Opportunities for Natural Language Processing Research in Education

Jill Burstein

Educational Testing Service
Rosedale Road, MS 12R
Princeton, New Jersey 08540,
USA
jburstein@ets.org

Abstract. This paper discusses emerging opportunities for natural language processing (NLP) researchers in the development of educational applications for writing, reading and content knowledge acquisition. A brief historical perspective is provided, and existing and emerging technologies are described in the context of research related to content, syntax, and discourse analyses. Two systems, e-rater® and *Text Adaptor*, are discussed as illustrations of NLP-driven technology. The development of each system is described, as well as how continued development provides significant opportunities for NLP research.

Keywords: Natural language processing, automated essay scoring and evaluation, text adaptation, English language learning, educational technology.

1 Introduction

Theoretically, opportunities for natural language processing (NLP) research have existed in education in the area of reading research since the 1940's; writing research since the 1960's, and in the teaching of content knowledge since the early 1970's. While opportunities have existed for several decades, the general lack of computer-based technology proved to be an obstacle early on. In later years, when computers became increasingly more available, the lack of well-instantiated technological infrastructure (especially in schools), where educational applications could be broadly distributed and used, presented yet another obstacle – however, we can still see where the opportunities existed and how they grew over time.

Research in the area of *readability*, or *text quality* investigates the linguistic aspects of text that make a text relatively easier or more difficult to comprehend or follow. Early research examined the effect of morphological and syntactic aspects that contributed to readability, and [1] reported that features such as syllable counts (of words in a text), and sentence length were predictors of readability (text difficulty). Research in the area of readability has continued [2],[3], and has included increasingly more NLP-based investigation toward predicting grade-level fo texts [4], [5], [6],[7], or text quality in terms of discourse coherence [8],[9],[10]. In the context of text quality research, this paper will discuss a relatively new research prototype, *Text Adaptor*, that employs NLP-based methods and systems to support the creation of linguistically-appropriate reading materials for English language learners.

A. Gelbukh (Ed.): CICLing 2009, LNCS 5449, pp. 6–27, 2009.

The first area of writing research related to NLP-based research was automated essay scoring with the development of Project Essay Grade [11]. While this early approach to automated essay scoring proposed in [11] was largely related to the number of words in an essay, newer methods examined linguistic aspects of text with the introduction of e-rater® [12], and Intelligent Essay Assessor [13]. E-rater examined lexical, syntactic and discourse-related text features and Intelligent Essay Asessor analyzed content through vocabulary usage with latent semantic analysis. Writer's Workbench was a pioneer in the development of automated editing and a proofreading tool [14]. The tool offered feedback primarily related to grammar and mechanics.

Intelligent tutoring systems are associated with support and evaluation of *content knowledge acquisition* [15],[16],[17],[18],[19]. The goal of intelligent tutoring is to help students work through problem sets in various domains (e.g, physics). This is another area that has involved increasingly more NLP for the purpose of evaluating students' responses as they work through problem sets in a subject area [20]. Intelligent tutoring applications make use of systems that use propositional information in responses to identify correct knowledge, given a particular problem. C-rater™ is an NLP-based system that was developed at ETS. The system is designed to evaluate the correct content of an open-ended response to a subject-area test question [21],[22]. The system generates tuple structures from sentences to examine word use, given a syntactic structure. C-rater identifies paraphrase, so that responses from different students that use different, but synonymous vocabulary, in alternate syntactic structures, can be identified as having simliar, correct meaning.

As both the availability and access to computer-based technology have become well-established, opportunities to build and distribute educational applications have vastly increased in a number of areas, including readability (text quality) research, evaluation of student writing, and assessment of student content knowledge. While the former two areas are strictly text-based, the latter, assessment of content knowledge, can be text- or speech-based [23],[24]. Text-based applications will be the focus in this paper. This paper will illustrate growth and opportunity for NLP with e-rater and *Text Adaptor* – two NLP-based, educational applications, developed to support writing and reading, respectively [12],[25].

2 E-rater®

2.1 Motivation

E-rater is an automated essay scoring system that was developed at ETS. The first version of e-rater was operationally deployed in February 1999 [12], and was used to provide one score for each of the two essays on the writing section of the Graduate Management Admissions Test (GMAT).[1] The second score for each essay was provided by an expert, human rater. Prior to e-rater deployment, GMAT used two human raters to score each essay. Since the first version of e-rater proved to be highly

[1] The GMAT is a high-stakes exam taken by individuals planning to apply to graduate business programs. E-rater no longer scores the writing section of the GMAT exam.

reliable, and e-rater was shown to agree with an expert rater as often as two expert raters agreed with each other, it made sense, at least from a cost perspective, to use e-rater for one of the two scores. So, essentially, the motivation for the first use of e-rater was a highly practical one.

While the initial motivation may have been practical, research and development around e-rater has *always* carefully attended to ensuring that the development of e-rater features reflect the *writing construct*, given a specific task.. Features that reflect the writing construct include aspects of essays that can be measured (evaluated), given a writing task. For instance, in expository writing tasks on assessments, readers might typically focus on features in writing that contribute to a high quality essay, including the writer's organization and development of ideas, the variety of syntactic constructions, the use of appropriate vocabulary, and the technical correctness of the writing in terms of its grammar, usage, and mechanics. All of these features are aspects of the writing construct that need to be evaluated for a reader to assign an appropriate rating to an essay. The linguistic nature of these features clearly illustrates that NLP methods offer natural approaches for the detection and evaluation of these features.

E-rater was designed specifically to assign a *holistic* rating. In a holistic essay scoring approach, readers take into account all aspects of writing as specified in the scoring guide, and assign a score based on their overall impression of an essay. For an automated system to simulate this approach, all feature values extracted from an essay text are combined to produce a single score that represents the overall quality of an essay. Holistic scoring often uses a six-point scale, where a score of "6" indicates the best quality essay, and a score of "1" indicates an essay of the lowest quality. E-rater scoring was modeled along a six-point scale; however, the system can be trained to score essays on a range of scoring scales.

As you will see in the remainder of this section, while the initial motivation for automated essay scoring systems was practical, research and development has always focused on capturing aspects of an essay that reflect the writing construct *and* that can be used for applications that provide diagnostic feedback and essay scoring.

2.2 E-rater v.1

For the first release of e-rater (e-rater v.1), a number of existing NLP capabilities were used that were readily available, and some new ones were developed [12]. Capabilities were used to identify and extract from essays, linguistic information related to content, syntax and discourse. In e-rater v.1 each of the *e-rater* modules identified features that corresponded to scoring guide criteria used in human scoring. These included features related to *content, syntactic variety, and organization and development (discourse structure).* These features were then used to build e-rater models for predicting essay score.

2.2.1 Content

2.2.1.1 Topical Analysis. To capture use of vocabulary, *e-rater* used content vector analyses based on the vector-space model [26]. Training essays were converted into vectors of word frequencies, and the frequencies were then transformed into word

weights, where the weight of a word was directly proportional to its frequency in the essay but inversely related to number of essays in which it appears. To calculate the topical analysis of a test essay, the essay was converted into a vector of word weights, and a search was conducted to find the training vectors most similar to it. Similarity was measured by the cosine of the angle between two vectors. For one feature, called *topical analysis by essay*, the test vector consists of all the words in the essay. The value of the feature is the mean of the scores of the most similar training vectors. The other feature, *topical analysis by argument*, evaluates vocabulary usage at the argument level. The discourse analysis (see Section 2.2.3) was used to partition the essay into its main discussion points, and a vector was created for each. These argument vectors were individually compared to the training set so that a topical analysis score could be assigned to each argument. The value for this feature was a mean of the argument scores [27].

2.2.1.2 Analysis of Lexical Complexity. While the topical analysis features compared the specific words of the test essay to the words in the scored training set, the lexical complexity features treated words more abstractly [28]. Each essay was described in terms of the number of unique words it contains, average word length, the number of words with 5 or more characters, with 6 or more characters, etc. These numerical values reflected the range, frequency, and morphological complexity of the essay's vocabulary. For example, longer words are less common than shorter ones, and words beyond 6 characters are more likely to be morphologically derived through affixation.

2.2.2 Syntactic Analysis

In order to evaluate this aspect of an essay, a shallow syntactic parser developed for *e-rater* identified several syntactic structures, such as subjunctive auxiliary verbs (e.g., would, should, might), and complex clausal structures, such as complement, infinitive, and subordinate clauses. The parsed sentences also provided the input for discourse analysis.

2.2.3 Discourse Analysis

E-rater contained a lexicon based on the conceptual framework of conjunctive relations [29] in which cue terms, such as "In summary," are classified as conjuncts used for summarizing ideas in the essay. These classifiers indicated whether or not the item was a discourse development term (e.g., "for example" and "because"), or whether it initiated a new discourse segment (e.g., "First," "Second," or "Third"). *E-rater* contained heuristics that denoted the syntactic structures in which these terms must appear to be considered discourse markers. For example, for the word "first" to be considered a discourse marker, it must not be a nominal modifier, as in the sentence, "*The first time that I read a very long essay, I thought that it was well-written*," in which "first" modifies the noun "time." Instead, "first" must occur as an adverbial conjunct, as in the sentence, "*First, it has often been noted that length is highly correlated with essay score." E-rater* uses a lexicon of cue terms and associated heuristics to automatically annotate a high-level discourse structure of each essay. The system used these annotations to partition essays into separate arguments, used these as input to the *topical analysis by argument* component (see Section 2.2.1.1).

2.2.4 Model Building and Score Prediction

In order to predict a score, *e-rater* measured more than 50 features in a training sample of approximately 270 human-scored essays. A stepwise linear regression was run to select the features that made significant contributions to the prediction of essay score. For each essay question, the result of training was a regression equation that was applied to the features of a new test essay to produce a predicted value. This value was rounded to the nearest whole number to yield the predicted score. Agreement between a human rater and e-rater, and two human raters was comparable – about 90% exact-plus-adjacent agreement. Agreement between human raters is typically measured in these terms for scoring standardized writing assessments. Only when two human raters disagreed by more than a single point was a third human rater introduced to adjudicate the score.

2.3 New NLP for *Criterion*^{SM} -- Diagnostic Feedback

In late 1999 and early 2000, there were issues with e-rater's ability to handle anomalous essays, such as off-topic essays [30]. We also began to talk to K-12 schools and community colleges about the use of e-rater in classroom settings. As an outcome of these discussions, and after a considerable development effort, the *Criterion*^{SM} online essay evaluation service was released in 2001.[2] It was intended for use in classroom settings. In the very first version of *Criterion*, teachers could select an essay topic and select a writing assignment for their students. Students could write an essay in *Criterion*, and receive an *essay score* in seconds (from e-rater v.1.), and, if they applied, *anomalous essay advisories* (described in the following section). *Criterion* embodies the *process writing* approach. This approach supports the idea that students should be able to write several drafts of a piece of writing. Consistent with this, *Criterion* allowed students to submit multiple revisions of essays, and receive a new score for each revision.

As we continued to interact and collaborate with teachers and school administrators for *Criterion* development, we consistently received feedback from teachers and school administrators who explained that in order to make the e-rater score more meaningful to students, it had to be accompanied by diagnostic feedback that more closely resembled the kind of feedback that teachers provide when grading students' writing assignments. This included information about grammar, usage, and mechanics errors, style advice, and essay-based organization and development. This led to a considerable amount of new development in the areas of detailed feedback related to students' essays. The feedback fed into a new incarnation of e-rater -- e-rater v.2, the foundation of the present version of e-rater. Both the feedback and e-rater v.2 are used in the current version of *Criterion* [31]. See Figure 1 for a screenshot of *Criterion* feedback.

2.3.1 Content-Related: Anomalous Essay Advisories

Capabilities needed to be developed to detect if an essay was anomalous. Specifically, such capabilities had to be able to determine if an essay was either *off-topic,* or the essay content was *overly repetitious* to the point where it was likely that the writer had copied-and-pasted either the text of the essay question, or sections of the essay,

[2] *Criterion* was developed and originally deployed by ETS Technologies, Inc. which was a wholly-owned subsidiary of ETS.

over-and-over again. Essentially, these advisories needed to be designed to catch e-rater user attempts to fool e-rater. It is interesting to note that the need for this arose more so due to newspaper journalists trying to fool the system than students. Most of these anomalous essays came from submissions to an e-rater demo that was released for public use. Over time, three methods were developed to detect *off-topic* and *overly repetitious* essays.

The *first method* deployed in *Criterion*, uses the follow approach to detect off-topic and overly repetitious essays.

For each essay, *z-scores* are calculated for two variables:

1. Relationship to words in a set of training essays written to an essay question
2. Relationship to words in the text of the essay question

The z-score value indicates a novel essay's relationship to the mean and standard deviation values of a particular variable based on a training corpus of human-scored essay data. The score range is usually 1 through 6, where 1 indicates a poorly written essay, and 6 indicates a well-written essay. To calculate a z-score, which ranges from 0 to 1, the mean value and the corresponding standard deviation (SD) for *maximum cosine* or *prompt cosine* are computed based on the human-scored training essays for a particular test question. The z-score indicates how many standard deviations from the mean our essay is on the selected dimension. The formula for calculating the z-score for an new novel essay is the following.

$$z\text{-}score = \frac{value - mean}{SD}$$

Z-scores are computed for the following.

1. The *maximum cosine*, which is the highest cosine value among all cosines between an unseen essay and all human-scored training essays, and
2. The *prompt cosine*, which is the cosine value between an essay and the text of essay question.

When a z-score exceeds a set threshold, it suggests that the essay is anomalous, since the threshold value indicates an acceptable distance from the mean.

A *second*, newer prompt-specific method was developed more recently and complements performance of the *maximum cosine* method. This new method helps to flag additional off-topic essays. It is based on calculating two rates for each word used in essays:

1. Proportion of word occurrences across many essay questions (generic, or essay question-independent, rate)
2. Proportion of word occurrences within a topic (essay question -specific rate).

The generic rate of occurrence for each word (G_i) across the large sample is calculated one time only from a large sample of essays across different essay questions from within one program, or within similar grade-levels. It is interpreted as the base-rate level of popularity of each word. The prompt-specific rate (S_i) is computed from

a training sample of essays that were written to the specific prompt for which an individual essay is to be compared. These two rates are used to compute an overall index for each individual essay.

$$\frac{1}{N}\sum_{i=1}^{n}\sqrt{S_i(1-G_i)}$$

Equivalently, in order to compute this index, we carry out the following steps.

1. Identify S_i and G_i values for all words in an essay based on pre-determined values from training sets.
2. For each word, compute $S_i(1-G_i)$ and take the square root. These square roots are summed over all words.
3. Multiply the sum of square roots by $1 \div N$, where N is the number of words in the essay, and the two rates are computed for all words in the essay.

A word in neither of the training samples will have a rate of 0, so a totally new word, not in either the generic or the specific sample, will also have a weight of zero ($0 \times (1 - 0) = 0$). This index can be interpreted as an average of word weights, where the weights are larger for words that appear more frequently in the prompt-specific essays, but at the same time are not frequent in other prompts. These are the words that would most contribute to the discrimination between on-topic essays and off-topic essays. The range of word weights is from 0 (when a word never appears in the specific sample and/or always appears in the generic sample of essays) to 1 (when it appears in every specific essay but never appears in the generic sample). The use of the square-root transformation in the weighting of words is designed to emphasize heavily-weighted words over low-weighted words. The classification of new essays as off- or on-topic is determined by setting a cutoff on the index values. This cutoff is based on the distribution of index values in the prompt-specific training sample.

In a *third method* for detecting off-topic essays, a training corpus *is not* required. The need for this method arose when *Criterion* began to allow teachers to introduce their own essay questions. We had no training data from teacher-created essay questions. Our topic-independent model for off-topic essay detection uses content vector analysis, and also relies on similarity scores computed between new essays and the text of the prompt on which the essay is supposed to have been written. Unlike the first and second method, this method does not rely on a pre-specified similarity score cutoff to determine whether an essay is on- or off-topic. Because this method is not dependent on a similarity cutoff, it also does not require any prompt-specific essay data for training in order to set the value of this parameter. Instead of using a similarity cutoff, our newer method uses a set of reference essay prompts, to which a new essay is compared. The similarity scores from all of the essay-prompt comparisons, including the similarity score that is generated by comparing the essay to the target prompt, are calculated and sorted. If the target prompt is ranked amongst the top few vis-a-vis its similarity score, then the essay is considered on topic. Otherwise, it is identified as off topic. This new method utilizes information that is available within *Criterion*, and does not require any additional data collection of student essays or test questions. Details for all three methods can be found in [30].

2.3.2 Diagnostic Feedback

Diagnostic feedback was developed to support teachers' requests that feedback was needed to support e-rater scores, to give the scores more meaning for students. Feedback is based on the detection of numerous errors in grammar, usage, and mechanics (*syntax*), highlights undesirable style (*content*), and identification of segments of essay-based discourse elements (*discourse*) for the student. This feedback was subsequently used to build a new version of e-rater (e-rater v.2).

2.3.2.1 Syntax

2.3.2.1.1 Grammar, Usage and Mechanics. The feedback capabilities identify five main types of errors – agreement errors, verb formation errors, wrong word use, missing punctuation, and typographical/proofreading errors. 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 [32] 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 [33]. 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 non-independent, dis-associated 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 English grammar, so filters are used to handle special conditions. Filters allow for low probability, but grammatical, sequences. With bigrams that detect subject-verb agreement, for example, filters check that the first element of the bigram is not part of a prepositional phrase or relative clause (e.g., *My friends in college assume.*) where the bigram *college assume* is not an error because the subject of *assume* is *friends*.

While the bigram method is used to handle a number of grammar, usage and mechanics error types, a number of error types in these categories are also implemented using a rule-based approach.

2.3.2.1.2 Confusable Words. Some of the most common errors in writing are due to the confusion of homophones, words that sound alike. In *Criterion*, we 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, a Bayesian classifier [34] is used 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.

2.3.2.2 Content

2.3.2.2.1 Undesirable Style. The feedback tool highlights 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. A 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 [35]. This is a feature that teachers consistently requested. The detection of overly repetitious words fits most appropriately

in the domain of content, i.e., vocabulary use. This feature is unique to Criterion, and does not exist in competitor systems.

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 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, 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.

2.3.2.3 Discourse

2.3.2.3.1 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 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. While the original version of e-rater took into account the discourse cue words and terms in a text, it did not handle discourse segments of an essay. Therefore, to enhance *Criterion's* ability to handle organization and development in an essay, a system was built that identified essay based discourse elements in text. No competitor system has this text analysis feature.

For a system to learn how to identify discourse elements, humans annotated a sample of about 1400 student essays with essay-based discourse elements, based on a written annotation protocol. The annotation schema reflected the discourse structure of essay writing genres, such as *persuasive* writing where a highly-structured discourse strategy is employed. 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. The system labels sentences according to the discourse element to which they belong: *Introductory Material*, *Thesis Statement, Main Point, Support, Conclusion* and *Other*. The category, *Other*, typically is used for opening and closing salutations in essays that are written in a letter format. Full system details can be found in [36].

Criterion offers additional feedback that indicates if critical discourse elements are missing (e.g., *Thesis Statement*), or if more elements are desirable (e.g., the essay contains only 1 *Main Point*, and 3 *Main Points* would be desirable.)

2.4 E-rater v.2

As has been mentioned throughout the section on e-rater, a critical goal in e-rater development has been to continue to enrich the system with new features that better reflect the writing construct. When e-rater v.1 was developed, ETS researchers had access to a small number of *freeware* tools, such as part-of-speech taggers and syntactic parsers that could be used to develop e-rater features. The first version, therefore, was to some extent limited to existing tools, and new tools that could be built given time constraints and available resources. With the development of the diagnostic feedback described in the sections above, e-rater could now be enhanced with features that were considered more central to the writing construct, including errors in grammar, usage and mechanics, style features, and essay-based discourse analysis. To this end, e-rater had a complete makeover, and now includes a standard set of 8-10 features that are believed to be more representative of features associated with the writing construct for expository and persuasive essay writing.

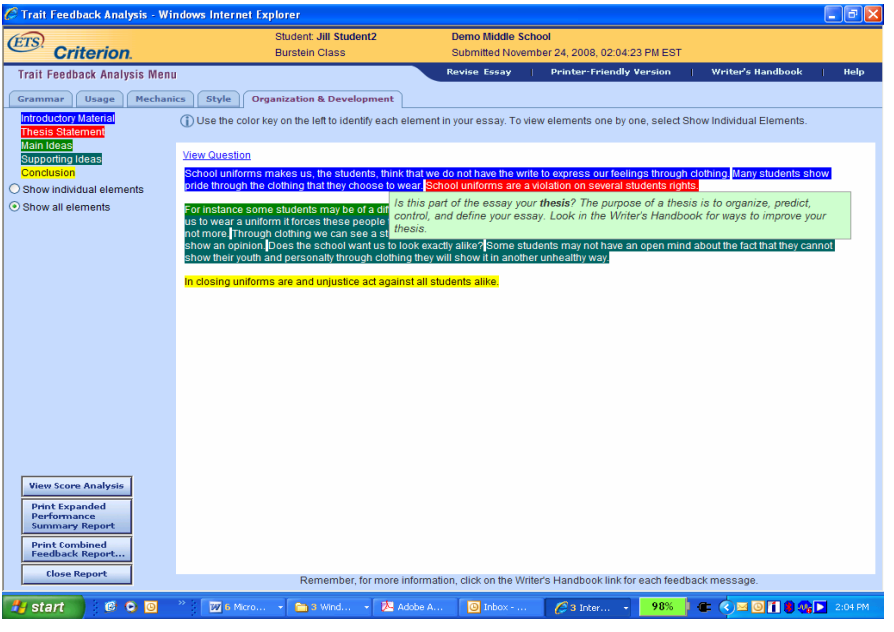

Fig. 1. *Criterion* Organization and Development Screen

E-rater v.2 forms the basis for all e-rater upgrades and uses the following information to create the standard feature set: (1) grammatical errors, (2) usage errors, (3) mechanics errors, (4) presence of discourse elements, (5) development of discourse elements, (6) style information, (7) a content vector analysis feature comparing an essay to the set of training essays (8) a content vector analysis feature comparing an essay to the set of training essays that received a score of 6, (9) average word length, and (10) a word frequency-based feature. Features (1) – (6) are derived from the

diagnostic feedback, so e-rater scores are aligned with *Criterion* feedback. The full feature set is run in a multiple regression to obtain feature weights for the final scoring model.

Two kinds of e-rater models are currently built. *Topic-specific models* are built using a set of human-scored essays on a given topic, and all ten features are typically used for these models. Topic-specific models can be built only when there are sufficient human-scored data for a topic. *Grade-level models* are built using a set of human-scored essay data written by students in a particular grade, across a number of essay topics. All features, except for the content-specific features created for (7) and (8), are used to build these models. These models can be applied to essay responses for any topic written by students at the specified grade level. No new training is required for new topics. [37] should be consulted for a full description of the e-rater v.2.

3 Text Adaptor[3]

3.1 Motivation

Subject-area academic vocabulary and general knowledge of English language skills can interfere with English language learners (ELLs) reading comprehension in K-12, content-area classrooms. Especially for ELLs beyond elementary school, large achievement gaps have been noted when the emphasis switches from *learning to read* to *reading to learn* [38]. At that point, ELLs must be able to understand grade-level, academic subject-area texts far beyond their English reading level [38],[39]. For example, social studies teachers use content materials from history, political science, sociology, geography, and economics, and each contains *specialized jargon* rooted in American culture [40]. ELLs must learn the specialized, academic vocabulary which often includes low-frequency, more difficult words. Therefore, the responsibility for educating ELLs rests not only with English language specialists or bilingual educators, but with *all* teachers. However, the number of teachers trained in effective instructional strategies to meet the needs of ELLs has not increased at the same pace as the increases in the population [41],[42].

Text adaptation [43],[44],[45] and *linguistically-targeted instruction* [38],[46] [47], [48],[49] are recommended instructional approaches that address the need to provide ELLs with improved access to academic content in text-based curriculum. *Text adaptation* involves the actual modification of a text, using techniques including, linguistic simplification (e.g., reducing complex sentence structure), elaboration of text (e.g., inserting an easier synonym adjacent to a difficult word), text summarization, and supplemental native language support. In developing *linguistically-targeted instruction,* teachers might emphasize specific linguistic features in the text, such as polysemous and morphologically-complex words, or complex syntactic structures, but do not necessarily modify the text. Using this approach, teachers might highlight a targeted linguistic feature in a text, and provide supplementary instruction about that feature via a related lesson. Both *text adaptation* and *linguistically-targeted instruction* require a strong linguistic awareness, specifically related to features that would interfere with ELL students' ability to comprehend text content.

[3] *Text Adaptor* core research and development, and the *Text Adaptor* vision presented in this paper were created in full collaboration with ETS researchers, Jane Shore and John Sabatini.

Teachers often lack training in identifying linguistic features of English that may be barriers for ELLs [50],[51],[52]. In addition, it takes considerable time to modify classroom texts, and to develop linguistically-targeted instruction for students with varying needs, even for teachers who are skilled in these practices. Consistent with the theme of this paper, effective strategies for text adaptation and linguistically-targeted instruction tie in directly to several NLP capabilities. Therefore, from *Text Adaptor's* inception, there has been considerable opportunity for NLP research. *Text Adaptor* currently uses a number of NLP capabilities to support text modification activities.

3.2 Text Adaptor and NLP

The motivation to develop *Text Adaptor* can be summarized by these three interrelated factors: (a) there is a large ELL population in K-12 classrooms in the U.S., (b) there is a lack of academic reading material where key content is accessible to ELLs, and (c) K-12 content-area teachers are not necessarily trained in methods to effectively communicate difficult academic content to ELLs.

Text Adaptor is a web-based system that was designed to provide linguistic guidance to teachers to help build their knowledge and skill, and facilitate actual development of *text adaptations* [25]. *Text Adaptor* feedback is designed to build teacher knowledge and foster the creation of more content-accessible reading materials for ELLs. The central idea is that *Text Adaptor* feedback offers linguistic insight related to linguistic features in a text that might be difficult for ELLs. Linguistic complexity can interfere with ELL students' content comprehension. The idea, then, is that teachers can use *Text Adaptor* feedback to make appropriate changes to a text, rendering the text more comprehensible to the student.

Text Adaptor's interface design and embedded functionality were inspired by academic research in the area of text adaptation, and through our early collaboration with teachers who participated in school-based pilot studies over the past three years [25]. See the screenshot of *Text Adaptor* Figure 2, below. *Text Adaptor* incorporates several NLP capabilities which were selected because (a) they are aligned with pedagogy in text adaptation practice and linguistically-targeted instructional approaches, and (b) these capabilities were immediately available and could be incorporated in the system given available resources. In light of the fact that all NLP capabilities incorporated into *Text Adaptor* were *off-the-shelf*, much opportunity still remains to build additional NLP capabilities that would enhance the current system. Our ideas for new capabilities are discussed later in this section.

Current *Text Adaptor* features highlight different kind of linguistic features in a text that address *content* and *syntax*. Given the following set of *Text Adaptor* features, (a) - (d) address text *content* issues, and (e) addresses *syntax*: (a) automated synonym detection to replace or supplement difficult words with synonyms [21], (b) antonym detection (using WordNet®) as a vocabulary lesson supplement, (c) automated text summarization [53], (d) English-to-Spanish machine translation[4], and (e) shallow parsing to identify complex sentence structures [12]. *Text Adaptor* also contains a

[4] Language Weaver's English-to-Spanish machine translation system is used: http://www.languageweaver.com.

Fig. 2. *Text Adaptor* main screen with pull-down menu of synonym options for the word *document*

hand-built, English-Spanish cognate dictionary, so that Spanish cognates can be used to supplement the English text. This is another feature that addresses text content, and currently supports the large population of native Spanish speakers in U.S. classrooms.

Users (teachers) can opt to accept *Text Adaptor* feedback to include in an adaptation. In addition, the system allows them to freely edit adaptations. They can incorporate their own ideas, and introduce new text, formatting, and highlighting into *text adaptations*. The system also retains images from the original text in the adapted version of the text. These can also be edited at the teacher's discretion. The system has a backend database where user-system interactions are stored. These data are used for research purposes, to gauge teacher performance, and to examine feature use, with system development in mind.

3.2.1 Current Text Adaptor Use

Text Adaptor is currently being piloted with two online teacher professional development programs for ELL teachers in the United States: one at a large, private university on the west coast, and another at a large, private university on the east coast. Approximately 120 teachers are participating in the pilot.

The goal of *Text Adaptor* use in teacher professional development settings is to expose teachers to linguistic complexity in text, and make them aware of how to identify and modify linguistic complexity so that text content is more accessible to an ELL

reader. Our hypothesis is that through continued exposure to linguistic complexity, teachers will develop a heightened awareness to linguistic complexity in texts which will better prepare them to develop reading materials for their ELLs. Linguistic sensitivity is one important factor that we believe will contribute to improved teacher quality.

In the scope of the pilot, teachers have completed the following activities in the following sequence: (a) a survey that elicits information about background, (b) manual adaptations based on example ELL student profiles (e.g., 7th grade native-Spanish speaker with intermediate English proficiency), (c) background readings about text adaptation, (d) *Text Adaptor* training, (e) two *Text Adaptor*-created adaptations, and (f) a perception survey to elicit teacher feedback about adaptation practice and use of *Text Adaptor*. There are control and treatment cohorts in the pilot. Control groups complete all adaptations manually, and do not receive information about the tool. Data collection and analyses are underway to evaluate if adaptation quality improves with *Text Adaptor* use. Teachers responses to the perception survey will inform future *Text Adaptor* development.

Teacher adaptations, created in the context of the pilot, will be rated by trained experts, according to a scoring protocol developed for the purpose of assigning numerical ratings to the adaptations for overall quality and for individual traits believed to contribute to overall text difficulty. *From an NLP perspective*, the relationships between the expert ratings and the linguistic features that can be automatically captured in texts (e.g., syntactic complexity, word choice, synonym use, idiom use) will be the first step toward building a model that defines "effective text adaptation" for English language learners. Continuing to build an increasingly larger corpus of rated adaptation data, created for different profiles of English language learners, could provide an extraordinary resource for NLP research related to text quality.

3.2.2 Thinking about NLP Capabilities in *Text Adaptor*

As *Text Adaptor* is currently a prototype, we see that there is opportunity to develop *Text Adaptor* even further to draw attention to potential obstacles in text in the areas of *content*, *syntax* and *discourse*. Our interactions with teachers in school settings (prior to the 2008 pilot) have informed our vision about what current opportunities related to NLP-based development may be.

3.2.2.1 Content. The presence of *polysemous words* in a text can contribute to overall text difficulty [54], [55], [56]. A polysemous word feature would help teachers become more sensitive to another aspect of vocabulary that can render a text more difficult. This is especially important with regard to academic vocabulary, since it is central to content learning. For instance, Social Studies and Science have their own sublanguages. To illustrate, let's take the senses of the word "plant." The sense, "crop," is more likely to be associated with the word "plant" in biology, and the sense "factory" is more likely to be associated with the word "plant" in Social Studies. Teachers can use this information to create an appropriate adaptation.

Several studies reviewed by [57] show a relationship between reading comprehension and knowledge of derivational morphology [43],[58],[59]. In derivational morphology, adding a suffix will change a word's part of speech (e.g., *information*

(noun) → *informational* (adjective)). In [57], the authors studied how ELLs' ability to break down words into meaningful units (popularity = "popular" + "ity") in 4th and 5th grade related to their vocabulary knowledge and reading comprehension. They reported that students who showed a greater understanding of morphology had high reading comprehension scores when holding constant their word reading fluency. They found a significant effect in the 4th grade, and a stronger effect in the 5th grade. Their findings suggest that teaching morphology might improve students' reading comprehension and language outcomes. The ability to identify and highlight *morphologically complex words* would further address vocabulary.

Non-literal expressions, such as idioms and fixed expressions can introduce difficulty into a text for ELLs. Teaching *multi-word expressions* (MWEs) is a critical part of teachers' reading comprehension curriculum [60]. MWEs can be divided into two categories: compositional and non-compositional. Let's use the expression, *red tape*, as an example. The compositional meaning of red tape is, "tape that is red," while the non-compositional meaning is "bureaucratic procedure." Highlighting MWEs in texts would point these terms out to teachers, and further draw their attention to linguistic aspects of a text that could pose difficulty.

3.2.2.2 Syntax. Sentence complexity contributes to text difficulty [1][2],[5],[6]. Further, the types of complex sentence structures that appear in texts can vary by subject domain [61]. *Text Adaptor* currently uses a shallow syntactic parser to capture the number of clauses that are in sentences. Using the parses, the system highlights passive sentences, and sentences with 1, 2 and 3 or more dependent clauses. Additional features related to syntactic complexity can be easily captured, and highlighted in texts, such as complex verb formation (e.g., past and progressive forms), complex noun phrases, and prepositions. While prepositions are not necessarily complex, they are abstract and may require further explanation, and teachers could provide this.

3.2.2.3 Discourse. Text difficulty exists when texts lack coherence [6],[62],[63], [64],[65],[66]. Specifically, if a text jumps from topic-to-topic, and/or the ideas in the text do not logically follow, this could confuse the reader. In terms of developing capabilities that predict text coherence, [13], and [67] have developed systems that examine coherence in student writing. Their systems measure lexical relatedness between text segments by using vector-based similarity between adjacent sentences. This approach to similarity scoring is in line with TextTiling [68],[69], an NLP approach used to identify the subtopic structure of a text. The issue of establishing the coherence of student essays, using the Rough Shift element (abruptness of topic shifts) of Centering Theory [70] has been addressed by [71].

In more recent work, developed a new approach for identifying text coherence using an entity-based representation that measures text coherence through the density of occurrence of vocabulary throughout a text [8],[9]. The approach is different from previous coherence algorithms as it takes into account the sentence position of vocabulary (i.e., subject, object, other). [9] applied their algorithm to relevant tasks in which the algorithm was able to predict text summary coherence and text readability (based on coherence factors). In this research, we will examine how well this new coherence algorithm predicts coherence in content-area classroom texts. Given the success of experiments in [9], useful measures might be derived to predict the relative

cohesiveness of classroom texts, and that these measures could be incorporated into *Text Adaptor* as useful indicators of the overall coherence of original and subsequent adaptations, created by teachers.

In addition, *transition words and phrases* can also contribute to text coherence. ETS has a capability that identifies transitional words and phrases in text. This capability is currently deployed in *Criterion.* In addition, it is critical from a coherence perspective that pronouns have appropriate and clear referents. *Pronoun* tools could be used to evaluate the appropriateness of co-reference in a text. Identification of unclear reference between nouns and pronouns and their locations in a text can help teachers resolve any unclear pronoun issues.

4 Discussion and Conclusions

To illustrate continued opportunities in the field of NLP and education, this paper has used system examples: e-rater®, a commercially-deployed application, and *Text Adaptor*, a research prototype. While both systems incorporate a significant number of NLP methods and tools, both systems have room to grow from an NLP perspective.

E-rater development began with a modest set of NLP-based tools that extracted content, syntactic, and discourse information from student essays to capture the kinds of text characteristics that reflect a range of writing quality. Since its initial development, new features have been developed and added to the system, always keeping in mind the goal of broadening and enriching the system's representation of the writing construct. New features in the past several years have included information about grammar, usage, and mechanics errors, style information, and essay-based discourse analysis. These features are closer to the kinds of information that teachers provide when grading papers, and the kinds of information that human raters consider when scoring writing assessments. All of this new information has been used to build and develop e-rater-produced feedback and scoring in *Criterion*, ETS' online essay evaluation service that has been used by over a million K-12 students in 3,200 schools in the United States. E-rater is also being used in other low-stakes practice settings, and in high-stakes assessments, including the Graduate Record Exam (GRE). Moving forward with e-rater development, considerable work is being done in the area of determiner and preposition error detection, especially to accommodate non-native English speakers who are more likely to make these kinds of errors [72], [73]. Both error types are implemented in *Criterion*. To further support English language learners, research is being done to identify collocation errors in student writing for inclusion in *Criterion* [74]. Additional research is also being pursued using entity-based approaches to develop discourse coherence measures that could be used to provide feedback, and to enhance e-rater scoring [8],[9].

With regard to *Text Adaptor* NLP-based development, continued collection of teacher adaptations created with *Text Adaptor* and then scored by experts, opens a couple of NLP research doors. First, increased numbers of human-rated data would give NLP researchers information about text quality that can be used to derive specific text quality measures which can then be associated with the appropriateness of a text for students at different levels of English proficiency. Second, the text quality features derived from the expert-rated adaptation data can be used to build an array of

models that reflect a range of ELL proficiency levels. While in our current research, we require *expert human raters* to evaluate teacher adaptations, over time, such models could be used to automatically rate the quality of teacher adaptations, in much the same way that students' essays are evaluated in *Criterion*. Certainly, this would be useful in teacher professional development settings for teacher instruction and assessment. It is also possible that a *Text Adaptor*-inspired test question might be used on a teacher certification exam, where teachers have to create adaptations in the context of a certification exam. These adaptations could then be scored automatically. Thinking about more pure NLP research, *Text Adaptor* can essentially be considered as an annotation tool. In our pilot studies, the multiple teacher-created adaptations of the same text that teachers create using *Text Adaptor* can be used for paraphrase research, since the texts are re-written (*paraphrased*) by different teachers. For a given text, teacher-created adaptations yield information about synonymous word use, sentence rewrites, and text summaries.

As indicated in the introduction to this paper, *only* two NLP-based educational applications were discussed in detail. Beyond these applications, however, there are a number of other NLP-based or –supported applications for education, across application domains, including, but not limited to, text- and dialogue-based intelligent tutoring systems, speech scoring systems, and readability or text quality applications. Data collected in the context of these applications can be used both for educational purposes, and for basic NLP research (e.g., paraphrase research).

Continued NLP research to develop educational applications could produce NLP-driven applications that make significant contributions to education, and to the populations (students or teachers) that they serve.

References

1. Flesch, R.: A new readability yardstick. Journal of Applied Psychology 32, 221–233 (1948)
2. MacGinitie, W.H., Tretiak, R.: Measures of sentence complexity as predictors of the difficulty of reading materials. In: Proceedings of the 77th Annual Convention of the APA (1969)
3. Chall, J.S., Dale, E.: Readability revisited: The new Dale-Chall readability formula, p. 159. Brookline Books, Cambridge (1995)
4. Collins-Thompson, K., Callan, J.: A language modeling approach to predicting reading difficulty. In: Proceedings of the HLT/NAACL (2004)
5. Schwarm, S., Ostendorf, M.: Reading level assessment using support vector machines and statistical language models. In: Proceedings of the ACL, Ann Arbor, MI, pp. 523–530 (2005)
6. Deane, P., Sheehan, K.M., Sabatini, J., Futagi, Y., Kostin, I.: Differences in text structure and its implications for assessment of struggling readers. Scientific Studies of Reading 10(3), 257–275 (2006)
7. Elhadad, N., Sutaria, K.: Mining a lexicon of technical terms and lay equivalents. In: Biological, translational, and clinical language processing, ACL, Prague, Czech Republic, pp. 49–56 (2007)
8. Barzilay, R., Lapata, M.: Modeling local coherence: An entity-based approach. In: Proceedings of the 43rd Annual Meeting of the ACL, Ann Arbor, MI, pp. 141–148 (2005)

9. Barzilay, R., Lapata, M.: Modeling local coherence: An entity-based approach. Computational Linguistics 34(1), 1–34 (2008)
10. Pitler, E., Nenkova, A.: Revisiting Readability: A Unified Framework for Predicting Text Quality. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (2008)
11. Page, E.B.: The imminence of grading essays by computer. Phi Delta Kappan 48, 238–243 (1966)
12. Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., Harris, M.D.: Automated scoring using a hybrid feature identification technique. In: Proceedings of the Annual Meeting of the ACL, Montreal, Canada (1998)
13. Foltz, P.W., Kintsch, W., Landauer, T.K.: Analysis of Text Coherence Using Latent Semantic Analysis. Discourse Processes 25(2-3), 285–307 (1998)
14. Macdonald, N.H., Frase, L.T., Gingrich, P.S., Keenan, S.A.: Writer's Workbench: Computer Aid for Text Analysis. IEEE Transactions on Communications, Special Issue on Communication in the Automated Office 30(1), 105–110 (1982)
15. Carbonell, J.R.: AI in CAI: An artificial intelligence approach to computer-assisted instruction. IEEE Transactions on Man-Machine Systems 11(4), 190–202 (1970)
16. Brown, J.S., Burton, R.R., Bell, A.G.: SOPHIE: A sophisticated instruction environment for teaching electronic troubleshooting (An example of AI in CAI). BBN Technical Report 2790. Bolt, Beranek, and Newman, Inc., Cambridge (1974)
17. Stevens, A.L., Collins, A.: The goal structure of a Socratic tutor. BBN Technical Report 351. Bolt, Beranek, and Newman, Inc., Cambridge (1977)
18. Burton, R.R., Brown, J.S.: An investigation of computer coaching for informal Activities. In: Sleeman, D.H., Brown, J.S. (eds.) Intelligent Tutoring Systems. Academic Press, New York (1982)
19. Clancy, W.J.: Knowledge-Based Tutoring: The GUIDON Program. MIT Press, Cambridge (1987)
20. VanLehn, K., Graesser, A.C., Jackson, G.T., Jordan, P., Olney, A., Rose, C.P.: When are tutorial dialogues more effective than reading? Cognitive Science 31(1), 3–52 (2007)
21. Leacock, C., Chodorow, M.: c-rater: Scoring of short-answer questions. Computers and the Humanities 37(4), 389–405 (2003)
22. Sukkarieh, J., Bolge, E.: Leveraging C-rater's automated scoring capability for providing instructional feedback for short constructed responses. In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 779–783. Springer, Heidelberg (2008)
23. Bernstein, J.: PhonePass testing: Structure and construct, Ordinate Corporation, Menlo Park, CA (May 1999)
24. Zechner, K., Higgins, D., Xi, X.: SpeechRater^{TM}: A Construct-Driven Approach to Scoring Spontaneous Non-Native Speech. In: SLaTE 2007 (2007)
25. Burstein, J., Shore, J., Sabatini, J., Lee, Y., Ventura, M.: Developing a reading support tool for English language learners. In: Demo Proceedings of NAACL-HLT 2007, Rochester, NY (2007)
26. Salton, G.: Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley, Reading (1989)
27. Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M.: Enriching automated scoring using discourse marking. In: Proceedings of the workshop on discourse relations & discourse marking in conjunction with the ACL, Montreal, Canada (1998)
28. Larkey, L.: Automatic Essay Grading Using Text Categorization Techniques. In: Proceedings of the 21st ACM-SIGIR Conference on Research and Development in Information Retrieval, Melbourne, Australia, pp. 90–95 (1998)

29. Quirk, R., Sydney, G., Leech, G., Svartik, J.: A Comprehensive Grammar of the English Language, Longman, New York (1985)
30. Higgins, D., Burstein, J., Attali, Y.: Identifying off-topic student essays without topic-specific training data. Natural Language Engineering 12(2), 145–159 (2006)
31. Burstein, J., Chodorow, M., Leacock, C.: Automated essay evaluation: The Criterion Online Writing Evaluation service. AI Magazine 25(3), 27–36 (2004)
32. Ratnaparkhi, A.: A Maximum Entropy Part-of- Speech Tagger. In: Proceedings of EMNLP, University of Pennsylvania (1996)
33. Chodorow, M., Leacock, C.: An Unsupervised Method for Detecting Grammatical Errors. In: Proceedings of NAACL, pp. 140–147 (2000)
34. Golding, A.: A Bayesian Hybrid for Context-Sensitive Spelling Correction. In: Proceedings of the 3rd Workshop on Very Large Corpora, Cambridge, MA, pp. 39–53 (1995)
35. Burstein, J., Wolska, M.: Toward evaluation of writing style: Finding overly repetitive word use in student essays. In: Proceedings of the 11th Conference of the EACL, Budapest, Hungary (2003)
36. Burstein, J., Marcu, D., Knight, K.: Finding the WRITE stuff: Automatic identification of discourse structure in student essays. In: Harabagiu, S., Ciravegna, F. (eds.) IEEE Intelligent Systems: Special Issue on Advances in Natural Language Processing, vol. 18(1), pp. 32–39 (2003)
37. Attali, Y., Burstein, J.: Automated essay scoring with e-rater v.2. Journal of Technology, Learning, and Assessment 4(3) (2006)
38. Francis, D., Rivera, M., Lesaux, N., Keiffer, M., Rivera, H.: Practical guidelines for the education of English language learners: Research based recommendations for instruction and academic interventions. Center on Instruction, Portsmouth (2006) (retrieved April 25, 2008), http://www.centeroninstruction.org/files/ELL1-Interventions.pdf
39. Calderón, M.: Curricula and methodologies used to teach Spanish-speaking limited English proficient students to read English. In: Slavin, R.E., Calderón, M. (eds.) Effective programs for Latino students, pp. 251–305. Lawrence Erlbaum, Mahwah (2001)
40. National Council for the Social Studies. Expectations of excellence: Curriculum standards for social studies, Washington, DC (1994)
41. Calderón, M., Minaya-Rowe, L.: Teaching reading, oral language and content to English language learners – How ELLs keep pace with mainstream students. Corwin Press, Thousand Oaks (2007)
42. Gándara, P., Maxwell-Jolly, J., Driscoll, A.: Listening to teachers of English language learners: A survey of California teachers' challenges, experiences, and professional development needs. The Regents of the University of California, Sacramento (2005) (retrieved September 14, 2007), http://www.tyg.jp/tgu/school_guidance/bulletin/K14/images/nishimura.pdf
43. Nagy, W.E., Berninger, V.W., Abbott, R.D.: Contributions of morphology beyond phonology to literacy outcomes of upper elementary and middle-school students. Journal of Educational Psychology 98(1), 134–146 (2006)
44. August, D.: Supporting the development of English literacy in English language learners: Key issues and promising practices (Report No. 61). The John Hopkins University, Center for Research on the Education of Students Placed at Risk, Baltimore (2003), http://www.cde.state.co.us/cdesped/download/pdf/ELL_SupportDevelopEngLangLit.pdf

45. Echevarria, J., Vogt, M., Short, D.: Making content comprehensible for English language learners: The SIOP model. Pearson Education, Inc., New York (2004)
46. Scarcella, R.: Academic English: A Conceptual Framework. University of California, Linguistic Minority Research Group, University of California, Irvine, Technical Report 2003-1 (2003)
47. Calderón, M.: Teaching Reading to English Language Learners, Grades 6-12: A Framework for Improving Achievement in the Content Areas. Corwin Press, Thousand Oaks (2007)
48. Schleppegrell, M.: The Linguistic Challenges of Mathematics Teaching and Learning: A Research Review. Reading and Writing Quarterly 23, 139–159 (2007)
49. Kieffer, M.J., Lesaux, N.K.: Breaking down words to build meaning: Morphology, vocabulary, and reading comprehension in the urban classroom. The Reading Teacher 61, 134–144 (2007)
50. Adger, C.T., Snow, C., Christian, D.: What teachers need to know about language. Center for Applied Linguistics, Washington (2002)
51. Calderón, M., August, D., Slavin, R., Cheung, A., Durán, D., Madden, N.: Bringing words to life in classrooms with English language learners. In: Hiebert, A., Kamil, M. (eds.) Research and development on vocabulary. Lawrence Erlbaum, Mahwah (2005)
52. Hiebert, A., Kamil, M. (eds.): Research and development on vocabulary. Lawrence Erlbaum, Mahwah (2005)
53. Marcu, D.: The Theory and Practice of Discourse Parsing and Summarization. MIT Press, Cambridge (2000)
54. Gernsbacher, M.A., Faust, M.: The mechanism of suppression: A component of general comprehension skill. Journal of Experimental Psychology: Learning, Memory, & Cognition 17, 245–262 (1991)
55. Kang, H.-W.: How can a mess be fine? Polysemy and reading in a foreign language. Mid-Atlantic Journal of Foreign Language Pedagogy 1, 35–49 (1993)
56. McNamara, D.S., McDaniel, M.: Suppressing irrelevant information: Knowledge activation or inhibition? Journal of Experimental Psychology: Learning, Memory, & Cognition 30, 465–482 (2004)
57. Kieffer, M.J., Lesaux, N.K.: Breaking down words to build meaning: Morphology, vocabulary, and reading comprehension in the urban classroom. The Reading Teacher 61, 134–144 (2007)
58. Carlisle, J.F.: Awareness of the structure and meaning of morphologically complex words: Impact on reading. Reading and Writing: An Interdisciplinary Journal 12, 169–190 (2000)
59. Freyd, P., Baron, J.: Individual differences in acquisition of derivational morphology. Journal of Verbal Learning and Verbal Behavior 21, 282–295 (1982)
60. Palmer, B.C., Brooks, M.A.: Reading until the cows come home: Figurative language and reading comprehension. Journal of Adolescent and Adult Literacy 47(5), 370–379 (2004)
61. Schleppegrell, M., de Oliveira, L.C.: An integrated language and content approach for history teachers. Journal of English for Academic Purposes 5(4), 254–268 (2006)
62. Kintsch, W.: The use of knowledge in discourse processing: A construction-integration model. Psychological Review 95, 163–182 (1988)
63. Kintsch, W.: Comprehension: A paradigm for cognition. Cambridge University Press, Cambridge (1998)
64. McNamara, D.S.: Reading both high and low coherence texts: Effects of text sequence and prior knowledge. Canadian Journal of Experimental Psychology 55, 51–62 (2001)
65. McNamara, D.S., Kintsch, W.: Learning from text: Effects of prior knowledge and text coherence. Discourse Processes 22, 247–287 (1996)

66. Van Dijk, T.A., Kintsch, W.: Strategies of discourse comprehension. Academic Press, New York (1983)
67. Weimer-Hastings, P., Graesser, A.: Select-a-Kibbutzer: A computer tool that gives meaningful feedback on student compositions. Interactive Learning Environments 8(2), 149–169 (2000)
68. Hearst, M.A., Plaunt, C.: Subtopic structuring for full-length document access. In: Proceedings of the Association of Computational Linguistics, Special Interest Group on Information Retrieval, pp. 59–68 (1993)
69. Hearst, M.A.: Text Tiling: Segmenting text into multi-paragraph subtopic passages. Computational Linguistics 23(1), 33–64 (1997)
70. Grosz, B., Joshi, A., Weinstein, S.: Centering: A framework for modeling the local coherence of discourse. Computational Linguistics 21(2), 203–226 (1995)
71. Miltsakaki, E., Kukich, K.: Automated evaluation of coherence in student essays. In: Proceedings of the 2nd International Conference on Language Resources & Evaluation, Athens, Greece (2000)
72. Han, N.-R., Chodorow, M., Leacock, C.: Detecting errors in English article usage by non-native speakers. Natural Language Engineering 12(2), 115–129 (2006)
73. Tetreault, J., Chodorow, M.: The Ups and Downs of Preposition Error Detection in ESL Writing. In: COLING, Manchester, UK (2008)
74. Futagi, Y., Deane, P., Chodorow, M., Tetreault, J.: A computational approach to detecting collocation errors in the writing of non-native speakers of English Computer Assisted Language Learning 21(4), 353–367 (2008)