grammar, usage, and mechanics, that identify discourse elements in the essay, and that recognize potentially undesirable elements of style, and (2) *e-rater* version 2.0, an automated essay scoring system. *Critique* and *e-rater* provide students with feedback that is specific to their writing in order to help them improve their writing skills and is intended to be used under the instruction of a classroom teacher. Both applications employ natural language processing and machine learning techniques. All of these capabilities outperform baseline algorithms, and some of the tools agree with human judges in their evaluations as often as two judges agree with each other.

The best way to improve one's writing skills is to write, receive feedback from an instructor, revise based on the feedback, and then repeat the whole process as often as possible. Unfortunately, this puts an enormous load on the classroom teacher, who is faced with reading and providing feedback for perhaps 30 essays or more every time a topic is assigned. As a result, teachers are not able to give writing assignments as often as they would wish.

With this in mind, researchers have sought to develop applications that automate essay scoring and evaluation. Work in automated essay scoring began in the early 1960s and has been extremely productive (Page 1966; Burstein et al. 1998; Foltz, Kintsch, and Landauer 1998; Larkey 1998; Rudner 2002; Elliott 2003). Detailed descriptions of most of these systems appear in Shermis and Burstein (2003). Pioneering work in the related area of automated feedback was initiated in the 1980s with the Writer's Workbench (MacDonald et al. 1982).

The *Criterion Online Essay Evaluation Service* combines automated essay scoring and diagnostic feedback. The feedback is specific to the student's essay and is based on the kinds of evaluations that teachers typically provide when grading a student's writing. *Criterion* is intended to be an aid, not a replacement, for classroom instruction. Its purpose is to ease the instructor's load, thereby enabling the instructor to give students more practice writing essays.

Criterion contains two complementary applications that are based on natural language processing (NLP) methods. *Critique* is an application that is comprised of a suite of programs that evaluate and provide feedback for errors in grammar, usage, and mechanics, that identify the essay's discourse structure, and that recognize potentially undesirable stylistic features. The companion scoring application, *e-rater* version 2.0, extracts linguistically-based features from an essay and uses a statistical model of how these features are related to overall writing quality to assign a holistic score to the essay. Figure 1 shows *Criterion*'s interface for submit-

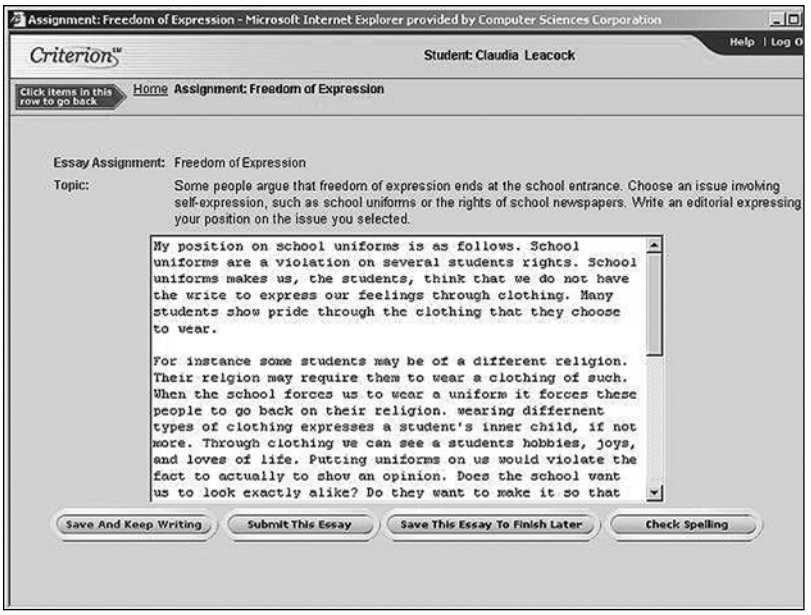

Figure 1. Entering an Essay into Criterion.

ting an essay, and figures 2 and 3 provide examples of its evaluations and feedback.

Critique Writing Analysis Tools

The *Critique* Writing Analysis Tools detect numerous errors under the broad headings of grammar, usage, and mechanics. The system also highlights potentially undesirable style—such as too much repetition. Finally, *Critique* identifies segments of essay-based discourse elements for the student. In this article, we describe those aspects of *Critique* that use NLP and statistical machine learning techniques.

Grammar, Usage and Mechanics

The writing analysis tools identify five main types of errors—agreement errors, verb formation errors, wrong word use, missing punctuation, and typographical/proofreading errors. Some examples are shown in table 1. The approach to detecting violations of general English grammar is corpus-based and statistical. The system is trained on a large corpus of edited text, from which it extracts and counts sequences of adjacent word and part-of-speech pairs called *bigrams*. The system then searches student essays for bigrams that occur much less often than is expected based on the corpus frequencies.

The expected frequencies come from a model of English that is based on 30-million words of newspaper text. Every word in the corpus is tagged with its part of speech using a version of the MXPOST (Ratnaparkhi 1996) part-of-speech tagger that has been trained on student essays. For example, the singular indefinite determiner *a* is labeled with the part-of-speech symbol AT, the adjective *good* is tagged JJ, the singular common noun *job* gets the label NN. After the corpus is tagged, frequencies are collected for each tag and for each function word (determiners, prepositions, etc.), and also for each adjacent pair of tags and function words. The individual tags and words are called *unigrams,* and the adjacent pairs are the bigrams. To illustrate, the word sequence, "a good job" contributes to the counts of three bigrams: *a*-JJ, AT-JJ, JJ-NN, which represent, respectively, the fact that the function word *a* was followed by an adjective, an indefinite singular determiner was followed by a noun, and an adjective was followed by a noun.

To detect violations of general rules of English, the system compares observed and expected frequencies in the general corpus. The statistical methods that the system uses are commonly used by researchers to detect combinations of words that occur more frequently than would be expected based on the assumption that the words are independent. These methods are usually used to find technical terms or collocations. *Criterion* uses the measures for the opposite purpose—to find combinations that occur *less often* than expected, and therefore might be evidence of a grammatical error (Chodorow and Leacock 2000). For example, the bigram for *this desks,* and similar sequences that show number disagreement, occur much less often than expected in the newspaper corpus based on the frequencies of singular determiners and plural nouns.

The system uses two complementary methods to measure association: pointwise mutual information and the log likelihood ratio. Pointwise mutual information gives the direction of association (whether a bigram occurs more often or less often than expected, based on the frequencies of its parts), but this measure is unreliable with sparse data. The log likelihood ratio performs better with sparse data. For this application, it gives the likelihood that the elements in a sequence are independent (we are looking for nonindependent, disassociated words), but it does not tell whether the sequence occurs more often or less often than expected. By using both measures, we get the direction and the strength of association, and performance is better than it would otherwise be when data are limited.

Of course, no simple model based on adjacency of elements is adequate to capture Eng-

Agreement Errors	Subject-verb: Friends is one thing I have learned from. Determiner-noun: This things would help us.
Verb Formation Errors	Their parents are expect good grades. Someone else could published a better book.
Wrong Words and Confusable Words	Verb instead of noun: Some have no chose. There/their confusion: Because of there different genres ...
Missing Punctuation	Missing apostrophe: It is a song about a mans love for a woman. Missing comma: To me community service is a student's choice
Typographical or Proofreading Errors	*The* instead of *They*: The would have happier employees. Two determiners in a row: Very often the a new head coach inherits a problem.

Table 1. Examples of Error Types that Are Identified by Criterion.

lish grammar. This is especially true when we restrict ourselves to a small window of two elements. For this reason, we needed special conditions, called *filters,* to allow for low probability, but nonetheless grammatical, sequences. The filters can be fairly complex. With bigrams that detect subject-verb agreement, filters check that the first element of the bigram is not part of a prepositional phrase or relative clause (for example, "My friends in *college assume....*") where the bigram *college assume* is not an error because the subject of *assume* is *friends*.

Confusable Words

Some of the most common errors in writing are due to the confusion of homophones, words that sound alike. The Writing Analysis Tools detect errors among *their/there/they're, its/it's, affect/effect* and hundreds of other such sets. For the most common of these, the system uses 10,000 training examples of correct usage from newspaper text and builds a representation of the local context in which each word occurs. The context consists of the two words and part-of-speech tags that appear to the left, and the two that appear to the right, of the confusable word. For example, a context for *effect* might be "a typical *effect* is found," consisting of a determiner and adjective to the left, and a form of the verb "BE" and a past participle to the right. For *affect,* a local context might be "it can *affect* the outcome," where a pronoun and modal verb are on the left, and a determiner and noun are on the right.

Some confusable words, such as *populace/ populous,* are so rare that a large training set cannot easily be assembled from published text. In this case, generic representations are used. The generic local context for nouns consists of all the part-of-speech tags found in the two positions to the left of each noun and in the two positions to the right of each noun in a large corpus of text. In a similar manner, generic local contexts are created for verbs, adjectives, adverbs, etc. These serve the same role as the word-specific representations built for more common homophones. Thus, *populace* would be represented as a generic noun and *populous* as a generic adjective.

The frequencies found in training are then used to estimate the probabilities that particular words and parts of speech will be found at each position in the local context. When a confusable word is encountered in an essay, the Writing Analysis Tools use a Bayesian classifier (Golding 1995) to select the more probable member of its homophone set, given the local context in which it occurs. If this is not the word that the student typed, then the system highlights it as an error and suggests the more probable homophone.

Undesirable Style

The identification of good or bad writing style is subjective; what one person finds irritating another may not mind. The Writing Analysis Tools highlight aspects of style that the writer may wish to revise, such as the use of passive sentences, as well as very long or very short sentences within the essay. Another feature of potentially undesirable style that the system detects is the presence of overly repetitious words, a property of the essay that might affect its rating of overall quality.

Criterion uses a machine learning approach to finding excessive repetition. It was trained on a corpus of 300 essays in which two judges had labeled the occurrences of overly repetitious words. A word is considered as being overused if it interferes with a smooth reading of the essay. Seven features were found to reliably predict which word(s) should be labeled as being repetitious. They consist of the word's total number of occurrences in the essay, its relative frequency in the essay, its average relative frequency in a paragraph, its highest relative

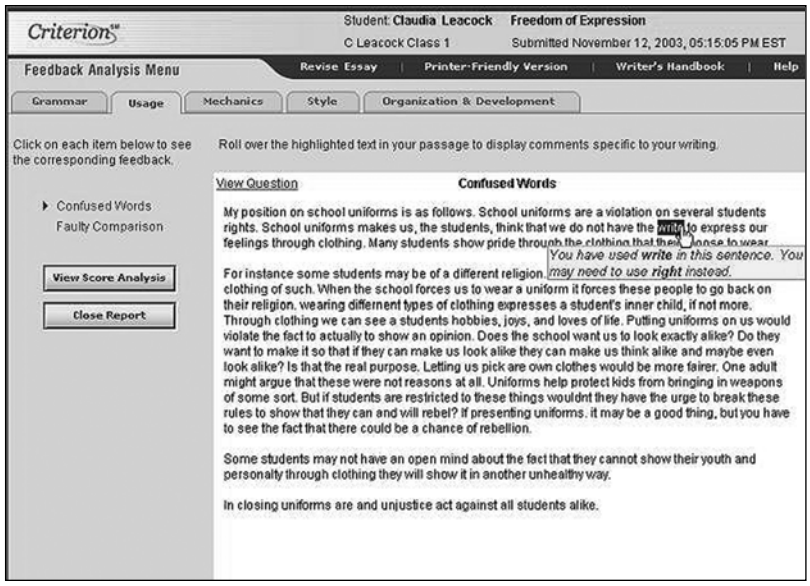

Figure 2. Confused Word Usage Error.

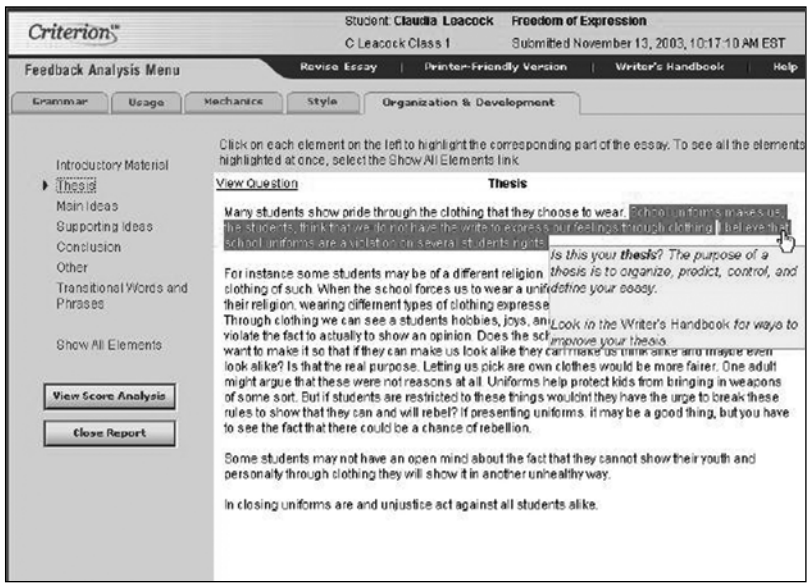

Figure 3. Identification of Thesis Statement in the Organization and Development Tab.

frequency in a paragraph, its length in characters, whether it is a pronoun, and the average distance between its successive occurrences. Using these features, a decision-based machine learning algorithm, C5.0 (www.rulequest.com), was used to model repetitious word use, based on the human judges' annotations. Some function words, such as prepositions and the articles *the* and *a,* were excluded from the model building. They are also excluded as candidates for words that can be assigned a repetition label. See Burstein and Wolska (2003) for a detailed description.

Essay-based Discourse Elements

A well-written essay generally should contain discourse elements, which include introductory material, a thesis statement, main ideas, supporting ideas, and a conclusion. For example, when grading students' essays, teachers provide comments on these aspects of the discourse structure. The *Critique* system makes decisions that simulate how teachers perform this task. Teachers may make explicit that there is no thesis statement, or that there is only a single main idea with insufficient support. This kind of feedback helps students to develop the discourse structure of their writing.

For *Critique* to learn how to identify discourse elements, humans annotated a large sample of student essays with essay-based discourse elements. The annotation schema reflected the discourse structure of essay writing genres, such as *persuasive* writing where a highly-structured discourse strategy is employed to convince the reader that the thesis or position that is stated in the essay is valid.

The discourse analysis component uses a decision-based voting algorithm that takes into account the discourse labeling decisions of three independent discourse analysis systems. Two of the three systems use probabilistic methods, and the third uses a decision-based approach to classify a sentence in an essay as a particular discourse element. Full details are presented in Burstein, Marcu, and Knight (2003).

The *e-rater* 2.0 Scoring Engine

The *e-rater* 2.0 scoring engine[1] is designed to identify some features in student essay writing that reflect characteristics that are specified in reader scoring guides. Readers are trained to read essays quickly in order to get a holistic impression of the writing sample, taking into account features, such as syntactic variety, use of grammar, mechanics, and style, organization and development, and vocabulary usage. The scoring scales often have a six-point range (1 is assigned to very poor quality writing; 6 is assigned to the highest quality writing). In *Criterion*, the description of an essay that gets a 6 is depicted in figure 4 while, by contrast, the description of lower scoring essays is shown in figure 5.

The *e-rater* 2.0 feature set includes features such as information about the thesis statement, corresponding to the first criterion for an essay receiving a score of 6 ("Clearly states

the author's position..."), and features that reflect errors in grammar, usage, and mechanics, which directly address the fourth criterion for a score of 1.

E-rater Features

E-rater 2.0 uses 12 features when scoring an essay. Eleven of these reflect essential characteristics in essay writing and are aligned with human scoring criteria. The importance of this alignment is that it increases the validity of the scoring system. *Validity* here refers to the degree to which the system actually does what is intended, in this case, measuring the quality of writing. For our users—writing teachers and the larger assessment community—validity is a crucial consideration.

E-rater 2.0's features are shown in table 2. The first six features are derived from the *Critique* writing analysis tools. Features 1–3 are based on the number of errors in grammar, usage, and mechanics that *Critique* has identified in the essay. Similarly, feature 4 derives from *Critique*'s style diagnostics. Features 5 and 6 are based on *Critique*'s analysis of the essay's organization and development. Feature 5 counts how many discourse elements are present in the essay relative to a typical eight units: a thesis, three main ideas, three supporting ideas, and a conclusion. If an essay has one thesis, four main points, and three supporting ideas, it gets credit for seven units, since it is missing a conclusion. An extra main idea does not contribute to the count as the program is looking for a particular development structure, not just identifiable discourse units in any category. The sixth feature is the average length of the discourse elements as a proportion of the total number of words in the essay. This provides an indication of the relative amount of discourse development.

To capture an essay's topical content, *e-rater* uses content vector analyses that are based on the vector-space model (Salton, Wong, and Yang 1975) that is often used in information retrieval. A set of essays that are used to train the model are converted into vectors of word frequencies. These vectors are transformed into word weights, where the weight of a word is directly proportional to its frequency in the essay but inversely related to the number of essays in which it appears. To calculate the topical analysis of a novel essay, *e-rater* represents each of the six score points with a vector of word weights based on the training essays. To calculate feature 7, *e-rater* converts the novel essay into a vector of word weights and conducts a search to find the training vectors that are most similar to it. Similarity is measured by the cosine of the angle between two vectors. The second topical content-based feature is the cosine between the vocabulary of the essay and the vocabulary of the very best training essays—those to which readers have assigned a score of 6.

- Clearly states the author's position, and effectively persuades the reader of the validity of the author's argument.
- Well organized, with strong transitions helping to link words and ideas.
- Develops its arguments with specific, well-elaborated support.
- Varies sentence structures and makes good word choices; very few errors in spelling, grammar, or punctuation.

Figure 4. Description of an Essay that Gets a "6."

- Little effort is made to persuade, either because there is no position taken or because no support is given.
- Lacks organization, and is confused and difficult to follow; may be too brief to assess organization.
- Lacks support.
- Little or no control over sentences, and incorrect word choices may cause confusion; many errors in spelling, grammar, and punctuation severely hinder reader understanding.

Figure 5. Description of an Essay that Gets a "1."

The remaining features are word based. For feature 9, *e-rater* computes the ratio of number of word types to tokens. The number of *word types* is the size of the vocabulary used in the essay (the number of different words it contains). The number of *word tokens* is the total number of word occurrences. For example, if a word appears three times in an essay it increases the type count by one and the token count by three. The type/token ratio can reveal a number of important characteristics of writing, including the level of repetitive word use.

Word frequency is closely associated with

1	Number of grammar errors ÷ essay length
2	Number of usage errors ÷ essay length
3	Number of mechanics errors ÷ essay length
4	Number of style diagnostics ÷ essay length
5	Number of required discourse elements
6	Average length of discourse elements ÷ essay length
7	Score assigned to essays with similar vocabulary
8	Similarity of vocabulary to essays with score 6
9	Number word types ÷ number of word tokens
10	Log frequency of least common words
11	Average length of words
12	Total number of words

Table 2. E-rater *2.0 Features.*
Essay length refers to the number of words in the essay.

word difficulty (Breland, Jones, and Jenkins 1994; Breland 1996), and word frequency information is commonly used to help develop assessments that evaluate verbal ability. To capture whether the writer is comfortable using relatively difficult words, *e-rater* incorporates an index based on word frequency as feature 10. These frequencies were collected from a general corpus of about 14 million words, and *e-rater* calculates the logarithm of the frequency of the least common words in the essay.

Feature 11 calculates the average word length in characters across all words in the essay as an additional indicator of the sophistication of vocabulary. Finally, feature 12 is a count of the total number of word tokens in the essay.

Model Building and Score Prediction

E-rater builds a model for each prompt or essay question. For each prompt, it trains on a sample of 200–250 essays that readers have scored and that represent the range of scores from 1 to 6. The feature set is standardized for each model, and each feature significantly contributes to the goal of predicting the human score. We use a multiple regression approach to generate weights for the 12-feature set, with the exception of the weight for word count.

Weights for any features may also be specified in advance, because it is important to be able to control feature weights when there are theoretical considerations related to components of writing ability. For example, the word count feature is highly predictive of essay score on its own. Therefore, we reduce this feature weight so that it does not dominate the final scoring model. This reduces the possibility that an essay's score will be artificially high simply because the essay is very long.

For each essay question, the result of training is a regression equation that can be applied to the features of a novel essay to produce a predicted score value. The last step in assigning an *e-rater* score is to convert the continuous regression value to a whole number along the six-point scoring scale.

Evaluation Criteria

We have described the computational approaches in the two applications in *Criterion*: *Critique* Writing Analysis Tools and *e-rater* 2.0. In this section we answer the question: "How do we determine that the system is accurate enough to provide useful feedback?" by discussing the approach we used to evaluate the capabilities before they were commercially deployed.

The purpose of developing automated tools for writing instruction is to enable the student to get more practice writing. At the same time, it is essential that students receive accurate feedback from the system with regard to errors, comments on potentially undesirable style, and information about discourse elements and organization of the essay. If the feedback is to help students improve their writing skills, then it should be similar to what an instructor's comments might be. With this in mind, we assess the accuracy of *e-rater* 2.0 scores and the writing analysis feedback by examining the agreement between people who perform these tasks. This inter-rater performance is considered to be the gold standard against which human-system agreement is compared. Additionally, where relevant, both inter-rater human agreement and human-system agreement are compared to baseline algorithms, when such algorithms exist. The performance of the baseline is considered the lower threshold. For a capability to be used in *Criterion* it must outperform the baseline measures and, in the best case, approach human performance.

For the different capabilities of *Critique*, we evaluate performance using precision and recall. Precision for a diagnostic *d* (for example, the labeling of a thesis statement or the labeling of a grammatical error) is the number of cases in which the system and the human judge (i.e., the gold standard) agree on the label *d,* divided by the total number of cases that the system labels *d.* This is equal to the number of the system's hits divided by the total of its hits and false positives. Recall is the number of cases in which the system and the human judge agree on the label *d,* divided by the total number of cases that the human labels *d.* This

is equal to the number of the system's hits divided by the total of its hits and misses.

Grammar, Usage, and Mechanics

For the errors that are detected using bigrams and errors caused by the misuse of confusable words, we have chosen to err on the side of precision over recall. That is, we would rather miss an error than tell the student that a well-formed construction is ill-formed. A minimum threshold of 90% precision was set in order for a bigram error or confusable word set to be included in the writing analysis tools.

Since the range for precision is between 90–100%, the recall varies from bigram to bigram and confusable word set to confusable word set. In order to estimate recall, 5,000 sentences were annotated to identify specific types of grammatical errors. For example, the writing analysis tools correctly identified 40% of the subject-verb agreement errors that the annotators identified and 70% of the possessive marker (apostrophe) errors. Precision for subject-verb agreement errors is 92% and for possessive marker errors is 95%. The confusable word errors were detected 71% of the time.

Repetitious Use of Words

Precision, recall, and the F-measure (the harmonic mean of precision and recall, which is equal to 2 * (precision * recall) / (precision + recall)) were computed to evaluate the performance of the repetitious word detection system. The total sample contained 300 essays where judges had labeled the words in the essay that they considered repetitious. Of the total sample, the two judges noted repetitious word use in only 74 of the essays, so the results are based on this subset.

A baseline was computed for each of the seven features used to build the final system. Of these, the highest baseline was achieved using the essay ratio feature that measures a word's relative frequency in an essay. For this baseline, a word was selected as repetitious if the proportion of that word's occurrences was greater than or equal to 5%. This resulted in a baseline precision, recall, and F-measure of 0.27, 0.54, and 0.36, respectively. The remaining six features are described above in the section on undesirable style. No single feature reached the level of agreement found between two judges (precision, recall, and F-measure of 0.55, 0.56, and 0.56, respectively). It is interesting to note that the judges showed considerable disagreement in this task, but each judge was internally consistent. When the repetitious word detection system, which combines all seven features, was trained on data of a single judge, it could accurately model that individual's performance (precision, recall, and F-measure of 0.95, 0.90, and 0.93, respectively).

Discourse Structure

To evaluate system performance, we computed precision, recall, and F-measure values for the system, the baseline algorithm, and also between the two judges. The baseline algorithm assigns a discourse label to each sentence in an essay based solely on the sentence position. An example of a baseline algorithm assignment would be that the system labels the first sentence of every paragraph in the body of the essay as a "Main Point."

The results from a sample of 1,462 human-labeled essays indicate that the system outperforms the baseline measure for every discourse category. Overall, the precision, recall, and F-measure for the baseline algorithm are 0.71, 0.70, and 0.70, respectively, while for the discourse analysis system, precision, recall, and F-measure are uniformly 0.85. For detailed results, see Burstein, Marcu, and Knight (2003). The average precision, recall, and F-measure are approximately 0.95 between two judges.

E-rater Performance Evaluation

The performance of *e-rater* 2.0 is evaluated by comparing its scores to those of human judges. This is carried out in the same manner that the scores of two judges are measured during reader scoring sessions for standardized tests such as the Graduate Management Admissions Test (GMAT). If two judges' scores match exactly, or if they are within one point of each other on the 6-point scale, an additional reader is not required to resolve the score discrepancy. When judges disagree by more than a single point, a third judge resolves the score. In evaluating *e-rater* 2.0, its score is treated as if it were one of the two judges' scores. A detailed description of this procedure can be found in Burstein, et al. (1998).

For a baseline, the agreement is computed based on the assignment of the modal (or most common) score to all essays in the cross-validation sample. Typical exact plus adjacent agreement between *e-rater* 2.0 and the human score is approximately 97%, which is comparable to that between two readers. Baseline agreement using the modal score is generally 75%-80%.

Application Use

Criterion with *Critique* Writing Analysis Tools and *e-rater*[2] was deployed in September 2002. As of February 2004, the application had been purchased by approximately 445 institutions,

and has approximately 500,000 users. Examples of the user population are: elementary, middle and high schools, public charter schools, community colleges, universities, military institutions (e.g., the United States Air Force Academy and The Citadel), and national job training programs (such as Job Corps). The system is being used outside of the United States in China, Taiwan, and Japan. The United Kingdom's Department of Education and Skills has also endorsed the application.

The strongest representation of users is in the K-12 market. Within K-12, middle schools have the largest user population. Approximately 10,000 essays are processed through *Criterion* each week and we anticipate increased usage as teachers become more familiar with the technology. Most of the usage is in a computer lab environment.

Criterion User Evaluation

As part of an ongoing study to evaluate the impact of *Criterion* on student writing performance, nine teachers in the Miami-Dade County Public School system, who used *Criterion* in the classroom once a week during the fall 2002 term, responded to a survey about their experience with *Criterion*. The questions elicited responses about *Criterion*'s strengths, weaknesses, and ease of use.

The teachers' responses indicate that *Criterion* provides effective help for students. All of the teachers stated that the strength of the application was that it supplies immediate scores and feedback to students. In terms of weaknesses, the responses primarily addressed technical problems that have since been fixed (for example, problems with the spell checker). In addition, all of the teachers maintained that learning how to use the system was, by and large, smooth.

The goal of this study, which Mark Shermis from Florida International University conducted independently, was to see whether access to the *Criterion* software would have a positive impact on a statewide writing assessment, specifically the Florida Comprehensive Assessment Test (FCAT). One thousand seventy-two tenth grade writing instruction students in an urban high school participated in the study. Half of the students were in the treatment group and the other half in the control group. In the treatment group, students had access to up to seven *Criterion* test questions over a period of 20 weeks. Most students wrote on only four or five of the prompts. Students in the control group wrote on other topics in a traditional classroom setting during times when treatment students were using *Criterion*. Though there were no significant differences in FCAT writing assessment scores between the treatment and control groups, the study indicated that students who used *Criterion* had significant increases in writing production, scored significantly better on later prompts, and significantly reduced a number of writing errors that are tracked by the *Criterion* software.

Application Development and Deployment

The *Criterion* project involved about 15 developers at a cost of over one million dollars. The team had considerable experience in developing electronic scoring and assessment products and services with regard to on-time delivery within the proposed budget. Members of the team had previously developed the Educational Testing Service's Online Scoring Network (OSN) and had implemented *e-rater* 1.3 within OSN for scoring essays for GMAT.

The project was organized into four phases: definition, analysis, development, and implementation. In the definition phase, we established the scope and depth of the project based on an extensive fact-finding process by a cross-disciplinary team that included researchers, content developers, software engineers, and project managers. This phase established the high-level project specifications, deliverables, milestones, timeline, and responsibilities for the project. In the analysis phase, the team developed detailed project specifications and determined the best approach to meeting the requirements set forth in the specifications. When necessary, storyboards and prototypes were used to communicate concepts that included interface, architecture, and processing steps. The development phase included the construction of the platform used to deliver the service, the development and modification of the tools used by the platform, and the establishment of connections to any external processes. The final implementation phase involved full integrated testing of the service and moving it into a production environment. Extensive tests were run to ensure the accuracy and scalability of the work that was produced.

The *Criterion* interface was developed by showing screen shots and prototypes to teachers and students and eliciting their comments and suggestions. The interface presented one of the larger challenges. A major difficulty was determining how to present a potentially overwhelming amount of feedback information in a manageable format via browser-based software.

Although a new version of the *Criterion* software is scheduled for release with the start of each school year, interim releases are possible. As new functionality is defined, it is evaluated and a determination is made as to a proper release schedule. *Criterion* was released in September 2002. Because the software is centrally hosted, updates are easily deployed and made immediately available to users. The software is maintained by an internal group of developers.

Conclusion

Criterion is fully deployed in classrooms, and is used by approximately 500,000 students internationally. We plan to continue improving the algorithms that are used, as well as adding new features. For example, we hope to implement the detection of grammatical errors that are important to specific native language groups, such as identifying when a determiner is missing (a common error among native speakers of East Asian languages and of Russian) or when the wrong preposition is used. The current system identifies discourse elements but does not evaluate their quality. We are extending the analysis of discourse so that the expressive quality of each discourse element will be assessed. This means, for example, not only telling the writer which sentence serves as the thesis statement but also indicating how good that thesis statement is.

Our goal is to talk to the teachers who use our system and, wherever possible, to use current NLP technology to incorporate their suggestions into *Criterion*.

Acknowledgements

We would like to thank Slava Andreyev, John Blackmore, Chi Lu, Christino Wijaya, and Magdalena Wolska for their intellectual contributions and programming support. We are especially grateful to Mark Shermis for sharing the teacher surveys from his user evaluation study. Any opinions expressed here are those of the authors and not necessarily of the Educational Testing Service.

Notes

1 *E-rater* 2.0 was invented by Yigal Attali, Jill Burstein, and Slava Andreyev.

2 In September 2001, *Criterion* was deployed with *e-rater* version 1.3 but without the *Critique* writing analysis tools. E-rater 1.3 and earlier versions have been used at Educational Testing Service to score GMAT Analytical Writing Assessment essays since February 1999. E-rater 2.0 is scheduled to be deployed in *Criterion* by September 2004.

References

Breland, H. M. 1996. Word Frequency and Word Difficulty: A Comparison of Counts in Four Corpora. *Psychological Science* 7(2): 96-99.

Breland, H. M.; Jones, R. L.; and Jenkins, L. 1994. The College Board Vocabulary Study. ETS Research Report No. 94-26. Princeton, N.J.: Educational Testing Service.

Burstein, J.; and Wolska, M. 2003. Toward Evaluation of Writing Style: Overly Repetitious Word Use in Student Writing. In Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics. East Stroudsburg, Penn.: Association for Computational Linguistics.

Burstein, J.; Marcu, D.; and Knight, K. 2003. Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing* 18(1): 32-39.

Burstein, J.; Kukich, K.; Wolff, S.; Lu, C.; Chodorow, M.; Braden-Harder, L.; and Harris M. D. 1998. Automated Scoring Using A Hybrid Feature Identification Technique. In Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics, 206-210. East Stroudsburg, Penn.: Association for Computational Linguistics.

Chodorow, M.; and Leacock, C. 2000. An Unsupervised Method for Detecting Grammatical Errors. In Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics, 140-147. East Stroudsburg, Penn.: Association for Computational Linguistics.

Elliott, S. 2003. Intellimetric: From Here to Validity. In *Automated Essay Scoring: A Cross-Disciplinary Perspective,* ed. M. Shermis and J. Burstein. Hillsdale, N. J.: Lawrence Erlbaum Associates.

Foltz, P. W.; Kintsch, W.; and Landauer, T. K. 1998. Analysis of Text Coherence Using Latent Semantic Analysis. *Discourse Processes* 25(2-3): 285-307.

Golding, A. 1995. A Bayesian Hybrid for Context-Sensitive Spelling Correction. Paper presented at the Third Workshop on Very Large Corpora, 39-53. 30 Cambridge, Mass., 30 June.

Larkey, L. 1998. Automatic Essay Grading Using Text Categorization Techniques. Proceedings of the Twenty-First ACM-SIGIR Conference on Research and Development in Information Retrieval, 90-95. New York: Association for Computing Machinery Special Interest Group on Information Retrieval.

MacDonald, N. H.; Frase, L. T.; Gingrich P. S.; and Keenan, S. A. 1982. The Writer's Workbench: Computer Aids for Text Analysis. *IEEE Transactions on Communications* 30(1): 105–110.

Page, E. B. 1966. The Imminence of Grading Essays by Computer. *Phi Delta Kappan,* 48:238-243.

Ratnaparkhi, A. 1996. A Maximum Entropy Part-of-Speech Tagger. In Proceedings of the Empirical Methods in Natural Language Processing Conference, University of Pennsylvania. East Stroudsburg, Penn.: Association for Computational Linguistics.

Rudner. L. M.; and Liang, T. 2002. Automated Essay Scoring Using Bayes' Theorem. *Journal of Technology, Learning, and Assessment,* 1(2).

Salton, G.; Wong, A.; and Yang, C. S. 1975. A Vector Space Model for Automatic Indexing. *Communications of the ACM* 18(11): 613–620.

Shermis, M.; and Burstein, J. eds. 2003. *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Hillsdale, N. J.: Lawrence Erlbaum Associates.

Jill Burstein is a principal development scientist at Educational Testing Service (ETS). She received her Ph.D. in linguistics from the City University of New York, Graduate Center. Her research focuses on the development of automated writing evaluation technology. She is a coinventor of *e-rater*®, an automated essay scoring system developed at ETS. She has collaborated on the research and development of capabilities that provide evaluative instructional feedback on student writing for grammar, usage, mechanics, style, and discourse analysis. Burstein is coeditor (with Mark Shermis) of the book *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Her e-mail address is jburstein at ets.org.

Martin Chodorow is a professor of psychology and linguistics at Hunter College and the Graduate School of the City University of New York. He received his Ph.D. in cognitive psychology from the Massachusetts Institute of Technology and has served as a researcher and consultant on numerous projects in natural language processing. His current research interests include automatic detection of grammatical errors and corpus-based representations of word meaning. His email address is martin.chodorow at hunter.cuny.edu.

Claudia Leacock is a principal development scientist at Educational Testing Service. She received a Ph.D. in linguistics from the Graduate Center of the City University of New York and was a postdoctoral fellow at IBM T. J. Watson Research Center. Since joining ETS in 1997, she has focused on using natural language processing tools for automated scoring and grammatical error detection. Prior to joining ETS, Claudia was a research staff member on the WordNet project at Princeton University's Cognitive Science Lab. Her e-mail address is cleacock at ets.org.